
The Evolving Clean Energy Landscape:

Drivers, Trends, and Future Prospects

By
YUQI ZHENG
Supervised by
PROF. BRAIN LUCEY



School of Business
TRINITY COLLEGE DUBLIN,
THE UNIVERSITY OF DUBLIN

A dissertation submitted to the University of Dublin
in accordance with the requirements of the degree of
DOCTOR OF PHILOSOPHY in the School of Business.

JANUARY 2026

DECLARATION

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work. I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Juré sur l'honneur en connaissance des conséquences juridiques.

SIGNED:

Yuqi Zheng

DATE: 19/01/2026.....

SUMMARY

The rapid growth of the clean energy sector raises critical questions about the factors shaping its financial performance. Prior studies have documented that clean energy stock returns are primarily driven by market uncertainty and financial distress, particularly in the context of small-cap stocks and emerging markets, rather than by conventional determinants such as oil prices (Ciner et al., 2023).

Building on this, the first study examines a broad set of financial drivers, including macroeconomic conditions, policy uncertainty, technological innovation, market volatility, global economic trends, oil prices, green stocks, and emerging market equities. Additionally, it incorporates insights from the GDELT database, capturing sentiment indicators from media reports and biases in political coverage. To preliminarily validate these factors, this study applies unsupervised learning techniques and employs Extreme Bound Analysis (EBA) to assess whether both established and novel factors significantly impact the clean energy market. This approach extends prior research and provides a foundation for future studies. The findings on traditional and emerging factors further inform clean energy market forecasting.

Given the widely use of machine learning, our second study integrates statistical, machine learning, and hybrid models to identify the most effective forecasting approach. Model performance is evaluated based on forecasting horizons, accuracy, and trend estimation differences. This comprehensive assessment aims to determine the most suitable model for predicting the clean energy market under various conditions.

The first study has primarily analyzed the relationship between media reporting preferences, sentiment values, and the clean energy market. In the third stage of research, the scope of data collection will be expanded to include models that measure the strength of policy commitments, introducing a novel perspective on the clean energy market from a political science standpoint. Specifically, our

third study intends to examine textual reports of speeches by central bankers and employ advanced language models to assess key factors such as urgency (how soon), intensity (how much), and timeliness (how long). By incorporating this policy commitment dimension, the research aims to identify additional driving forces influencing the clean energy market.

This thesis advances understanding of the clean energy market by combining traditional and emerging drivers with insights from media coverage and political discourse, addressing a dimension that has received limited scholarly attention. It shows how positive narratives and credible policy commitments enhance transparency, attract responsible investors, and shape portfolio decisions, creating a self-reinforcing cycle that accelerates the clean energy transition. Using machine learning and statistical models, it also provides a methodological basis for forecasting market volatility, offering practical insights for investors, policymakers, and researchers.

DEDICATION

To whom it may concern.

ACKNOWLEDGMENTS

Time really flies — I am now 26 years old and about to complete my PhD. While this marks the end of my years as a student, it also opens a brand-new chapter in life. I believe it is time to step into the next chapter of my growth and keep contributing to academia and the global economy. Looking back on 26 years of growth, I am genuinely proud of who I have become and I truly enjoy the journey of constant progress. This journey would not have been possible without the support of many people. Their help has shaped my commitment to lifelong learning and laid a solid foundation for my career.

First, I want to thank my family members. They always supported me in everything I wanted to do, which gave me the courage to face challenges. I also want to thank my supervisor, Prof. Brian Lucey. He accepted me when I felt lost about my life goals and hesitant to try new things, and he helped me become someone who can meet any challenge with confidence. Since my bachelor's and master's degrees were in accounting, working with PySpark code for GDELT once felt impossible to me. With his guidance and encouragement, however, I not only developed new code for data collection but also created my own GitHub. He has supported me not only in research but also in life, helping me integrate fully into the academic community.

I would like to thank my undergraduate thesis supervisor, Prof. Zhenkun Wang(Felix), who inspired my passion for research and encouraged me to pursue a PhD. His guidance and support have been invaluable, giving me confidence and a steady platform for growth. I am grateful to my senior colleagues for their advice on writing and career planning, and to my panel members, Prof. Jenny Berrill and Prof. Aleksandar Sevic, whose constructive feedback greatly strengthened my PhD thesis.

Finally, I would like to thank everyone in my life who has cared for and supported me, as well as Trinity College Dublin. I wish all of them continued success and happiness, and I hope the university will keep rising in the QS rankings.

TABLE OF CONTENTS

	Page
List of Tables	xv
List of Figures	xvii
1 Introduction	1
1.1 Background and Motivation	1
1.2 Aims and Objectives	2
1.3 Contribution	3
1.3.1 Chapter 2: What REALLY drives clean energy stocks - Fear or Fundamentals?	4
1.3.2 Chapter 3: Supervised Learning Models, Statistical Models or Hybrid Models? A Prediction of Clean Energy Stock Based on Fear and Fundamental Factors	5
1.3.3 Chapter 4: Central Bankers' Political Discourse as a Driver of Clean Energy Markets	5
1.4 Structure	6
1.5 Publications from this Doctoral Thesis	7
1.6 Variable Definitions	8
2 What <i>REALLY</i> drives clean energy stocks - Fear or Fundamentals?	13
2.1 Introduction	1
2.2 Literature Review	2
2.2.1 Previous Studies on Media, Sentiment and Clean Energy Transition	2
2.2.2 Media Preferences in Global Key Clean Energy Leaders	3
2.2.3 Traditional Factors Influencing the S&P Clean Energy Index	5
2.3 Methodology	7

TABLE OF CONTENTS

2.3.1	Validating the Reliability of Factors—Isolation Forest	7
2.3.2	Data Collection	10
2.3.2.1	GDELT Variables Source	12
2.3.3	EBA with All Combinations of Doubtful Variables	14
2.3.4	A More Sophisticated EBA Analysis	16
2.3.5	Discussions and Results	17
2.4	Conclusion and Future Research	23
2.4.1	Conclusion and Contribution	23
2.4.2	Limitation of GDELT Database and Future Research	25
3	Supervised Learning Models, Statistical Models or Hybrid Models?	
	A Prediction of Clean Energy Stock Based on Fear and Fundamen-	
	tal Factors	27
3.1	Introduction	28
3.2	Literature Review	30
3.2.1	Tree-based Models	30
3.2.2	Classification: K-Nearest Neighbors(KNN)	30
3.2.3	Prophet and its Hybrid Model	31
3.2.4	Long Short-Term Memory(LSTM), Statistical and Hybrid Model	31
3.2.5	Supervised Autoencoder (SAE)	33
3.2.6	Theoretical Motivation for Model Selection	33
3.3	Methodology	35
3.3.1	Data Selection and Pre-processing	35
3.3.1.1	Data Source and Variables Explanation	35
3.3.1.2	Correlation of Vairables	36
3.3.1.3	Feature Importance of Vairables	38
3.3.1.4	SHAP Value Validation and PCA technique	39
3.3.1.5	Processing the Missing Values	42
3.3.1.6	Feature Data Processing and Overall Model Approach	43
3.3.1.7	Stochastic Simulated Data: Random Forest and LSTM	44
3.3.2	Statistical Models Forecasting	52
3.3.2.1	ARIMA-Univariate Linear Model	52
3.3.2.2	VAR-Multivariate linear Model	52
3.3.3	Supervised Learning and Hybrid Models	53
3.3.3.1	LSTM-Multivariate Nonlinear Model	53

3.3.3.2	LSTM_Autoregressive-Multivariate Nonlinear with regression trends Model	53
3.3.3.3	SAE-Multivariate Nonlinear Model	54
3.3.3.4	Prophet Model Based on Historical Trends-Univariate Nonlinear Model	54
3.3.3.5	Prophet Model Based on Generated Features-Multivariate Nonlinear with regression trends Model	54
3.3.4	Results and Comparative Analysis	55
3.3.4.1	Linear Trends Results	55
3.3.4.2	Non-linear Trends with Supervised Learning Models	57
3.3.4.3	Non-linear Trends with Hybrid Models	61
3.3.4.4	Prediction Comparisons-Whole Prediction Period . .	64
3.3.4.5	Prediction Comparisons-Short and Medium-term Period	68
3.4	Conclusion and Contribution	71
3.4.1	Conclusion and Future Research	71
3.4.2	Technical Contribution	74
3.4.3	Theoretical Contribution	74
3.5	Limitations	75
A	Equations and Parameters Tuning	77
A.1	XGBoost	77
A.2	Random Forest	78
A.3	KNN	79
A.4	LSTM Simulated Data	79
A.5	ARIMA	80
A.6	VAR	81
A.7	LSTM	84
A.8	LSTM_Autoregressive	87
A.9	SAE	88
A.10	Prophet	89
A.11	Prophet Features	91
4	Central Bankers' Political Discourse as a Driver of Clean Energy Markets	93
4.1	Introduction	94

TABLE OF CONTENTS

4.2	Literature Review	95
4.2.1	Authenticity, Performance, and Priorities in Political Discourse	95
4.2.2	China–U.S. Dominance in the Clean Energy Market and Their Different Expressions	98
4.2.3	The Relationship Between Central Bankers’ Speeches, Sus- tainability and Financial Market Dynamics	100
4.3	Methodology	102
4.3.1	Data Collection	102
4.3.1.1	Data Selection and Deduplication	102
4.3.1.2	In the Absence of Political Priorities	103
4.3.1.3	In the Presence of Political Priorities	103
4.3.2	Extreme Bound Analysis, SHAP Value and Results	104
4.3.2.1	In the Absence of Political Priorities	104
4.3.2.2	In the Presence of Political Priorities	105
4.3.2.3	Combining Traditional, Media-driven and Political Text-based Effects	107
4.3.2.4	SHAP Value-Supplementary Evidence	109
4.4	Discussion and Conclusion	109
4.4.1	In the Absence of Political Priorities	109
4.4.2	In the Presence of Political Priorities	111
4.4.3	Combining Traditional, Media-driven and Political Text-based Effects	112
4.4.4	SHAP Value-Supplementary Evidence	113
4.5	Contribution and Limitation	113
4.5.1	Contribution	113
4.5.2	Limitation and Future Research	115
B	Data Collection	117
B.1	In the Absence of Political Priorities	117
B.2	In the Presence of Political Priorities	117
B.3	Iterative Matching	117
C	Extreme Bound Analysis-Steps	121
C.1	Absence	121
C.1.1	All Combinations of Doubtful Variables	121
C.1.2	Verifying the Accuracy of Variable Classification	122

C.1.3	More Sophisticated	123
C.2	Presence	123
C.2.1	All Combinations of Doubtful Variables	123
C.2.2	Verifying the Accuracy of Variable Classification	124
C.2.3	More Sophisticated	125
C.3	Combining Traditional, Media-driven and Political Text-based Effects	125
C.3.1	All Combinations of Doubtful Variables	125
C.3.2	Verifying the Accuracy of Variable Classification	127
C.3.3	More Sophisticated	128
D	Extreme Bound Analysis-Some Results	129
D.1	Absence	129
D.1.1	All Combinations of Doubtful Variables	129
D.1.2	Verifying the Accuracy of Variable Classification	131
D.2	Presence	131
D.2.1	All Combinations of Doubtful Variables	131
D.2.2	Verifying the accuracy of variable classification	134
D.3	Combining Traditional, Media-driven and Political Text-based Effects	134
D.3.1	All Combinations of Doubtful Variables	134
D.3.2	Verifying the Accuracy of Variable Classification	136
E	SHAP Value-Supplementary Evidence	141
5	Conclusion	143
5.1	Summary of Results	143
5.1.1	Methods and Data Engineering	143
5.1.2	Correlation Findings	144
5.1.2.1	Media, Sentiment, and Traditional Factors	144
5.1.2.2	Speech-Based Drivers(Central Bank Communica- tions)	145
5.1.3	Forecasting Results	145
5.1.3.1	Different Time Periods	145
5.1.3.1.1	Short term (1 week).	145
5.1.3.1.2	Medium term (1 month).	145
5.1.3.1.3	Long term (6 months).	145
5.1.3.2	Model Interpretation	145

TABLE OF CONTENTS

5.1.4 Media Perspectives and Implications	146
5.2 Future Research and Outlook	146
Bibliography	147

LIST OF TABLES

TABLE	Page
2.1 Traditional Influencing Factors and Source	6
2.2 Isolation Forest Results	9
2.3 Sala-i-Martin's EBA (All Combinations of Doubtful Variables)	18
2.4 Sala-i-Martin's EBA (More Sophisticated)	20
2.5 Sala-i-Martin's EBA (Enhance the Robustness)	21
3.1 Variables Explanation	36
3.2 RMSE for Each Feature	46
3.3 Model Comparison Metrics (Random Forest)	66
3.4 Model Comparison Metrics (LSTM)	67
3.5 Comparison of Different Models with Their Strengths, Weaknesses, and Accuracy	73
A.1 ADF Test Results(Random Forest)	82
A.2 ADF Test Results(LSTM)	82
A.3 Summary of VAR Model(Random Forest)	83
A.4 Summary of VAR Model(LSTM)	83
4.1 Definitions and Interpretations of Key Textual Variables	97
4.2 Sala-i-Martin's EBA (In the Absence of Political Priorities-More Sophisticated)	104
4.3 Sala-i-Martin's EBA (In the Presence of Political Priorities-More Sophisticated)	105
4.4 Sala-i-Martin's EBA (More Sophisticated)	107
D.1 Sala-i-Martin's EBA(In the Absence of Political Priorities-All Combinations of Doubtful Variables)	129

LIST OF TABLES

D.2	Sala-i-Martin's EBA (In the Absence of Political Priorities–Verifying the Accuracy of Variable Classification)	131
D.3	Sala-i-Martin's EBA (In the Presence of Political Priorities-All Combinations of Doubtful Variables)	132
D.4	Sala-i-Martin's EBA (In the Presence of Political Priorities-Verifying the Accuracy of Variable Classification)	134
D.5	Sala-i-Martin's EBA (All-All Combinations of Doubtful Variables)	135
D.6	Sala-i-Martin's EBA (All-Verifying the Accuracy of Variable Classification)	137
5.1	Comparative Overview	144

LIST OF FIGURES

FIGURE	Page
1.1 Research Roadmap	7
2.1 GDELT Data Collection Flow Chart	13
2.2 GDELT Variables Source	14
2.3 All Combinations of Doubtful Variables	17
2.4 A More Sophisticated EBA Analysis	20
2.5 A More Sophisticated EBA Analysis (Enhance the Robustness)	21
3.1 Correlation Heatmap	37
3.2 Feature Importance of Variables	39
3.3 SHAP Value of Variables	40
3.4 Feature Data Processing and Overall Model Approach	44
3.5 Cluster Evaluation	47
3.6 Cluster 0	48
3.7 Cluster 1	49
3.8 Cluster 2	50
3.9 Cluster 3	51
3.10 S&P Global Clean Energy Index Forecast_ARIMA	55
3.11 S&P Global Clean Energy Index Forecast_VAR(Random Forest)	56
3.12 S&P Global Clean Energy Index Forecast_VAR(LSTM)	56
3.13 S&P Global Clean Energy Index Forecast_LSTM(Random Forest)	57
3.14 S&P Global Clean Energy Index Forecast_LSTM(LSTM)	58
3.15 S&P Global Clean Energy Index Forecast_LSTM_Autoencoder(Random Forest)	58
3.16 S&P Global Clean Energy Index Forecast_LSTM_Autoencoder(LSTM)	59
3.17 S&P Global Clean Energy Index Forecast_Prophet_without_other_features	60
3.18 Prophet Time Series Forecasting Plot	60

LIST OF FIGURES

3.19	S&P Global Clean Energy Index Forecast_LSTM_Autoregressive(Random Forest)	61
3.20	S&P Global Clean Energy Index Forecast_LSTM_Autoregressive(LSTM)	62
3.21	S&P Global Clean Energy Index Forecast_Prophet_Features (Random Forest)	62
3.22	Prophet Time Series Forecasting Plot_Random Forest	63
3.23	S&P Global Clean Energy Index Forecast_Prophet_Features_LSTM)	63
3.24	Prophet Time Series Forecasting Plot_LSTM	64
3.25	Prediction Comparison Plot(Random Forest)	65
3.26	Prediction Comparison Plot(LSTM)	65
3.27	Prediction Comparison Plot(Random Forest_7)	69
3.28	Prediction Comparison Plot(Random Forest_30)	69
3.29	Prediction Comparison Plot(LSTM_7)	70
3.30	Prediction Comparison Plot(LSTM_30)	70
A.1	ARIMA Parameters Selection	80
4.1	In the Absence of Political Priorities-More Sophisticated	104
4.2	In the Presence of Political Priorities-More Sophisticated	106
4.3	All-More Sophisticated	108
4.4	SHAP	110
B.1	Iterative Matching	119
D.1	In the Absence of Political Priorities-All Combinations of Doubtful Variables	130
D.2	In the Absence of Political Priorities-Verifying the Accuracy of Variable Classification	132
D.3	In the Presence of Political Priorities-All Combinations of Doubtful Variables	133
D.4	In the Presence of Political Priorities-Verifying the Accuracy of Variable Classification	136
D.5	All-All Combinations of Doubtful Variables	138
D.6	All-Verifying the Accuracy of Variable Classification	139

INTRODUCTION

1.1 Background and Motivation

Concerns over fossil fuels and their contribution to climate change have intensified global efforts to seek cleaner energy alternatives. Technological advancements have accelerated the adoption of renewable energy sources, such as solar and wind power, which are valued for their sustainability and minimal environmental impact. This transition has not only heightened attention on the clean energy industry but has also driven governments and policymakers to enact eco-friendly policies and legislation aimed at curbing greenhouse gas emissions and mitigating climate change. As a result, the growing shift toward renewable energy has fueled the rapid expansion of the clean energy sector, making it increasingly attractive to environmentally friendly investors (Bilgen et al., 2004; Obama, 2017; Liang et al., 2019; Pham, 2019; Mangla et al., 2020; Dawar et al., 2021; Farouq et al., 2021). The rapid expansion of the clean energy sector raises critical questions about its relationship with existing markets, particularly the factors driving its financial performance. Therefore, in addition to the impact of traditional factors on the clean energy market, such as market volatility and oil prices, the influence of policy orientation and media reporting preferences on clean energy index has gradually provided a new perspective for studying the clean energy market. Although previous studies have examined the influence of media coverage and specific policies on the clean energy transition, no research has systematically analysed this issue

through the lens of five years of media reports and policy commitments. This study primarily focuses on data from the world's two leading contributors to clean energy index, China and the United States, utilizing extensive datasets to refine certain traditional factors and introduce new variables that influence the clean energy market.

Another essential dimension in the study of the clean energy market is its pricing dynamics. Stock price prediction studies revolve around two perspectives: the Efficient Market Hypothesis (EMH), which argues that prices are unpredictable (Fama, 1965), and the contrasting perspective that historical data can provide insights into future trends (Lo and Hasanhodzic, 2010). Recently, researchers primarily use supervised learning methods such as Support Vector Machines (SVM), decision trees, and Artificial Neural Networks (ANN) (Ślepaczuk and Zenkova, 2018). Machine learning, including deep learning, has become a key tool for classification and prediction, with recent studies shifting from regression to classification models to better predict price trends and generate investment signals (Bieganowski and Slepaczuk, 2024). Matsubara et al. (2018) propose a generative model based on media reports that integrates price movements as conditions to reduce overfitting risk. While these models perform well for long-term forecasts, short-term predictions remain a challenge. Besides, with the rise of machine learning models and hybrid models, integrating these novel factors into such models to predict both the long-term and short-term volatility, as well as specific values of the clean energy market, presents a highly innovative research approach.

This thesis integrates the policy commitment dimension and applies Extreme Bounds Analysis (EBA) and SHAP values to evaluate the relative importance of both novel and traditional determinants, including market volatility and oil prices (Zheng and Lucey, 2025c), while identifying additional drivers of clean energy market dynamics in China and the United States. It further assesses whether policymakers' commitments genuinely reflect their intentions, evaluates the feasibility of implementation, and examines whether market movements are driven solely by policy announcements or also by traditional and media factors.

1.2 Aims and Objectives

This thesis studies the main factors and prediction of the S&P Clean Energy Index within a multifactor informational framework that links financial market, media,

and policy. It uses data from the GDELT Database, including media confidence, sentiment and coverage in the US and China, together with traditional factors such as oil prices and volatility. Machine learning methods such as Isolation Forest are used to validate key variables, and SHAP values test their feature importance. Statistical models like ARIMA and VAR, along with machine learning models such as LSTM and SAE, are applied for forecasting. Extreme Bounds Analysis is used to test the stability of media and policy effects. This framework demonstrates the ways in which information and communication influence market dynamics and investor behavior. Although the clean energy sector has grown rapidly, existing research tends to address its financial, informational, and policy aspects separately. This thesis fills this gap by developing a multifactor framework that captures the informational architecture of clean energy market, which links traditional fundamentals with new media and policy-based sources of information. Within this framework, each chapter examines one analytical layer of the architecture:

Market Fundamentals Layer: How strongly do traditional market factors, such as oil prices and volatility, influence the clean energy markets?

Media Information Layer: How do media-based factors, including confidence, media sentiment, and coverage preferences in the US and China, affect market behavior and perception?

Modeling and Prediction Layer: How do statistical, supervised, and hybrid models perform under different forecasting horizons in terms of suitability, accuracy, and lag effects?

Policy and Communication Layer: To what extent do policymakers' statements and commitments reflect true intent, and how do central-bank communication and media signals compare with traditional market indicators?

1.3 Contribution

This thesis makes several contributions to both theory and practice. First, it demonstrates how global media narratives and timely policy commitments shape the clean energy market by enhancing transparency, raising public awareness, and attracting responsible investors. These interactions foster a self-reinforcing cycle that accelerates the clean energy transition and supports environmental objectives.

Second, the research advances market prediction methodologies by integrating statistical and machine learning models with both historical and simulated

data. This improves forecasting accuracy for the clean energy index across short-, medium-, and long-term horizons, thereby offering practical tools for investors to make more informed decisions and optimize investment strategies.

Third, the thesis contributes to the understanding of how corporate and policymaker discourse translates into tangible commitments within the clean energy market. By evaluating the authenticity, timing, and credibility of these commitments, it sheds light on the hidden drivers of market reactions beyond official announcements.

Finally, this thesis bridges perspectives on incentives and communication by relating clean energy market dynamics to signals identified in media and political narratives. By reducing information asymmetry and supporting credible messaging, it helps align the interests of market participants and lays a technical basis for future work on market volatility and financial forecasting. The parameter-tuning routines and hybrid modeling approaches introduced here are readily transferable to other financial indices, offering lasting value to both academia and industry.

1.3.1 Chapter 2: What REALLY drives clean energy stocks - Fear or Fundamentals?

This chapter examines the relationship between the S&P Global Clean Energy Index and two groups of factors: traditional fundamentals and fear factors derived from the GDELT database. The fundamentals, including market volatility, the Global Financial Stress Indicator, the Green Bond Index, and the US Dollar Index, show only conditional or unstable effects, suggesting that their influence is weaker than previously thought. In contrast, the fear factors have strong and consistent effects on the index. Negative U.S.–China reporting also correlates with the index but seems to encourage clean energy growth as both countries compete for leadership. Methodologically, this chapter combines Extreme Bounds Analysis with Isolation Forest and develops daily GDELT pre-processing routines, improving the robustness and clarity of the results.

Overall, positive media narratives enhance transparency, attract investors, and reinforce supportive policies, creating a self-reinforcing cycle that accelerates the clean energy transition.

1.3.2 Chapter 3: Supervised Learning Models, Statistical Models or Hybrid Models? A Prediction of Clean Energy Stock Based on Fear and Fundamental Factors

The main technical contribution of this chapter is using media-related factors to predict the S&P Clean Energy Index, improving accuracy for stakeholders who rely on news. By incorporating historical data, I identify optimal time-series models for short-, medium-, and long-term forecasts, offering both more precise predictions and insights into lag-related challenges.

This work also lays the groundwork for forecasting the volatility of other financial indices. My parameter-tuning techniques, combined with machine learning and statistical models, advance financial prediction beyond traditional reliance on historical data. By integrating both historical and forward-looking factors, my models help investors make better-informed short- and long-term decisions, enhancing strategy effectiveness.

Besides, the S&P Clean Energy Index predictions provide a clearer picture of global clean energy trends, highlighting drivers of sustainable growth and supporting environmental governance. These insights benefit not only investors but also policymakers, enabling more effective strategies for the global energy transition and long-term sustainability.

1.3.3 Chapter 4: Central Bankers' Political Discourse as a Driver of Clean Energy Markets

This chapter contributes by introducing political text-based indicators as novel drivers of the clean energy markets, focusing on central bank governors' speeches. By systematically examining dimensions such as "Commitment", "Intensity", "Timing", and "Urgency", I show that stronger and more urgent statements not only convey credible policy intentions but also directly shape investor confidence and expectations in clean energy markets. Compared with traditional financial indicators and media-based variables, these political text indicators demonstrate distinct explanatory power, further validated through machine learning techniques and robustness checks across multiple time periods.

Methodologically, I develop an innovative text-event alignment strategy that links unstructured political discourse to policy event dates, improving temporal

precision and strengthening the reliability of my conclusions. Together, these contributions highlight the crucial role of political communication in driving clean energy investment, offering valuable insights for both researchers and policymakers.

1.4 Structure

This thesis is organized into five chapters. Chapter 1 (Integrative Introduction) defines the overarching research gap through the concept of a multifactor informational architecture of clean energy markets, positioning each chapter as one analytical layer within this framework.

Chapter 2 (Empirical Identification) establishes which factors matter by testing the influence of both traditional fundamentals and fear-related informational factors based on the GDELT database. It finds that traditional drivers such as volatility and the US Dollar Index show limited robustness, while fear factors—such as media sentiment, exert consistent and significant effects.

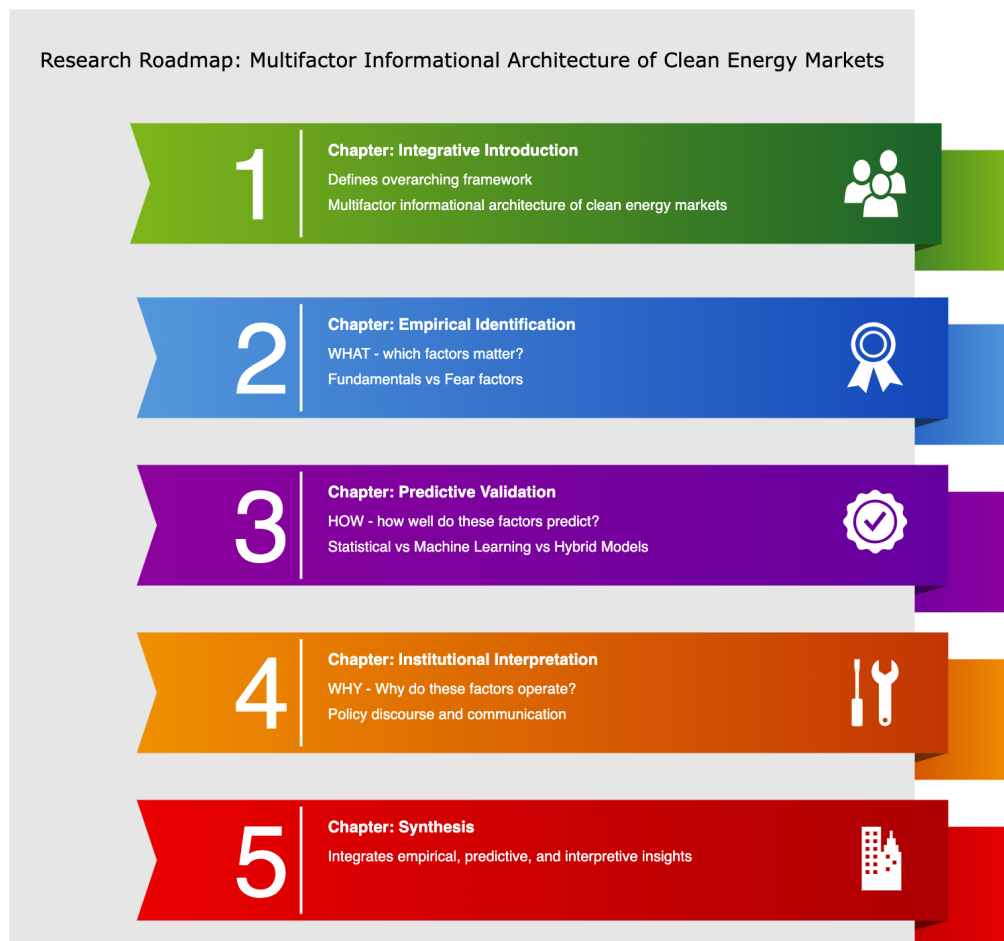
Chapter 3 (Predictive Validation) examines how well those factors predict future market movements. It integrates fundamentals and fear variables using feature selection methods like SHAP value and XGBoost, and evaluates model performance through statistical and machine learning frameworks (ARIMA, VAR, LSTM, SAE, and hybrid models).

Chapter 4 (Institutional Interpretation) explores why those factors operate by analyzing policy discourse and communication channels. It studies how the timing and tone of policymakers' statements affect market confidence, showing that credible and urgent policy commitments accelerate investor response and clean energy transition.

Chapter 5 (Synthesis) links empirical, predictive, and interpretive findings, highlighting theoretical and practical implications and proposing future research directions.

The structure of this thesis is shown in Figure 1.1.

Figure 1.1: Research Roadmap



1.5 Publications from this Doctoral Thesis

Large parts of Chapter 2 are published in *Energy Economics* as:

Zheng, Y. and Lucey, B. (2025c).

What really drives clean energy stocks – fear or fundamentals?

Energy Economics, 148:108558

Large parts of Chapter 3 are published in *Finance Research Open* as:

Zheng, Y. and Lucey, B. (2025b).

Supervised learning models, statistical models or hybrid models? a prediction of clean energy stock based on fear and fundamental factors.

Finance Research Open, 1(2):100006

Large parts of Chapter 4 are published in *Research in International Business and Finance* as:

Zheng, Y. and Lucey, B. (2025a).

Central bankers' political discourse as a driver of clean energy markets.

Research in International Business and Finance, 80:103126

1.6 Variable Definitions

CRI: Climate Risk Indexes. Data Source: Economic Policy Uncertainty

TRI: Transition Risk Index. Data Source: Economic Policy Uncertainty

PRI: Physical Risk Index. Data Source: Economic Policy Uncertainty

GBI: Green Bond Index. Data Source: Bloomberg

Oil_Close: Crude Oil WTI Futures. Data Source: Bloomberg

VIX_Close: CBOE Volatility Index. Data Source: Bloomberg

GFSI: Global Financial Stress Indicator. Data Source: Bloomberg

CHN_market: China Shanghai Composite Stock Market Index. Data Source: Bloomberg

US_market: S&P 500 Index. Data Source: Bloomberg

SP_IT: S&P 500 Information Technology (sector) Index. Data Source: Bloomberg

DXY: US Dollar Index. Data Source: Bloomberg

USD_CNY: United States Dollar / Chinese Renminbi Cross. Data Source: Bloomberg

MSCIEF: MSCI Emerging Markets. Data Source: Bloomberg

SP_GCEI: S&P Global Clean Energy Index. Data Source: Bloomberg

RFE: Recursive Feature Elimination.

GDELT: Global Database of Events, Language, and Tone.

GKG: GDELT 2.0 Global Knowledge Graph.

CSV: Comma-Separated Values.

EBA: Extreme Bounds Analysis, an econometric robustness check used to assess the sensitivity of coefficient estimates to alternative model specifications.

ReportCount_env: the daily count of environmental-themed news reports. Data Source: GDELT database; author's calculations

ReportCount_whole: the daily count of all news reports. Data Source: GDELT database; author's calculations

average_goldstein_scale: the daily average Goldstein Scale. Data Source: GDELT database; author's calculations

average_confidence: the daily average confidence of news reports. Data Source: GDELT database; author's calculations

AverageTone: the daily average tone of environmental reports. Data Source: GDELT database; author's calculations

ENV.number: the ratio of *ReportCount_env* to *ReportCount_whole*. Data Source: GDELT database; author's calculations

VIF: variance inflation factor.

LRI: likelihood ratio index.

EMH: Efficient Market Hypothesis.

SVM: Support Vector Machines, a class of margin-based supervised learning models.

ANN: Artificial Neural Networks, a class of nonlinear supervised learning models.

RF: Random Forest, a class of ensemble-based supervised learning models.

GBDT: Gradient Boosting Decision Trees, an ensemble-based supervised learning model that builds decision trees sequentially using gradient boosting.

XGBoost: Extreme Gradient Boosting, a boosting-based ensemble learning model that extends gradient boosting with regularization.

LightGBM: Light Gradient Boosting Machine, a class of gradient boosting-based ensemble learning models designed for computational efficiency.

CatBoost: Categorical Boosting, a gradient boosting-based ensemble learning model designed to handle categorical features.

KNN: K-Nearest Neighbors, a distance-based supervised learning method.

ARIMA: AutoRegressive Integrated Moving Average, a class of statistical, linear time-series models.

LSTM: Long Short-Term Memory, a class of supervised neural network models for sequence data.

LSTM-RNN: Long Short-Term Memory Recurrent Neural Network, a class of supervised learning models for sequential data.

RNN: Recurrent Neural Networks, a class of supervised neural network models for sequential data.

AR: AutoRegressive, a class of statistical, linear time-series models.

VAR: Vector Autoregression, a class of statistical, linear multivariate time-series models.

SAE: Supervised Autoencoder, a class of supervised neural network models; *LSTM_Autoencoder* used in this thesis follows the same framework.

PCA: Principal Component Analysis, an unsupervised dimensionality reduction method.

LSTM_Autoregressive: Autoregressive Long Short-Term Memory, a class of hybrid models combining neural networks with autoregressive structures.

Prophet Features: Features extracted from the Prophet model, representing hybrid time-series components based on trend, seasonality, and holiday effects.

GBRT: Gradient Boosting Regression Trees, a gradient boosting-based ensemble learning model for regression.

DNN: Deep Neural Network, a class of supervised neural network models with multiple hidden layers.

RMSE: Root Mean Squared Error, a standard metric used to evaluate prediction accuracy.

MSE: Mean Squared Error, a standard loss-based metric measuring the average squared prediction error.

MAE: Mean Absolute Error, a standard loss-based metric measuring the average absolute prediction error.

SHAP: SHapley Additive exPlanations, a model-agnostic method for interpreting

machine learning predictions.

NLP: Natural Language Processing.

CBDCs: Central Bank Digital Currencies.

IMF: International Monetary Fund.

UA: Urgency Analysis.

BIS: Bank for International Settlements.

ISO: International Organization for Standardization codes used to identify countries and geographic units.

FIPS: Federal Information Processing Standards codes used to identify countries, U.S. states, and counties.

CPU: Central Processing Unit.

GPU: Graphics Processing Unit.

WHAT *REALLY* DRIVES CLEAN ENERGY STOCKS - FEAR OR FUNDAMENTALS?

Abstract

This study investigates the relationship between the S&P Global Clean Energy Index and novel factors based on the GDELT Database, such as global news confidence levels, environmental sentiment, media coverage preferences in the US and China, and the ratio of environmental to overall reporting. I identify variables sourced from recent literature. Using the Isolation Forest method to select potential explanatory variables, Extreme Bounds Analysis reveals that "Fear" factors such as media sentiment and confidence show consistent and significant correlations with the S&P Global Clean Energy Index. These findings highlight the influential role of media sentiment in driving market confidence and industry growth. In contrast, some traditionally popular "Fundamental" factors, such as the Global Financial Stress Indicator, Green Bond Index, and US Dollar Index, lack robustness. While they appear reliable under normal distribution models, they exhibit substantial uncertainty under alternative models, limiting their explanatory power.

2.1 Introduction

Concerns about fossil fuels and their role in climate change have intensified global efforts to find cleaner alternatives. Advances in technology have accelerated the adoption of renewable energy sources like solar and wind power, valued for their low environmental impact and sustainability. This transition has not only brought increased attention to the clean energy industry but also prompted governments and policymakers to implement eco-friendly agreements and legislation aimed at reducing greenhouse gas emissions and addressing climate change. This movement toward renewable energy has spurred the rapid growth of the clean energy sector, making it highly appealing to environmentally responsible investors (Bilgen et al., 2004; Obama, 2017; Liang et al., 2019; Pham, 2019; Mangla et al., 2020; Dawar et al., 2021; Farouq et al., 2021). The clean energy industry is expected to experience rapid expansion in the coming years. This growth is anticipated to be driven by government financial incentives, advancements in clean energy technologies, and the increasing consumer demand for environmentally friendly and sustainable goods and services. Furthermore, the projected increase in environmental, social, and governance (ESG) investments is expected to further accelerate the growth of the clean energy industry (Dias et al., 2023). Su et al. (2024) state that researchers have gradually increased their focus on clean energy in financial studies, reflecting the close alignment of this field with the Sustainable Development Goals.

The rapid growth of the clean energy sector raises important questions about its connection to existing areas, particularly the factors that influence its financial performance. Ciner et al. (2023) highlights that returns on clean energy stocks are more significantly affected by market variables associated with uncertainty and financial distress, such as small-cap stocks and emerging markets, rather than traditional factors like oil prices. Building on this, the present study seeks to explore not only a broad range of traditional factors identified in the literature, such as macroeconomic conditions, policy uncertainty, technological innovation, market volatility, global economic conditions, oil prices, green stocks, and emerging market equities, but also novel factors drawn from the GDELT database. By employing Extreme Bounds Analysis (EBA), this research aims to identify both traditional and novel variables that have a significant and robust impact on the S&P clean energy index. The findings will contribute to a deeper understanding of the complex dynamics shaping the clean energy sector and its broader economic

and financial implications. Another highlight of this study is the use of Isolation Forests, a method commonly used for anomaly detection, to help identify key drivers of the clean energy market. The application of this innovative approach not only extends beyond traditional analytical tools but also provides a novel perspective for uncovering significant variables related to the clean energy index, enhancing the scientific rigor and robustness of the research findings (Xu et al., 2021). Accordingly, this chapter focuses on addressing the following two core research questions. First, are traditional explanatory factors robust across the various regression specifications employed in the Extreme Bounds Analysis (EBA)? Second, are media- and sentiment-related factors robust across the same set of regression models within the EBA framework?

2.2 Literature Review

2.2.1 Previous Studies on Media, Sentiment and Clean Energy Transition

El Ghouli et al. (2019) show strong evidence that the media plays a crucial role in corporate social responsibility by influencing reputational capital. Through its reporting of irresponsible corporate behavior, the media directly impacts a company's perceived value. The study found a positive correlation between corporate social responsibility involvement and press freedom. Media coverage plays an important role in the energy transition by helping people understand the big trends shaping energy systems. It can influence public and government agendas while also being shaped by them. At the niche level, media serves as a conduit or filter for information, encouraging or discouraging actors from adopting innovative technologies and practices. Research has shown that the future of niche-level energy technologies, such as biogas, can be significantly shaped by information flows, public perceptions, and expectations, with media coverage being a key contributing factor. In Finland, newspaper coverage of the energy transition has been mixed, showing both supportive and critical views (Lyytimäki et al., 2018). Shang et al. (2023) uses predictive simulations and empirical results to show that sentiment indicators are more suitable than technical indicators for forecasting WTI spot prices using machine learning methods. Interestingly, they also found that sentiment indicators provide better predictive performance than using lagged crude oil prices. Therefore,

the *sentiment* conveyed in media reports can reasonably be expected to play a critical role in shaping corporate social responsibility and influencing corporate value. While every country has its media outlets, the variations in reporting styles across nations result in differences in objectivity, which is reflected in the confidence level of the reports. These differences in media confidence can significantly impact how information about clean energy companies is perceived and, in turn, influence decision-making processes in both policy formulation and investment.

In recent years, there has been a growing global emphasis on environmental protection, with a significant focus on the transition to clean energy. The adoption of cleaner energy sources, such as solar power, biomass, and wind energy, is seen as a key strategy for reducing fossil fuel emissions and mitigating environmental degradation. This approach not only addresses the pressing necessity for sustainability but also aligns with international efforts to meet environmental targets and combat climate change. In particular, in 2022, with the outbreak of the Ukraine war and a reduction in natural gas supply, gas prices surged rapidly, highlighting the increasing scarcity of energy resources. As a result, economies worldwide intensified their investments in clean energy (Dias et al., 2023).

The primary source of information on clean energy for various stakeholders, including governments, investors, and consumers, is often social media platforms. Despite the growing influence of media in shaping public perceptions, there is limited academic research connecting media and sentiment factors with the proxy for clean energy here, the S&P Global Clean Energy Index. Previous studies, such as the qualitative analysis by Roche et al. (2023), focused on assessing and categorizing the social impact of 15 key resources in the clean energy transition through media analysis. This prior work has provided inspiration and direction for the research presented in this study.

2.2.2 Media Preferences in Global Key Clean Energy Leaders

The world faces a significant challenge in turning the goal of achieving net-zero emissions by 2050 from a narrow possibility into a practical reality. As economies recover from the impacts of the pandemic, global carbon dioxide emissions have surged again. This makes it crucial for governments to take decisive action and accelerate the shift to clean energy (International Energy Agency, 2021). According

to S&P Dow Jones Indices (2024), the contribution rates of various countries to the Clean Energy Index were highlighted, with the USA contributing 37%, ranking first globally, and China contributing 8.3%, ranking third globally.

As the world's two largest economies and major contributors to the Clean Energy Index, the development of these two countries will have a substantial influence on the global trends in the clean energy industry. Therefore, studying media coverage in both the U.S. and China becomes increasingly valuable. Additionally, over the past five years, the U.S.-China trade war has resulted in significant differences in the media reporting preferences between the two nations. Hollihan et al. (2021) examine how the media in the U.S. and China portrayed and discussed trade war. The book explores the different narratives and arguments presented by each country's media, revealing how each side's political and economic interests shaped their coverage, it highlights the role of media in shaping public opinion and the broader implications of these divergent narratives for international relations and trade policies. Similarly, Awad AlAfnan (2020) show that media bias operates through three key mechanisms: gatekeeping bias, coverage bias, and statement bias. Ideological bias is reflected in the choice of topics covered and the tone used to present them. Power dynamics are expressed through how information is framed and contextualized. This comparative study also highlights that topics emphasized in U.S. newspapers are often reinterpreted or filtered by Chinese newspapers, while topics contextualized by Chinese newspapers tend to be spotlighted in U.S. coverage.

This is the primary reason for adopting the S&P Global Clean Energy Index in this study. The index is widely used in both investment practice and academic research as a benchmark for clean energy market performance (S&P Dow Jones Indices, 2024; Broadstock et al., 2012). In this thesis, the clean energy industry is defined as comprising firms whose core business activities are directly related to the production, development, or deployment of renewable and low-carbon energy technologies, including solar, wind, hydro, energy storage, and related enabling technologies, consistent with the existing literature (Henriques and Sadorsky, 2008). The S&P Global Clean Energy Index is compiled using a transparent and rules-based methodology, with constituent firms selected based on their primary exposure to clean energy activities, as well as market capitalization and liquidity criteria. The index is periodically reviewed and rebalanced to ensure its continued representativeness of the global clean energy sector (S&P Dow Jones Indices, 2024).

Besides, firms from China and the United States account for a substantial share of the index constituents, reflecting the dominant role of these two markets in the global clean energy industry. These features make the index a suitable and widely recognized benchmark for the empirical analysis conducted in this chapter.

The level of positivity in media coverage from China and the United States, specifically the sentiment of reports from both countries, will be measured using the Goldstein Scale as derived from the GDELT database. A detailed explanation of the GDELT database and the Goldstein Scale, including their relevance for capturing media sentiment, will be provided in *Data Collection*.

2.2.3 Traditional Factors Influencing the S&P Clean Energy Index

Khalfaoui et al. (2022) argue that clean energy, climate policy uncertainty, and innovation in renewable energy are key drivers of risk spillover effects, particularly under extreme market conditions such as booms and busts. Additionally, uncertainty indices and global economic conditions are significant predictors of volatility in both clean energy and natural gas markets, with real economic activity having a greater impact on volatility than text-based uncertainty indicators, especially in the energy sector. This highlights the critical importance of global economic conditions for policymakers and investors (Wang et al., 2022).

Therefore, to effectively identify the traditional factors influencing the clean energy index from the literature, this study conducted a search on Scopus for highly cited and recently published papers. From these sources, quantifiable factors related to the clean energy index were extracted. These factors will serve as explanatory variables in the subsequent EBA analyses. The variables, along with their sources, have been organized into a table, as shown in Table 2.1.

CHAPTER 2. WHAT *REALLY* DRIVES CLEAN ENERGY STOCKS - FEAR OR FUNDAMENTALS?

Table 2.1: Traditional Influencing Factors and Source

Variable	Index	Source	Study
Economic / Climate Policy Uncertainty	Climate Risk Indexes (CRI), includes Transition Risk Index (TRI) and Physical Risk Index (PRI)	Economic Policy Uncertainty	(Ren et al., 2023; Zhang et al., 2023a; Khalfaoui et al., 2022)
Green Bond Stock	Green Bond Index (GBI)	Bloomberg	(Ren et al., 2023; Saeed et al., 2021)
Oil Price	Crude Oil WTI Futures (Oil_Close)	Bloomberg	(Ren et al., 2023; Zhang et al., 2023a; Saeed et al., 2021)
Market Volatility	CBOE Volatility Index (VIX_Close)	Bloomberg	(Wang et al., 2022; Saeed et al., 2021)
Global Economic Conditions	Global Financial Stress Indicator (GFSI), China Shanghai Composite Stock Market Index (CHN_market), S&P 500 Index (US_market)	Bloomberg	(Wang et al., 2022)
Technological Development	S&P 500 Information Technology (sector) Index (SP_IT)	Bloomberg	(Ghosh and Jana, 2024; Ciner et al., 2023)
Macroeconomic Factors	US Dollar Index (DXY), United States Dollar / Chinese Renminbi Cross (USD_CNY)	Bloomberg	(Saeed et al., 2021)
Emerging Market Equities	MSCI Emerging Markets (MSCIEF)	Bloomberg	(Ciner et al., 2023)

Note: The Transition Risk Index (TRI) and Physical Risk Index (PRI) are text-based climate risk measures developed by Bua et al. (2024). The indices are constructed using relevance-ranked

climate risk vocabularies and Reuters News articles from January 2005 onward. TRI captures innovations in climate transition risk related to policy, regulatory, and technological changes, while PRI captures innovations in physical climate risk associated with climate-related physical events. Both indices are updated on a semi-annual basis (Economic Policy Uncertainty).

The first column of Table 2.1, "Variable," lists the factors influencing the Clean Energy Index, which have been compiled from the fourth column, "Study". The "Study" column represents the papers selected from Scopus and recent literature, and each "Variable" is associated with its corresponding source(s). Some "Variables" are linked to multiple studies, indicating that these factors were mentioned in more than one paper. The second column, "Index", contains the quantitative indicators used to measure each "Variable". All indicators are updated daily, with monthly or annual data excluded to maintain consistency with the data from the GDELT database, as mentioned in the *Data Collection* section. Sometimes, a single "Variable" is associated with multiple indicators, as using several indicators can provide a more robust analysis. Global economic conditions and macroeconomic factors are represented by stock market indicators from both the U.S. and China, reflecting the two countries' significant contributions to the clean energy transition. These same countries were also used in calculating the Goldstein scale from the GDELT database in the *Data Collection* process. Except for the CRI, all other Indexes are sourced from Bloomberg, with the source for the CRI specified in Table 2.1.

The existing literature has largely emphasized qualitative investigations of media effects in clean energy markets, as well as the influence of traditional factors. This chapter addresses this gap by providing a quantitative analysis of emerging media- and sentiment-related factors, with a particular focus on the Chinese and U.S. clean energy markets, which play a dominant role in the global clean energy transition.

2.3 Methodology

2.3.1 Validating the Reliability of Factors—Isolation Forest

To validate the reliability of the summarized variables, I combined the previous study with the literature found by searching "factors influencing the clean energy index" on Google Scholar and placed all the relevant papers into a folder. I then iterated through all the files in this folder to extract the text. After processing the

data by removing numbers and punctuation, I extracted the top 100 most important keywords.

Building on the previously identified keywords, I expanded the original keyword list from Table 2.1. After obtaining the new keyword variables, I counted their frequencies in the literature. This data will serve as the raw training set for further analysis. I opted for the Isolation Forest method from the ensemble package in sklearn due to the absence of a defined "y" value. Since it is unclear whether these factors significantly influence the clean energy index, this approach allows us to detect patterns and anomalies without relying on predefined outcomes.

The Isolation Forest algorithm was chosen because it is specifically designed for anomaly detection by isolating outliers rather than modeling the distribution of data. It achieves this by constructing random partitions to isolate anomalous points more quickly than normal points (Kareem, 2024). As demonstrated by Salmimaa (2024), combining semantic feature extraction with the Isolation Forest method effectively enhances the accuracy of outlier detection by leveraging the semantic characteristics of the data. This approach enables the model to more precisely define the baseline class and identify observations that deviate significantly from typical patterns. Therefore, Isolation Forest significantly improves both the accuracy and interpretability of anomaly detection at the semantic level. This provides theoretical support for detecting anomalies in keyword semantic frequencies in this study. Bagi (2022) use the Isolation Forest method to detect abnormal trading behavior in the Bitcoin market based on volume and price, providing technical support for its application in this study. Extending this approach to the field of green finance represents a methodological innovation. Since I aim to detect unusual patterns in keyword frequency that may indicate highly significant or irrelevant factors, the strengths of Isolation Forest make it particularly well-suited to my dataset.

After multiple iterations, I identified the optimal model parameters: Contamination (0.05): This parameter specifies the proportion of the dataset expected to be anomalies. The value of 0.05 was selected to account for the fact that the majority of keywords are expected to be relevant while flagging a small portion (5%) as potential outliers. This aligns with the objective of identifying exceptional keywords without overwhelming the results with false positives. Max_samples (0.6): This parameter determines the fraction of the data used to build each decision tree. A value of 0.6 was chosen to ensure that the model can capture the underlying variability in the dataset without overfitting. This smaller subset size provides a

robust estimate of the feature importance and anomaly scores while maintaining computational efficiency. **N_estimators (50)**: This parameter sets the number of trees in the forest. A value of 50 was selected to balance performance and computational cost. The ensemble of 50 trees ensures that the model is sufficiently robust and capable of capturing subtle patterns without introducing unnecessary complexity. These settings enhance the model’s sensitivity to anomalies while minimizing the risk of overfitting, ensuring it doesn’t become dependent on specific data points. The parameters were chosen after iterative experimentation, aiming to achieve a balance between precision and recall in detecting anomalies.

After setting up the model, I will use the frequency of these keywords as input features for training. The model will then calculate the importance of each feature and output the results, similar to those shown in Table 2.2. In this context, frequency refers to how many times a keyword appears in the text, **Anomaly_Score** is the outlier score assigned to each keyword by the Isolation Forest model, and **Feature_Importance** is the absolute value of the **Anomaly_Score**, which reflects the importance of the feature.

Table 2.2: Isolation Forest Results

Keyword	Frequency	Anomaly_Score	Feature_Importance
Green Bond	297	-1	1
Oil Price	216	1	1
Renewable Energy	170	1	1
Technology	129	1	1
Climate Risk	64	1	1
Climate Policy Uncertainty	58	1	1
Global Economic Conditions	42	1	1
Macroeconomic	42	1	1
Economic Policy Uncertainty	23	1	1
Market Volatility	23	1	1
Clean Energy Investments	15	1	1
Emerging Market Equities	6	1	1
Financial Risk	6	1	1

Green Bond was flagged as an anomaly due to its **Anomaly_Score** of -1, but this was because its frequency was exceptionally high. Given its significance, it should still be considered one of the most relevant variables. Guyon et al. (2002) suggest that the input corresponding to the smallest difference in feature importance

should be removed. This procedure can be iterated to perform Recursive Feature Elimination (RFE), whereby the model gradually removes the least important features based on their importance scores, ultimately retaining only the most relevant variables. This is largely consistent with the results of anomaly detection. By integrating these findings with the anomaly detection outcomes, I obtained the final set of key variables.

The final set of key variables includes Green Bond, Oil Price, Climate Policy Uncertainty, Global Economic Conditions, Macroeconomic, Economic Policy Uncertainty, Market Volatility, Emerging Market Equities, Financial Risk, Clean Energy Investments, Renewable Energy, Climate Risk and Technology.

These variables are largely consistent with those listed in Table 2.1. Additionally, GFSI measures financial risk, CRI serves as a proxy for climate risk, and renewable energy and clean energy investments closely align with the clean energy index, which is the dependent variable of this study.

2.3.2 Data Collection

The Global Database of Events, Language, and Tone (GDELT) is a comprehensive resource that tracks global news coverage, analysing language and sentiment across a vast range of events worldwide. The most comprehensive and up-to-date datasets utilized in this chapter, encompassing the GKG, Mention, and Event tables, have been continuously available since 2015. The specific names of these datasets are the GDELT 2.0 Event Database, the GDELT 2.0 Global Knowledge Graph (GKG), and the GDELT 2.0 Mention Table. The Global Database of Events, Language, and Tone (GDELT) is compiled and maintained by the GDELT Project, which was created and is led by Kalev H. Leetaru. According to the official GDELT documentation, Leetaru has over 25 years of experience studying large-scale digital information systems and the evolution of the web, with a particular focus on understanding how global information flows reflect and shape societal dynamics. During his tenure at the National Center for Supercomputing Applications, he co-developed one of the earliest web-scale data mining platforms designed to monitor and analyze global open information streams, including early forms of online and social media. Building on this foundation, the GDELT Project applies large-scale automated text analysis and natural language processing techniques to global news media in more than 100 languages, systematically extracting information on events, actors, locations, and sentiment in near real time (The GDELT Project, 2022).

GDELT covers news in over 100 languages, spanning print, broadcast, and online media, with data available from January 1, 1979. Television news data collection began in July 2009, encompassing reports from over 160 U.S. national and local TV stations, as well as several international broadcasters (Ahmad et al., 2024). As an open global news platform, GDELT is updated daily and includes information on actors, locations, organizations, topics, and sources reported in the news. It identifies thousands of sentiments, organizations, places, quantities, news sources, global events, and analyses the average tone of news articles (Hartvig et al., 2023; Shen et al., 2022). GDELT has been widely used in academic research across economics and finance to study media coverage, sentiment, and their economic implications. In particular, prior studies have employed GDELT-based news data to analyze geopolitical risk and financial market volatility (Leetaru and Schrodtt, 2013; Nyman et al., 2021). Given its comprehensive global coverage, high-frequency structure, and established use in the literature, the GDELT database provides a suitable and well-recognized data source for measuring media attention and sentiment in this thesis.

According to the The GDELT Project (2022), each dataset is updated at 15-minute intervals. This study draws on data covering the past five years, from 00:00 on June 30, 2019, to 00:00 on June 30, 2024. Given the sheer volume of data, it is not feasible to perform computations, including filtering and merging operations, on the entire dataset locally, as local memory resources are inadequate for these tasks. To address this limitation, I looked into leveraging Microsoft Azure’s cloud computing services to manage and process the data efficiently. The need for efficient distributed processing to handle the complexity and size of the dataset made leveraging Microsoft Azure’s cloud computing services essential. Azure provides scalable computing resources, enabling the management of large-scale data operations and ensuring seamless processing across multiple nodes, which would be impractical using local infrastructure. Additionally, Azure Storage Accounts were utilized to securely store and manage large volumes of data. They provide high availability, durability, and scalability, ensuring efficient access to data for distributed processing tasks (Microsoft, 2025b,a). As for Spark and Databricks, Apache Spark was utilized for its capability to efficiently process large-scale datasets in a distributed manner. Its in-memory computation capabilities enabled faster execution of iterative tasks such as data filtering, transformation, and aggregation, all of which were critical for this study (Iomete, 2025). To enhance productivity and streamline

operations, Databricks, a cloud-based platform optimized for Spark, was chosen. Databricks simplifies cluster management, offers seamless scalability, and integrates effectively with Azure’s storage and analytics services. This combination of tools provided a robust framework for scalable data processing and ensured the workflow could handle the substantial computational demands of analysing five years of GDELT data (Consultancy, 2025).

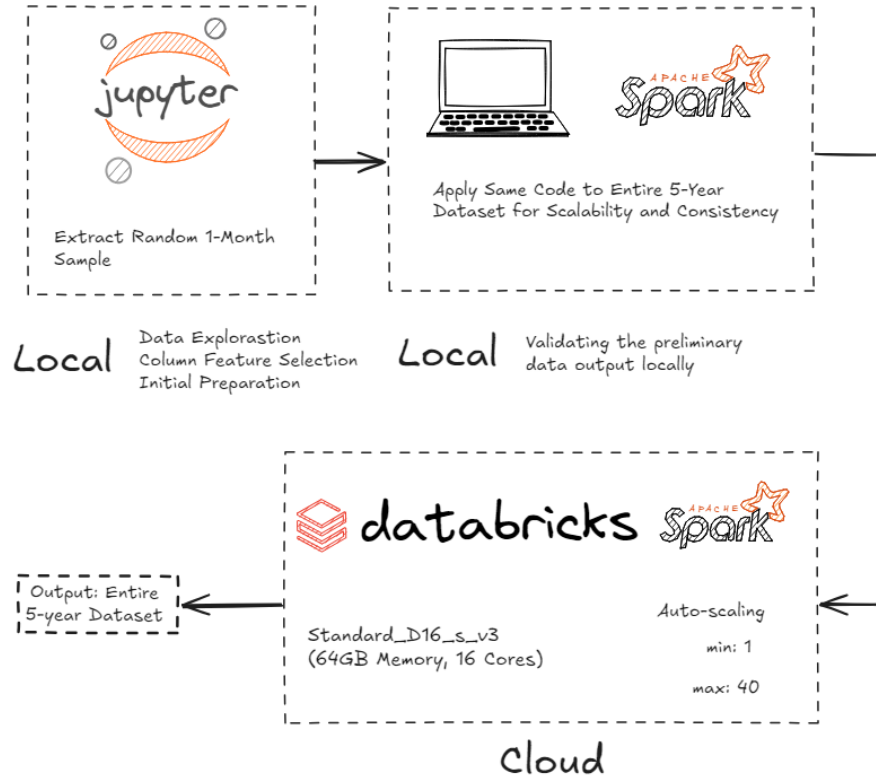
The data collection process followed these steps: 1. A random one-month sample data was extracted from the full five-year dataset to serve as the basis for preliminary processing and validation. 2. Using a local one-node Spark environment, the dataset (in CSV format) extracted from step 1 was explored and underwent pre-processing. This includes data exploration, column feature selection, and initial preparation for subsequent model construction. The pre-processed data was then used for applying EBA analyses. 3. After validating the preliminary data output locally, the full five-year dataset was processed using a dedicated Databricks cluster on Azure (Standard_D16_s_v3, 64GB Memory, 16 Cores, with auto-scaling enabled from 1 to 40 nodes). This leveraged the power of distributed computing to extract the complete dataset, which was then stored as CSV files in the Azure storage account, organized in a yyyy/mm/dd hierarchy. 4. The same code was applied to the entire five-year dataset, ensuring scalability and consistency in the analysis. In order to make the data collection process more intuitive, this chapter provides the flowchart shown in Figure 2.1.

2.3.2.1 GDELT Variables Source

In this study, political sentiment or tone of international events was assessed using a factor derived from the Goldstein scale (Leetaru, 2015), which was computed for China and the United States—two key nations with substantial contributions to the global clean energy index. The data was sourced from the GDELT 2.0 Event Database by filtering news reports concerning interactions between China and the U.S. over the past five years. The daily average of the Goldstein scale was then calculated to quantify the political sentiment or tone of international events for both countries. This offers a more macro-level perspective on assessing the positivity of media coverage.

To evaluate the credibility of media coverage (Leetaru, 2015), the Confidence metric was extracted from the GDELT 2.0 Mention Table. Average daily sentiment values over the five years were calculated by filtering for environment-related

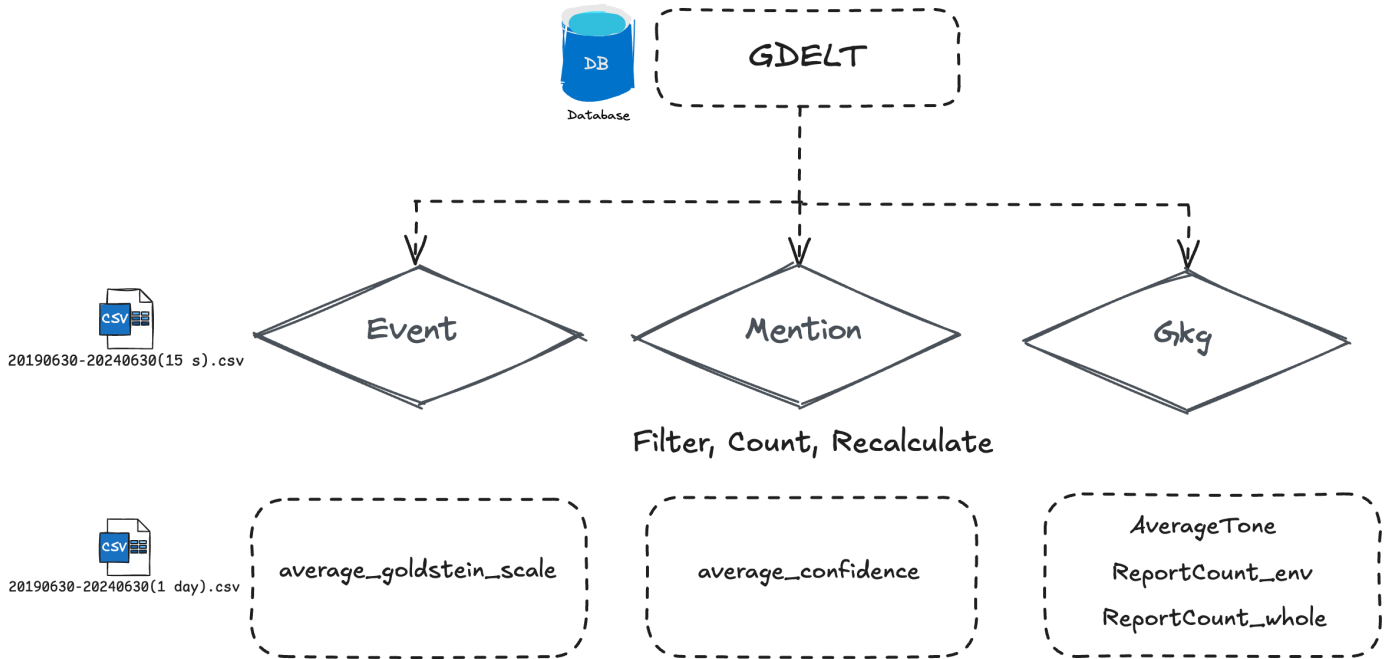
Figure 2.1: GDELT Data Collection Flow Chart



topics from GDELT GLOBAL KNOWLEDGE GRAPH 2.1. Specifically, entries in the "V1THEME" column containing "ENV" were selected, and the corresponding values in the "Tone" field of the "V1.5TONE" column were averaged. Additionally, the total number of daily news reports and the count of environment-related reports were tallied and exported using cloud-based resources for efficient processing. All relevant variables obtained from GDELT Database will be as shown in Figure 2.2.

The global clean energy index utilized in this study is the S&P Global Clean Energy Index(SP_GCEI). Along with the indices presented in Table 2.1, the data collection period spans from June 30, 2019, to June 30, 2024. However, the CRI index is currently updated only until December 29, 2023, and therefore, missing data has been temporarily denoted as "NaN". This small subset of missing values will be excluded during the analysis. Since they will not be classified as Focus variables in the EBA analyses, their exclusion is expected to have a negligible impact on the overall results. All data related to market transactions are represented using the close price.

Figure 2.2: GDELT Variables Source



2.3.3 EBA with All Combinations of Doubtful Variables

After data collection, a raw DataFrame and corresponding variables were obtained. This study firstly filters out significant variables with relatively concentrated coefficient values to serve as Focus variables. The current raw DataFrame contains the following variables: the S&P Global Clean Energy Index (SP_GCEI), the daily count of environmental-themed news reports (ReportCount_env), the daily count of all news reports (ReportCount_whole), the daily average Goldstein Scale (average_goldstein_scale), and the daily average confidence of news reports (average_confidence), Transition Risk Index (TRI), Physical Risk Index (PRI), Green Bond Index (GBI), Crude Oil WTI Futures (Oil_Close), CBOE Volatility Index (VIX_Close), Global Financial Stress Indicator (GFSI), China Shanghai Composite Stock Market Index (CHN_market), S&P 500 Index (US_market), S&P 500 Information Technology (sector) Index (SP_IT), US Dollar Index (DXY), United States Dollar / Chinese Renminbi Cross (USD_CNY), MSCI Emerging Markets (MSCIEF). To distinguish between the average "Tone" assigned to each news report in the GDELT database and the daily average tone of environmental reports calculated in this study, I will label the latter as "AverageTone". Additionally, to observe the media's emphasis on environmental reporting, this study introduces the variable

"ENV.number", which is derived from the ratio of ReportCount_env to ReportCount_whole.

This study treats the Isolation Forest and Extreme Bounds Analysis (EBA) as complementary rather than substitutive methods. Since the research focuses on correlation rather than causality, EBA is applied to test the robustness of relationships between variables (Hlavac, 2016), particularly between Fear and Fundamentals. It examines whether the estimated coefficients remain stable under different model specifications, making it well suited for robustness analysis. In contrast, the Isolation Forest is used for anomaly detection and feature validation, identifying outliers and ensuring that the variables included in the EBA are statistically reliable (Salmimaa, 2024). In this framework, the Isolation Forest serves as a pre-validation step, while EBA provides the coefficient evaluation. The variable Fear captures media-based information reflecting collective sentiment and uncertainty, like media sentiment, confidence level, and preferences in reporting. Together, these elements represent the informational dimension of Fear, which complements traditional Fundamentals such as oil prices and market volatility in explaining movements in the clean energy index.

Referring to the experimental steps of Hlavac (2016), I first conducted the following EBA analysis. Since the Focus variables were not confirmed, all variables were initially classified as doubtful variables. The mathematical expression constructed is shown as Equation 2.1:

$$(2.1) \quad Y = \beta_0 + \sum_{i=1}^n \beta_i X_i + \epsilon$$

Where:

- Y represents the dependent variable SP_GCEI;
- X_i represents the set of doubtful variables, which includes ReportCount_env, ReportCount_whole, ENV.number, average_goldstein_scale, average_confidence, AverageTone, TRI, PRI, GBI, Oil_Close, VIX_Close, GFSI, CHN_market, US_market, SP_IT, DXY, USD_CNY, MSCIEF;
- β_0 is the intercept term;
- β_i is the coefficient of the doubtful variables;
- ϵ is the error term.

For ease of interpretation, the final results are presented visually. The corresponding output is illustrated in Figure 2.3 and detailed in Table 2.3.

2.3.4 A More Sophisticated EBA Analysis

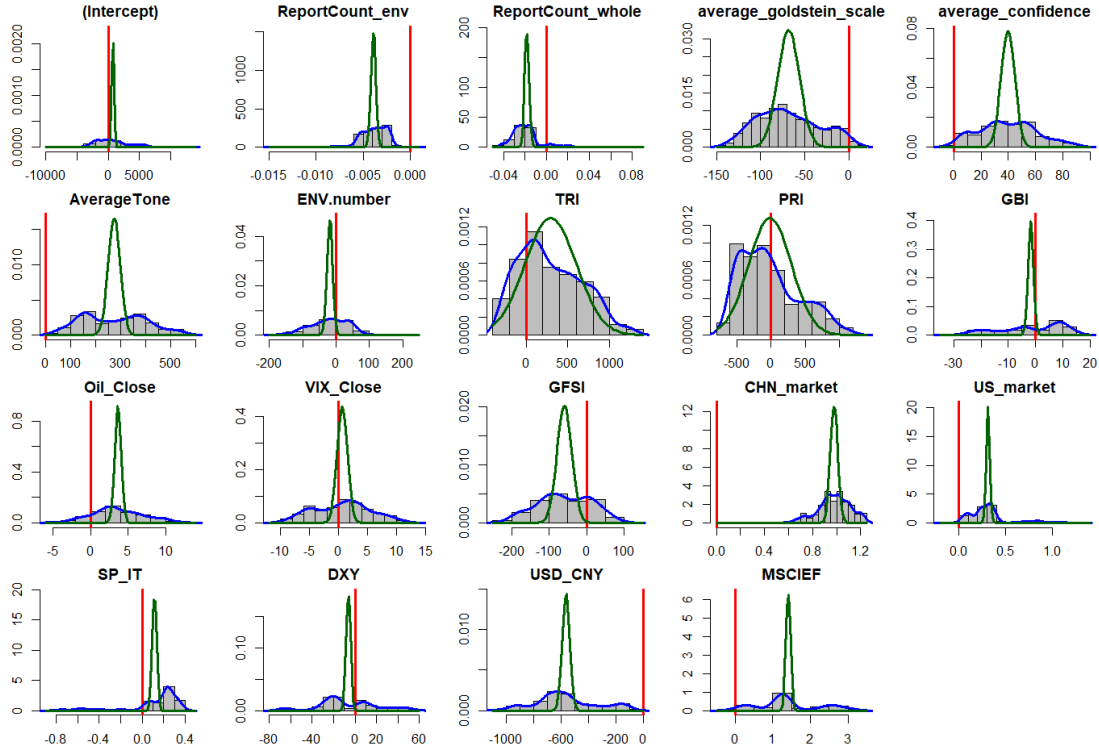
Building on the results of 2.3.3, this study proceeded to estimate Equation 2.2, incorporating the focus variables ENV.number, AverageTone, average_goldstein_scale, average_confidence, ReportCount_whole, and ReportCount_env. These variables, extracted from the GDELT database, represent the media and sentiment-related factors that I want to closely examine.

$$(2.2) \quad Y = \beta_0 + \sum_{i=1}^m \beta_i X_i + \sum_{j=1}^n \gamma_j Z_j + \sum_{k \in S_d} \delta_k W_k + \epsilon$$

Where:

- Y represents the dependent variable SP_GCEI;
- X_i are the free variables ReportCount_env, AverageTone, average_goldstein_scale, average_confidence, ReportCount_whole, Oil_Close, CHN_market, US_market, SP_IT, USD_CNY, MSCIEF;
- Z_j represents the set of focus variables which includes ENV.number, AverageTone, average_goldstein_scale, average_confidence, ReportCount_whole, ReportCount_env;
- W_k are the doubtful variables, which include GFSI, DXY, TRI, PRI, GBI, VIX_Close;
- β_i is the coefficient of the free variable;
- γ_j is the coefficient of the focus variables;
- δ_k represents the coefficient for the doubtful variable;
- β_0 is the intercept term;
- ϵ is the error term;
- S_d is the subset of doubtful variables selected (with a maximum of 3 doubtful variables).

Figure 2.3: All Combinations of Doubtful Variables



I assume that a maximum of 3 doubtful variables is allowed in the model ($k=3$). The final results are presented in Figure 2.4 and Table 2.4. To enhance the robustness of the model, adjustments were made to address potential heteroscedasticity, following the methodology of White (1980), by defining a custom heteroscedasticity-robust standard error function. A variance inflation factor (VIF) threshold of 7 was established to address potential multicollinearity. Additionally, likelihood ratio index (LRI) weighting was applied to improve model fit (McFadden, 1974).

All other conditions remain unchanged, the constructed mathematical expression remains consistent with Equation 2.2, and the experimental results are presented in Figure 2.5 and Table 2.5.

2.3.5 Discussions and Results

Table 2.3 presents the statistical results of the EBA analysis. The columns "N: CDF ($\beta \leq 0$)" and "N: CDF ($\beta > 0$)" assume that the regression model follows a normal distribution, representing the probability that the regression coefficient is either negative or positive under this assumption. A higher likelihood that the data

CHAPTER 2. WHAT *REALLY* DRIVES CLEAN ENERGY STOCKS - FEAR OR FUNDAMENTALS?

Table 2.3: Sala-i-Martin's EBA (All Combinations of Doubtful Variables)

Variable	N: CDF($\beta \leq 0$)	N: CDF($\beta > 0$)	G: CDF($\beta \leq 0$)	G: CDF($\beta > 0$)
(Intercept)	0.080	99.920	45.621	54.379
ReportCount_env	100.000	0.000	99.854	0.146
ReportCount_whole	100.000	0.000	89.145	10.855
average_goldstein_scale	100.000	0.000	96.472	3.528
average_confidence	0.000	100.000	1.822	98.178
AverageTone	0.000	100.000	0.172	99.828
ENV.number	97.198	2.802	60.285	39.715
TRI	17.505	82.495	30.340	69.660
PRI	51.805	48.195	55.615	44.385
GBI	93.894	6.106	51.235	48.765
Oil_Close	0.000	100.000	14.507	85.493
VIX_Close	24.141	75.859	42.401	57.599
GFSI	99.844	0.156	75.290	24.710
CHN_market	0.000	100.000	0.000	100.000
US_market	0.000	100.000	0.920	99.080
SP_IT	0.000	100.000	15.576	84.424
DXY	99.341	0.659	58.933	41.067
USD_CNY	100.000	0.000	99.967	0.033
MSCIEF	0.000	100.000	1.865	98.135

conforms to a normal distribution suggests greater robustness for the variable. The columns "G: CDF ($\beta \leq 0$)" and "G: CDF ($\beta > 0$)" represent the cumulative distribution of the regression coefficients being negative or positive under the Generic model, which makes no distributional assumptions. This model captures a broader range of scenarios. The same analytical framework is applied consistently across all subsequent histograms and tables.

Figure 2.3 presents overlaid histograms with density plots for all variables, summarizing the distribution of beta coefficients for each explanatory variable across different models. The red line represents the null hypothesis ($\beta = 0$). If the beta coefficients for a given variable consistently lie either to the left or right of the red line, the variable can be considered significant and robust. The green and blue lines illustrate the distribution of beta coefficients under different model assumptions, with greater deviation from the red line indicating a more robust effect. The green line represents the probability density function based on the assumption of a normal distribution, while the blue line shows the actual distribution of the variables.

The coefficients for ReportCount_env, ReportCount_whole, average_goldstein_scale, average_confidence, and AverageTone are consistently either positive or negative, both under the assumption of a normal distribution

and in the Generic model. This study focuses primarily on the actual distribution of these coefficients, while the normal distribution serves as a reference point to assess robustness based on the normality assumption. To determine statistical significance, I follow the guidelines outlined in Hlavac (2016) and consider variables to be reliable in the model if their coefficients have more than an 80% chance of being either positive or negative in the Generic model.

Similarly, the histogram distributions demonstrate that these variables tend to cluster either to the left or right of the null hypothesis line ($\beta = 0$), indicating strong robustness across all models. In contrast, traditional factors such as Oil_Close, CHN_market, US_market, SP_IT, USD_CNY, and MSCIEF exhibit similarly stable coefficient distributions, suggesting that these variables can be considered free variables in the subsequent EBA analysis.

For ENV.number, there is a 97.178% probability of a positive coefficient under the Normal distribution model, but this drops to 60.285% in the Generic model. Given its importance to media and sentiment variables, further investigation into its effects within the model is needed, and it will be treated as a Focus variable in the subsequent EBA analysis.

On the other hand, variables such as TRI, PRI, and VIX_Close show more inconsistent positive and negative effects on their regression coefficients across both the Normal distribution and Generic models. Recently, many studies have looked into how market volatility, climate policy uncertainty, and the energy market are connected, raising more doubts about these traditional factors. As a result, they will be classified as doubtful variables in the upcoming analysis. Similarly, GBI, GFSI, and DXY exhibit behavior similar to ENV.number, showing relative robustness under the Normal distribution model but more variability in the Generic model, which also justifies their classification as doubtful variables.

Figure 2.4 and Table 2.4 align with the results shown in Figure 2.3 and Table 2.3, further affirming the appropriateness of the designation of Focus and Free variables. Besides, the Focus variables exhibit a consistent and robust impact on the dependent variable, even when less stable traditional factors are introduced into the model combinations. Therefore, to further analyse the effects of the Focus variables, and consider the potential for high multicollinearity (Hlavac, 2016), this study moves forward with a more detailed and robust EBA analysis.

Figure 2.5 and Table 2.5 demonstrate that most coefficients for ENV.number are negative. The CDF values indicate a higher probability of negative coefficients

CHAPTER 2. WHAT *REALLY* DRIVES CLEAN ENERGY STOCKS - FEAR OR FUNDAMENTALS?

Figure 2.4: A More Sophisticated EBA Analysis

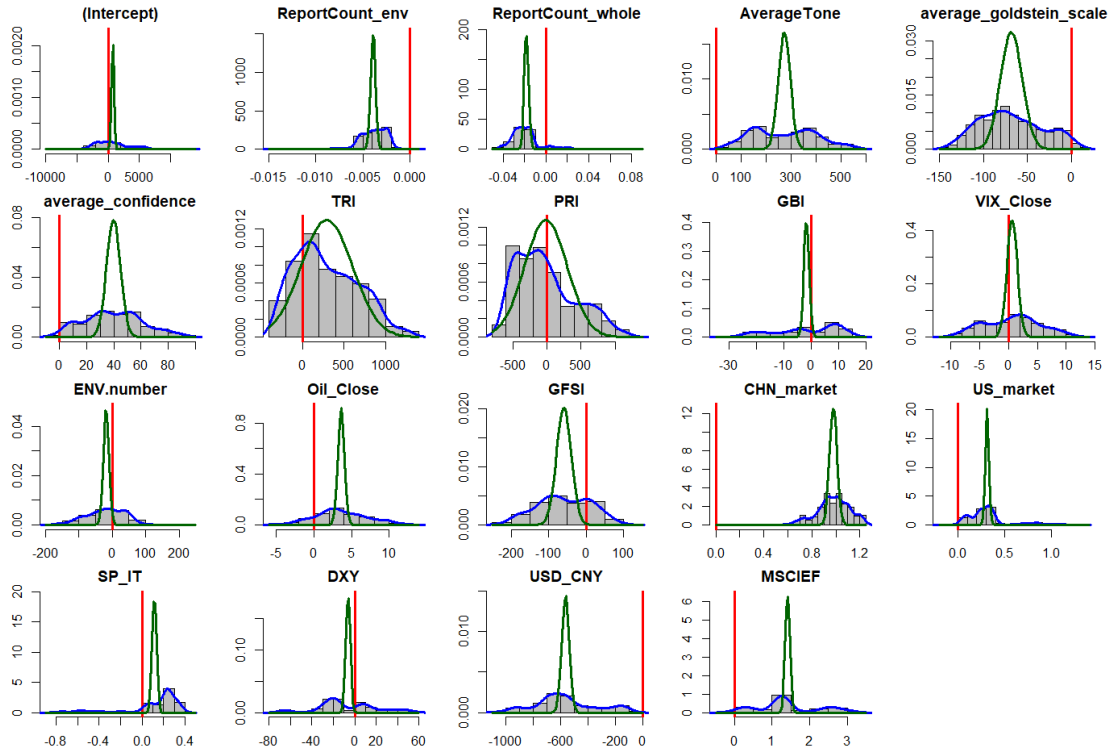


Table 2.4: Sala-i-Martin's EBA (More Sophisticated)

Variable	N: $\text{CDF}(\beta \leq 0)$	N: $\text{CDF}(\beta > 0)$	G: $\text{CDF}(\beta \leq 0)$	G: $\text{CDF}(\beta > 0)$
(Intercept)	0.080	99.920	45.621	54.379
ReportCount_env	100.000	0.000	99.854	0.146
ReportCount_whole	100.000	0.000	89.145	10.855
AverageTone	0.000	100.000	0.172	99.828
average_goldstein_scale	100.000	0.000	96.472	3.528
average_confidence	0.000	100.000	1.822	98.178
TRI	17.505	82.495	30.340	69.660
PRI	51.805	48.195	55.615	44.385
GBI	93.894	6.106	51.235	48.765
VIX_Close	24.141	75.859	42.401	57.599
ENV.number	97.198	2.802	60.285	39.715
Oil_Close	0.000	100.000	14.507	85.493
GFSI	99.844	0.156	75.290	24.710
CHN_market	0.000	100.000	0.000	100.000
US_market	0.000	100.000	0.920	99.080
SP_IT	0.000	100.000	15.576	84.424
DXY	99.341	0.659	58.933	41.067
USD_CNY	100.000	0.000	99.967	0.033
MSCIEF	0.000	100.000	1.865	98.135

Figure 2.5: A More Sophisticated EBA Analysis (Enhance the Robustness)

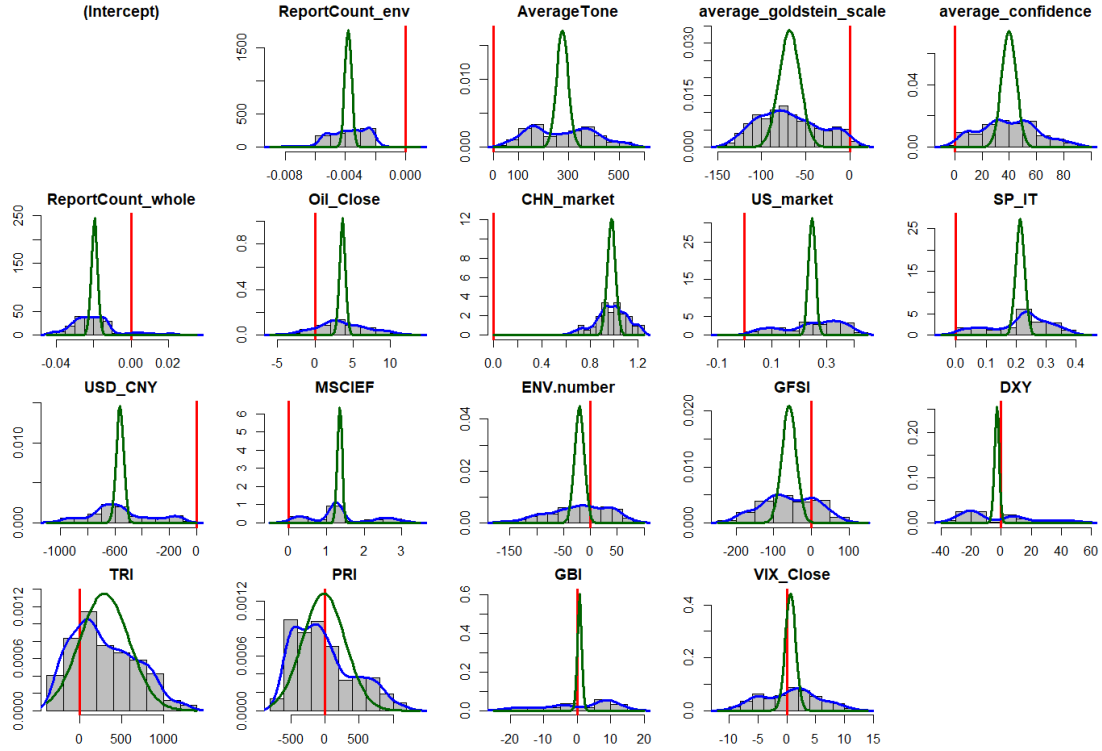


Table 2.5: Sala-i-Martin's EBA (Enhance the Robustness)

Variable	N: $\text{CDF}(\beta \leq 0)$	N: $\text{CDF}(\beta > 0)$	G: $\text{CDF}(\beta \leq 0)$	G: $\text{CDF}(\beta > 0)$
(Intercept)	3.911	96.089	53.644	46.356
ReportCount_env	100.000	0.000	99.856	0.144
AverageTone	0.000	100.000	0.112	99.888
average_goldstein_scale	100.000	0.000	94.021	5.979
average_confidence	0.000	100.000	2.999	97.001
ReportCount_whole	100.000	0.000	91.165	8.835
Oil_Close	0.000	100.000	15.712	84.288
CHN_market	0.000	100.000	0.000	100.000
US_market	0.000	100.000	0.883	99.117
SP_IT	0.000	100.000	2.113	97.887
USD_CNY	100.000	0.000	99.966	0.034
MSCIEF	0.000	100.000	2.334	97.666
ENV.number	93.480	6.520	57.374	42.626
GFSI	98.854	1.146	68.968	31.032
DXY	0.368	99.632	39.550	60.450
TRI	27.735	72.265	39.445	60.555
PRI	66.289	33.711	64.460	35.540
GBI	99.998	0.002	63.077	36.923
VIX_Close	1.442	98.558	31.250	68.750

under the Normal distribution model; however, the probability under the Generic model is only 57.374%, suggesting that this variable lacks robustness. In contrast, AverageTone and ReportCount_env exhibit a higher likelihood of having coefficients either greater than or less than zero across the Generic model, indicating their robustness. Similarly, average_goldstein_scale, average_confidence, and ReportCount_whole also display robustness in their effects.

The histogram for average_goldstein_scale reveals a bimodal distribution, with most coefficients being negative, suggesting that this variable largely contributes negatively to the model. ReportCount_whole and ReportCount_env are tightly distributed around zero but show a slight bias towards negative values. In contrast, AverageTone and average_confidence display a rightward skew in their distributions, indicating a substantial number of positive beta coefficients. This suggests a robust positive contribution of these variables to the overall model.

While the coefficients for GBI, DXY, VIX_Close and GFSI show relative robustness, with more than 60% probability of $\beta > 0$ or $\beta \leq 0$ in the Generic model, their robustness is slightly weaker compared to other variables. TRI and PRI show relatively high probabilities for both positive and negative coefficients, suggesting they are non-robust.

Taken together, the sequential EBA analyses provide a detailed and systematic evaluation of the robustness of both media-related and traditional influencing factors in explaining movements in the S&P Global Clean Energy Index. Across all stages of the analysis, media and sentiment-related variables, including AverageTone, average_confidence, average_goldstein_scale, ReportCount_env, and ReportCount_whole, consistently exhibit stable coefficient distributions and high cumulative probabilities away from the null hypothesis under the Generic model. Importantly, this robustness persists even as additional doubtful and traditional variables are introduced into the model combinations, indicating that these media-based measures represent core explanatory factors rather than artifacts driven by specific model specifications. In contrast, several traditional factors display notable sensitivity to distributional assumptions and model configurations. Variables such as TRI, PRI, and VIX_Close exhibit substantial variation in coefficient signs across both the Normal and Generic models, suggesting that their effects on clean energy markets are unstable and highly dependent on model assumptions. Other variables, including GBI, GFSI, and DXY, demonstrate moderate levels of robustness but remain weaker relative to the primary media-related variables, supporting

their classification as doubtful rather than core drivers. Meanwhile, a subset of traditional control variables, such as Oil_Close, CHN_market, US_market, SP_IT, and MSCIEF, maintain consistently signed coefficients. Overall, the progressive refinement of the EBA framework confirms that media sentiment and information intensity play a dominant and robust role in shaping clean energy market dynamics, while many commonly cited traditional risk and macro-financial indicators exhibit limited or conditional explanatory power. These results provide a clear empirical justification for focusing on media-related variables in subsequent analyses and establish a solid foundation for the forecasting and mechanism-based investigations developed in the following chapters.

2.4 Conclusion and Future Research

2.4.1 Conclusion and Contribution

This chapter examines the relationship between the S&P Global Clean Energy Index and various factors from the GDELT database. Multiple EBA analyses reveal that media and sentiment-related variables consistently and significantly impact the index, contributing to the model's robustness across different regressions. In contrast, some traditional factors previously identified in the literature as influential fail to demonstrate robustness in these analyses when examined using the unsupervised method, Isolation Forest. Market volatility remains unstable without adjustments for heteroscedasticity and likelihood ratio index weighting in the Normal distribution model, indicating that its robustness depends on specific model conditions. In the model Global Financial Stress Indicator is not as robust compared to other stable factors. Although the Green Bond Index and US Dollar Index initially appear robust under Normal distribution models, they later show substantial uncertainty. Therefore, previous studies Ren et al. (2023); Zhang et al. (2023a); Khalfaoui et al. (2022); Saeed et al. (2021); Wang et al. (2022) which suggest that these variables play a role in the clean energy stock determination may need to be reassessed. Other traditional factors in Table 2.1 maintain relative stability in robustness.

During the data collection phase, I developed custom code for data pre-processing and calculated the daily averages of Focus variables within the GDELT database. This approach lays a strong foundation for future research utilizing

cloud computing with the GDELT database and supports subsequent analyses of media topics. The incorporation of the EBA, Isolation Forest, and the utilization of valuable data from the GDELT database significantly enhances the methodological originality and contribution of this study.

I find that the daily average Goldstein Scale for the U.S. and China in the S&P Global Clean Energy Index shows a significantly negative correlation, indicating that although the news coverage between the U.S. and China has been relatively negative over the past five years, the overall impact on the S&P Global Clean Energy Index remains positive. The statement that more negative reports between the U.S. and China lead to better market performance for the clean energy industry cannot be supported. Instead, it can be said that negative events, such as the trade war, have forced both countries to increasingly focus on the clean energy sector to enhance their competitive advantage, driving strong industry growth and increasing investor confidence. As a result, the global transition to clean energy has become a dominant trend.

The S&P Global Clean Energy Index shows a clear positive correlation with the daily average confidence across global news reports, indicating that the higher the confidence in news reporting, the higher the S&P Global Clean Energy Index. This encourages more reliable global reporting to promote the development of emerging technologies and the clean energy market.

There is also a significant positive correlation between the daily average sentiment of environmental-themed news and the S&P Global Clean Energy Index, suggesting that the more positive the sentiment in environmental news, the larger the clean energy index. While there is a statistically significant relationship between the S&P Global Clean Energy Index and the volume of news reporting, the effect size is too small to provide meaningful insights into the development of the clean energy sector. The positive or negative effect of the variable concerning the proportion of environmental reporting has always been unstable, so its effect is not considered.

Therefore, when global media positively covers the clean energy market, it increases transparency, raises public awareness, and attracts responsible investors. These developments have significant implications for various stakeholders. Influenced by media narratives, investors may adjust their portfolios to align with clean energy trends, maximizing their returns. In turn, governments and policymakers may introduce clean energy-friendly policies, such as a "carbon credit" system for

businesses, to drive market growth. This sustained media engagement further promotes transparent and constructive reporting on clean energy initiatives, creating a reinforcing cycle that accelerates the clean energy transition.

The findings of this chapter suggest that media and sentiment channels play a substantive role in shaping clean energy market dynamics, with important policy implications. The robustness of media-related variables implies that information transparency and news credibility should be considered integral components of clean energy policy design, rather than peripheral factors. Policymakers may enhance market confidence and investment efficiency by promoting clear, consistent, and credible communication surrounding climate objectives, regulatory pathways, and transition timelines. Moreover, the limited robustness of several traditional macro-financial indicators indicates that conventional policy monitoring frameworks may be insufficient to capture risks and expectations in clean energy markets, highlighting the value of incorporating media- and sentiment-based indicators into policy assessment and early-warning systems. The evidence from U.S.–China media sentiment further suggests that geopolitical frictions do not necessarily impede the clean energy transition, but may instead intensify strategic investment and innovation in clean energy technologies. Taken together, these results indicate that effective clean energy governance requires an integrated approach that combines financial regulation with strategic information management and policy communication to support a stable and sustained transition toward clean energy.

In addition, this chapter provides the empirical foundation for forecasting in Chapter 3, which tests how well these factors predict future outcomes using different models, and for Chapter 4, which explores how these factors operate through policy communication channels and explains their mechanisms within policy discourse.

2.4.2 Limitation of GDELT Database and Future Research

The GDELT database does not allow direct access to environment-related topics, such as "V1THEME", in the Event and Mention datasets. Previously, an attempt was made to filter environment-related fields in the "URL" column of both datasets, but the filtered data could not cover the entire data range. This makes it difficult to obtain data related to environmental factors that impact the clean energy index. For instance, the average Goldstein Scale suggests that by filtering its values for China and the United States in the Event Table, along with data related to environmental

topics, I may identify more precise variables that significantly influence the clean energy market.

Fortunately, factors such as the positivity of U.S.-China media reports are not directly tied to the necessity of filtering environmental topics, these types of factors are more closely related to the reporting preferences between countries, and I also introduced the variable "ENV.number" to address this issue, but the results of the EBA analyses suggest that its impact on the clean energy index is not key. In contrast, the Goldstein Scale has consistently proven robust across multiple EBA analyses, indicating that this variable can be reliably used in the model to explore its relationship with the S&P Global Clean Energy Index.

Future studies could explore economic and financial issues from the perspective of media coverage, employing more sophisticated models to investigate the relationship between novel media and sentiment factors and the S&P Global Clean Energy Index. Additionally, these approaches could be extended to examine other indices in sustainability, such as ESG scores, Energy Consumption and explore their interactions with these factors. Beyond correlation analysis, future research could further investigate the causal relationships among media coverage and clean energy market outcomes by adopting explicit causal inference frameworks, thereby clarifying the direction and underlying mechanisms through which information-driven factors influence clean energy markets.

SUPERVISED LEARNING MODELS, STATISTICAL MODELS OR HYBRID MODELS? A PREDICTION OF CLEAN ENERGY STOCK BASED ON FEAR AND FUNDAMENTAL FACTORS

Abstract

Building on the key drivers identified in the previous chapter, this chapter tests their predictive capacity using several time series and machine learning models. It first examines the main market and information factors that influence the S&P Clean Energy Index and applies XGBoost and SHAP values to measure their importance. Future feature data are simulated using Random Forest and LSTM models, while the K-Nearest Neighbors algorithm is employed to enhance prediction accuracy. Statistical models such as ARIMA and VAR, and machine learning models such as LSTM, SAE, and Prophet, are applied to forecast both index trends and values. The results show that Prophet-related models provide the most accurate and stable forecasts, especially when fitted on Random Forest simulated data. Prophet-related models perform well for green financial indices because it can clearly separate seasonal and trend patterns, improving both the understanding and prediction of movements in the clean energy index.

3.1 Introduction

Previous studies on stock price fluctuations have presented two prevailing perspectives. The first is the Efficient Market Hypothesis (EMH), which fundamentally rejects the assumption that stock prices are predictable (Fama, 1965). However, a substantial number of technical analysts and practitioners argue that historical price data can be utilized to forecast future stock prices (Lo and Hasanhodzic, 2010). In the context of predicting historical prices and asset trends, previous researchers have mainly used supervised learning methods, including Support Vector Machines (SVM), decision trees, and Artificial Neural Networks (ANN) (Ślepaczuk and Zenkova, 2018). When I predict the direction of price movements, stock prices have been shown to be more effective input features compared to return values (Bieganowski and Slepaczuk, 2024). Therefore, using machine learning models effectively and appropriately can be a valuable tool for investors to accurately predict price trends and seize profitable opportunities in the future.

Machine learning can be categorized into supervised learning, unsupervised learning, and reinforcement learning based on the learning approach. Similarly, machine learning can be structurally divided into single-layer and multilayer structures, with the multilayer structure more commonly known as deep learning (Mitchell, 1997; Sutton and Barto, 1998; Hinton et al., 2006; Bengio, 2009). As machine learning continues to develop, its application in models for classification and prediction has become increasingly widespread. Bieganowski and Slepaczuk (2024) state that they transformed their regression model into a classification model, focusing on predicting the trend of price changes rather than specific prices. In this case, the choice of the loss function becomes particularly important. This transformation is more likely to effectively generate investment signals. Similarly, Matsubara et al. (2018) propose a generative model based on media reports, which incorporates price movements as conditions, thus mitigating the risk of overfitting commonly associated with the nature of generative models. Previous predictive models have generally achieved success in long-term forecasting, but often face limitations in short-term predictions.

This chapter focuses on trend and value analysis, with the goal of developing a more accurate prediction model for the S&P Clean Energy Index. Previous studies have largely ignored forecasting in the clean energy market, which is due to the market emerging later than traditional financial markets and the fact

that financial researchers often overlook media factors when predicting trends in this sector. But the importance of the clean energy market cannot be ignored. Govindarajan and Ganesh (2022) find that energy policies impact a country's energy economy and whole economy because of the close relationship between energy and economics. Policymakers in different countries can use these frameworks to help guide the shift to a low-carbon economy through renewable energy. The research on environmental justice also needs to pay more attention to new energy governance issues, particularly around ensuring global fairness in the transition to low-carbon energy through the use of renewable energy systems. Previous research has shown that most governance measures have a positive effect in easing the connection between environmental regulations and clean energy consumption. To reduce greenhouse gas emissions, governments should encourage the use of clean energy in their countries. Policymakers can promote clean energy consumption through strict environmental regulations and effective governance (Shahzad et al., 2024). Therefore, predicting the clean energy index can be useful not only for investors but also for policymakers—government officials can use this information to guide their decisions. Understanding these trends is essential for making effective policies.

Against this background, this chapter seeks to address several key research questions related to forecasting the clean energy market. First, can media- and sentiment-related factors serve as effective predictors of clean energy market dynamics beyond traditional financial and macroeconomic variables. Second, does incorporating these feature-based indicators improve forecasting performance across different time horizons in a market undergoing rapid structural change and development. Third, how do different classes of models, including statistical, supervised learning, and hybrid approaches, compare in their ability to capture nonlinear relationships and evolving information flows in the clean energy market, and how does their predictive performance vary across short-, medium-, and long-term forecasting horizons. By addressing these questions, this chapter aims to clarify the role of information-driven features in clean energy market forecasting and to explore their implications for both investors and policymakers.

To fill this gap, this chapter builds on the factors identified by Zheng and Lucey (2025c) and tests their effectiveness as features for forecasting the dynamics of the clean energy market using several types of models. The theoretical foundations of the relevant statistical, supervised learning, and hybrid models are discussed in 3.2.

3.2 Literature Review

3.2.1 Tree-based Models

Tree models are widely used for interpreting feature importance. Tree-based ensembles are machine learning models that combine decision tree weak learners to improve performance. These ensembles can be broadly categorized into two types based on their combination strategies: bagging and boosting. Bagging-based models, such as Decision Tree Bagging and Random Forest (RF), are classic examples of tree-based ensembles (Liu et al., 2023). Polamuri et al. (2019) state that random forest refers to ensembles of estimators for data. They fit individual decision trees on various subsamples of the given dataset. This method controls overfitting and improves predictive accuracy.

On the other hand, boosting-based methods, particularly Gradient Boosting Decision Trees (GBDT), have gained popularity due to their scalability and exceptional performance. This has led to the development of advanced GBDT-based models like XGBoost, LightGBM, and CatBoost. These tree-based ensembles provide feature importance based on each feature's contribution to the final prediction outcome (Liu et al., 2023).

3.2.2 Classification: K-Nearest Neighbors(KNN)

In pattern recognition, the KNN algorithm classifies objects by comparing them to the closest training examples in the feature space. KNN is a simple, instance-based method where computations are delayed until classification. It works well when there is little or no prior knowledge about the data. The algorithm keeps the entire training set and assigns a class to each query based on the majority class of its k-nearest neighbors. Essentially, a sample is classified based on how similar it is to its nearby samples. If a sample's class is unclear, it can be predicted by looking at the classes of its nearest neighbors (Imandoust and Bolandraftar, 2013).

Data mining, which helps discover useful patterns and relationships, offers unique advantages in financial modeling. Similar to other methods, many data mining techniques have been applied in finance. These include linear and nonlinear models, neural networks, k-means and hierarchical clustering, k-nearest neighbors, decision trees, regression (like logistic and multivariate regression), ARIMA, principal component analysis, and Bayesian learning (Imandoust and Bolandraftar,

2013).

3.2.3 Prophet and its Hybrid Model

The Prophet model is an open-source automated forecasting tool developed by Facebook in 2017. It is specifically designed for modeling nonlinear time series data with yearly, weekly, and daily seasonality, as well as holiday effects. It also handles missing data and outliers, making the fitting process almost entirely automatic and outperforming other methods in many cases (Facebook, 2017; Taylor and Letham, 2017).

Ning et al. (2022) find that the Prophet model is convenient due to its fixed structure, with only two hyperparameters needing further adjustment for oil production forecasting. Besides cross-validated parameter values, nearby values can also produce good results. More importantly, Prophet can identify implicit past seasonal effects.

Building on this, Wei (2025) uses the Prophet model to forecast endogenous variables, leveraging its strengths in capturing trends and seasonality in time series data. For aspects of the endogenous variables that cannot be fully predicted, this study seeks effective exogenous variables for regression analysis. This dual approach enhances the model's interpretability and robustness by incorporating additional relevant factors that may influence the outcomes. By integrating exogenous variables, the model not only improves prediction accuracy but also provides a clearer understanding of the underlying mechanisms driving the data.

3.2.4 Long Short-Term Memory(LSTM), Statistical and Hybrid Model

Afzali et al. (2021) state that in traditional neural networks, the model generates outputs by multiplying inputs with weights and applying an activation function. However, Recurrent Neural Networks (RNNs) process future inputs by utilizing results obtained from the hidden layers. LSTM networks are a specific type of RNN architecture that can capture long-term dependencies for predicting future events. This architecture enhances the memory capacity of standard RNNs by integrating long-term memory, enabling the model to retain and utilize information over extended periods.

CHAPTER 3. SUPERVISED LEARNING MODELS, STATISTICAL MODELS OR HYBRID MODELS? A PREDICTION OF CLEAN ENERGY STOCK BASED ON FEAR AND FUNDAMENTAL FACTORS

The performance of the LSTM trained predictive models exhibited randomness and the number of training epochs did not substantially impact their performance. The primary factors influencing the effectiveness of LSTM in algorithmic investment strategies include hyperparameter tuning methods, model architecture, and estimation process (Bieganski and Slepaczuk, 2024). Lim and Zohren (2021) state that the latest trend in deep learning is the development of hybrid models to address these limitations, where hybrid models outperform purely statistical or machine learning models in various applications. Hybrid methods combine well-established quantitative time series models with deep learning - using deep neural networks to generate model parameters at each time step. On the one hand, hybrid models allow domain experts to use prior information to guide neural network training, reducing the hypothesis space of the network and improving generalizability. On the other hand, this is particularly useful for small datasets, as deep learning models are more prone to overfitting on small datasets. For example, Shah et al. (2018) highlight the potential of LSTM-RNN in forecasting non-stationary data. By incorporating additional features, LSTM-RNN can identify underlying trends and generate long-term forecasts, particularly for volatile stock datasets. Zhang et al. (2023b) introduce a novel hybrid model that leverages the strengths of highly structured autoregressive (AR) models and LSTM models for COVID-19 case prediction tasks. Through extensive numerical experiments, their results demonstrated that the hybrid model achieves superior predictive performance compared to using AR or LSTM models individually. In principle, the hybrid model fully capitalizes on the respective advantages of its two components: the expressive capability of LSTM in capturing nonlinear patterns in the data and the interpretability of the simple structure in AR, which better captures the trend dynamics of time series.

In time series forecasting, the AutoRegressive Integrated Moving Average (ARIMA) model is widely recognized for its maturity and reliability. Siامي-Namini et al. (2018) compare the effectiveness of deep learning-based time series forecasting algorithms, including Long Short-Term Memory (LSTM), with traditional methods like ARIMA. The results demonstrated that LSTM and other deep learning-based techniques outperformed ARIMA, achieving significantly lower average error rates. Although the ARIMA model incorporates autoregressive characteristics, it is primarily designed for time series with trends and seasonality, especially those with periodic and seasonal fluctuations (Afzali et al., 2021). Besides, the Vector Autoregression (VAR) model is a widely used, flexible, and efficient tool for analysing

multivariate time series. It extends the univariate autoregressive model to capture the dynamics of multiple time series and has proven particularly valuable in modeling and forecasting economic and financial data (Zivot and Wang, 2003).

However, the VAR model has its limitations, particularly in its inability to effectively capture nonlinear trends. The advantage of this hybrid approach lies in its ability to estimate optimal parameters without relying on the assumptions required by traditional statistical methods (Afzali et al., 2021).

3.2.5 Supervised Autoencoder (SAE)

A large amount of labeled data is essential for supervised learning, with labels typically provided by domain experts. However, models that rely only on input data for training, without the necessity of labeled data, provide considerable convenience. This is the key advantage of unsupervised models. An autoencoder consists of two parts: the encoder and the decoder. The encoder compresses and encodes the input data, transforming it into a different representation space for feature extraction. The decoder then reconstructs the data from this encoded form, aiming to minimize the error between the original and reconstructed data. Autoencoders are effective for feature selection and extraction, with the algorithm performing best when the reconstruction error is minimal (Li et al., 2023b).

SAE is a type of autoencoder designed for regression and classification tasks. In SAE, the encoded features are combined with the original inputs to train a supervised model on the given labels. Its performance is measured by how well it reconstructs the input data (unsupervised part) and how accurate its predictions are (supervised part). SAE is especially effective with noisy data, making it a great choice for algorithmic trading models (Bieganowski and Slepaczuk, 2024).

3.2.6 Theoretical Motivation for Model Selection

Building on the literature reviewed above, the selection of the seven forecasting models in this chapter is primarily guided by theoretical considerations rather than reliance on existing empirical applications. Specifically, the models are chosen to represent distinct theoretical perspectives on financial market dynamics and information processing. Statistical time series models, such as ARIMA and VAR, are motivated by classical economic theory that emphasizes linear dependence, persistence, and intertemporal interactions in financial data.

CHAPTER 3. SUPERVISED LEARNING MODELS, STATISTICAL MODELS OR HYBRID MODELS? A PREDICTION OF CLEAN ENERGY STOCK BASED ON FEAR AND FUNDAMENTAL FACTORS

At the same time, financial markets are widely recognized as complex and nonlinear systems, in which price dynamics may reflect heterogeneous beliefs, adaptive behavior, and nonlinear interactions among market participants. As argued by Farmer and Lo (1999), such complexity limits the explanatory power of purely linear models and motivates the use of flexible, nonlinear approaches. Accordingly, tree-based models and KNN are included to capture nonlinear relationships and local dependence structures implied by behavioral finance and heterogeneous agent frameworks.

Deep learning models, particularly LSTM and supervised autoencoders, are theoretically motivated by their ability to model complex nonlinear dynamics, long-term dependencies, and high-dimensional information environments. These features are especially relevant in clean energy markets, where price formation is influenced not only by historical fundamentals but also by evolving media coverage and sentiment.

Hybrid models, including Prophet-based regressions and LSTM–statistical combinations, are selected to bridge theory-driven structure and data-driven flexibility. These models reflect the theoretical view that financial time series exhibit both systematic components, such as trends and seasonality, and irregular components driven by changing information flows. While prior empirical studies inform the practical feasibility of these approaches, the primary rationale for model selection lies in their theoretical capacity to capture complementary aspects of clean energy market dynamics. Together, the seven models provide a coherent and theoretically grounded framework for evaluating predictive performance under different assumptions about market behavior and information transmission.

Inspired by these models, this study introduces the seven types of time series models, combining statistical and supervised approaches to predict the S&P Clean Energy Index, and finally select the model that best aligns with the prediction trends for the current period.

3.3 Methodology

3.3.1 Data Selection and Pre-processing

3.3.1.1 Data Source and Variables Explanation

Media coverage plays a crucial role in financial market forecasting. Investors analyse all publicly available information, such as news reports, to understand past and future events, and subsequently react by making decisions regarding financial instruments. The prices of financial instruments respond rapidly to these events (Matsubara et al., 2018). GDELT is a highly significant database related to media. Based on the GDELT database, previous studies such as Cherief et al. (2025) construct a national-level biodiversity risk index, allowing us to isolate the dates of urgent biodiversity events. Compared to long-term declines, such events are easier to identify and are more likely to have unforeseen impacts on financial markets. Therefore, using GDELT data to study financial market fluctuations is reasonable.

The data used in this study are sourced from the GDELT, Bloomberg databases and Economic Policy Uncertainty website, and the data collection method follows the approach used in previous studies, the variables from the GDELT database require attention. The daily average of the Goldstein scale was used to measure the political sentiment of international events involving China and the United States—two key contributors to the global clean energy index. The `average_confidence` was calculated using the Confidence metric from the GDELT 2.0 Mention Table to assess media credibility. `AverageTone` represents the average daily sentiment over five years for environment-related topics, filtered from the GDELT GLOBAL KNOWLEDGE GRAPH 2.1. This was done by selecting entries with "ENV" in the "V1THEME" column and averaging the corresponding values in the "Tone" field of the "V1.5TONE" column (Zheng and Lucey, 2025c). The full name of all variables will be explained by Table 3.1.

The historical data for all variables in this study spans from June 30, 2019, to June 30, 2024. However, the updates for TRI and PRI on the Economic Policy Uncertainty website are only available up to December 29, 2023. Therefore, the imputation process for the missing data between December 30, 2023, and June 30, 2024, will be detailed in 3.3.1.7. The prediction period is set from July 1, 2024, to December 31, 2024, with all data displayed in this study is on a daily basis.

CHAPTER 3. SUPERVISED LEARNING MODELS, STATISTICAL MODELS OR HYBRID MODELS? A PREDICTION OF CLEAN ENERGY STOCK BASED ON FEAR AND FUNDAMENTAL FACTORS

Table 3.1: Variables Explanation

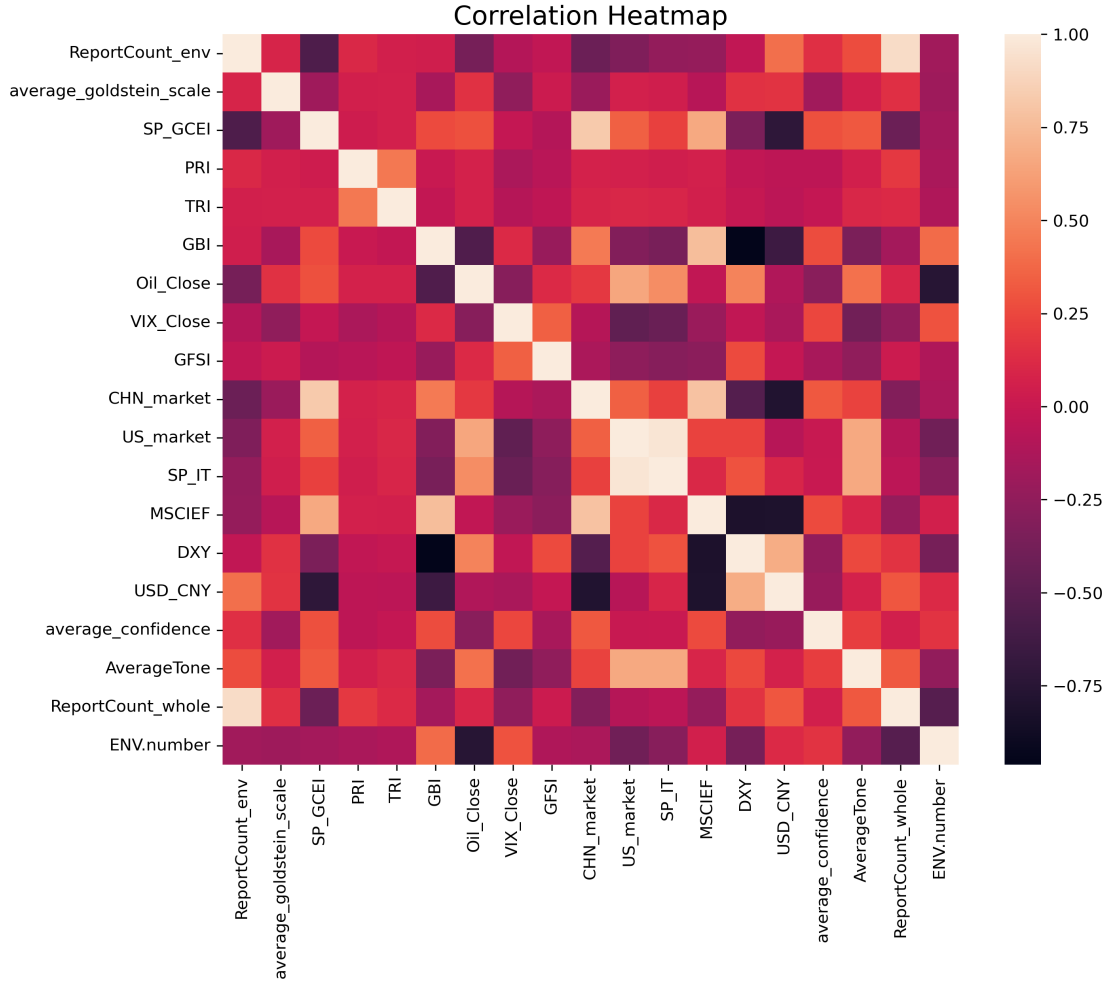
Variables	Explanations
SP_GCEI	S&P Global Clean Energy Index
ReportCount_env	The daily count of environmental-themed news reports
ReportCount_whole	The daily count of all news reports
average_goldstein_scale	The daily average Goldstein Scale
average_confidence	The daily average confidence of news reports
TRI	Transition Risk Index
PRI	Physical Risk Index
GBI	Green Bond Index
Oil_Close	Crude Oil WTI Futures
VIX_Close	CBOE Volatility Index
GFSI	Global Financial Stress Indicator
CHN_market	China Shanghai Composite Stock Market Index
US_market	S&P 500 Index
SP_IT	S&P 500 Information Technology (sector) Index
DXY	US Dollar Index
USD_CNY	United States Dollar / Chinese Renminbi Cross
MSCIEF	MSCI Emerging Markets
AverageTone	The daily average tone of environmental reports calculated
ENV.number	ReportCount_env / ReportCount_whole

3.3.1.2 Correlation of Vairables

To identify the correlations between variables and avoid multicollinearity among variables, which could affect the subsequent results of feature importance selection using the XGBoost method, I first plot the correlations of all variables, obtaining the results shown in Figure 3.1.

As shown in Figure 3.1, USD_CNY, ReportCount_env, ReportCount_whole, and DXY demonstrate a strong negative correlation with the clean energy index, while CHN_market and MSCIEF exhibit a strong positive correlation. In contrast, variables such as average_goldstein_scale, VIX_Close, PRI, TRI, Oil_Close, and AverageTone show relatively weak correlations with the clean energy index. Morrow-Howell (1994) suggest that in order to simplify the model, it is generally best to reduce the number of independent variables before multicollinearity becomes an issue. If only one variable is highly correlated, the easiest solution is to remove that variable, which might be preferable if researchers find a simpler model more appealing. However, it is important to note that a weaker correlation with

Figure 3.1: Correlation Heatmap



the SP_GCEI does not necessarily indicate a loss in the model's predictive power. Methodologically, the goal should be to retain as many relevant explanatory variables as possible, while systematically removing those with high multicollinearity, to ensure both the stability and interpretability of the model.

Therefore, it is crucial to check for significant correlations among the variables, excluding the target variable, SP_GCEI. Since the variables analysed in this study are features selected from various dimensions to predict the S&P clean energy index, the heatmap reveals significant high correlations between CHN_market and US_market, as well as between ReportCount_whole and ENV.number, DXY, and other variables. The dimensions measured by these highly correlated variables are similar, so I can remove one of them. However, deciding which variable to

CHAPTER 3. SUPERVISED LEARNING MODELS, STATISTICAL MODELS OR HYBRID MODELS? A PREDICTION OF CLEAN ENERGY STOCK BASED ON FEAR AND FUNDAMENTAL FACTORS

eliminate and how to address the other correlated variables still requires further evaluation. Based on my previous analysis in Zheng and Lucey (2025c), which examined the stability of each variable's influence on the S&P clean energy index, I found that the impact of ENV.number is relatively insignificant. Additionally, with ReportCount_whole and ReportCount_env serving as supplementary variables, excluding ENV.number at this stage appears reasonable. However, correlation and feature contribution in a model are distinct concepts. Variables that are weakly correlated or exhibit high multicollinearity cannot be excluded outright without further validation. Therefore, I will employ the XGBoost model for additional feature selection.

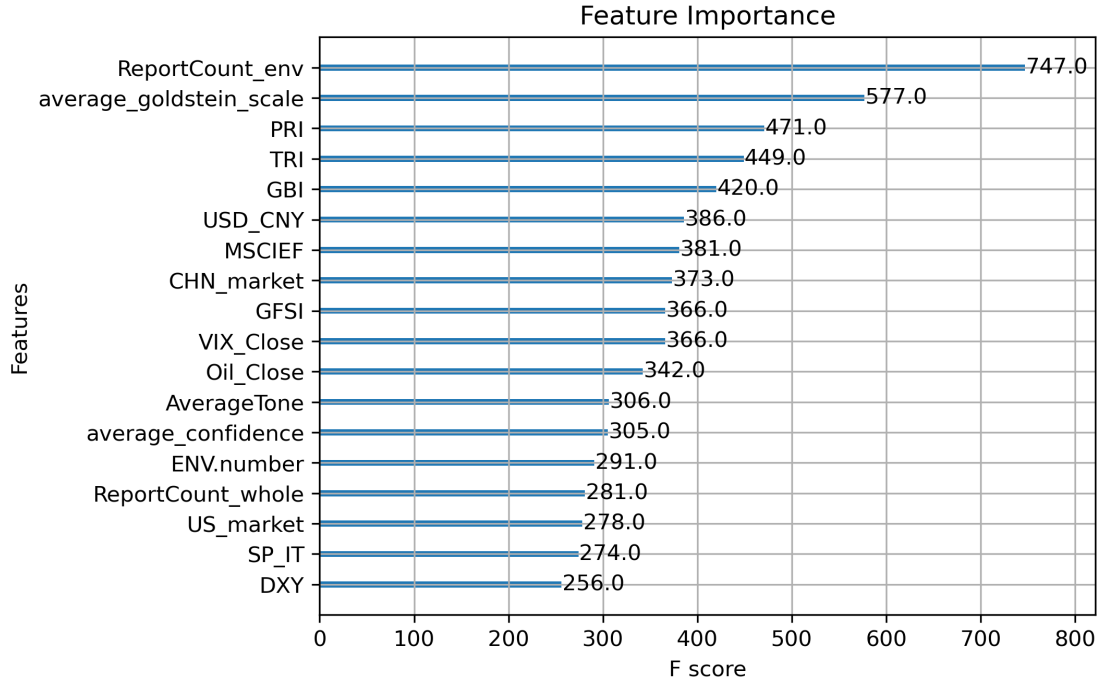
3.3.1.3 Feature Importance of Variables

After loading the data from 3.3.1.1, I first address the missing values. The raw dataset contains some missing figures, the reasons for which will be discussed in 3.3.1.5. Instead of using average values to fill in these gaps, I retain the missing values to preserve the original contributions of each variable to the S&P clean energy index. Global importance offers a comprehensive interpretation of a machine learning model. It ranks the significance of each feature based on its contribution scores or weights for the model's input variables, helping to identify the key features or metrics that significantly influence the target variable (Liu et al., 2023). Most importantly, the features used for prediction in each phase might differ, so I need to use a similar approach to reselect them (Liu et al., 2023). Next, I define the target variable as the S&P Clean Energy Index, which becomes the Y dataset, while the remaining variables constitute the X dataset. The formulas and parameter tuning sections of the XGBoost will be placed in A.1.

Through the calculation of the optimal parameters, I obtain an R^2 value of 0.98 for the model, indicating a strong explanatory power on the test data. With such a highly fitted model, I can evaluate the contribution of different input variables to the final prediction of the target variable and use this to select features to predict the clean energy market. As shown in Figure 3.2, feature importance is calculated using the F score from the XGBoost model, which represents the cumulative number of times a feature is used as a splitting node. The higher the F score, the more frequently the feature is used in the model and the greater its impact.

Based on the results of the XGBoost model, correlation analysis, and previous studies on the impact of media factors, I retained variables with feature importance

Figure 3.2: Feature Importance of Variables



scores greater than 300 for subsequent predictions first. However, directly removing variables with contributions below 300 is subjective. Liu et al. (2023) suggest that an approach based on cooperative game theory can be used to calculate the optimal distribution of contributions, which helps identify statistically significant predictors related to the target variable. Therefore, I will conduct further validation.

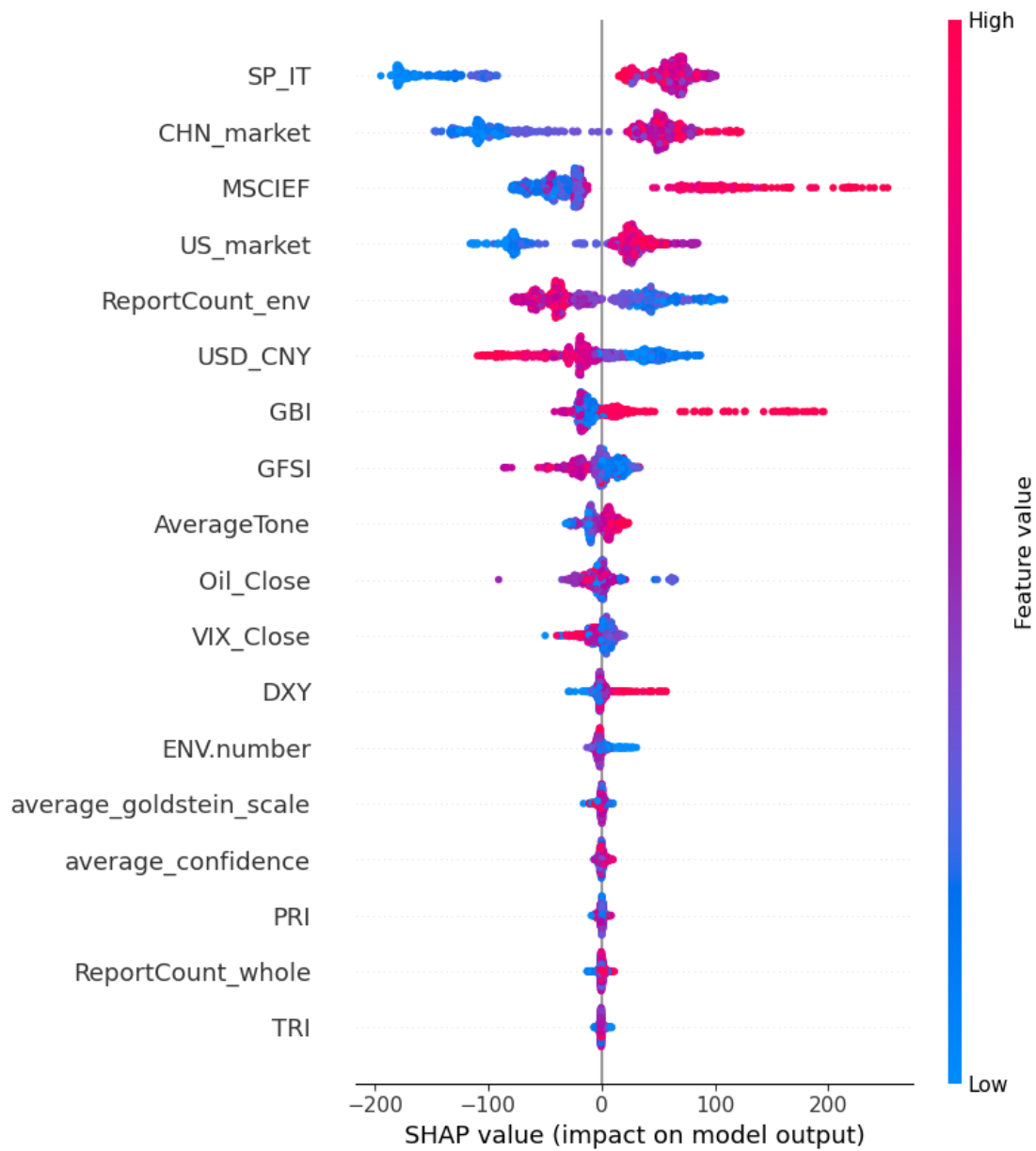
3.3.1.4 SHAP Value Validation and PCA technique

SHAP stands for Shapley Additive Explanations, which combines Shapley values with local explanation methods. The SHAP value method resolves the inconsistency between various feature attribution methods when there are many features, ensuring consistent feature importance (Liu et al., 2023).

Since the R^2 of the model is relatively high, indicating a good fit, I used the same parameters as 3.3.1.3, following the approach of Liu et al. (2023), the SHAP values are determined for each tree in the boosted ensemble model, allowing us to make long- and medium-term warning preparations and rank feature importance using SHAP techniques. Then I could obtain the following results:

SHAP values are a local metric that measures the impact of each feature on

Figure 3.3: SHAP Value of Variables



an individual prediction, while feature importance is a global metric that reflects how much each feature contributes to the overall model training process (Liu et al., 2023). This also leads to inconsistencies in the conclusions drawn from Figure 3.2 and Figure 4.4. Features like PRI and TRI may have a smaller contribution to individual predictions, reflected by lower SHAP values; however, due to their interactions with other features, they may still appear frequently in the model, thus ranking higher in feature importance.

On the other hand, features like SP_IT, which are less correlated with other features, may have a higher contribution to individual predictions, as indicated by higher SHAP values, but since they are used less frequently in the model's decision-making, their feature importance ranking may be lower. Furthermore, given SP_IT's high correlation with US_market, it might be worth considering its removal.

While average_goldstein_scale and average_confidence show relatively small impacts in the SHAP values, they perform well in terms of feature importance and are not highly correlated with other features. Furthermore, as media-related factors, they are key variables of focus for us. Therefore, these features should be retained. Given that US_market performs worse than CHN_market in both Figure 3.2 and Figure 4.4, CHN_market should be retained based on this consideration.

DXY exhibits poor performance in both Figure 3.2 and Figure 4.4, and is highly correlated with the better-performing GBI, warranting its exclusion. While USD_CNY is highly correlated with MSCIEF and CHN_market, all three variables demonstrate strong feature importance and SHAP values, thus they will be retained. To address multicollinearity among these highly correlated features, Principal Component Analysis (PCA) can be employed (Kyriazos and Poga, 2023). Given the presence of missing values in the dataset, PCA will be applied after the missing data have been imputed. It is important to clarify that PCA, together with SHAP values and feature importance measures, is used exclusively for pre-processing and screening explanatory variables by addressing multicollinearity among features, rather than for modeling nonlinear relationships with the S&P Global Clean Energy Index. The nonlinear dynamics between the selected features and the target variable are explicitly captured at a later stage through nonlinear forecasting models, including statistical models and LSTM-based frameworks.

The approach employed is similar to that of Elsayed et al. (2021): the multivariate time series forecasting setup describes a scenario where multiple features

CHAPTER 3. SUPERVISED LEARNING MODELS, STATISTICAL MODELS OR HYBRID MODELS? A PREDICTION OF CLEAN ENERGY STOCK BASED ON FEAR AND FUNDAMENTAL FACTORS

are provided locally in the dataset, but only a single target variable needs to be predicted. Based on this approach and the selection results, I will use the following variables as features for subsequent predictions: ReportCount_env, average_goldstein_scale, PRI, TRI, GBI, USD_CNY, MSCIEF, CHN_market, GFSI, VIX_Close, Oil_Close, AverageTone, and average_confidence.

3.3.1.5 Processing the Missing Values

Some variables sourced from the Bloomberg database contain missing values over this five-year period. These missing values are understandable in the context of financial markets, often due to factors such as holiday closures, trading suspensions (Javed Awan et al., 2021; Toms, 2011). The way missing prices are approximated clearly impacts the estimated abnormal returns and the subsequent test statistics (Kallunki, 1997), making it essential to properly handle missing values before making predictions.

I primarily adopt the approach from Javed Awan et al. (2021), and apply the following methods to handle the missing values: 1. If the missing value occurs for only one day, it is replaced with the average of the values from the preceding and following days for the same variable. 2. For data missing for more than two consecutive days, the missing values are filled using the last available value before the gap or the first available value after the gap, depending on the direction of the missing data. 3. When data is missing for five consecutive days, the filling strategy is as follows: the first two missing days are replaced with the last available value, the last two missing days are replaced with the subsequent available value, and the middle day is filled with the average of the values from the preceding and following days. 4. After filling the missing data for trading days (a relatively small proportion over the five years), weekend values are replaced using the moving average of the preceding five trading days.

In this study, I used the concept of the "moving average", where the moving average crossover rule helps reduce sensitivity to noise by smoothing the price (Toms, 2011). This was applied to all data except for the last six months of historical data for TRI and PRI. However, because some missing values were irregular and unpredictable, I was concerned that using automated imputation methods through code might introduce biases or inaccuracies. To ensure the reliability and accuracy of the data preprocessing, I chose to manually handle the missing values.

3.3.1.6 Feature Data Processing and Overall Model Approach

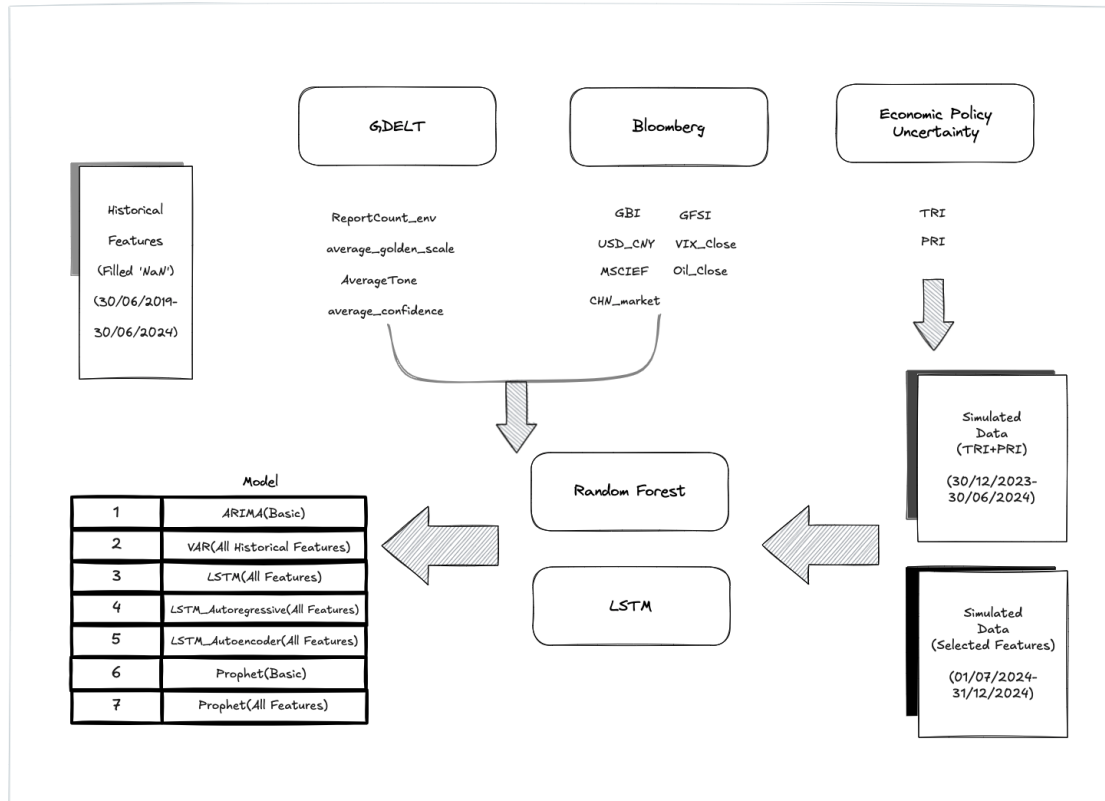
To enhance the clarity of the structure of this study, I have visualized the data processing workflow discussed and the comprehensive process required for model prediction in a flowchart. The Figure 3.4 illustrates the data sources and preprocessing steps for the selected features. In conventional time series analysis models, the features typically comprise the historical values of the target variable and their associated timestamps. Historical market data represents various scenarios, helping to identify time series patterns with statistically significant predictive power. This increases the likelihood of profitable trades and returns in competitive business investments (Polamuri et al., 2019). The "Basic" label in the Figure 3.4 denotes these standard time series models. In contrast, the "All Features" concept in this study encompasses not only the conventional features used in time series models but also all the past features and future simulated data generated from the selected features, reflecting a more comprehensive feature set. "All Historical Features" refers to the features based on a basic time series model, supplemented with all past features that have been fully imputed to address missing values. "Selected Features" represents the variables filtered out in the previous 3.3.1.4. This also includes the simulated data for TRI and PRI from the last six historical months.

It is important to note that the clean energy market is currently undergoing a period of rapid structural change and development, driven by technological innovation, climate policies, geopolitical dynamics, and evolving investor preferences. In such a transitional environment, relying solely on historical index values may be insufficient to capture shifts in market dynamics. This motivates the use of a feature-based forecasting framework, in which economically and information-relevant variables are treated as leading indicators rather than passive controls. By incorporating media-related factors, climate risk indicators, and macro-financial variables as features, the proposed approach is better suited to capture forward-looking signals that reflect changing expectations and structural adjustments in the clean energy market. From a forecasting perspective, this framework allows the model to respond more flexibly to regime changes and emerging trends, thereby enhancing its relevance for both investors and policymakers operating in a rapidly evolving clean energy landscape.

I will follow the procedure outlined in Figure 3.4 to complete the subsequent simulation data generation and prediction processes across various models. To

CHAPTER 3. SUPERVISED LEARNING MODELS, STATISTICAL MODELS OR HYBRID MODELS? A PREDICTION OF CLEAN ENERGY STOCK BASED ON FEAR AND FUNDAMENTAL FACTORS

Figure 3.4: Feature Data Processing and Overall Model Approach



simulate future feature data, I employ two supervised learning models: Random Forest and LSTM. This approach facilitates the imputation of missing values in feature data and ensures that the feature data meets the requirements for future predictions in subsequent models. To capture the complex dynamics of the S&P Clean Energy Index, this study utilizes a variety of models, ranging from statistical methods (ARIMA, VAR) to supervised learning techniques (LSTM, SAE, Prophet), as well as hybrid models (LSTM_Autoregressive, Prophet Features). Each method serves a distinct purpose: statistical models address linear relationships, with ARIMA focusing on univariate linear dependencies, while VAR captures the historical interactions among multiple variables. In contrast, machine learning and hybrid models are designed to model nonlinear and multivariable relationships.

3.3.1.7 Stochastic Simulated Data: Random Forest and LSTM

To predict the changes in the S&P Clean Energy Index over the next six months, starting from July 1, 2024, I utilize not only the past five years of time-series data

for the index but also time-series data for these features over the past five years and the upcoming six months. This is because I have already validated through various methods in 3.3.1.4 that these features influence the index's prediction.

However, as of June 30, 2024, I am unable to access future data for these features. Consequently, I adopt the approach proposed by Afzali et al. (2021), dividing the data into real and simulated datasets. Given the uncertainty surrounding the most effective method for generating this data, I plan to perform separate predictions using the data generated by each model, allowing us to evaluate which method best supports accurate forecasting. Additionally, due to the absence of historical data for TRI and PRI over the past six months, the gap is too large for the imputation methods outlined in 3.3.1.5 to be applicable. Therefore, I will use the simulated data to fill in this six-month gap, treating it as historical data. Data generated after July 1, 2024, will be considered as future data.

Although deep learning models are now widely used for prediction tasks, experiments show that simpler models like GBRT can perform just as well as advanced DNN models when effective feature engineering is applied to their input and output. In some cases, they can even outperform DNNs. More broadly, the results highlight that simpler machine learning models should not be ignored. Instead, they should be carefully tuned to ensure meaningful progress in time series forecasting (Elsayed et al., 2021). Thus, the simulated data employs two algorithms: a simple machine learning model, Random Forest, and a more complex deep learning model, LSTM.

The random forest algorithm is less prone to overfitting, can capture non-linearity, and has a small number of model parameters. It is easy to implement and can predict the average output for new input data using a set of observed input-output training data (Schoppa et al., 2020). Therefore, I use the Random Forest regression method to generate simulated data. Before applying Random Forest, I need to impute the missing historical data for the last six months of TRI and PRI using the moving average method. Without this pre-imputation, the time span of the data generated by Random Forest would be inconsistent, failing to cover the entire prediction period. Even if I specify the generation time, it would still not function correctly.

The moving average method I apply uses the past five data points, with the two most recent points given a weight of 2, the most distant point assigned a weight of -1, and the weighted average used to estimate the future value. I chose this method because some features in the dataset are missing data for weekends, but weekday

CHAPTER 3. SUPERVISED LEARNING MODELS, STATISTICAL MODELS OR HYBRID MODELS? A PREDICTION OF CLEAN ENERGY STOCK BASED ON FEAR AND FUNDAMENTAL FACTORS

data (Monday to Friday) are available. Once the simulated data are generated by Random Forest, it will overwrite the previously imputed TRI and PRI data, allowing the process to function correctly.

I can then use Random Forest regression to predict each variable separately and generate the simulated data. The formulas and parameter tuning sections for the models will be included in A.2.

After selecting the optimal model, I print the final RMSE for each variable. A smaller RMSE indicates that the model has been trained more effectively for that variable. The RMSE for these features is shown in Table 3.2.

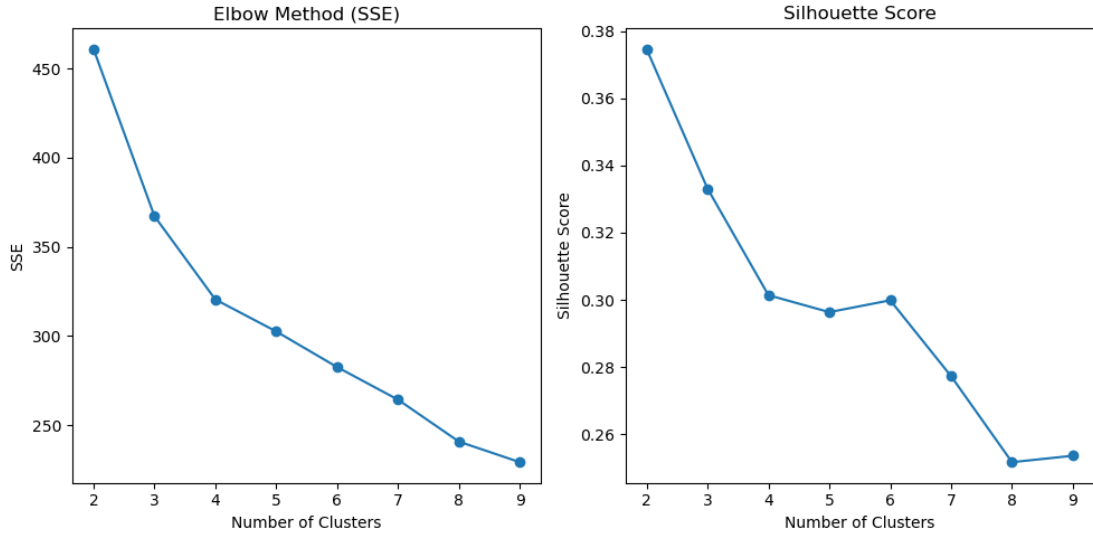
Table 3.2: RMSE for Each Feature

Variable	RMSE
ReportCount_env	14921.5349
average_goldstein_scale	0.4715
PRI	0.0197
TRI	0.0200
GBI	0.5983
Oil_Close	1.7094
VIX_Close	2.0147
GFSI	0.0587
CHN_market	33.2577
USD_CNY	0.0216
MSCIEF	1.1559
average_confidence	1.1559
AverageTone	0.2247

It can be observed that, with the exception of the high RMSE values for ReportCount_env and CHN_market, the RMSE values for the other variables are relatively small. The high RMSE for ReportCount_env is attributable to the generally large values in its data, while the elevated RMSE for CHN_market is likely due to its high volatility. Since PCA is applied for dimensionality reduction on the multicollinear variable, CHN_market, at the end of this section, its RMSE can be excluded from consideration in my analysis. Additionally, the simulated data is generated using Random Forest regression.

Next, I used the LSTM model to generate simulated data. Initially, I applied the KNN algorithm to classify the historical dataset. For missing TRI and PRI values in the historical six months, I use the same method as the one used to fill in missing values in the earlier stage with Random Forest. To address the sensitivity

Figure 3.5: Cluster Evaluation



of the KNN algorithm to different choices of neighborhood size k , several weighted voting methods have been developed (Imandoust and Bolandraftar, 2013). The importance of each variable, as measured by Figure 3.2, was used as a weight in the classification process. After normalizing the data, I performed the classification. Since I was unsure of the optimal number of clusters, I plot the Elbow Method (SSE) and the Silhouette Score to determine the appropriate number of clusters (Avcı Azkeskin and Aladağ, 2025). I do this because if K is too small, the data may be sparse and include noisy or mislabeled points. If K is too large, the neighborhood may contain outliers from other classes (Imandoust and Bolandraftar, 2013).

From Figure 3.5, it can be seen that when $SSE=4$, the decline in the magnitude decreases, indicating a turning point. In the Silhouette Score plot, the clustering results are quite good and the scores are both higher and more stable when $K = 3$ or $K = 4$. Therefore, dividing into 4 clusters seems to be the most appropriate choice, which means the parameter K is set to 4, so the classification of an unclear sample is based on the 4 nearest samples (Imandoust and Bolandraftar, 2013). The formulas of KNN will be included in A.3.

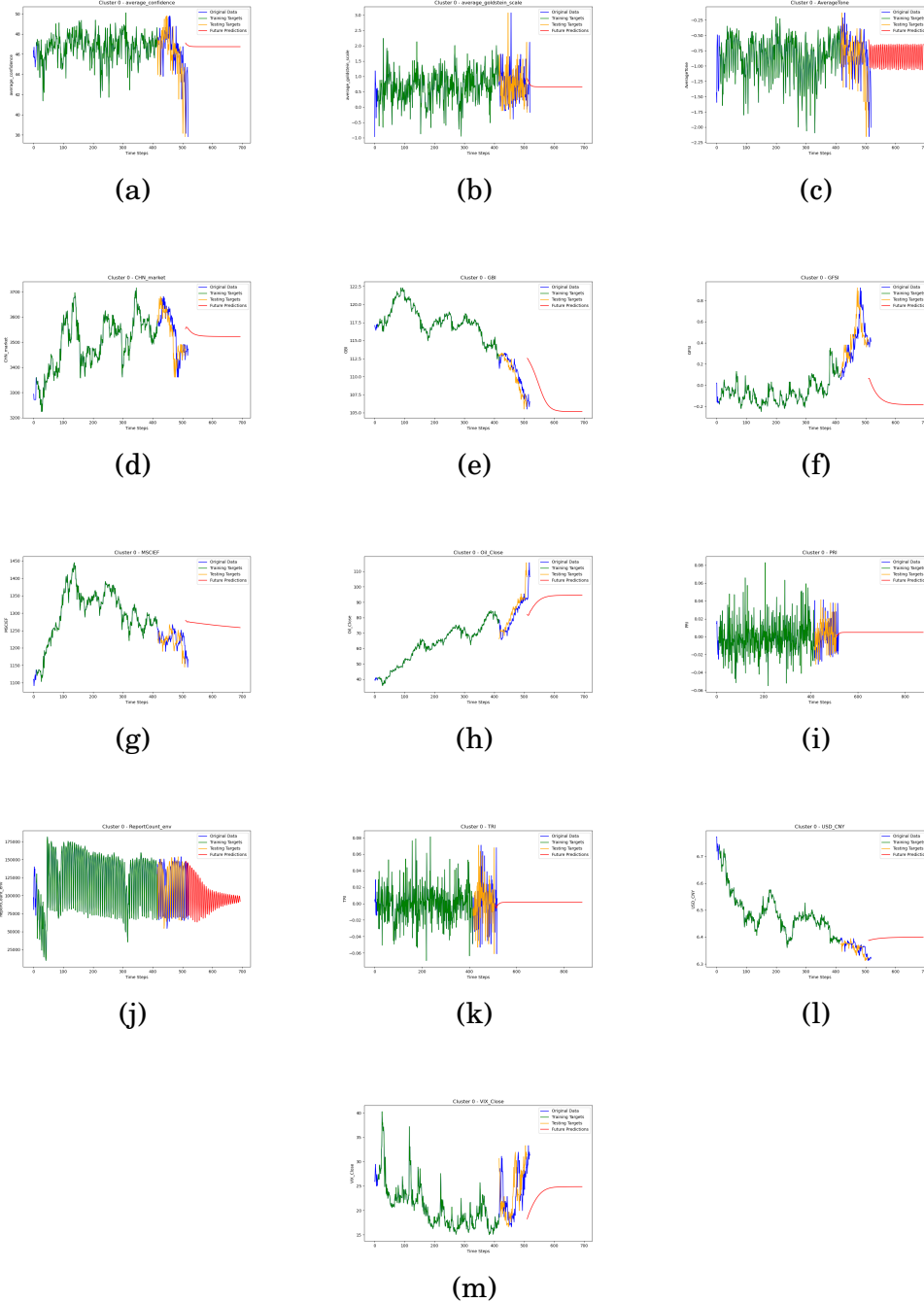
After the entire historical dataset is divided into 4 clusters, I use the LSTM model for each feature to generate simulated data for each cluster. The formulas and parameter tuning sections for the models will be included in A.4.

The basic LSTM model will be discussed in detail in A.7. Based on the para-

CHAPTER 3. SUPERVISED LEARNING MODELS, STATISTICAL MODELS OR HYBRID MODELS? A PREDICTION OF CLEAN ENERGY STOCK BASED ON FEAR AND FUNDAMENTAL FACTORS

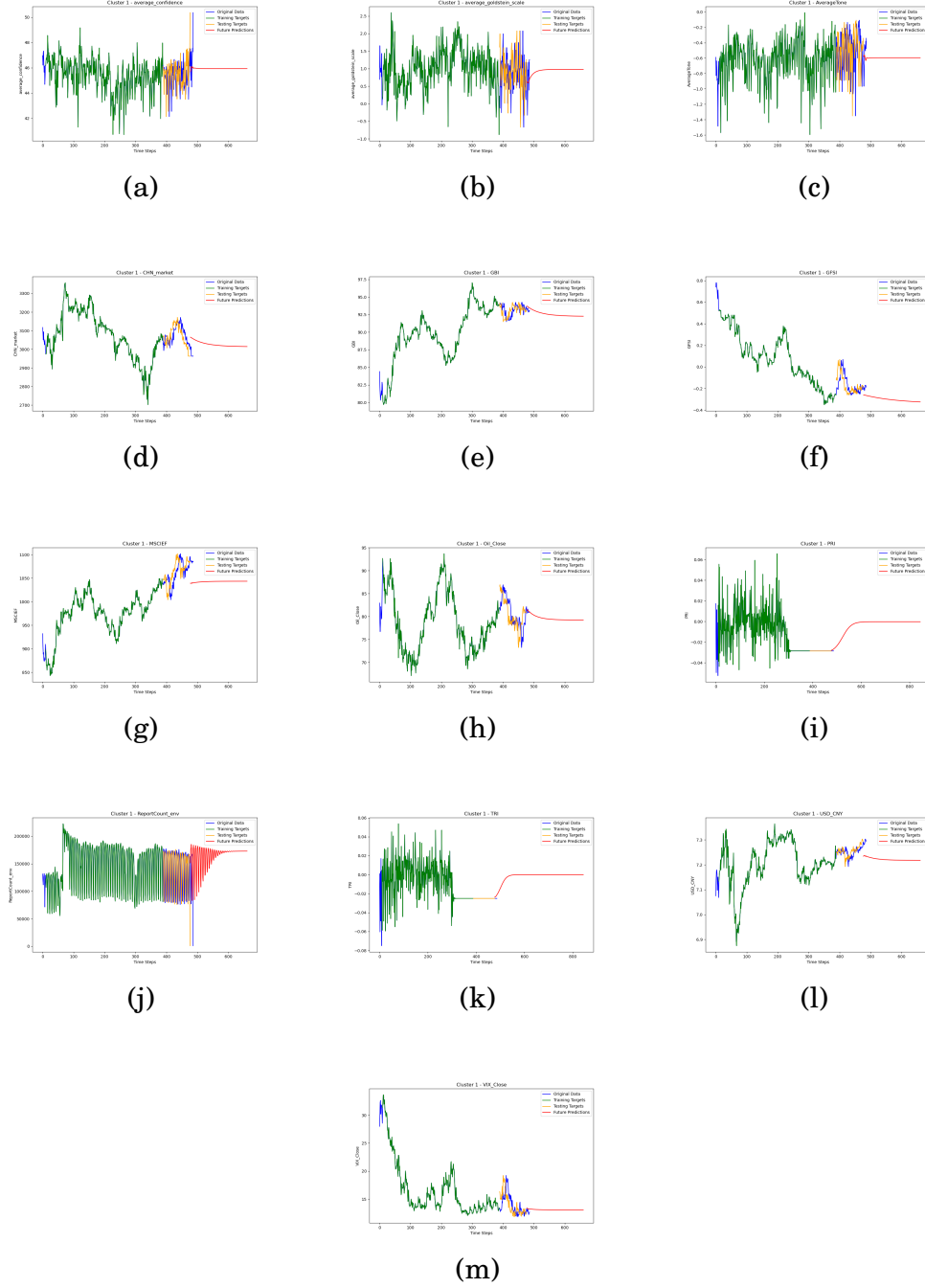
meters in Equation A.7, I will plot the results of these features across the 4 clusters (Figure 3.6 - Figure 3.9) to visualize the model's training performance.

Figure 3.6: Cluster 0



The trends of the training set and test set are quite close, which proves that the model has a good fit. However, there are still slight signs of overfitting, such as

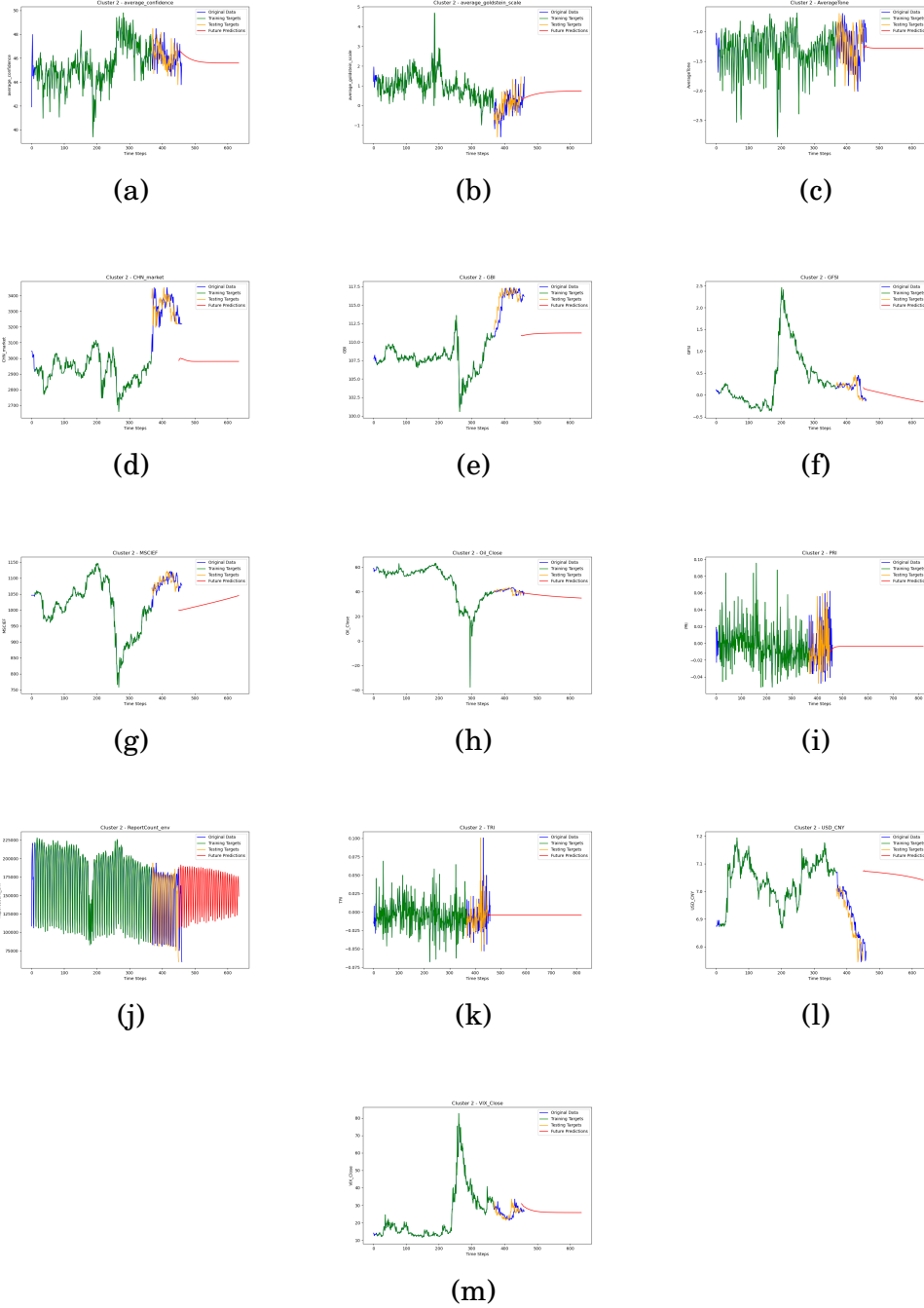
Figure 3.7: Cluster 1



the simulated predicted values of TRI and PRI being almost a flat line. However, since the values of these two variables are not particularly large, the results of this simulation can still be accepted. In order to combine the results from all clusters, I perform an averaging process on the simulated results of each variable across

CHAPTER 3. SUPERVISED LEARNING MODELS, STATISTICAL MODELS OR HYBRID MODELS? A PREDICTION OF CLEAN ENERGY STOCK BASED ON FEAR AND FUNDAMENTAL FACTORS

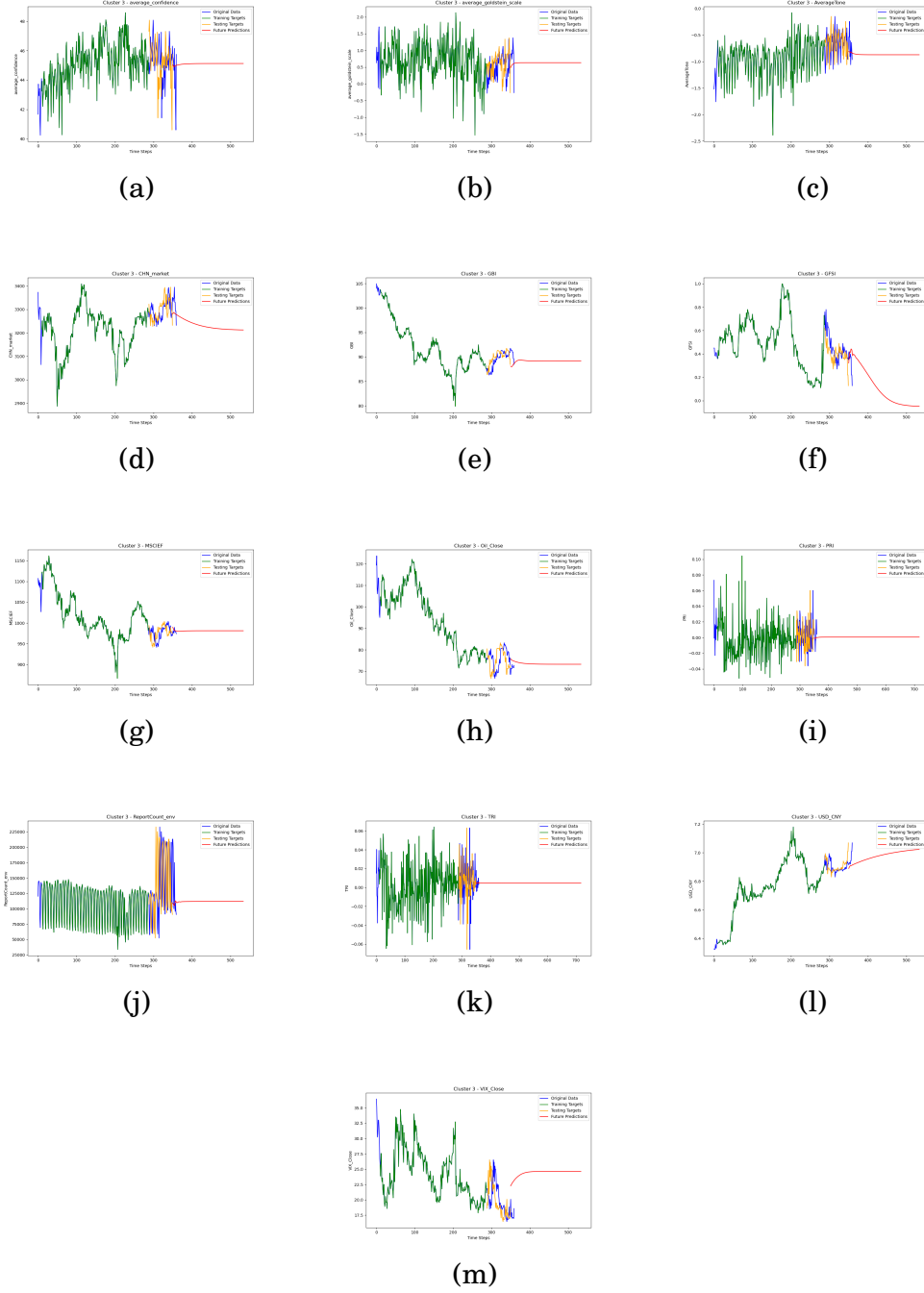
Figure 3.8: Cluster 2



different clusters. The averaged values were then used for subsequent predictions. Finally, the simulated data for each feature is the average value of the data from the same day across the 4 clusters.

Using the these two methods, I generated two sets of simulated data. After

Figure 3.9: Cluster 3



filling in all the missing values, I standardized the variables USD_CNY, MSCIEF, and CHN_market, which had multicollinearity issues, and used PCA to reduce them to two dimensions. The new variables will be renamed as PC1 and PC2. I will use this reduced data in my models when necessary. Next, when using these

CHAPTER 3. SUPERVISED LEARNING MODELS, STATISTICAL MODELS OR HYBRID MODELS? A PREDICTION OF CLEAN ENERGY STOCK BASED ON FEAR AND FUNDAMENTAL FACTORS

features to predict the S&P Clean Energy Index, I will split the data into two groups for prediction, and finally evaluate the performance of the target variable prediction.

3.3.2 Statistical Models Forecasting

3.3.2.1 ARIMA-Univariate Linear Model

I start by using a statistical model to make a simple prediction of the S&P clean energy index. For the ARIMA model, it is not possible to include additional features in forecasting; instead, the model relies solely on the historical five-year fluctuations of the S&P clean energy index to predict future trends. This makes ARIMA a solid "baseline" model for my study. If subsequent models that incorporate features or employ more complex techniques fail to outperform the ARIMA model, I would conclude that those models are not effective in predicting the S&P clean energy index. The formulas and parameter tuning sections of ARIMA will be placed in A.5.

The model uses the past 6 values of S&P Clean Energy Index, first-order differencing and a first-order lag, I predicted the clean energy index for the next six months. The results are shown in Figure 3.10.

3.3.2.2 VAR-Multivariate linear Model

Since ARIMA does not account for the feature variables used in forecasting, I consider the VAR model in similar linear models based on time series forecasting. The VAR model can incorporate all historical feature variables used to predict the S&P Clean Energy Index, due to missing values in the last six months of historical data for TRI and PRI, which were generated using simulated data, I divided the VAR model's predictions into two scenarios: one based on Random Forest and the other based on LSTM. Although there are two scenarios, the methods for data processing and modeling are the same. Therefore, I only present the results separately.

When preprocessing the data, a log transformation was first applied to the variables to address nonstationarity. Specifically, the variables ReportCount_env and S&P Global Clean Energy Index were logtransformed because of their large values. Subsequently, drawing on the experimental methodology outlined in Afzali et al. (2021), an ADF test was performed on all columns, with the results presented

in Table A.1 and A.2. The formulas and parameter tuning sections of VAR will be placed in A.6.

After ensuring that the processed data table contains all the historical features used in model training, the VAR model was executed. I aim to use the VAR model to predict the changes in the S&P Global Clean Energy Index over the next six months, which is equivalent to 184 periods. After setting the forecast steps to 184 and outputting the forecast results into a new Data Frame, the log-differencing of the $D(\log(\text{SP_GCEI}))$ is reversed, and the original values are restored by undoing the log transformation. Finally, the predicted S&P Global Clean Energy Index for past 5 years and the next six months are plotted, and the resulting figures are shown in Figure 3.11 and 3.12.

3.3.3 Supervised Learning and Hybrid Models

3.3.3.1 LSTM-Multivariate Nonlinear Model

Since VAR model is primarily used for linear time series analysis, and some of the previous data showed non-stationarity, I seek a model that can capture the dynamics of the S&P Clean Energy Index without the need for differencing or log transformation of non-stationary variables. Therefore, all historical and future features collected in the 3.3.1.7 are input into the LSTM model. The target variable defined in this context, representing the dependent data I aim to forecast, corresponds to the S&P Clean Energy Index.

Guo and Lin (2018) propose an interpretable multivariate LSTM for time series with exogenous variables. Based on the tensorized hidden states in MV-LSTM, they developed a hybrid temporal and variable attention mechanism, which can infer and quantify the importance of variables for the target sequence. The formulas and parameter tuning sections for LSTM will be included in the A.7.

3.3.3.2 LSTM_Autoregressive-Multivariate Nonlinear with regression trends Model

This hybrid model first defines an Autoregressive layer outside the LSTM model, LSTM already has the ability to remember and capture long-term dependencies. Adding an autoregressive (AR) layer is an additional adjustment to the LSTM output. It doesn't change the LSTM's core functionality. In addition to using historical data, the AR layer also takes the previous prediction as input. This way, the model

CHAPTER 3. SUPERVISED LEARNING MODELS, STATISTICAL MODELS OR HYBRID MODELS? A PREDICTION OF CLEAN ENERGY STOCK BASED ON FEAR AND FUNDAMENTAL FACTORS

relies not only on past data but also on previous predictions, combining the memory capability of LSTM with the autoregressive nature of AR (Zhang et al., 2023b). The formulas and parameter tuning sections of the models will be placed in A.8.

After building the LSTM_Autoregressive model, I obtained the results for both the Figure 3.19 and Figure 3.20 scenarios.

3.3.3.3 SAE-Multivariate Nonlinear Model

Previously, I employed statistical models based on linear relationships and supervised learning approaches with predefined target variables to forecast the S&P Clean Energy Index. AutoEncoder, in particular, is well suited for forecasting such financial market data. In deep learning, the autoencoder automatically extracts target features, addressing the limitations of traditional manual feature extraction methods. It also helps to effectively prevent overfitting (Li et al., 2023b). The formulas and parameter tuning sections for SAE will be included in A.9.

The final output layer gives a single predicted value for the sequence. I plot the obtained results and display them as Figure 3.15 and Figure 3.16.

3.3.3.4 Prophet Model Based on Historical Trends-Univariate Nonlinear Model

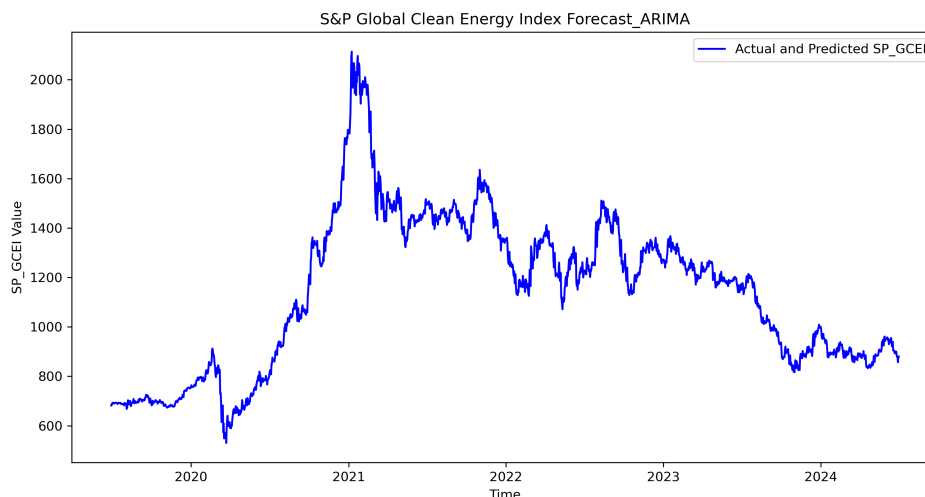
The Prophet model is mostly used for prediction based on the historical values of the target variable. It cannot directly handle the complex interdependencies between multiple variables, but it can make relatively accurate predictions of values and trends, and also generate more accurate graphs estimating the general range of fluctuations (Wei, 2025). The formulas and parameter tuning sections of the models will be placed in A.10.

After obtaining the optimal parameters for this round of the model, the results I obtained are shown in Figure 3.17 and 3.18.

3.3.3.5 Prophet Model Based on Generated Features-Multivariate Nonlinear with regression trends Model

Although previous studies have not explored the complex interdependencies between multiple variables, I can still incorporate each historical and future variable as features in the Prophet model. This approach allows Prophet to account for their individual effects (Wei, 2025). For multi-factor forecasting, if there are complex in-

Figure 3.10: S&P Global Clean Energy Index Forecast_ARIMA



interactions or feedback effects between the factors, models like VAR or LSTM might be better options. In my case, I use a method similar to LSTM_Autoregressive by adding a regression layer on top of the Prophet model. Regression in this context means treating each variable separately as an external factor to fine-tune the model. The formulas and parameter tuning sections for the models will be included in A.11.

The final results are presented in two scenarios, as shown in Figures 3.21 to 3.24.

3.3.4 Results and Comparative Analysis

3.3.4.1 Linear Trends Results

From the forecast for the next six months, I can see that the S&P Clean Energy Index shows a consistent downward trend, aligning with its fluctuations over the past five years. Toward the end of the forecast period, the ARIMA model predicts values that begin to stabilize. While this stabilization may not be clearly visible in Figure 3.10, it is evident in Figure 3.25.

As shown in Figure 3.11 and 3.12, The S&P Global Clean Energy Index initially shows a declining trend, followed by a period of recovery and steady growth. Toward the end of the year, the index stabilizes and gradually decreases, ending at a value

CHAPTER 3. SUPERVISED LEARNING MODELS, STATISTICAL MODELS OR HYBRID MODELS? A PREDICTION OF CLEAN ENERGY STOCK BASED ON FEAR AND FUNDAMENTAL FACTORS

Figure 3.11: S&P Global Clean Energy Index Forecast_VAR(Random Forest)

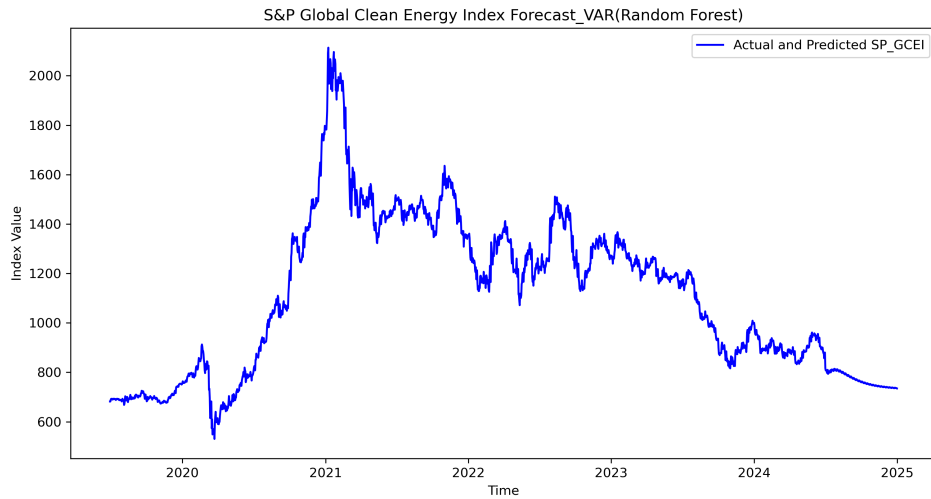


Figure 3.12: S&P Global Clean Energy Index Forecast_VAR(LSTM)

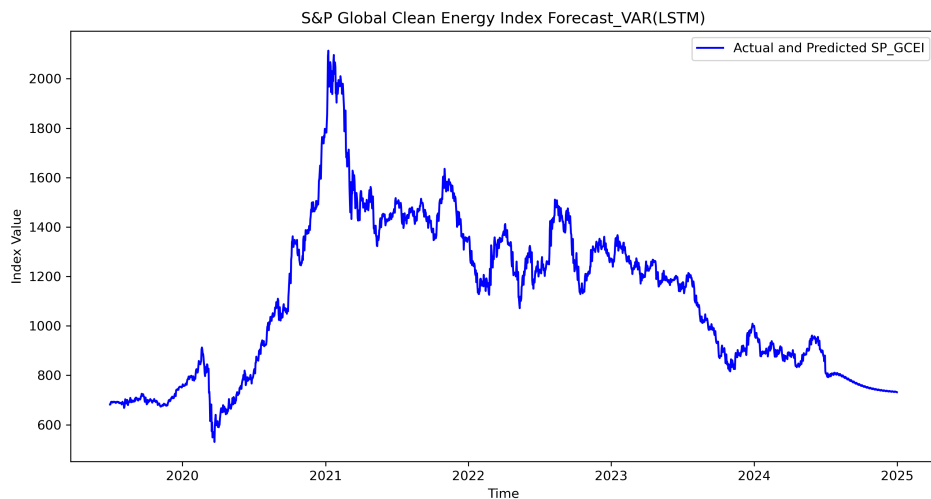
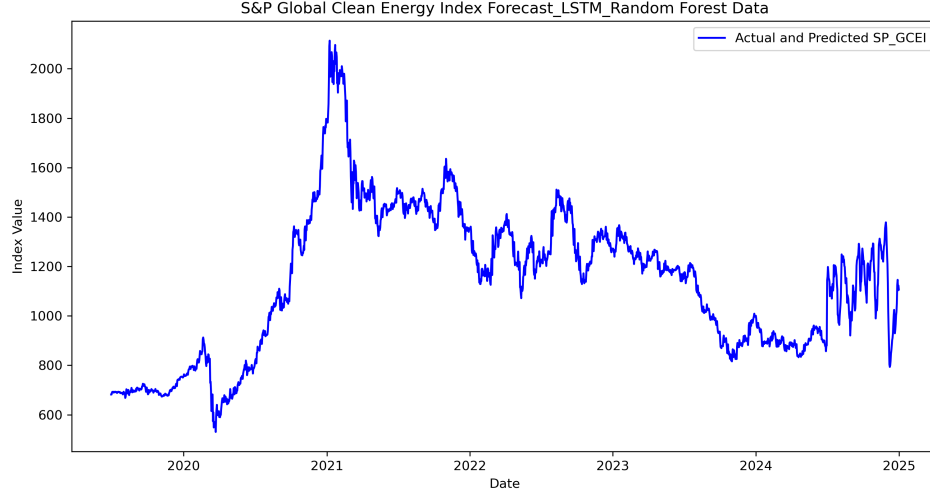


Figure 3.13: S&P Global Clean Energy Index Forecast_LSTM(Random Forest)



of 734.67 and 729.26 by the end of 2024.

Overall, ARIMA and VAR provided a foundational understanding of linear trends but were insufficient for capturing the non-linear behavior observed in the data.

3.3.4.2 Non-linear Trends with Supervised Learning Models

Building on this, LSTM-related models demonstrated an ability to model non-linear relationships.

I could obtain the LSTM performance based on the previous two sets of simulated data, as shown in the Figure 3.13 and 3.14.

It can be observed that the simulated data based on the Random Forest model seems to be better suited for this LSTM model. The prediction shows that the S&P Clean Energy Index fluctuates significantly during the forecast period, which may be helpful for predicting the trend of the S&P Clean Energy Index. However, the simulated data based on the LSTM model presents a risk of overfitting, which could be due to the fact that LSTM was already used to predict these features earlier, and the current data scale may be too small for the LSTM model. Regardless of the group of simulated data, there is an issue with the predicted values being overly large. To address the issues of overfitting and the adaptability of financial market data (Li et al., 2023b), I introduced the SAE model and obtained the results shown

CHAPTER 3. SUPERVISED LEARNING MODELS, STATISTICAL MODELS OR HYBRID MODELS? A PREDICTION OF CLEAN ENERGY STOCK BASED ON FEAR AND FUNDAMENTAL FACTORS

Figure 3.14: S&P Global Clean Energy Index Forecast_LSTM(LSTM)

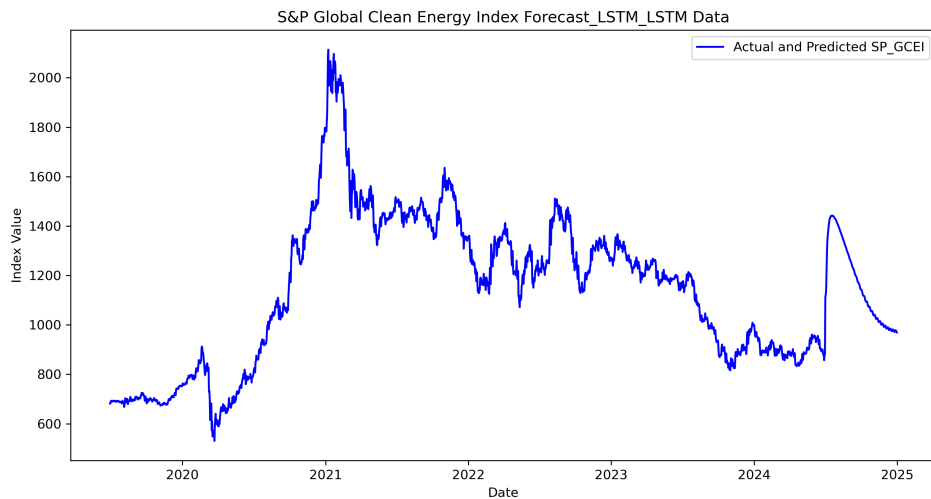
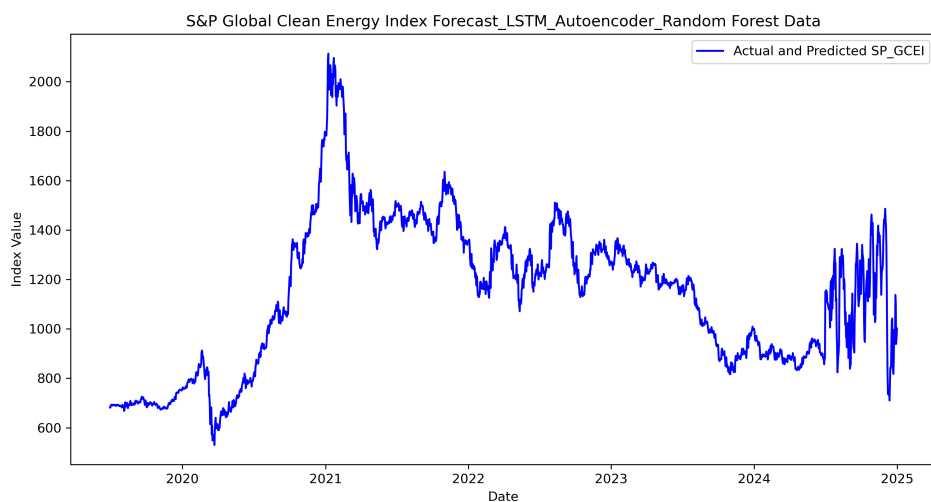


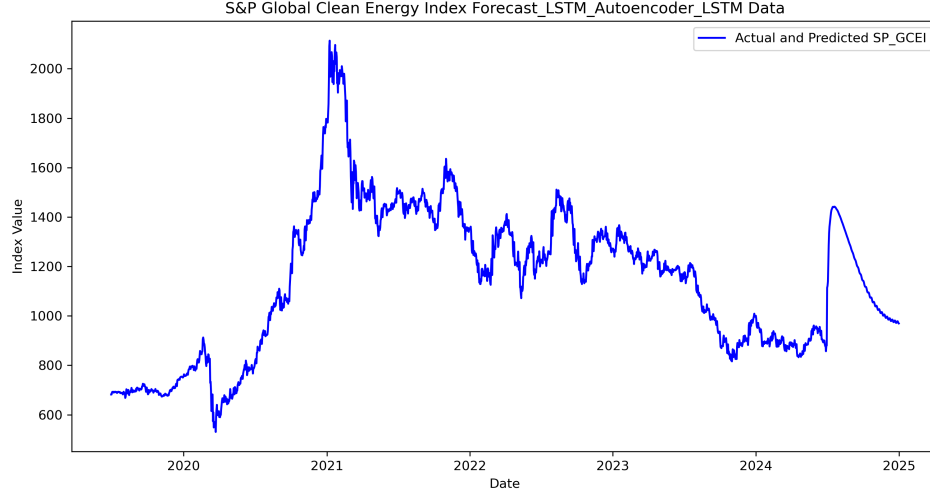
Figure 3.15: S&P Global Clean Energy Index Forecast_LSTM_Autoencoder(Random Forest)



in Figure 3.15 and 3.16.

The results appear to be largely consistent with a standard two-layer LSTM model, and the data simulated by Random Forest is better trained in the LSTM-based models. Currently, both the encoder and decoder processes consist of only two LSTM layers. When I attempt to add additional layers to both the encoder

Figure 3.16: S&P Global Clean Energy Index Forecast_LSTM_Autoencoder(LSTM)



and decoder processes, the model's performance worsened, likely due to overfitting caused by the limited size of the dataset. Thus, I could reasonably doubt that when using daily average media-related data and key energy market-related variables to predict the specific value of the S&P Clean Energy Index, LSTM-based models exhibit poor predictive performance. However, when the objective is to predict trends, incorporating Random Forest simulated data with a basic LSTM model may be a valuable choice.

Considering that a small dataset scale can easily lead to overfitting in LSTM, I try the Prophet model and get Figure 3.17 and 3.18, which can similarly capture nonlinear trends.

The prediction results show that the S&P Clean Energy Index follows an overall downward trend during the forecast period. In Figure 3.18, the black dots represent the original time series data points, the blue line indicates the predicted values of the model, and the light blue shaded area represents the uncertainty range of the predictions. The red dashed lines mark key events or trend points, while the red line segments are fitted curves based on fluctuations. The chart also extends to show the fluctuation range for a short period beyond 2025, suggesting that the S&P Clean Energy Index is expected to be around 750 by the end of 2024 and is likely to continue its gradual decline in the future.

Figure 3.17: S&P Global Clean Energy Index
Forecast_Prophet_without_other_features

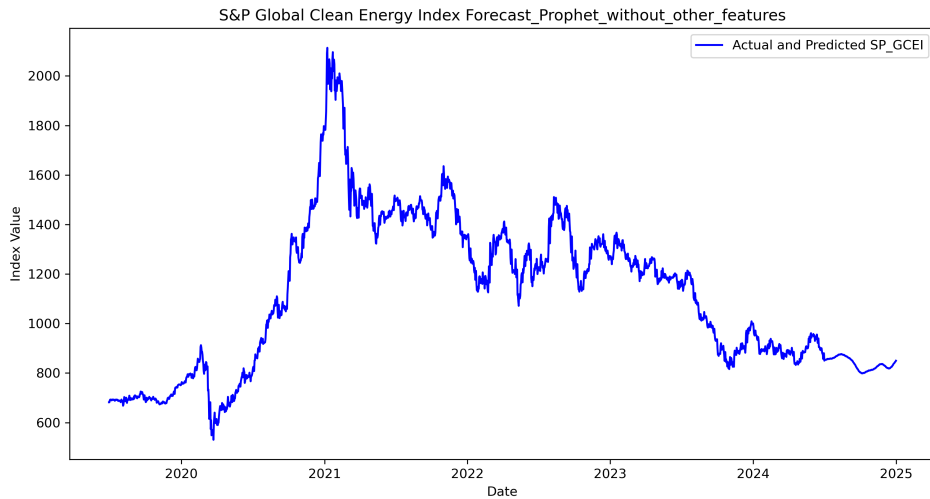


Figure 3.18: Prophet Time Series Forecasting Plot

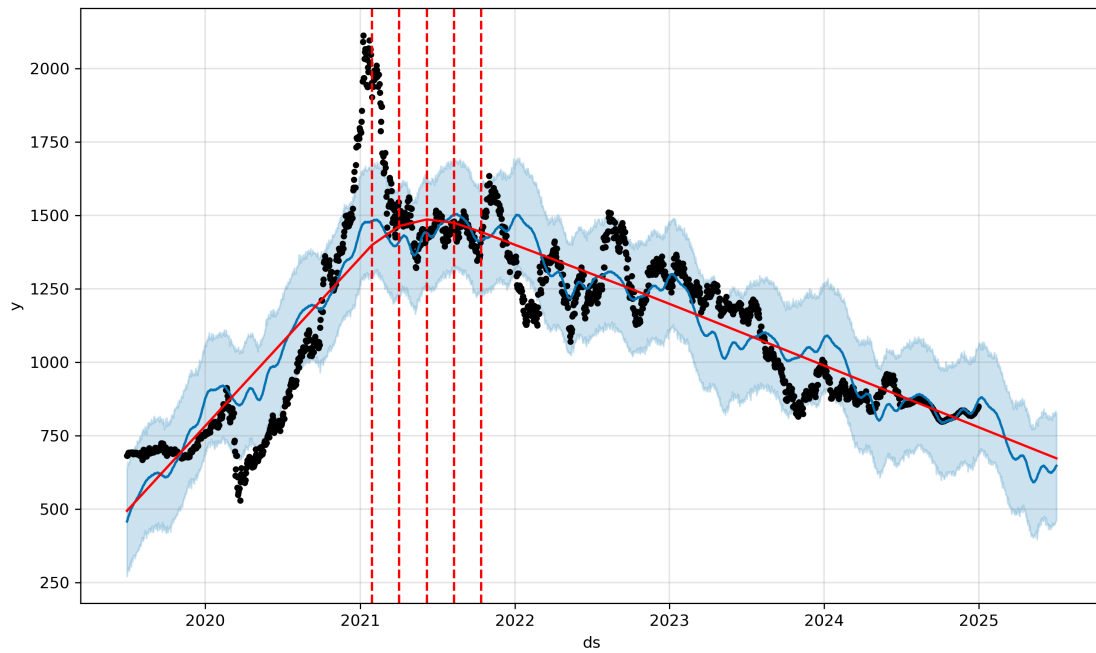
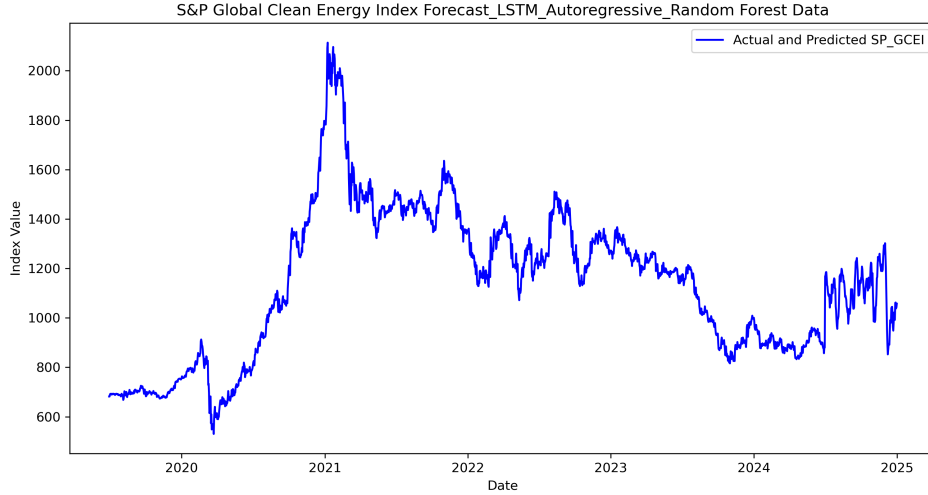


Figure 3.19: S&P Global Clean Energy Index
Forecast_LSTM_Autoregressive(Random Forest)



3.3.4.3 Non-linear Trends with Hybrid Models

To address the risk of overfitting and integrate the regression trends, I plan to add an Autoregressive layer on top of the LSTM model, creating a hybrid LSTM_Autoregressive model, and see if the prediction performance improves.

From the results, it appears that the LSTM_Autoregressive hybrid model performs similarly, whether the simulation data are based on Random Forest or LSTM. Compared to Figure 3.13 and 3.14, the predicted values in this hybrid model have generally decreased. The model still overfits when using the data generated based on LSTM simulations. Since the basic Prophet results were quite good previously, I want to explore whether combining both historical and future features could improve the model's performance.

From Figure 3.21, there are several outliers in the predictions based on the Random Forest simulated dataset (although extreme outliers have been partially handled during the forecasting process), while the overall values predicted using the LSTM simulated data (Figure 3.23) tend to be lower than expected. Besides, the trends for the S&P Clean Energy Index predicted in both scenarios are quite similar, with both showing an overall downward trend, despite a brief period of increase.

Looking at Figure 3.22 and Figure 3.24, the confidence interval in Figure 3.24

Figure 3.20: S&P Global Clean Energy Index
Forecast_LSTM_Autoregressive(LSTM)

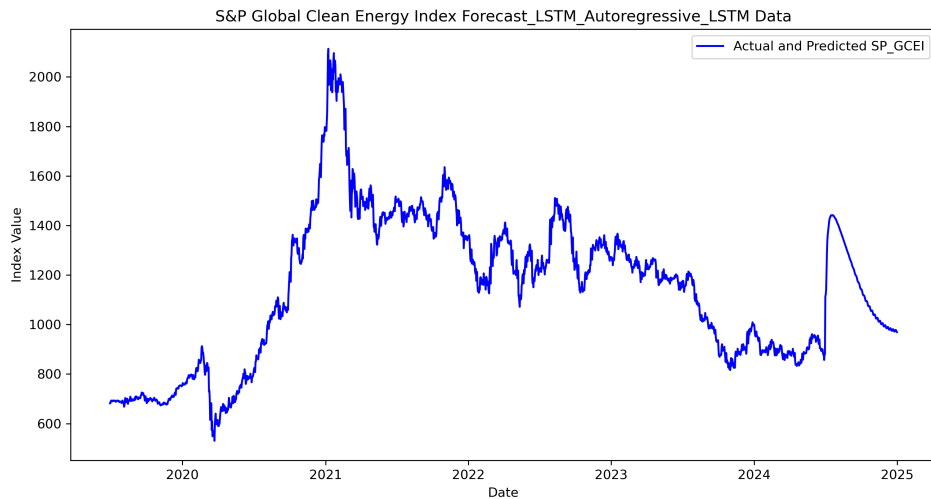


Figure 3.21: S&P Global Clean Energy Index Forecast_Prophet_Features (Random Forest)

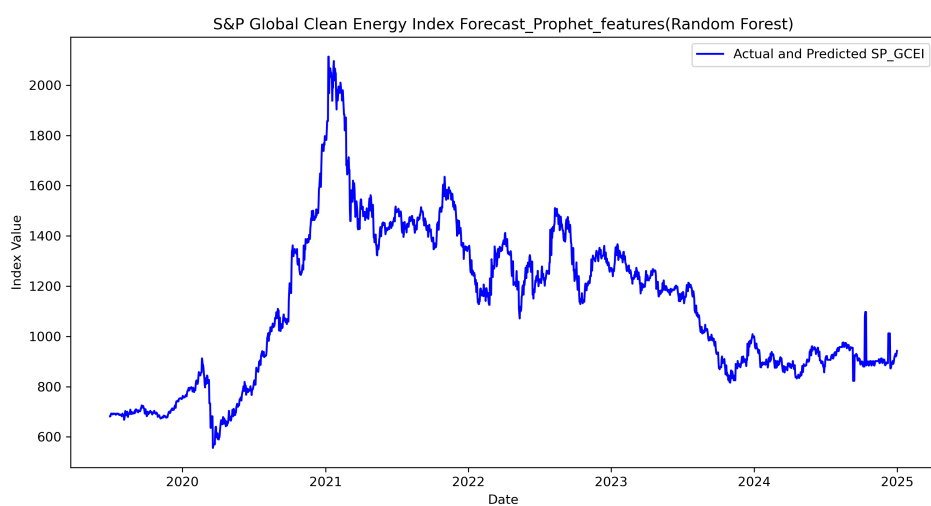


Figure 3.22: Prophet Time Series Forecasting Plot_Random Forest

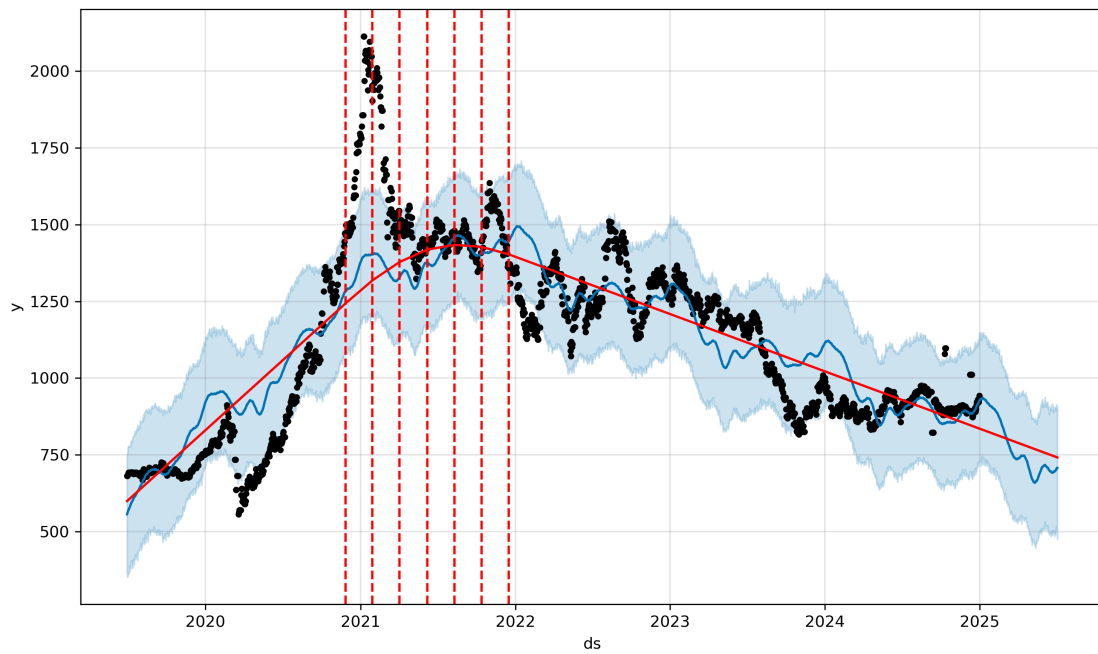


Figure 3.23: S&P Global Clean Energy Index Forecast_Prophet_Features_LSTM)

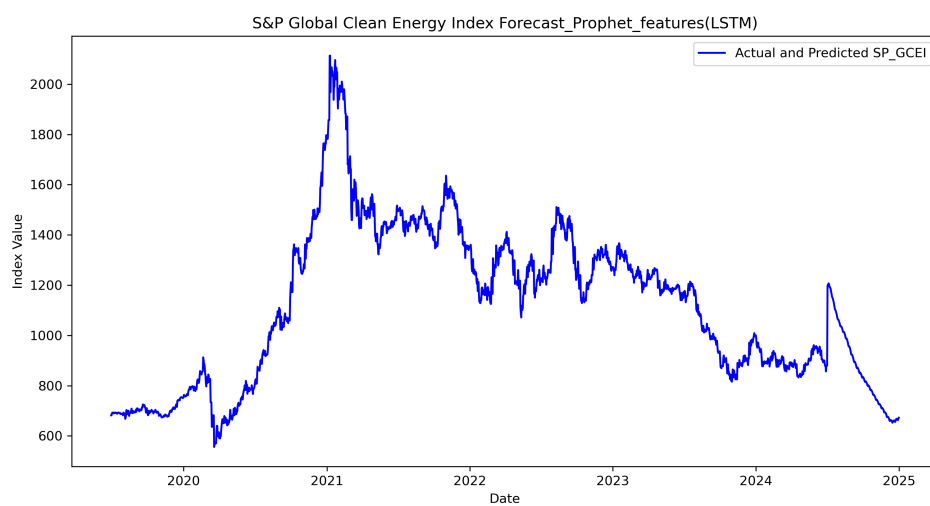
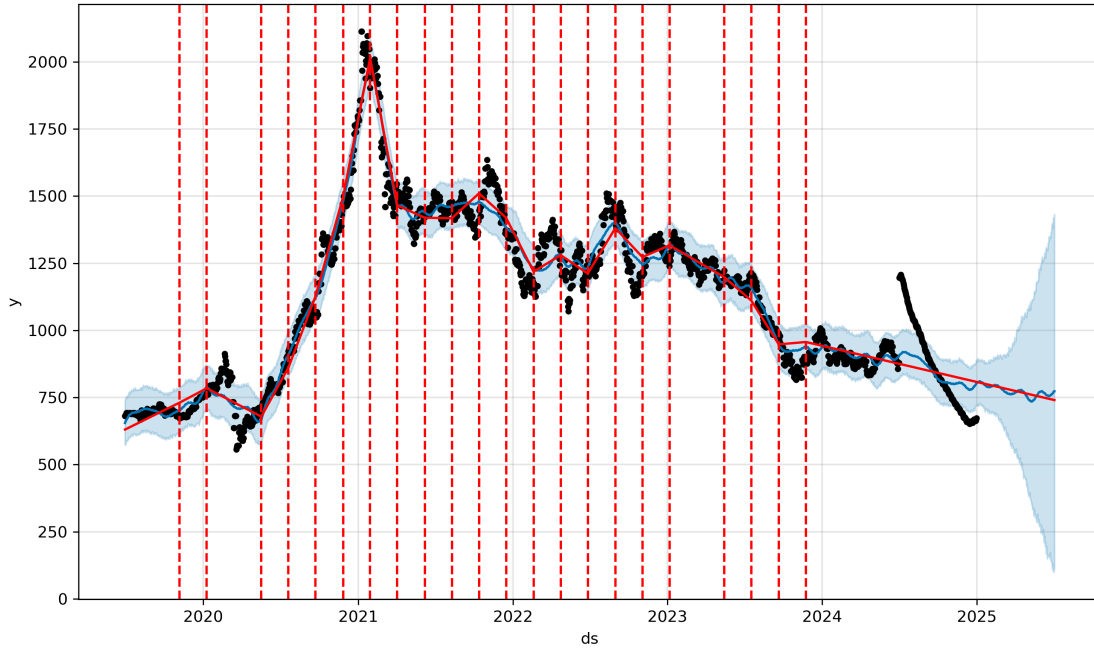


Figure 3.24: Prophet Time Series Forecasting Plot_LSTM



has a very large fluctuation range. This overfitting issue has been discussed in 3.3.4.2, although the value of the blue line around early 2025, or late 2024, their predicted values are near 750.

When comparing Figure 3.18 and Figure 3.22, the predicted trend and value range are nearly identical, indicating that the impact of the characteristics in the Prophet model is not particularly significant. This is also related to the characteristics of the Prophet model, as it cannot capture the interdependencies between multiple variables (Wei, 2025).

3.3.4.4 Prediction Comparisons-Whole Prediction Period

Subsequently, I compare the prediction results of all the prediction models by evaluating their performance using MSE, RMSE, and MAE over the prediction period. Additionally, I visualize the trends and compare them against the actual values. The ARIMA and Prophet basic models are applied to both simulated datasets to facilitate a more comprehensive examination of the role of features and to assess the performance of each model in forecasting the S&P Clean Energy Index. The performance of different models under the two sets of simulated data are Tables 3.3 and 3.4:

Figure 3.25: Prediction Comparison Plot(Random Forest)

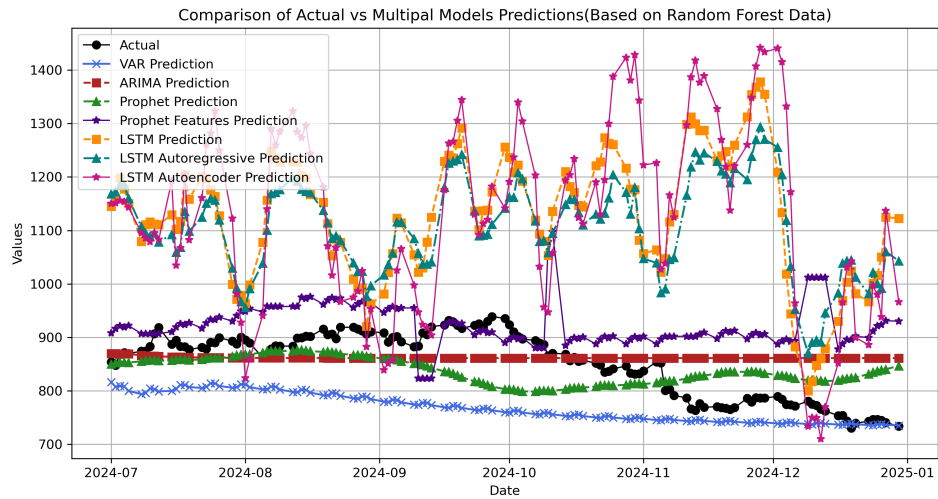
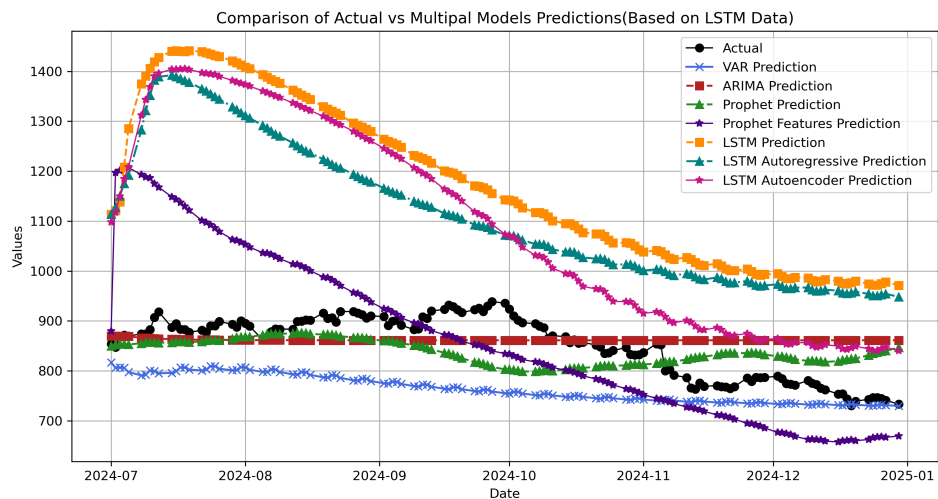


Figure 3.26: Prediction Comparison Plot(LSTM)



CHAPTER 3. SUPERVISED LEARNING MODELS, STATISTICAL MODELS OR HYBRID MODELS? A PREDICTION OF CLEAN ENERGY STOCK BASED ON FEAR AND FUNDAMENTAL FACTORS

Table 3.3: Model Comparison Metrics (Random Forest)

Panel A: Model Comparison Metrics for Random Forest							
Metric	VAR	ARIMA	Prophet	Prophet Features	LSTM	LSTM Autoregressive	LSTM Autoencoder
MSE	8954.97	3693.89	3495.17	9735.44	91555.12	75405.90	111967.86
RMSE	94.63	60.78	59.12	98.67	302.58	274.60	334.62
MAE	83.21	50.36	50.74	78.94	274.11	254.76	284.95

Panel B: Model Comparison Metrics for Random Forest_7							
Metric	VAR	ARIMA	Prophet	Prophet Features	LSTM	LSTM Autoregressive	LSTM Autoencoder
MSE	3984.05	125.23	214.81	2504.31	80152.58	84317.64	73077.06
RMSE	63.12	11.19	14.66	50.04	283.11	290.37	270.33
MAE	61.10	8.17	13.44	48.66	280.14	287.61	268.18

Panel C: Model Comparison Metrics for Random Forest_30							
Metric	VAR	ARIMA	Prophet	Prophet Features	LSTM	LSTM Autoregressive	LSTM Autoencoder
MSE	6055.00	614.06	663.92	2681.65	66543.03	54061.55	73787.29
RMSE	77.81	24.78	25.77	51.78	257.96	232.51	271.64
MAE	76.19	21.16	22.31	47.52	244.03	220.62	247.04

I first discuss the prediction performance of simulating this dataset using the Random Forest method. Table 3.3 presents the MSE, RMSE, and MAE, which are commonly used metrics to assess the error between predicted and actual values. These metrics cover the errors between the predictions and actual values for all models over a forecast period of 184 days. Overall, the basic Prophet model exhibits the smallest error, followed by ARIMA and VAR, while the LSTM-related models perform the worst.

However, evaluating the performance of a model solely based on overall error is not completely comprehensive. It is also essential to examine the trend graphs and final predicted values during the forecast period. At the end of 2024, the VAR model's predictions closely align with the actual values, indicating that the features play a crucial role in the model's accuracy. In contrast, for the Prophet model, the predictions remain fairly accurate whether or not features are included. However, to assess its forecast for the next six months, I primarily rely on the prediction interval Figure 3.18 and Figure 3.22.

An interesting observation arises from the Figure 3.25: When I shift the Prophet model with features slightly to the right by a few days, its trend aligns closely with the actual values. This indicates that the model can predict market dynamics ahead of time, likely due to the inclusion of media factors, as market reactions often

Table 3.4: Model Comparison Metrics (LSTM)

Panel A: Model Comparison Metrics for LSTM							
Metric	VAR	ARIMA	Prophet	Prophet Features	LSTM	LSTM Autoregressive	LSTM Autoencoder
MSE	9753.45	3693.89	3495.17	17104.65	116327.99	78405.12	84256.08
RMSE	98.76	60.78	59.12	130.78	341.07	280.01	290.27
MAE	87.81	50.36	50.74	105.53	318.56	259.59	243.22
Panel B: Model Comparison Metrics for LSTM_7							
Metric	VAR	ARIMA	Prophet	Prophet Features	LSTM	LSTM Autoregressive	LSTM Autoencoder
MSE	4259.77	125.23	214.81	94157.18	145518.22	112602.96	119484.10
RMSE	65.27	11.19	14.66	306.85	381.47	335.56	345.66
MAE	63.07	8.17	13.44	287.48	367.54	329.05	336.36
Panel C: Model Comparison Metrics for LSTM_30							
Metric	VAR	ARIMA	Prophet	Prophet Features	LSTM	LSTM Autoregressive	LSTM Autoencoder
MSE	6676.72	614.06	663.92	55550.19	250036.98	184138.04	216019.68
RMSE	81.71	24.78	25.77	235.69	500.04	429.11	464.78
MAE	80.03	21.16	22.31	223.59	492.45	423.51	458.01

exhibit a delay. Although the Prophet Features models still slightly overestimate, the forecast interval Figure 3.18 and Figure 3.22 provide a clearer understanding of the future trend of the S&P Clean Energy Index. Certainly, if there are concerns about significant deviations in the predictions, the forecast can be refined by incorporating insights from the ARIMA and VAR models.

Although the LSTM model demonstrates relatively poor performance in predicting the S&P Clean Energy Index, as discussed in 3.3.4.2, it is able to effectively capture market volatility trends when simulating data with Random Forest. In this context, it provides valuable insights and serves as a useful reference for predicting market trends. Whether it is LSTM, LSTM_Autoregressive, or Autoencoder, all three models show similar prediction curves, with their forecasts being relatively close to one another. However, in terms of the accuracy of the predicted values, LSTM_Autoregressive seems to perform better.

However, when applied to the LSTM-simulated data, the predictive performance is weaker, especially when using the LSTM model and its variants. In this case, overfitting becomes more evident, greatly reducing the reliability of the predictions for the S&P Clean Energy Index. Interestingly, even with the overfitting, the LSTM-related models still show a general downward trend, which matches the fluctuations in the actual values well.

CHAPTER 3. SUPERVISED LEARNING MODELS, STATISTICAL MODELS OR HYBRID MODELS? A PREDICTION OF CLEAN ENERGY STOCK BASED ON FEAR AND FUNDAMENTAL FACTORS

The performance of the VAR model is similar across both datasets, as VAR model only requires the last six months of missing historical data for the TRI and PRI variables.

The Prophet model, incorporating features as external factors, performs bad in LSTM simulated data because the predicted uncertainty intervals were too wide. When looking at the prediction line(blue) in Figure 3.24, the differences between the two sets of simulated data are small, suggesting that both can serve as useful points for future predictions.

3.3.4.5 Prediction Comparisons-Short and Medium-term Period

The 3.3.4.4 presents the results of various time series models based on a six-month forecast period. This forecast period is relatively long, but when evaluating the model's long-term forecasting ability, it is equally important to consider its short and medium-term forecasting performance. Therefore, I selected the first 7 steps and the first 30 steps of the forecast period to compare the short and medium-term forecasting abilities of the different models. It is important to note that the 7 steps and 30 steps here do not refer to the first 7 days or 30 days of the forecast period, as actual data is typically missing on weekends. As a result, the time labels in the figure are adjusted accordingly. Another reason for comparing short-term predictions is that LSTM-related models do not suffer from such severe overfitting in the early stages of the forecast period. The results will be shown in Table 3.3, Figure 3.27, Table 3.3, Figure 3.28, Table 3.4, Figure 3.29, Table 3.4, Figure 3.30:

First, I analysed the predictions based on the Random Forest simulated data. For the first 7 time steps, the best-performing model is ARIMA, followed by the Prophet-related models. In Figure 3.27, the predicted values of these three models are quite close to the actual values. However, when the forecast period is extended to 30 time steps, the Prophet-related models gradually catch up with ARIMA, while the predicted values from the LSTM-related models remain inaccurate.

When comparing the prediction trends of the models, I observed that, when the forecast period is less than 30 time steps, the Prophet model combined with features seems more suited to predicting trend changes. This challenges the previous understanding that many scholars aimed to address, such as the prediction lag issue mentioned by Li et al. (2023a). At this longer forecast horizon, the LSTM-related models fail to provide effective predictions.

Next, I analysed the predictions based on the LSTM simulated data. For the

Figure 3.27: Prediction Comparison Plot(Random Forest_7)

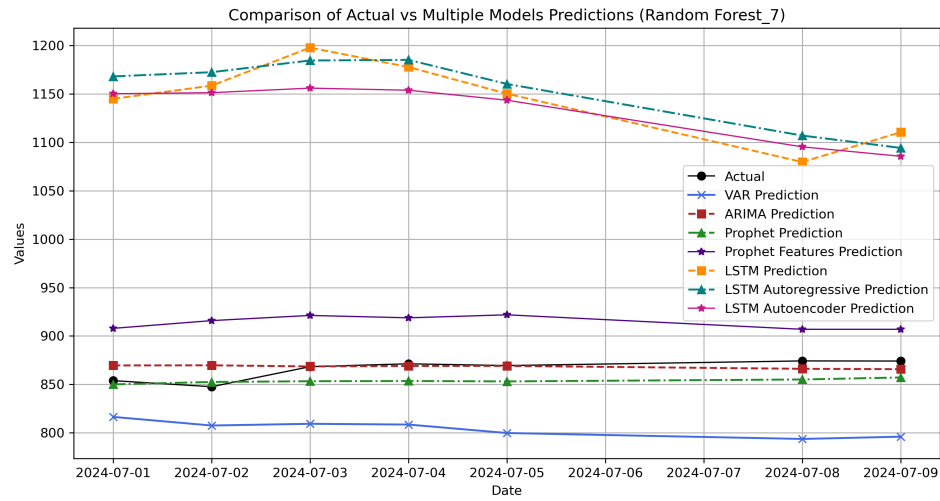
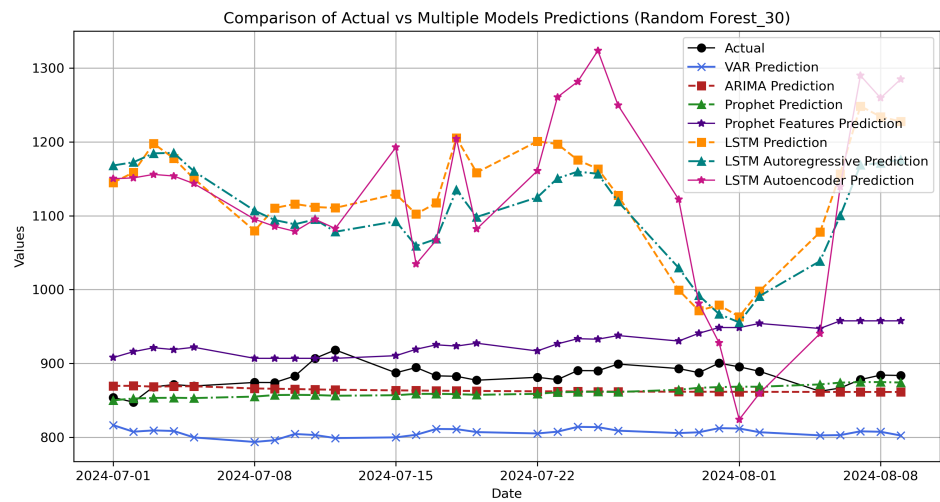


Figure 3.28: Prediction Comparison Plot(Random Forest_30)



CHAPTER 3. SUPERVISED LEARNING MODELS, STATISTICAL MODELS OR HYBRID MODELS? A PREDICTION OF CLEAN ENERGY STOCK BASED ON FEAR AND FUNDAMENTAL FACTORS

Figure 3.29: Prediction Comparison Plot(LSTM_7)

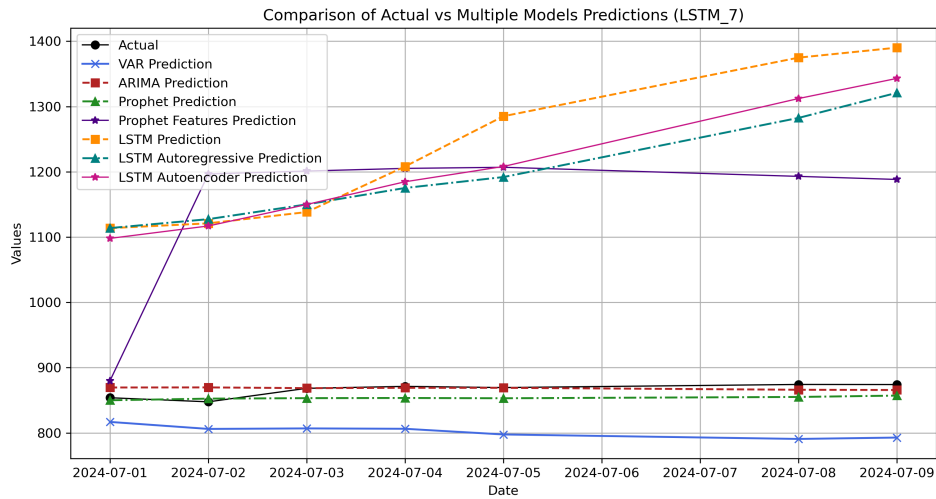
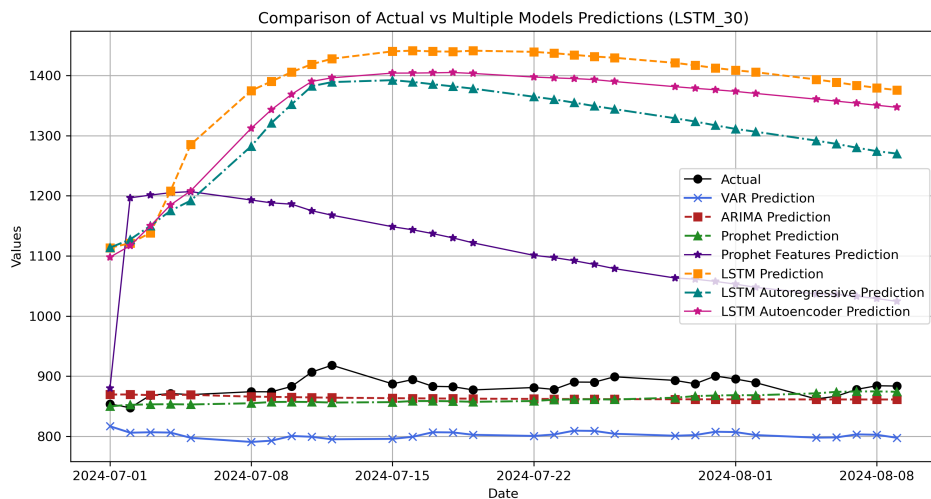


Figure 3.30: Prediction Comparison Plot(LSTM_30)



7 and 30-step forecast, the best-performing models remain ARIMA and Prophet. However, the Prophet model combined with features does not show significant improvement, indicating that, at the scale of this dataset, the data quality simulated by Random Forest is superior to that simulated by LSTM.

3.4 Conclusion and Contribution

3.4.1 Conclusion and Future Research

I discussed and compared seven time series models and two sets of generated simulation data. Based on the data scale available in this study, I draw the following conclusions, stakeholders can use the following strategies to make short-term, medium-term, and even long-term predictions for the clean energy market.

Short-term Forecasting (around one week): If the goal is to predict the S&P Clean Energy Index for the short term (about a week), ARIMA and Prophet are the most reliable models. In this case, it may not even be necessary to include additional features; the historical data of the S&P Clean Energy Index itself is sufficient to produce a good prediction range. However, if the aim is to predict trends, Prophet Features based on Random Forest simulated data provides more valuable insights.

Medium-term Forecasting (around one month): For a one-month forecast of the S&P Clean Energy Index, ARIMA can first be used to establish a "baseline" for the index, and then Prophet-related models can be combined to refine the prediction range. For this, the simulation data generated by Random Forest produces better results. If the primary goal is to predict trends, Prophet Features based on Random Forest-simulated datasets offer promising results.

Long-term Forecasting (around six months): For longer-term forecasts, such as the six-month forecast examined in this study, simulation data generated by Random Forest leads to higher overall prediction accuracy. The VAR model produces more accurate final predictions, while Prophet, when combined with additional features, offers a more forward-looking perspective, making it better suited for predicting trends.

Overall, if the goal is to forecast the S&P Clean Energy Index using this dataset, Prophet-related models are the most efficient and effective choice, as they can provide both a prediction range and an accurate representation of trends. If the

CHAPTER 3. SUPERVISED LEARNING MODELS, STATISTICAL MODELS OR HYBRID MODELS? A PREDICTION OF CLEAN ENERGY STOCK BASED ON FEAR AND FUNDAMENTAL FACTORS

objective is to generate simulated feature datasets, Random Forest emerges as the better option.

Furthermore, I have included a comprehensive summary table in Table 3.5, outlining the advantages, limitations, and overall accuracy of each model in predicting the S&P Clean Energy Index using this dataset, providing information for the reader to reference. Since it has been demonstrated that Random Forest-simulated data is more suitable for this context, the reported accuracy reflects the results obtained under the Random Forest-simulated data scenario.

From a forward-looking perspective, the forecasting framework proposed in this chapter has important future implications. Although the clean energy market is expected to continue evolving in terms of policy, technology, and market structure, the overall methodology remains applicable, as it is built around a flexible feature selection and preprocessing strategy rather than a fixed set of predictors. As new information sources and market-relevant variables emerge, the same feature screening procedures, including XGBoost model and SHAP-based validation, can be applied to update and refine the input feature set. This ensures that the forecasting models can adapt to structural changes in the clean energy market while maintaining interpretability and predictive reliability. Consequently, the proposed framework provides a scalable and robust approach for future clean energy market forecasting, even as the underlying drivers and information environment continue to change.

3.4. CONCLUSION AND CONTRIBUTION

Table 3.5: Comparison of Different Models with Their Strengths, Weaknesses, and Accuracy

Model	Strengths	Weaknesses	Accuracy
ARIMA	Simple, interpretable	Misses non-linear patterns, cannot use multiple variables	93.86%
VAR	Incorporates multi-variable data, final predicted value is the most accurate	Misses non-linear patterns	90.54%
LSTM	Incorporates multi-variable data, handles non-linearity well	Overfitting in small datasets	67.24%
LSTM_Autoregressive	Incorporates multi-variable data, handles non-linearity well, combine regression	Overfitting in small datasets	69.50%
LSTM_Autoencoder	Incorporates multi-variable data, handles non-linearity well	Overfitting in small datasets	65.87%
Prophet	Nonlinear time series data with yearly, weekly, daily seasonality, holiday effects	Cannot use multiple variables	93.99%
Prophet Features	Incorporates multi-variable data, handles non-linearity well, combine regression	Only apply the individual effect of each feature to the model	90.27%

3.4.2 Technical Contribution

The main technical contribution of this study is using media-related factors to predict the S&P Clean Energy Index. This approach improves prediction accuracy, especially for stakeholders who make decisions based on news. By including historical data of the index, I identify the best time series models for short-, medium-, and long-term forecasts. These models not only make predictions more accurate but also provide useful insights into future trends. I also provided new insights for analysts facing the "lag" issue when predicting indices. Through this approach in 3.4.1, they may achieve better solutions for lag-related feature or target variables.

Additionally, this study lays the groundwork for future research on the volatility of other financial market indices. The parameter-tuning techniques I introduced, along with various machine learning and statistical models, are expected to be more widely used in financial forecasting, marking an important step forward for the financial industry.

3.4.3 Theoretical Contribution

From a theoretical perspective, this study improves existing methods for predicting stocks and other markets by using statistical and machine learning models, which traditionally rely only on historical data. By incorporating both historical and future data, my models make more accurate predictions for the stock and energy markets. This helps investors make better decisions in both the short and long term, ultimately improving the quality and effectiveness of their investment strategies.

The S&P Clean Energy Index I forecast provides a clear picture of global clean energy trends. By understanding these patterns, I can focus on the key drivers of the clean energy market, ensuring steady and sustainable growth. This approach not only supports more effective use of clean energy but also plays a vital role in advancing environmental governance and long-term sustainability.

At the same time, my predictive insights can guide policymakers in designing more effective macro-level policies, thereby supporting the global energy transition and broader environmental objectives.

For policy implications, the evidence that media-related factors significantly enhance the predictive accuracy of the S&P Clean Energy Index highlights the importance of incorporating information-based indicators into policy monitoring and decision-making frameworks. For policymakers and regulatory agencies, enhanced

forecasting accuracy can support more timely and informed interventions in clean energy markets, particularly in periods of heightened uncertainty. The ability to distinguish short-, medium-, and long-term dynamics further enables policymakers to align policy instruments with different stages of the energy transition. Besides, the methodological advances in addressing lag-related prediction issues suggest that forward-looking policy analysis can be improved by integrating media signals alongside traditional historical data. At a broader level, more accurate forecasts of clean energy market trends can help policymakers better anticipate investment needs, design effective climate and energy policies, and support stable progress toward environmental governance and long-term sustainability goals.

3.5 Limitations

However, the dataset used in this study is limited to daily granularity and cannot be refined to hourly, minute, or second-level precision due to the varying update frequencies of each feature. If the dataset were available at a finer granularity or on a larger scale, the prediction accuracy could potentially be further improved. In addition, the LSTM model might demonstrate better performance when applied to larger-scale datasets.



EQUATIONS AND PARAMETERS TUNING

The equations and parameter tuning sections of all the models mentioned will be placed in the Appendix.

A.1 XGBoost

The XGBoost will be expressed by:

$$(A.1) \quad \mathcal{L} = \sum_{i=1}^n \ell(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

Where:

- $\ell(y_i, \hat{y}_i)$ is the loss function for the i -th sample.
- $\Omega(f_k)$ is the regularization term for the tree, which is typically defined as a function of tree complexity:

$$(A.2) \quad \Omega(f_k) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2$$

Where:

- T is the number of leaf nodes in the tree.
- w_j is the weight of the j -th leaf node.

- γ and λ are the regularization hyperparameters.

After defining the XGBoost model, I create parameter lists for `colsample_bytree`, `learning_rate`, `max_depth`, `min_child_weight`, `n_estimators`, and `subsample`. GridSearchCV is then employed for hyperparameter tuning to identify the optimal parameters for training.

A.2 Random Forest

First, I normalize these variables and then input them into the Random Forest model for simulation. To prepare the data for regression, I use a sliding window approach over the time series. The original time series be $S = [s_1, s_2, \dots, s_T]$, and the window size be L . The input-output pairs (X_i, Y_i) are generated as follows:

$$(A.3) \quad X_i = [s_i, s_{i+1}, \dots, s_{i+L-1}], \quad Y_i = s_{i+L}, \quad \forall i \in [1, T-L].$$

This results in the regression dataset:

$$(A.4) \quad \{(X_i, Y_i)\}_{i=1}^{T-L}.$$

For each decision tree, the training set is randomly selected from the dataset. In this study, decision trees use the technical indicators in the feature matrix x to make predictions independently (Zheng et al., 2024). The random forest then combines these predictions by voting, and the option with the most votes becomes the final prediction. For an input x , the prediction of the random forest is given by:

$$(A.5) \quad \hat{y} = \frac{1}{N} \sum_{j=1}^N T_j(x),$$

Where:

- N is the number of decision trees in the random forest.
- $T_j(x)$ is the prediction of the j -th decision tree for input x .

Since the optimal parameters for the Random Forest model are unclear, I perform hyperparameter tuning. I define a list of candidate parameters and use GridSearchCV to identify the best model.

A.3 KNN

I followed the mathematical expressions from Zhou et al. (2009), given a data point x_{new} , the goal is to classify it based on the labels of its $K = 4$ nearest neighbors.

$$d(x_{\text{new}}, x_i) = \sqrt{\sum_{j=1}^m (x_{\text{new},j} - x_{i,j})^2}$$

$$(A.6) \quad \hat{y} = \text{mode}(y_1, y_2, \dots, y_4)$$

Where:

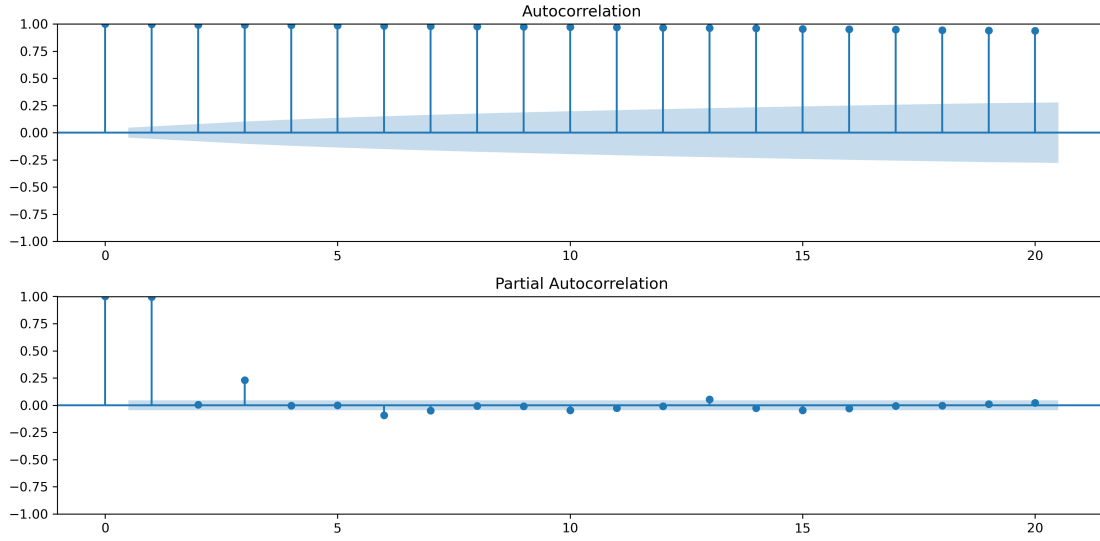
- x_{new} is the data point to classify,
- x_i are the training data points,
- $y_1, y_2, \dots, y_4$ are the labels of the nearest neighbors.

A.4 LSTM Simulated Data

In line with the methodology described in Afzali et al. (2021), I partition the dataset utilized for both the LSTM and hybrid models into 80% for training and 20% for testing, a standard procedure employed to assess model performance. This approach, which is widely adopted in the literature, ensures a robust evaluation of the model's generalization capabilities and is well-suited for training datasets. To prevent data leakage that could lead to unrealistic performance and affect the predictions of subsequent time series models, I first split the dataset and then normalize the data.

Since LSTM is more suitable for models with large datasets, If there is enough data, deep neural networks are more likely to improve accuracy due to their larger and deeper structure. However, if the data is insufficient, large networks are prone to overfitting and may not perform well, to avoid overfitting, I set early stopping and added dropout layers to mitigate this effect (Baek and Kim, 2018). The time range for data generation is the same as that used for the Random Forest model. The parameters for this LSTM model will be presented, as I observe that when these parameters are used, the model fits both the training and test sets for each feature relatively well:

Figure A.1: ARIMA Parameters Selection



$$(A.7) \quad \mathcal{M} = \text{create_lstm_model}(\text{look_back}, u, d, \eta)$$

Where:

- $\mathcal{M}$: The LSTM model created by the function `create_lstm_model`,
- `look_back`: The size of the look-back window, which is 10,
- $u = 50$: The number of LSTM units,
- $d = 0.2$: The dropout rate,
- $\eta = 0.001$: The learning rate.

A.5 ARIMA

Since I am not yet fully certain about the historical volatility and stationarity of SP_GCEI, I first plot the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF)—refer to Figure A.1—to determine the appropriate parameters for the ARIMA model (Ning et al., 2022).

The ACF plot indicates that the S&P Clean Energy Index exhibits non-stationarity over the past five years, suggesting that differencing may be required

to achieve stationarity. Meanwhile, the PACF plot shows a significant correlation at a lag of 1, with subsequent lag values gradually decreasing. Therefore, selecting a first-lag model appears to be suitable.

Subsequently, I utilized the "auto_arima" function to automatically select the optimal parameters for the ARIMA model, based on the AIC criterion. The results indicated that the optimal parameters for the ARIMA model are (6, 1, 1) (Akaike, 1974). The mathematical expression will be used as follows (Afzali et al., 2021):

$$(A.8) \quad Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_6 Y_{t-6} + \theta_1 \epsilon_{t-1} + \epsilon_t,$$

Where:

- Y_t : S&P Clean Energy Index at time t .
- c : The constant term or intercept.
- ϕ_i : The coefficients of the autoregressive (AR) terms, where $i = 1, 2, \dots, 6$ represents the lag order.
- Y_{t-i} : The lagged values of the target variable.
- θ_1 : The coefficient of the moving average (MA) term.
- ϵ_t : The error term at time t .
- ϵ_{t-1} : The lagged error term, capturing the moving average effect.

A.6 VAR

It can be observed that the p-values for Log_SP_GCEI, GBI, PC1, Oil_Close are greater than 0.05, indicating that these variables are non-stationary and require differencing. After differencing, the first row, which contains "NaN" values due to the differencing process, is removed from the data. Additionally, the six non-stationary variables identified in the ADF test are excluded from the Unprocessed Data table. Instead, their differenced results are retained and, together with the remaining original variables, are incorporated into the VAR model for prediction. Y_t represents all the variables included in the model. Before constructing the mathematical expression, it was necessary to select the optimal lag order. This

Table A.1: ADF Test Results(Random Forest)

Variable	ADF Statistic	p-value
average_goldstein_scale	-6.996747	0.0000000007499577
PRI	-9.558678	0.000000000000002459270
TRI	-7.797892	0.000000000007630617
GBI	-0.736772	0.8369934
Oil_Close	-1.680196	0.4413436
VIX_Close	-4.127630	0.0008715387
GFSI	-3.536743	0.007092107
average_confidence	-4.178804	0.0007155012
AverageTone	-3.235467	0.01802135
PC1	-1.553116	0.5070914
PC2	-3.249619	0.01728631
Log_ReportCount_env	-3.196386	0.02019485
Log_SP_GCEI	-1.885711	0.3388128

Table A.2: ADF Test Results(LSTM)

Variable	ADF Statistic	p-value
average_goldstein_scale	-6.996747	0.0000000007499577
PRI	-6.631907	0.000000005687093
TRI	-7.771645	0.00000000008888456
GBI	-0.736772	0.8369934
Oil_Close	-1.680196	0.4413436
VIX_Close	-4.127630	0.0008715387
GFSI	-3.536743	0.007092107
average_confidence	-4.178804	0.0007155012
AverageTone	-3.235467	0.01802135
PC1	-1.553116	0.5070914
PC2	-3.249619	0.01728631
Log_ReportCount_env	-3.196386	0.02019485
Log_SP_GCEI	-1.885711	0.3388128

study primarily used the AIC criterion, with the BIC values also considered, and determined the optimal lag order to be 8 periods, and the summaries output of the optimal lag order are presented in Table A.3 and A.4, and the expressions will be shown in Equation A.9 (Zivot and Wang, 2003):

$$(A.9) \quad Y_t = c + A_1 Y_{t-1} + A_2 Y_{t-2} + A_3 Y_{t-3} + \cdots + A_8 Y_{t-8} + \epsilon_t$$

Table A.3: Summary of VAR Model(Random Forest)

Optimal Lag Order:	8		
Summary of Regression Results			
Model:	VAR		
Method:	OLS		
Date:	Sun, 19, Jan, 2025		
Time:	20:22:56		
No. of Equations:	12.0000	BIC:	-37.5725
Nobs:	1819.00	HQIC:	-39.7960
Log likelihood:	7568.13	FPE:	1.42169e-18
AIC:	-41.0959	Det(Omega_mle):	7.62172e-19

Table A.4: Summary of VAR Model(LSTM)

Optimal Lag Order:	8		
Summary of Regression Results			
Model:	VAR		
Method:	OLS		
Date:	Sun, 19, Jan, 2025		
Time:	20:59:34		
No. of Equations:	12.0000	BIC:	-37.6336
Nobs:	1819.00	HQIC:	-39.8571
Log likelihood:	7623.69	FPE:	1.33744e-18
AIC:	-41.1570	Det(Omega_mle):	7.17000e-19

Where:

Y_t is the vector of variables at time t :

$$Y_t = \begin{pmatrix} D(\log(\text{SP_GCEI}(t))) \\ \log(\text{ReportCount_env}(t)) \\ \text{average_goldstein_scale}(t) \\ \text{PRI}(t) \\ \text{TRI}(t) \\ D(\text{GBI}(t)) \\ D(\text{PC1}(t)) \\ (\text{PC2}(t)) \\ \text{GFSI}(t) \\ \text{VIX_Close}(t) \\ D(\text{Oil_Close}(t)) \\ \text{AverageTone}(t) \\ \text{average_confidence}(t) \end{pmatrix}$$

c is the 13-dimensional vector of constant terms:

$$c = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_{13} \end{pmatrix}$$

$A_1, A_2, \dots, A_8$ are the 13×13 coefficient matrices for the lags, where:

$$A_i = \begin{pmatrix} a_{11}^{(i)} & a_{12}^{(i)} & \cdots & a_{1,13}^{(i)} \\ a_{21}^{(i)} & a_{22}^{(i)} & \cdots & a_{2,13}^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ a_{13,1}^{(i)} & a_{13,2}^{(i)} & \cdots & a_{13,13}^{(i)} \end{pmatrix}$$

i represents the i -th lag of each variable, with $i \in \{1, 2, \dots, 8\}$,
 ϵ_t is the 13×1 vector of error terms.

A.7 LSTM

I utilize the characteristics of the multivariate LSTM. In a typical LSTM model, the number of layers is usually set to two, so I follow this standard practice as

well. The first LSTM layer produces a sequence of hidden states, and the second LSTM layer produces a single vector. The minimum and maximum values for each unit, including the step, are set to commonly used parameters: 32, 256, and 32, respectively. Since the dataset consists of daily updated values and is relatively small, and given that LSTM models are typically more suitable for large-scale datasets, I incorporated a Dropout layer to mitigate overfitting. Additionally, to further reduce the risk of overfitting, I restricted the learning rate within an appropriate range and implemented early stopping to optimize the training process and enhance the model's generalization ability. Following some of the experimental procedures from Larsson and Ling (2021) and Afzali et al. (2021), after dividing the data and normalizing it according to the same steps as in the A.4, the following mathematical expressions are constructed (Guo and Lin, 2018):

Input sequence:

$$(A.10) \quad X = \{x_t\}_{t=1}^T, \quad x_t \in \mathbb{R}^d$$

The LSTM layer is defined by the following core equations:

Forget Gate:

$$(A.11) \quad f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

Input Gate:

$$(A.12) \quad i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$(A.13) \quad \tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

Cell State Update:

$$(A.14) \quad C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$$

Output Gate:

$$(A.15) \quad o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$(A.16) \quad h_t = o_t \odot \tanh(C_t)$$

Where:

- W_f, W_i, W_C, W_o : Weight matrices.
- b_f, b_i, b_C, b_o : Bias vectors.
- σ : Sigmoid activation function, defined as:

$$(A.17) \quad \sigma(x) = \frac{1}{1 + e^{-x}}$$

- $\odot$: Element-wise multiplication.

The dropout operation is represented as:

$$(A.18) \quad \mathbf{H}'_t = \mathbf{H}_t \cdot \mathbf{D}$$

Where: $\mathbf{D}$ is a mask vector with some entries set to zero, simulating the dropout.
The dense layer produces the output as:

$$(A.19) \quad y_t = \mathbf{W}_d \cdot \mathbf{H}'_t + \mathbf{b}_d$$

Where: $\mathbf{W}_d$ and $\mathbf{b}_d$ are the weights and bias of the dense layer.

The loss function is the Mean Squared Error (MSE), which are calculated by the predicted and true values, respectively. The hyperparameters are tuned by optimizing the validation loss, I will use this method to tune the optimal parameters, ensuring that it captures the iteration with the smallest loss and utilizes it for prediction. To predict future values, the model uses the last sequence from the training data and recursively generates future predictions:

$$(A.20) \quad \hat{y}_{\text{future},i} = f_{\text{model}}(\mathbf{X}_{\text{last}})$$

Where $\mathbf{X}_{\text{last}}$ is the most recent sequence, I previously set the sequence length to 10, which is also a very common practice, and the prediction is updated by shifting the sequence and appending new inputs. The results are clipped to the range $[-1, 1]$ and then inversely transformed.

A.8 LSTM_Autoregressive

The output of the autoregressive layer is given by (Zhang et al., 2023b):

$$(A.21) \quad y = X \cdot W + b$$

Where:

- $X \in \mathbb{R}^{N \times d}$: Input matrix, where N is the number of samples, and d is the number of features, which incorporates the values of all variables from both past and future data.
- $W \in \mathbb{R}^{d \times u}$: Weight matrix of the autoregressive layer, where u is the output dimension.
- $b \in \mathbb{R}^u$: Bias vector of the autoregressive layer.
- $y \in \mathbb{R}^{N \times u}$: Output of the autoregressive layer.

I used the same LSTM model and parameters as in the A.7, but with an additional Autoregressive layer added to the final output.

The autoregressive layer maps the hidden state h_T from the LSTM to the output:

$$(A.22) \quad y_{\text{ar}} = h_T \cdot W + b$$

Where:

- h_T : The last hidden state of the LSTM layer.
- W, b : The weights and bias of the autoregressive layer.

The dense layer maps the output of the autoregressive layer to the final prediction:

$$(A.23) \quad y_{\text{final}} = y_{\text{ar}} \cdot W_{\text{dense}} + b_{\text{dense}}$$

The overall model can be represented as the composition of functions:

$$(A.24) \quad y_{\text{final}} = f_{\text{dense}}(f_{\text{ar}}(f_{\text{dropout}}(f_{\text{lstm2}}(f_{\text{lstm1}}(X)))))$$

Where:

- f_{lstm1} : The first LSTM layer.
- f_{lstm2} : The second LSTM layer.
- f_{dropout} : The dropout regularization.
- f_{ar} : The autoregressive layer.
- f_{dense} : The dense (fully connected) layer.

A.9 SAE

The features, dataset splitting, normalization, LSTM model, and parameter settings are the same as in A.7. The main difference is that both the encoder and decoder parts contain two LSTM layers:

According to the figure in Li et al. (2023b), I can write the following mathematical expression, the encoder encodes the input sequence X into a latent vector:

$$(A.25) \quad h_1 = \text{LSTM}_1(X, U_1)$$

Where U_1 is the number of units in the first LSTM layer.

$$(A.26) \quad h_2 = \text{LSTM}_2(h_1, U_2)$$

Where U_2 is the number of units in the second LSTM layer.

The decoder reconstructs the sequence from the encoded latent vector:

$$(A.27) \quad h_3 = \text{LSTM}_3(h_2, U_3)$$

Where U_3 is the number of units in the third LSTM layer.

$$(A.28) \quad h_4 = \text{LSTM}_4(h_3, U_4)$$

Where U_4 is the number of units in the fourth LSTM layer.

The model generates predictions from the decoder's output:

$$(A.29) \quad \hat{X} = \text{TimeDistributed}(\text{Dense}(1))(h_4)$$

This is for predicting the output for each time step.

$$(A.30) \quad \hat{y} = \text{Dense}(1)(h_4)$$

A.10 Prophet

I first establish the following mathematical expression Ning et al. (2022); Taylor and Letham (2017) based on the historical data of the S&P Clean Energy Index and make predictions:

$$(A.31) \quad y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

Where:

- $y(t)$: Observed value at time t
- $g(t)$: Trend component (capturing the growth pattern)
- $s(t)$: Seasonal component (capturing periodic fluctuations)
- $h(t)$: Holiday effect (impact of special events)
- ϵ_t : Noise (residual error)

When the data exhibits a natural limit, a logistic growth model can be used:

$$(A.32) \quad g(t) = \frac{C}{1 + \exp(-\beta(t - \tau))}$$

Where:

- C : Carrying capacity (maximum value)
- β : Growth rate

- τ : The point where the growth slows down

The seasonal component is modeled using a Fourier series, which captures periodic fluctuations like daily, weekly, or yearly cycles:

$$(A.33) \quad s(t) = \sum_{p=1}^P \left[\alpha_p \cos\left(2\pi p \frac{t}{m}\right) + \beta_p \sin\left(2\pi p \frac{t}{m}\right) \right]$$

Where:

- P : Number of Fourier terms (determined by ‘seasonality_prior_scale’)
- m : Period of the seasonality (e.g., yearly, weekly, seasonality)
- α_p, β_p : Fourier coefficients (determined during training)

The holiday effect models the impact of specific holidays or special events:

$$(A.34) \quad h(t) = \sum_{k=1}^M \gamma_k I(t \in \text{Holiday}_k)$$

Where:

- γ_k : Effect of holiday k on the time series
- $I(t \in \text{Holiday}_k)$: Indicator function, 1 if t is a holiday, 0 otherwise

The noise term ϵ_t represents the residual error and is assumed to be Gaussian:

$$(A.35) \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2)$$

Where σ^2 is the variance of the residual error.

The final prediction $\hat{y}(t)$ is obtained by summing the trend, seasonal, and holiday components:

$$(A.36) \quad \hat{y}(t) = g(t) + s(t) + h(t)$$

I also adopted the concept of "hyperparameter tuning", assigning a scale to each parameter of the Prophet model. By adjusting the parameters, I observe the best loss from each run.

The best loss could be defined as Mean Absolute Percentage Error (MAPE):

$$(A.37) \quad \text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_{\text{true},i} - y_{\text{pred},i}}{y_{\text{true},i}} \right|$$

Where:

- $y_{\text{true},i}$ is the actual value at time step i ,
- $y_{\text{pred},i}$ is the predicted value at time step i ,
- n is the number of data points in the forecast horizon.

If the best loss stabilizes in the final steps, I should appropriately reduce the Number of Evaluations. It also takes into account the seasonal effects and holiday effects of the data, both of which can be adjusted through parameters. In the Prophet model, two hyperparameters are adjusted in addition to defaults like growth trend and seasonality. The "changepoint_prior_scale" controls trend flexibility, with lower values causing underfitting and higher values leading to overfitting. The "seasonality_prior_scale" regulates seasonality flexibility, with smaller values limiting seasonality amplitude and larger values allowing for more changes (Ning et al., 2022).

A.11 Prophet Features

I have already discussed the basic Prophet model in A.10. I used the same parameters and tuning methods and I only need to add an additional layer of external factor regression outside the basic model. The forecast $\hat{y}(t)$ for any future time point t is computed by the Prophet model:

$$(A.38) \quad \hat{y}(t) = \text{ProphetModel}(t, \text{regressors})$$

Where:

- regressors are the external variables (such as "ReportCount_env", "average_goldstein_scale", etc.),
- t represents a future time point for which the forecast is being made.

CENTRAL BANKERS' POLITICAL DISCOURSE AS A DRIVER OF CLEAN ENERGY MARKETS

Abstract

Following the analysis in Chapter 1, this chapter completes the research arc by introducing the policy commitment dimension to the previously identified media sentiment factors. I examine how the political language used in central bank governors' speeches influences the dynamics of clean energy markets in China and the United States. Using advanced textual analysis (poldis) and a lexical matching algorithm, I measure key aspects of policy discourse such as "frequency", "urgency", "timing", "intensity" and "commitment". The indicators are merged with traditional market and media variables and analyzed using Extreme Bounds Analysis (EBA) and SHAP values. This ensures comparability with earlier results while allowing consistent interpretation of new policy-related effects. I find that speeches expressing stronger commitment and intensity are associated with higher clean energy stock performance, whereas greater urgency and shorter timing intervals tend to signal weaker outcomes, reflecting perceived risks.

4.1 Introduction

The rapid growth of the clean energy sector raises important questions about its interaction with existing markets, particularly the key drivers of its financial performance. I have previously investigated the influence of the media as an emerging factor, along with traditional drivers, on the clean energy market (Zheng and Lucey, 2025c). However, media is not the only emerging dimension; I aim to explore additional dimensions of novel factors to enhance my understanding.

Language serves as a crucial medium for political activities, from election campaigns to news reporting. Natural Language Processing (NLP) leverages computational techniques to analyse text and extract key information essential for policymaking. As a fundamental data form in politics, language captures voters' opinions and demands, often expressed through social media posts, survey responses, and other textual sources. Politicians rely on debates, policy statements, and social media to conduct their campaigns, while government officials must compile and synthesize information from various documents to support decision-making (Bertoni et al., 2023). Ofer and Siegel (1987) examine the impact of unexpected financial policy announcements on unforeseen changes in firm performance, highlighting their significant influence on financial markets. This offers an alternative perspective for analysing the disconnect between political documents and financial dynamics, particularly in sectors such as the clean energy market. Textual analysis is becoming an essential tool in financial research, offering new ways to measure economic variables that traditional quantitative methods often cannot capture (Loughran and McDonald, 2020). Especially, Sposito (2024, 2021) develop a framework to identify and compare authenticity in political discourse across time and contexts, offering a valuable foundation for analysing both the overall content of political texts and specific political priorities, as well as their relationship to financial market behavior.

This study seeks to systematically analyse textual reports of central bankers' speeches by leveraging poldis R package (Sposito, 2021), an advanced language model, to quantify key dimensions of central bank governors' speeches. It separately examines cases that include the full content of political texts. When political priorities are identified, an iterative process is applied to align political expressions with specific policy event dates using a pattern-based lexical matching algorithm, thus enhancing the temporal precision of the analysis.

By integrating the policy commitment dimension, I employ Extreme Bounds Analysis (EBA) and SHAP values to rigorously evaluate and verify the relative importance of both novel and traditional determinants, such as market volatility and oil prices as identified in (Zheng and Lucey, 2025c), while also uncovering additional drivers from new dimensions that shape the financial dynamics of the clean energy market in two representative economies: China and the United States. I investigate whether central bank governors' statements and commitments genuinely reflect their policy intentions, and propose a systematic approach for evaluating such alignment. It further examines the extent to which these policy commitments are realistically implementable, and how strongly central bank governors are genuinely committed to their realization. Finally, I investigate whether movements in the clean energy market are driven exclusively by official policy announcements, or whether they are also shaped by traditional and media factors.

4.2 Literature Review

4.2.1 Authenticity, Performance, and Priorities in Political Discourse

Political authenticity is critically important. Perceptions of authenticity inform voters about how politicians are likely to act in the absence of public scrutiny, providing them with compelling reasons to support certain candidates (Jones, 2016). While different countries share certain geographical, historical, and cultural similarities, their distinct political and economic histories have given rise to important societal differences with significant implications for political discourse (Sposito, 2024). Authenticity in politics is neither static nor fixed. Instead, it is a flexible performance that requires constant negotiation, adaptation, and repetition (Fordahl, 2018). Consequently, authenticity becomes an integral part of successful political expression and is continuously enacted in the political sphere (Sposito, 2024).

The frequency of authenticity performances is reflected in the number of texts collected across specific contexts, years, settings, or politicians. These frequencies are normalized by the total word count of the texts in which they appear. In the analysis, the normalized scores represent the number of authenticity-related words per 1,000 words, calculated annually and by setting to enhance interpretability. Politicians' roles, settings, and timing shape expectations for authenticity per-

formances and guide the analysis of patterns within political discourse (Sposito, 2024).

Sposito (2021) suggest a way to measure urgency by first adding up how strong the commitment and intensity are—basically how necessary the action is—and then multiplying that by the timing and frequency, which reflect the immediacy of action. Within this framework, political priorities are identified as statements in which actors articulate their intent or commitment to future political action. Accordingly, urgency scores for priorities are calculated by multiplying commitment scores with all other relevant dimensions, providing a quantitative measure of the strength and immediacy of political intent. This offers a valuable analytical dimension for assessing the content of central bank governors' speeches. I list the definitions of each variable based on the descriptions provided in the literature, as shown in Table 4.1.

In sum, the contribution of the political discourse literature is to provide measurable textual dimensions that can be mapped into market-relevant signals, enabling an empirical test of whether central bank political discourse affects clean energy market dynamics.

To identify future-oriented political goals in text, the `select_priorities()` function helps pick out statements where politicians express plans or commitments to take action. After the text is broken down by sentence, the function looks for clues like modal verbs (such as *must*) and phrases that suggest future action (like *need to* or *going to*), using a custom-made dictionary. This makes it easier to systematically spot political intent in large volumes of text and analyse how strongly and how urgently certain policies are being emphasized (Sposito, 2021). Recent policy developments across major economies indicate broad and sustained governmental support for the expansion of clean energy markets. In particular, international policy frameworks emphasize accelerated investment in renewable energy, clean technologies, and energy transition pathways as central components of climate strategies (International Energy Agency, 2021). This policy environment suggests that publicly communicated commitments related to clean energy are increasingly aligned with actual policy implementation and market expectations. Based on the implications of these policy indicators, I propose Hypothesis 1 (H1): Central bank governors' publicly declared commitments related to clean energy development are positively associated with clean energy market performance, regardless of political prioritization in textual contexts. Stronger policy intensity (Intensity) and higher

Table 4.1: Definitions and Interpretations of Key Textual Variables

Variable	Definition and Interpretation
Frequency	Measures the number of authenticity-related expressions per 1,000 words, based on the dictionary provided in Sposito (2021). While labeled as capturing "truthful" or "authentic" language, this variable may more plausibly reflect realism or problem-acknowledgement, particularly in periods of market stress. For example, a high-frequency score may correspond to statements like "I acknowledge significant downside risks", while a low-frequency score may align with vague or overly optimistic language.
Timing	Reflects the temporal proximity of the speech to major political or macroeconomic events. A speech issued immediately after a policy change or crisis would score higher on this dimension than one issued during routine periods.
Intensity	Captures the emotional force or emphasis in the language. For instance, a statement such as "I will take all necessary measures without hesitation" implies high intensity, whereas "some measures may be considered if needed" suggests low intensity.
Commitment	Measures the degree of declared intent or resolve to take action. Strong expressions like "I am committed to ensuring stability" indicate high commitment, while weaker forms like "I am monitoring the situation" indicate low commitment.
Urgency	A multidimensional composite indicator that combines intensity, timing, frequency, and commitment to capture how immediate or critical the issue at hand is presented in the discourse. This measure is derived using the Urgency Analysis (UA) method embedded in Sposito (2021).

Note: These variables are constructed using the *poldis* package, which offers validated tools for processing and annotating political discourse in diverse formats including speeches, debates, and interviews. Beyond text wrangling, *poldis* incorporates a survey-backed, multidimensional framework for Urgency Analysis, combining NLP and dictionary approaches to quantify how immediate or critical political statements appear. These textual dimensions support Hypotheses 1 through 4, with Intensity and Commitment expected to exert particularly strong influences on market effects.

commitment feasibility (Commitment) are expected to elicit more positive market responses.

Nevertheless, the selection of which countries' central bank governors to include constitutes a critical methodological consideration. I therefore use existing literature to support my focus on China and the United States.

The S&P Global Clean Energy Index mainly comprises firms with substantial exposure to renewable energy generation, clean energy technologies, and related infrastructure, with U.S. and Chinese companies accounting for a significant share of market capitalization and industry influence (Henriques and Sadorsky, 2008; S&P Dow Jones Indices, 2024). By contrast, the European Union consists of multiple member states and operates under a supranational institutional framework. Climate-related communication by the European Central Bank reflects collective policy positions rather than the political discourse of individual central bank governors. Given this institutional difference, and the more concentrated role of U.S. and Chinese firms in the S&P Global Clean Energy Index, focusing on China and the United States allows for a more coherent and comparable empirical analysis. The role of the European Union is therefore beyond the scope of this chapter and is left for future research.

4.2.2 China–U.S. Dominance in the Clean Energy Market and Their Different Expressions

Reaching net-zero emissions by 2050 is a major global challenge, especially as post-pandemic economic recovery drives a renewed increase in carbon emissions. This highlights the critical need for governments to expedite clean energy adoption (International Energy Agency, 2021). S&P Dow Jones Indices (2024) points out that the USA leads the Clean Energy Index with a 37% contribution, while China ranks third at 8.3%. Since these are the two biggest players in the global economy, investigating media narratives in the U.S. and China becomes increasingly important for understanding global trends in clean energy. Chorzempa (2021) provides an overview of the motivations behind central banks around the world in evaluating Central Bank Digital Currencies (CBDCs), and analyses the reasons for the significant divergence between the world's two largest economies in developing CBDCs. As early as the beginning of 2016, China committed to launching a CBDC and has since taken more concrete steps than the Federal Reserve in piloting and issuing

a digital currency. In contrast, the Federal Reserve has not committed to issuing a CBDC. This highlights a significant policy divergence between the Chinese and U.S. central banks—one that also reflects differing political expressions shaped by their distinct political systems.

Cho et al. (2024) show that cultural differences significantly influence how emotions are expressed on social media. Comparing posts from Twitter in the U.S. and Weibo in China, the study focuses on two emotional dimensions: valence and arousal. The authors find that the relationship between these dimensions differs across cultures—Twitter users show a smoother emotional range, while Weibo users exhibit sharper changes in arousal with emotional intensity. Moreover, emotional expressions on Twitter tend to reflect personal life, whereas those on Weibo often relate to sociopolitical topics. These findings highlight a culturally driven East-West difference in emotional expression patterns online. Therefore, I consider the speeches of central bank governors from these two major economies to be highly representative. However, establishing a meaningful correlation between central bank governors' speeches and the clean energy index requires grounding in existing literature that demonstrates the impact of such speeches on financial market dynamics and sustainability. The following section strengthens the theoretical foundations and enhances the practical applicability of my study.

While China and the United States differ substantially in political institutions, central bank mandates, and communication norms, these differences are explicitly incorporated into the empirical design of this study. The textual indicators employed are derived from a validated political discourse analysis framework that has been widely used in studies of central bank communication, including applications to European Central Bank contexts as well as U.S. presidential press conferences. Rather than relying on language-specific sentiment scores, the approach focuses on structural dimensions of political discourse, such as policy intensity, commitment feasibility, timing, and urgency, which capture relative shifts in policy signaling within each institutional setting (Sposito, 2021, 2024). By emphasizing within-country variation and comparable textual constructs, this study does not assume linguistic or institutional homogeneity across China and the United States. Instead, it leverages these differences to examine whether central bank governors' political discourse exerts market effects under distinct cultural and policy environments.

4.2.3 The Relationship Between Central Bankers' Speeches, Sustainability and Financial Market Dynamics

Central banks face the complex task of determining how to communicate effectively about climate change and what role they should assume in addressing it. To explore this issue, Arseneau and Osada (2023) examine how the scope of a central bank's mandate influences its climate-related communication. They find that the degree to which sustainability objectives are embedded within central bank mandates varies significantly across countries. These mandate differences are shown to affect not only the frequency of climate-related discourse but also the framing and context in which such issues are presented to the public.

Dikau and Volz (2021) use the IMF's Central Bank Legislation Database to examine how climate-related goals are reflected in the mandates of 135 central banks. Their findings indicate that over half incorporate sustainability goals, either directly or indirectly by aligning with broader government policy priorities. These mandate differences are linked to the implementation of green financial policies and may also shape how central banks communicate about climate change. Such perspectives offer important analytical insights for connecting developments in the clean energy market with central bank communication, and for discerning how political priorities are articulated within policy-related texts.

Prior research has also employed central bank speeches to study financial market dynamics. Park et al. (2023) demonstrate that such speeches can serve as a valuable source for identifying market risk factors. This research contributes to a deeper understanding of emerging financial market trends, enabling better monitoring of latent risks and aiding the formulation of preventive strategies against potential financial crises. On the other hand, it is widely recognized that central banks play a crucial role in keeping economies stable and promoting growth, and this has caught the attention of many researchers. Consequently, speeches by central bank governors are closely followed by market participants. Emerging evidence suggests that, given the unique characteristics of these speeches and the shortcomings of conventional sentiment measures, speech-based sentiment provides a more robust explanatory variable for asset return dynamics (Anand et al., 2020).

Overall, the existing literature offers three key insights that motivate this chapter. First, extensive research in finance and macroeconomics shows that cen-

tral bank communication, particularly governors' speeches, influences asset prices by shaping market expectations, risk perceptions, and uncertainty. Speech-based measures of tone and sentiment have been shown to explain financial market movements beyond traditional quantitative policy indicators, highlighting communication as an independent policy transmission channel. Second, a growing literature in sustainable finance documents the increasing involvement of central banks in climate-related objectives through mandate design, climate communication, and green financial policies. Differences in institutional mandates are associated with variation in the intensity and framing of climate discourse, suggesting that climate-related communication can influence market beliefs about policy credibility and transition pathways. Third, the clean energy market is widely viewed as a transition market undergoing structural change, where policy signals and information flows play a particularly important role in valuation. In this setting, political discourse in central bank speeches may affect clean energy markets by shaping perceptions of policy commitment, implementation timing, and uncertainty. Together, these findings motivate the central research question of this chapter: whether political discourse dimensions extracted from central bank speeches provide additional explanatory power for clean energy market dynamics beyond traditional market and media-related factors.

In my previous study Zheng and Lucey (2025c), I have demonstrated the relationships among the clean energy market, media-related factors, and several traditional factors. Building on previous research and the new `poldis` R package, I develop Hypothesis 2 (H2) and Hypothesis 3 (H3):

Hypothesis 2 (H2). Political commitments expressed by central bank governors, as reflected in policy intensity (Intensity) and commitment feasibility (Commitment), have a significant impact on clean energy market dynamics beyond the effects of traditional market determinants, such as market volatility and energy prices.

Hypothesis 3 (H3). Movements in the clean energy market are jointly influenced by official policy communications, traditional market factors, and media coverage. In particular, political factors derived from central bank speeches (All Political Factors in Table 4.1) exert a significant effect on market dynamics alongside traditional and media-related determinants.

This comparison is made at the same time point. Therefore, I should also consider whether these factors robustly influence the clean energy market across

different time dimensions, leading us to propose Hypothesis 4 (H4): These driving factors (Financial Indices, Transition and Physical Risk Indices, Media and Sentiment Measures, Speech-Based Textual Indicators) are not strongly associated with the temporal dimension.

Hence, in the following section 4.3, I will start by collecting data on central bank governors' speeches. This will be followed by full-text analysis and a focus on political text priority analysis, to comprehensively prepare data for the correlation analysis.

4.3 Methodology

4.3.1 Data Collection

4.3.1.1 Data Selection and Deduplication

As an initial step, I programmatically accessed the BIS open-source library *gingado* to collect central bankers' speeches from China and the United States (Araujo, 2023), covering the period from June 30, 2019, to December 31, 2024. This process yielded a corpus of 3,183 speeches, forming the textual foundation for subsequent analysis. However, since the dataset does not include a dedicated country-level column, I implemented a method to extract country information from the "description" column, which typically includes the information of describing the country, state, or city where the speech occurred. For Chinese data, I filtered records by identifying entries where the "description" column contains the keyword "China". For the United States, where entries often mention specific states or cities without explicitly stating the country name, I utilized the Python library *geonamescache*. This library provides access to names and codes (ISO and FIPS) of continents, countries, U.S. states, and counties in the form of a Python dictionary, enabling us to map these descriptions to their respective countries. Using this approach, I created a new column that categorizes each record as either "USA" or "China", thereby facilitating the identification of key drivers in the clean energy market (Gómez, 2023).

After initially collecting political texts and their corresponding dates from the BIS website, I found that some entries contained duplicate content—specifically, identical values in the "description", "title", and "text" fields appeared multiple times. To ensure data quality and eliminate redundancy, duplicates were removed

based on identical values in the "description", "title", and "text" fields, resulting in a refined dataset of 3,160 unique entries. These entries span 1,225 distinct days, acknowledging that multiple speeches may occur on the same day, while certain dates may have no associated speeches. Recent studies suggest that speech-based sentiment, given its unique context and the limits of conventional measures, provides a stronger explanation for asset return dynamics (Anand et al., 2020). Therefore, I divide the text data into political priorities and non-Political priorities to compare their effects.

4.3.1.2 In the Absence of Political Priorities

I proceeded to analyse the refined dataset, focusing primarily on the "text" and "date" columns. The packages required for this section are listed in B.1.

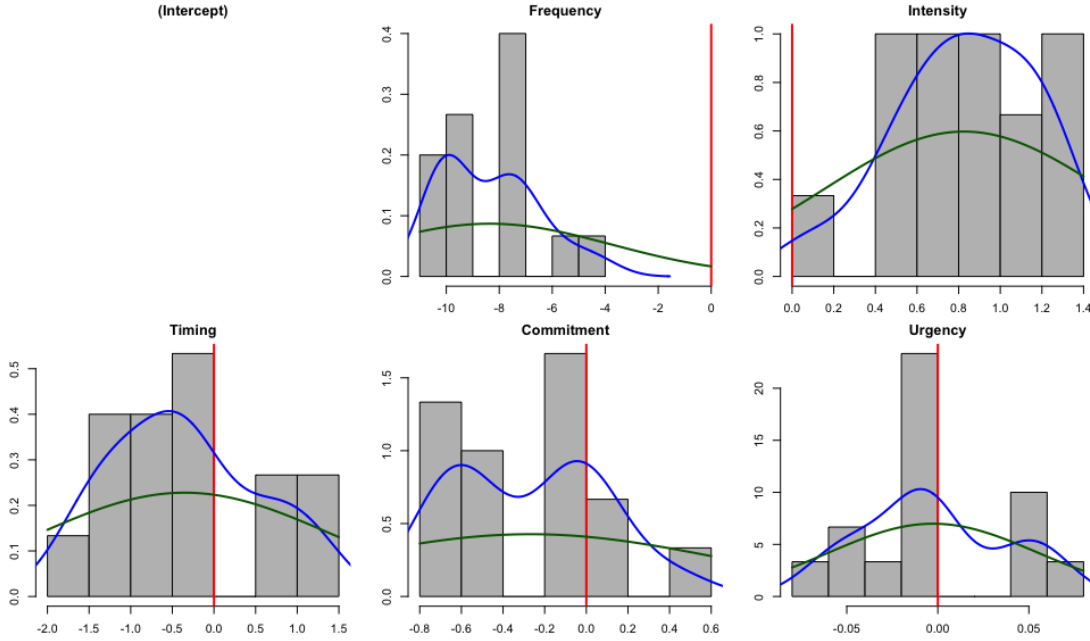
Since the ultimate goal is to examine the impact of these variables on the clean energy index, I required daily average values. After calculating the daily averages of these variables and aligning them with the S&P clean energy index over the same time period, I obtained a final dataset consisting of 1,112 entries. This because this index contains missing values, which are understandable in the context of financial markets and are often caused by factors such as holiday closures or trading suspensions (Javed Awan et al., 2021; Toms, 2011).

4.3.1.3 In the Presence of Political Priorities

The packages required for this section are listed in B.2. Running these codes generated 77,314 results. This is because the political-priority-related texts were segmented into multiple sentence fragments for analysis. However, after segmentation, the corresponding dates were no longer retained, and both the text identifiers and the result order were disorganized. As a result, the only viable anchor for alignment was the sentence fragments themselves. To address this, I matched the "sentence" column in the output to the dated "text" entries referenced in 4.3.1.1. Specifically, I employed a hybrid matching strategy to enhance text-to-event alignment. Each sentence in the target dataset was first tokenized into individual words. This is a robust, pragmatic matching procedure using standard n-gram and skip-gram techniques for temporal alignment. To reduce redundancy in the main text, I have included it in B.3, where the algorithm is described in detail.

Finally, I aggregated the data on a daily basis, calculating both the daily average values and the total sum of the variables "Frequency", "Timing", "Intensity",

Figure 4.1: In the Absence of Political Priorities-More Sophisticated



"Commitment", and "Urgency" for political priority texts occurring on the same day. This aggregation process yielded 308 data entries. After merging these entries with the S&P clean energy index based on matching dates, the final dataset consisted of 290 entries.

4.3.2 Extreme Bound Analysis, SHAP Value and Results

4.3.2.1 In the Absence of Political Priorities

-More Sophisticated

The process and formula for EBA can be found in C.1.3.

Table 4.2: Sala-i-Martin's EBA (In the Absence of Political Priorities-More Sophisticated)

Variable	N: $CDF(\beta \leq 0)$	N: $CDF(\beta > 0)$	G: $CDF(\beta \leq 0)$	G: $CDF(\beta > 0)$
(Intercept)	0.000	100.000	0.000	100.000
Frequency	97.074	2.926	96.674	3.326
Intensity	6.965	93.035	7.482	92.518
Timing	55.469	44.531	54.198	45.802
Commitment	65.077	34.923	63.214	36.786
Urgency	44.558	55.442	44.814	55.186

Figure 4.1 and Table 4.2 indicate that applying stricter model conditions to boost the credibility of the results consistently leads to clearer findings. Specifically, the coefficients for "Frequency" and "Intensity" display increased stability, with distributions clearly shifting toward one side of zero.

This supports the first half of Hypothesis 1. However, the positive correlation doesn't seem very precise, because I observed that the coefficient distributions for "Intensity" and "Commitment" aren't consistently positioned to the right of zero. Hypothesis 2 cannot be validated here because, in this section, I analyse all conversation content rather than focusing exclusively on political priorities text.

4.3.2.2 In the Presence of Political Priorities

-More Sophisticated

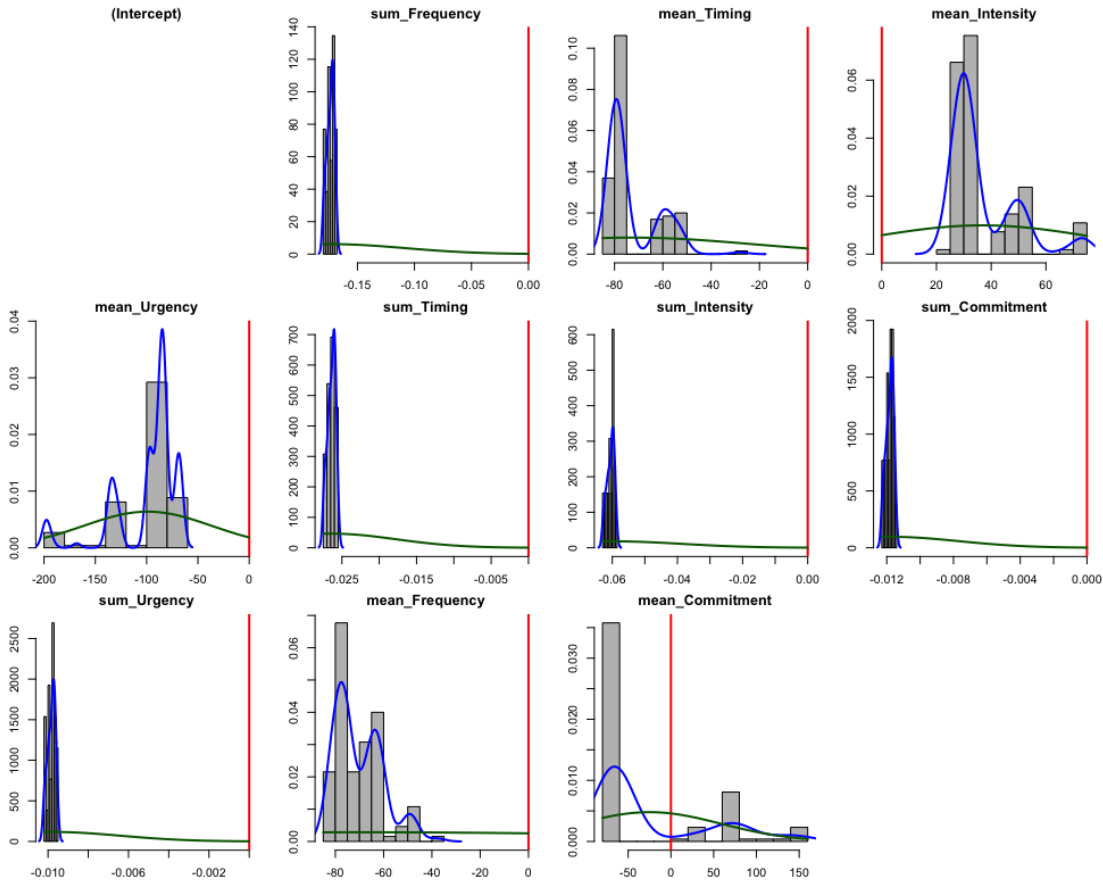
The process and formula for EBA can be found in C.2.3. I obtain the results shown in Figure 4.2 and Table 4.3.

Table 4.3: Sala-i-Martin's EBA (In the Presence of Political Priorities-More Sophisticated)

Variable	N: CDF($\beta \leq 0$)	N: CDF($\beta > 0$)	G: CDF($\beta \leq 0$)	G: CDF($\beta > 0$)
(Intercept)	0.000	100.000	0.000	100.000
sum_Frequency	99.621	0.379	99.620	0.380
mean_Timing	92.683	7.317	92.143	7.857
mean_Intensity	16.673	83.327	17.935	82.065
mean_Urgency	93.514	6.486	94.128	5.872
sum_Timing	99.888	0.112	99.888	0.112
sum_Intensity	99.718	0.282	99.717	0.283
sum_Commitment	99.786	0.214	99.786	0.214
sum_Urgency	99.802	0.198	99.801	0.199
mean_Frequency	68.870	31.130	68.839	31.161
mean_Commitment	57.108	42.892	64.742	35.258

After stabilizing the model by resolving multicollinearity and heteroskedasticity, I observe that besides the previously robust variables ("sum_Frequency", "mean_Timing", "mean_Intensity", "mean_Urgency"; Figure D.4, Table D.4), additional measures such as "sum_Timing", "sum_Intensity", "sum_Commitment", and "sum_Urgency" also show strong and robust impacts on the clean energy market. Each of these additional variables displays coefficient distributions with probabilities exceeding 99% of being either strictly positive or strictly negative, highlighting both their statistical and economic significance.

Figure 4.2: In the Presence of Political Priorities-More Sophisticated



These findings suggest that, under the condition of politically prioritized speech content, both the daily average and cumulative levels of "Timing" and "Urgency" are strongly negative with clean energy market. These variables are conceptually linked to the temporal urgency in central bank governors' speeches. Besides, the cumulative frequency of truthful statements and the total daily expression of "Commitment" emerge as additional robust drivers, both reflecting aspects closely tied to the credibility and intensity of policy commitments. While both daily average and daily cumulative measures of "Intensity" are robust, their differing impacts on the clean energy market suggest caution before making firm conclusions about this variable's overall impact.

Although the positive effect described in Hypothesis 1 still does not hold under the condition of politically prioritized texts, Hypothesis 2 is partially supported: political factors indeed significantly influence the clean energy market. However, the relevant comparisons still need further verification in 4.3.2.3.

4.3.2.3 Combining Traditional, Media-driven and Political Text-based Effects

-More Sophisticated

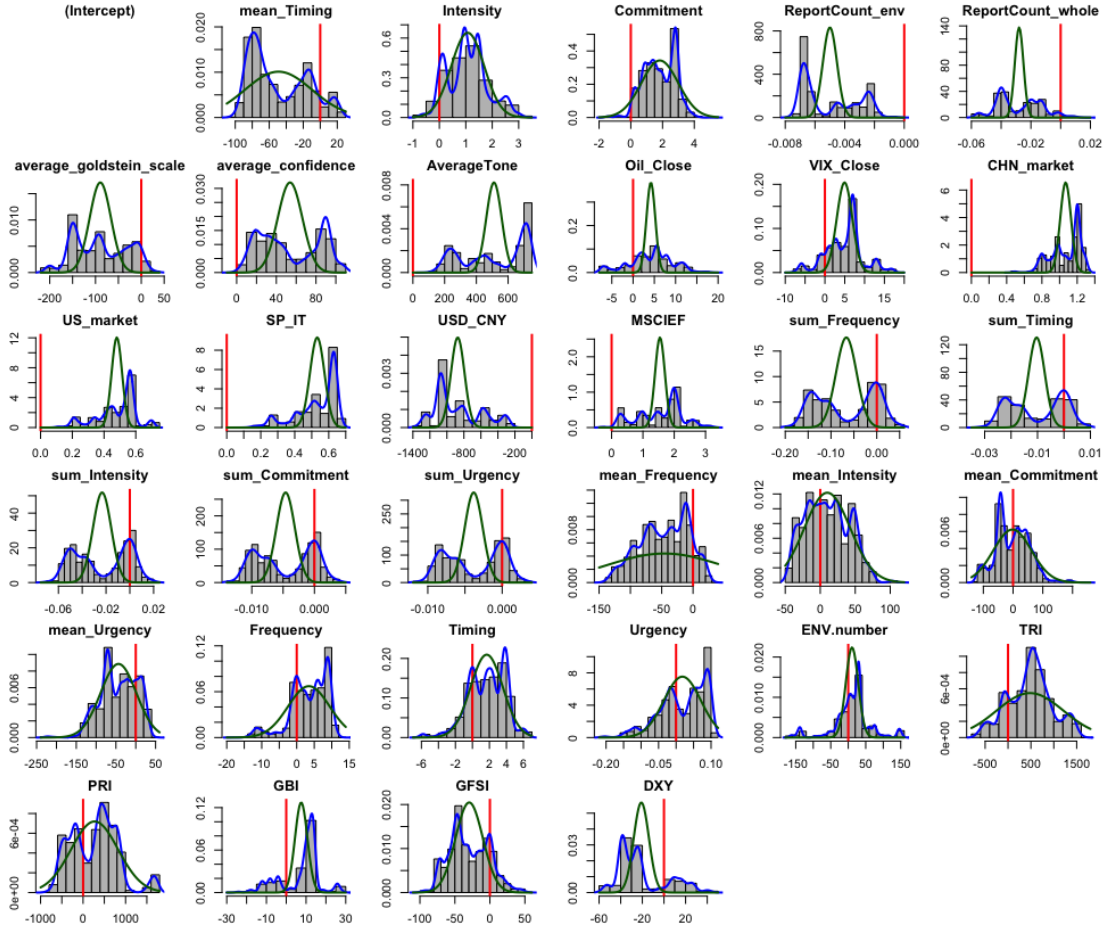
The process and formula for EBA can be found in C.3.3. This allows us to obtain more robust results, as presented in the corresponding Figure 4.3 and Table 4.4.

Table 4.4: Sala-i-Martin's EBA (More Sophisticated)

Variable	N: CDF($\beta \leq 0$)	N: CDF($\beta > 0$)	G: CDF($\beta \leq 0$)	G: CDF($\beta > 0$)
(Intercept)	4.589	95.411	55.634	44.366
mean_Timing	81.734	18.266	70.853	29.147
Intensity	11.319	88.681	20.405	79.595
Commitment	7.571	92.429	11.885	88.115
ReportCount_env	100.000	0.000	99.909	0.091
ReportCount_whole	100.000	0.000	96.796	3.204
average_goldstein_scale	99.824	0.176	86.055	13.945
average_confidence	0.053	99.947	2.490	97.510
AverageTone	0.000	100.000	0.002	99.998
Oil_Close	3.499	96.501	33.314	66.686
VIX_Close	0.166	99.834	15.932	84.068
CHN_market	0.000	100.000	0.000	100.000
US_market	0.000	100.000	0.004	99.996
SP_IT	0.000	100.000	0.003	99.997
USD_CNY	100.000	0.000	99.988	0.012
MSCIEF	0.000	100.000	2.045	97.955
sum_Frequency	97.458	2.542	69.136	30.864
sum_Timing	98.665	1.335	68.189	31.811
sum_Intensity	97.627	2.373	68.900	31.100
sum_Commitment	97.918	2.082	68.609	31.391
sum_Urgency	98.026	1.974	68.555	31.445
mean_Frequency	66.579	33.421	64.324	35.676
mean_Intensity	69.516	30.484	69.782	30.218
mean_Commitment	38.618	61.382	40.291	59.709
mean_Urgency	70.340	29.660	63.026	36.974
Frequency	27.380	72.620	30.169	69.831
Timing	30.366	69.634	34.888	65.112
Urgency	44.405	55.595	47.782	52.218
ENV.number	21.341	78.659	29.043	70.957
TRI	26.739	73.261	31.556	68.444
PRI	49.476	50.524	57.022	42.978
GBI	7.044	92.956	37.995	62.005
GFSI	94.268	5.732	77.565	22.435
DXY	97.459	2.541	65.405	34.595

Figure 4.3 and Table 4.4 present the relationship between each variable and the S&P Clean Energy Index, based on the subset of dates for which data are available across all variables. The following variables exhibit relatively robust

Figure 4.3: All-More Sophisticated



effects on the clean energy index during these periods: "Intensity"(approximately 80% of coefficients are positive, rounded), "Commitment", "ReportCount_env", "ReportCount_whole", "average_goldstein_scale", "average_confidence", "AverageTone", "VIX_Close", "CHN_market", "US_market", "SP_IT", "USD_CNY", and "MSCIEF". Some variables exhibit deviations from previous estimations, primarily due to discrepancies in sample size. The simultaneous availability of data across all variables is highly limited, which represents a key limitation of this study. This constraint has the greatest impact on variables such as the daily average and cumulative levels of "Timing" and "Urgency". After including traditional and media-related factors in this section, I observe that factors associated with political commitment consistently exhibit a positive effect on the clean energy market within the same time frame. This effect appears more robust compared to certain traditional factors, though less robust than media-related factors. To some extent, this confirms

Hypothesis 2 and Hypothesis 3. Although my earlier analysis in 4.3.2.1, Zheng and Lucey (2025c), I utilize the full five-year time span, it does not alter the conclusion that these variables robustly influence the clean energy market, thus partially supporting Hypothesis 4.

To address this issue and ensure the robustness of my findings, I proceed by applying machine learning interpretability methods in the subsequent steps. This enables a secondary filtering of variable importance and, when combined with econometric methods, helps enhance the accuracy of the final results.

4.3.2.4 SHAP Value-Supplementary Evidence

The process of obtaining the SHAP value can be found in E.

The absolute value of SHAP reflects the extent to which the model relies on a given feature for prediction—i.e., the larger the SHAP value, the more important the variable. Among the traditional financial factors, the following variables in Figure 4.4 exhibit strong explanatory power for the clean energy market: "SP_IT", "CHN_market", "USD_CNY", "MSCIEF", "US_market", "GBI", "GFSI", "Oil_Close", and "VIX_Close", which is largely consistent with the findings from the EBA analysis.

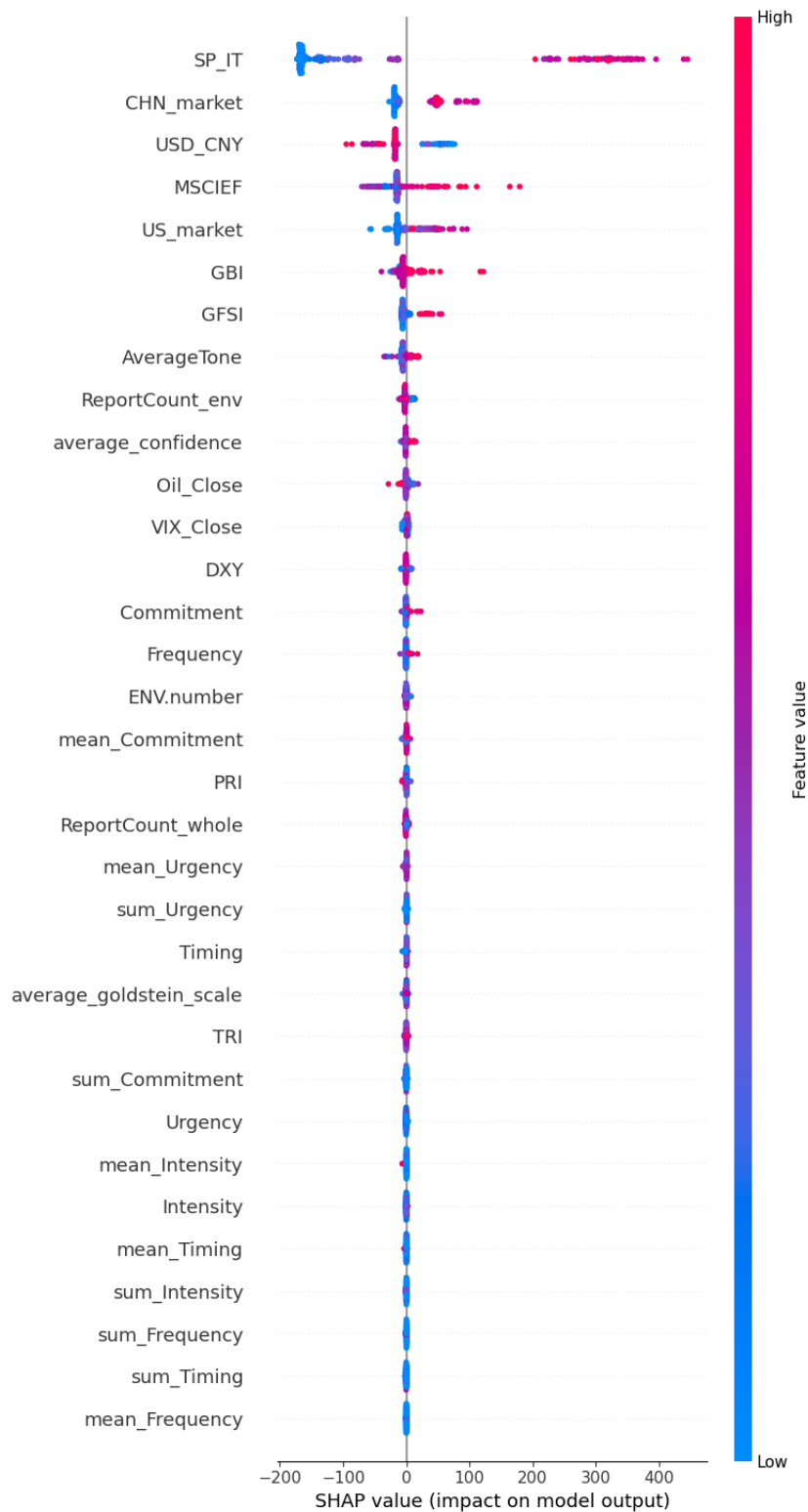
4.4 Discussion and Conclusion

4.4.1 In the Absence of Political Priorities

I find that, in the absence of constraints prioritizing political textual variables, the frequency of truthful expressions in central bank governors' speeches is significantly negatively correlated with the clean energy market index. This finding is intriguing, especially when compared with media signals. It prompts questions about whether central banks intentionally obscure truthful information, perhaps releasing overly positive or misleading messages to guide investor perceptions. This highlights why it is crucial to carefully separate out political content. Further research is needed to clarify if and how these policy signals directly affect clean energy market movements, providing greater understanding of institutional communication dynamics.

The variable "Intensity", which captures the perceived necessity of action, exhibits a significant positive correlation with prospect in the clean energy market.

Figure 4.4: SHAP



This suggests that the central bank governor's intention to act—and the strength or urgency of that intention—can positively influence market expectations, even when the speech content is not explicitly tied to political textual themes. This finding emphasizes the critical role of central bank communication in shaping investor expectations. It also implies that markets may be highly sensitive to affective cues embedded in central bank discourse, suggesting that in emerging sectors such as clean energy, the perceived credibility and strength of policy support are central to investment decisions.

4.4.2 In the Presence of Political Priorities

After filtering for politically relevant textual content, I turn to the interpretation of the variable "Frequency". I find that the clean energy index is negatively associated with the total daily volume of truthful expressions, while its relationship with the daily average is statistically insignificant. This indicates that "Frequency" functions as a cumulative signal—its market impact emerges only when the accumulation of truthful content crosses a certain threshold. Once that threshold is perceived by stakeholders, a stronger link with clean energy market movements is triggered. Importantly, this effect remains robust both in the presence and absence of political textual content, highlighting the consistent role of accumulated truth-related signals in shaping market expectations.

Regarding the relationship between "Urgency" and the clean energy index, both Table 4.3 and Figure 4.2 show that "Urgency" is negatively related to market performance. I use the average-based measure of "Urgency" because it better reflects the overall tone of speeches rather than short-term fluctuations. A higher level of "Urgency" often suggests that central bank governors are facing serious challenges and may soon introduce policy changes. However, investors may see this as a signal of unexpected negative news or delayed action. Market reactions to central bank communication depend not only on the content of the message, but also on whether the information is unexpected (Ehrmann and Fratzscher, 2007). When central banks sound urgent, investors may think the situation is worse than expected. This can lower confidence in the market and reduce investment in the clean energy sector. The results also show a clear difference between "Intensity" and "Urgency". "Intensity" reflects how strongly a message is expressed. "Urgency" includes "Intensity" but also considers the timing and content of the message. Markets seem to react well to strong and clear messages (high Intensity), but they

respond negatively to signs of stress or pressure (high Urgency). This helps explain why the two variables have opposite effects on market outcomes.

The variable "Timing" exhibits a similar effect to "Urgency". The key difference between these two variables lies in their measurement dimensions, as "Urgency" encompasses more dimensions than "Timing" following its definition (Sposito, 2021). Consequently, the interpretability of "Timing" is relatively straightforward, with daily averages being more representative. A shorter "Timing" indicates quicker policy responses aligned more closely with market dynamics, thus positively influencing the development of the clean energy market.

Although the variable "Intensity" significantly influences the clean energy market both in terms of daily averages and cumulative daily values, these two measures exhibit opposite effects on the clean energy index—one positively correlated, and the other negatively correlated. Furthermore, both measures are statistically significant only under the "More Sophisticated" scenario. As a result, I exclude "Intensity" from consideration in scenarios with political textual prioritization.

4.4.3 Combining Traditional, Media-driven and Political Text-based Effects

On dates where political textual expressions are prioritized, I observe that the variables "Intensity" and "Commitment" are significantly and positively correlated with the clean energy index, consistent with the findings presented in 4.3.2.1, except for the significance level of "Commitment". Nevertheless, since "Intensity" and "Commitment" closely align in capturing rhetorical strength and urgency, this minor difference in significance doesn't significantly affect my earlier conclusions. Meanwhile, the variable "mean_Timing" exhibits a significant negative correlation with the clean energy index, which aligns well with my previous results.

My analysis reveals that, within the same temporal framework as the Speech-Based Textual Indicators, the following traditional variables exert significant influence on the clean energy market: "VIX_Close", "CHN_market", "US_market", "SP_IT", "USD_CNY", and "MSCIEF". Additionally, media-related variables, including "ReportCount_env", "ReportCount_whole", "average_goldstein_scale", "average_confidence", and "AverageTone", also demonstrate significant impacts. These findings align closely with my previous research (Zheng and Lucey, 2025c), indicating that the choice of temporal dimension does not substantially affect conclusions

about the significance of media-related variables in the clean energy market.

Therefore, I can conclude that while certain traditional factors exhibit relatively weak effects on the clean energy market, others demonstrate impacts comparable to media-related variables, with significance levels exceeding 90%. Political textual factors from central bank speeches show weaker effects, although SHAP values further confirm the positive explanatory power of "Commitment". Regarding the variable "Frequency", I suggest that its effect on the clean energy market would be better understood by considering the broader economic and political contexts.

4.4.4 SHAP Value-Supplementary Evidence

Figure 4.4 illustrates that in terms of media-related variables, "AverageTone", "ReportCount_env", and "average_confidence" show relatively high importance. For political textual variables, "Commitment" and "Frequency" emerge as key contributors. Combined with previous EBA results in Table 4.4, I can reasonably confirm the importance of the "Commitment" variable. In terms of directional impact, the SHAP distribution for "Commitment" is skewed towards the positive side, indicating that stronger expressions of policy commitment are associated with more optimistic market expectations for clean energy. In light of the empirical findings in Section 4.3.2.1, I contend that the relationship between commitment and the clean energy index is significant only across certain time dimensions.

However, it is important to note that SHAP primarily reflects the model's predictive contribution of each feature, rather than statistical correlations or causal relationships. Therefore, these results should be interpreted as supplementary evidence supporting the robustness of the EBA-based conclusions, rather than as independent validation.

4.5 Contribution and Limitation

4.5.1 Contribution

The primary contribution of this chapter is identifying novel drivers of the clean energy market linked to central bank governors' speeches. My findings clearly address several critical research questions. Additionally, I find that Hypotheses 1, 2, 3, and 4 are supported to some extent, depending on specific conditions. Firstly, by systematically evaluating variables such as "Frequency", "Commitment" and "In-

tensity", I demonstrate whether central bank governors' statements and expressed commitments genuinely reflect their underlying intentions. Through robust analyses, particularly focusing on "Commitment" and "Intensity", I confirm that stronger and more emphatic statements are positively associated with market perceptions and expectations regarding clean energy investments, thereby affirming the authenticity and credibility of policy intentions. Secondly, by examining the dimensions of "Timing" and "Urgency", I clarify the anticipated timeline for policy commitments to materialize, as well as the degree to which central bank governors are genuinely committed to timely implementation. Specifically, shorter timing intervals and higher levels of urgency in central bank governors' speeches suggest clearer policy pathways and swifter responsiveness to economic conditions, directly influencing the clean energy market by shaping investor confidence and expectations. In this way, central bank governors can leverage central bank communications to assess whether policy initiatives are genuinely being implemented. This understanding can support the formulation of more feasible and effective measures to promote the development of clean energy markets in both of the world's largest economies, as well as across the global landscape, ultimately benefiting a wide range of stakeholders. Overall, my empirical results systematically address these research questions and provide nuanced insights into how central bank speeches, incorporating both political and full textual signals, shape investment behavior and clean energy market effects, with a particular focus on the policy implications for China and the United States.

By comparing the effects of these newly identified factors with traditional market indicators and media-related variables, I further assess and differentiate their relative explanatory power and driving forces within the market. The effects of these variables were further validated using machine learning techniques. Additionally, empirical analyses conducted across different time periods, based on data availability over the five-year span, lend further robustness to my findings and provide additional support for the conclusions presented in this study.

In terms of technical innovation, I contribute by developing an innovative text-event alignment strategy to accurately identify politically prioritized textual content. Specifically, I employ a hybrid matching approach that effectively links unstructured political text fragments to their corresponding policy event dates. This enhances the temporal precision of my analysis and improves the reliability of my conclusions about the effects of political text.

The findings of this chapter indicate that central bank communication plays a meaningful role in shaping clean energy market expectations. Clear and credible policy signals, reflected in the frequency, commitment, and intensity of central bank governors' statements, can enhance investor confidence and support clean energy investment decisions. Moreover, time-specific and urgent policy communication helps reduce uncertainty regarding policy implementation and strengthens market alignment with policy objectives. From a policy perspective, this suggests that central banks can strategically use communication frameworks to reinforce clean energy goals, particularly during periods of structural transition. In the context of China and the United States, effective policy communication may further contribute to more orderly and transparent clean energy market dynamics.

4.5.2 Limitation and Future Research

The primary limitation of this study stems from the limited number of observations with complete, non-missing data across all analysed variables. Expanding the temporal scope would not adequately resolve this issue, as it is inherently tied to the nature of the available data. Nonetheless, considering that my sample size aligns with those reported in Zheng and Lucey (2025c) and that no substantial effect on the significance of my findings was observed, I deem this limitation acceptable.

Future research could broaden the application of factors from central bank governors' speeches to examine additional indicators of financial market dynamics, particularly within the field of green finance. There also remains considerable scope to identify and analyse other potential driving factors in this domain. In future research, I can also examine central bank speeches from a broader range of countries. Beyond representative economies, the perspectives and market dynamics of emerging markets and their central bank communication also deserve greater scholarly attention.

DATA COLLECTION

B.1 In the Absence of Political Priorities

The data were organized chronologically, and the `poldis` R package, Sposito (2021) was employed to process the "text" column. This analysis generated values for five key variables: "Frequency", "Timing", "Intensity", "Commitment", and "Urgency" across all 3,160 entries.

B.2 In the Presence of Political Priorities

The text processing based on political priorities differs slightly from 4.3.1.2 in terms of code implementation—specifically, it requires an additional step to select priority-related content and the installation of the "en_core_web_sm" model for the `spacyr` package prior to calculating the urgency score. However, there are more substantial differences in the subsequent stages of data processing.

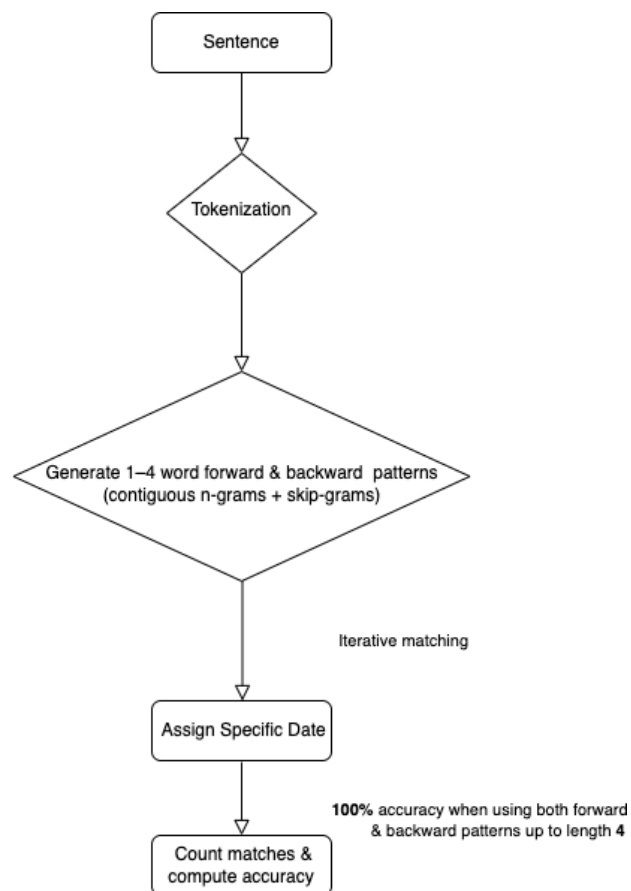
B.3 Iterative Matching

From these tokens, I generate both forward and backward phrase patterns, including contiguous sequences (n-grams) ranging from one to four words, as well as skip-grams (e.g., first and last word pairs), to capture meaningful lexical fragments. These patterns are then iteratively matched against the source corpus, starting

from unigrams (single-word patterns) and progressing incrementally up to four-word patterns. Upon finding a match at a specific n-gram length, the corresponding earliest date from the source dataset was assigned to the sentence, after which the search is terminated for that sentence to ensure mutually exclusive counting. This approach enable robust linkage of unstructured textual fragments to specific policy events, even in the absence of exact string matches, thereby significantly improving the temporal accuracy of my political text analysis. I record successful matches at each n-gram level, allowing us to compute match accuracy as the ratio of successful matches to the total number of sentences for each pattern length. This provides insights into the trade-off between pattern specificity and recall. Notably, when employing both forward and backward n-gram patterns up to a length of four words, I achieved 100% accurate date matching on my internal dataset; that is, every sentence fragment in the political priorities dataset could be successfully re-associated with a date from the source corpus. This level of accuracy reflects an internal validation check specific to my dataset and matching procedure, rather than a universal guarantee across other corpora. Given the relatively high computational demand of this approach, processing is performed using Google Colab, a cloud-based platform providing free GPU/CPU resources. Its accessibility, efficiency, and suitability for my dataset’s modest storage needs made it an effective practical solution for this analytical stage.

While this matching procedure proved highly effective for my dataset, its performance may vary for other corpora with different language structures or noise levels. Further validation would be required to generalize this approach beyond the present context.

Figure B.1: Iterative Matching



EXTREME BOUND ANALYSIS-STEPS

C.1 Absence**C.1.1 All Combinations of Doubtful Variables**

After the steps of 4.3.1.2, I obtained a dataframe containing the following variables: the S&P Clean Energy Index, "Frequency", "Timing", "Intensity", "Commitment" and "Urgency". Following the experimental procedures outlined in Hlavac (2016), I first conducted the EBA analysis by treating all variables as doubtful variables, as the focus variables had not been identified. The corresponding mathematical formulation is presented in Equation C.1.

$$(C.1) \quad Y = \beta_0 + \sum_{i=1}^n \beta_i X_i + \epsilon$$

Where:

- Y represents the dependent variable S&P Clean Energy Index;
- X_i represents the set of doubtful variables, which include Frequency, Timing, Intensity, Commitment, and Urgency, without political priority text processing;
- β_0 is the intercept term;

- β_i is the coefficient of the doubtful variables;
- ϵ is the error term.

C.1.2 Verifying the Accuracy of Variable Classification

Drawing on the preliminary diagnostic results, I classify the variables exhibiting greater robustness—specifically "Frequency" and "Intensity"—as free variables, whereas the remaining variables are treated as doubtful variables. Importantly, all of these variables are extracted from the speeches of central bank governors, which represent the focus variables of this research.

$$(C.2) \quad Y = \beta_0 + \sum_{i=1}^m \beta_i X_i + \sum_{j=1}^n \gamma_j Z_j + \sum_{k \in S_d} \delta_k W_k + \epsilon$$

Where:

- Y represents the dependent variable S&P Clean Energy Index;
- X_i are the free variables Frequency, Intensity;
- Z_j represents the set of focus variables which include Frequency, Timing, Intensity, Commitment and Urgency;
- W_k are the doubtful variables, which include Timing, Commitment and Urgency;
- β_i is the coefficient of the free variables;
- γ_j is the coefficient of the focus variables;
- δ_k represents the coefficient for the doubtful variables;
- β_0 is the intercept term;
- ϵ is the error term;
- S_d is the subset of doubtful variables selected (with a maximum of 3 doubtful variables).

Building on these specified conditions of Equation C.2, I extend the analysis by including up to three combinations of doubtful variables.

C.1.3 More Sophisticated

To improve model robustness, heteroscedasticity was addressed using a custom robust standard error function White (1980), with a VIF threshold of 7 set to control multicollinearity. LRI weighting was applied to enhance model fit (McFadden, 1974). All other settings remained unchanged, and the model specification aligns with Equation C.2.

C.2 Presence

C.2.1 All Combinations of Doubtful Variables

After I have collected the political priority texts following the steps outlined in 4.3.1.3, my dataframe contains the following variables: the S&P Clean Energy Index, "sum_Frequency", "sum_Timing", "sum_Intensity", "sum_Commitment", "sum_Urgency", "mean_Frequency", "mean_Timing", "mean_Intensity", "mean_Commitment" and "mean_Urgency". The prefix "sum_" represents the total value of the political priority texts extracted from all central bank governors' speeches on a given day, while the prefix "mean_" represents the average value of the political priority texts extracted from all speeches on that day. The experimental procedure is similar to that in 4.3.2.1, resulting in a mathematical expression similar to Equation C.1, but with different variables and a different dataframe.

$$(C.3) \quad Y = \beta_0 + \sum_{i=1}^n \beta_i X_i + \epsilon$$

Where:

- Y represents the dependent variable S&P Clean Energy Index;
- X_i represents the set of doubtful variables, which include sum_Frequency, sum_Timing, sum_Intensity, sum_Commitment, sum_Urgency, mean_Frequency, mean_Timing, mean_Intensity, mean_Commitment, and mean_Urgency, with political priority text processing;
- β_0 is the intercept term;
- β_i is the coefficient of the doubtful variables;
- ϵ is the error term.

C.2.2 Verifying the Accuracy of Variable Classification

Based on preliminary results, I classify the more robust variables—"sum_Frequency", "mean_Timing", "mean_Intensity", and "mean_Urgency"—as free variables, while treating the rest as doubtful variables. All are extracted from the politically prioritized content of central bank governors' speeches, which form the focus variables of this study. The detailed mathematical formulation is provided in Equation C.4.

$$(C.4) \quad Y = \beta_0 + \sum_{i=1}^m \beta_i X_i + \sum_{j=1}^n \gamma_j Z_j + \sum_{k \in S_d} \delta_k W_k + \epsilon$$

Where:

- Y represents the dependent variable S&P Clean Energy Index;
- X_i are the free variables sum_Frequency, mean_Timing, mean_Intensity, mean_Urgency;
- Z_j represents the set of focus variables which include sum_Frequency, sum_Timing, sum_Intensity, sum_Commitment, sum_Urgency, mean_Frequency, mean_Timing, mean_Intensity, mean_Commitment, mean_Urgency;
- W_k are the doubtful variables, which include sum_Timing, sum_Intensity, sum_Commitment, sum_Urgency, mean_Frequency, mean_Commitment;
- β_i is the coefficient of the free variables;
- γ_j is the coefficient of the focus variables;
- δ_k represents the coefficient for the doubtful variables;
- β_0 is the intercept term;
- ϵ is the error term;
- S_d is the subset of doubtful variables selected (with a maximum of 3 doubtful variables).

C.2.3 More Sophisticated

Next, I further assess the robustness of the results by addressing multicollinearity and correcting for heteroskedasticity across observations, hence improving the reliability of inference and enhancing model fit. The empirical procedures follow those outlined in 4.3.2.1.

C.3 Combining Traditional, Media-driven and Political Text-based Effects

C.3.1 All Combinations of Doubtful Variables

Following the preliminary findings discussed, I draw a connection to previous research (Zheng and Lucey, 2025c), which also explored the influence of media-related factors and traditional determinants on the clean energy market. To integrate insights from that earlier study with the newly introduced political-text dimensions in this chapter, I combine the dataset used in the prior work with the one constructed in the current study. Given that the two datasets are largely aligned in terms of the time dimension (with the TRI and PRI indices updated only until December 29, 2023), I ultimately obtain a merged dataset comprising 241 observations with complete coverage across all dimensions and matched by date. I have previously tested these variables individually, but now I need to analyse them collectively. This allows us to better compare their impacts on the clean energy market by evaluating all variables simultaneously within the same timeframe, provided data is available.

After I have merged all the datasets, my dataframe contains the following variables:

- **Financial Indices:**
 - S&P Clean Energy Index
 - Green Bond Index (GBI)
 - Crude Oil WTI Futures (Oil_Close)
 - CBOE Volatility Index (VIX_Close)
 - Global Financial Stress Indicator (GFSI)
 - China Shanghai Composite Index (CHN_market)

- S&P 500 Index (US_market)
- S&P 500 Information Technology Index (SP_IT)
- US Dollar Index (DXY)
- USD/CNY Exchange Rate (USD_CNY)
- MSCI Emerging Markets Index (MSCIEF)
- **Transition and Physical Risk Indices:**
 - Transition Risk Index (TRI)
 - Physical Risk Index (PRI)
- **Media and Sentiment Measures:**
 - Daily count of environmental-themed news (ReportCount_env)
 - Daily count of all news (ReportCount_whole)
 - Daily average Goldstein scale (average_goldstein_scale)
 - Daily average news confidence (average_confidence)
 - AverageTone
 - Number of environment-related articles (ENV.number)
- **Speech-Based Textual Indicators (original, sum, and mean):**
 - Frequency, Timing, Intensity, Commitment, Urgency
 - sum_Frequency, sum_Timing, sum_Intensity, sum_Commitment, sum_Urgency
 - mean_Frequency, mean_Timing, mean_Intensity, mean_Commitment, mean_Urgency

Using the same steps as in 4.3.2.1, I obtain the mathematical formulation for Equation C.5, with results shown in Figure D.5 and Table D.5.

$$(C.5) \quad Y = \beta_0 + \sum_{i=1}^n \beta_i X_i + \epsilon$$

Where:

- Y represents the dependent variable S&P Clean Energy Index;

- X_i represents the set of doubtful variables, which include all variables from dataframe, with the exception of the S&P Clean Energy Index;
- β_0 is the intercept term;
- β_i is the coefficient of the doubtful variables;
- ϵ is the error term.

C.3.2 Verifying the Accuracy of Variable Classification

My final benchmark results will be based on estimations that incorporate VIF, LRI, and a custom robust standard error function. At this stage, I proceed with the steps outlined in Equation C.6 to ensure the appropriate classification of doubtful variables, focus variables, and free variables. As this study primarily concentrates on political text-related factors, all of my focus variables are drawn from the dimension of Speech-Based Textual Indicators.

$$(C.6) \quad Y = \beta_0 + \sum_{i=1}^m \beta_i X_i + \sum_{j=1}^n \gamma_j Z_j + \sum_{k \in S_d} \delta_k W_k + \epsilon$$

Where:

- Y represents the dependent variable S&P Clean Energy Index;
- X_i are the free variables mean_Timing, Intensity, Commitment, ReportCount_env, ReportCount_whole, average_goldstein_scale, average_confidence, AverageTone, Oil_Close, VIX_Close, CHN_market, US_market, SP_IT, USD_CNY, MSCIEF;
- Z_j represents the set of focus variables which include sum_Frequency, sum_Timing, sum_Intensity, sum_Commitment, sum_Urgency, mean_Frequency, mean_Timing, mean_Intensity, mean_Commitment, mean_Urgency, Frequency, Timing, Intensity, Commitment, Urgency;
- W_k are the doubtful variables, which include sum_Timing, sum_Intensity, mean_Intensity, sum_Commitment, mean_Commitment, sum_Urgency, mean_Urgency, sum_Frequency, mean_Frequency, Frequency, Timing, Urgency, ENV.number, TRI, PRI, GBI, GFSI, DXY;

- β_i is the coefficient of the free variables;
- γ_j is the coefficient of the focus variables;
- δ_k represents the coefficient for the doubtful variables;
- β_0 is the intercept term;
- ϵ is the error term;
- S_d is the subset of doubtful variables selected (with a maximum of 3 doubtful variables).

C.3.3 More Sophisticated

I extend the naïve EBA framework by defining a fixed set of free variables(including speech variables, media indicators, and market factors), and allowing up to 3 additional variables from a set of 15 "doubtful" variables to vary across model specifications. I proceed to perform steps similar to those in 4.3.2.2, incorporating additional controls such as the removal of multicollinearity and other external influences.



EXTREME BOUND ANALYSIS-SOME RESULTS

D.1 Absence

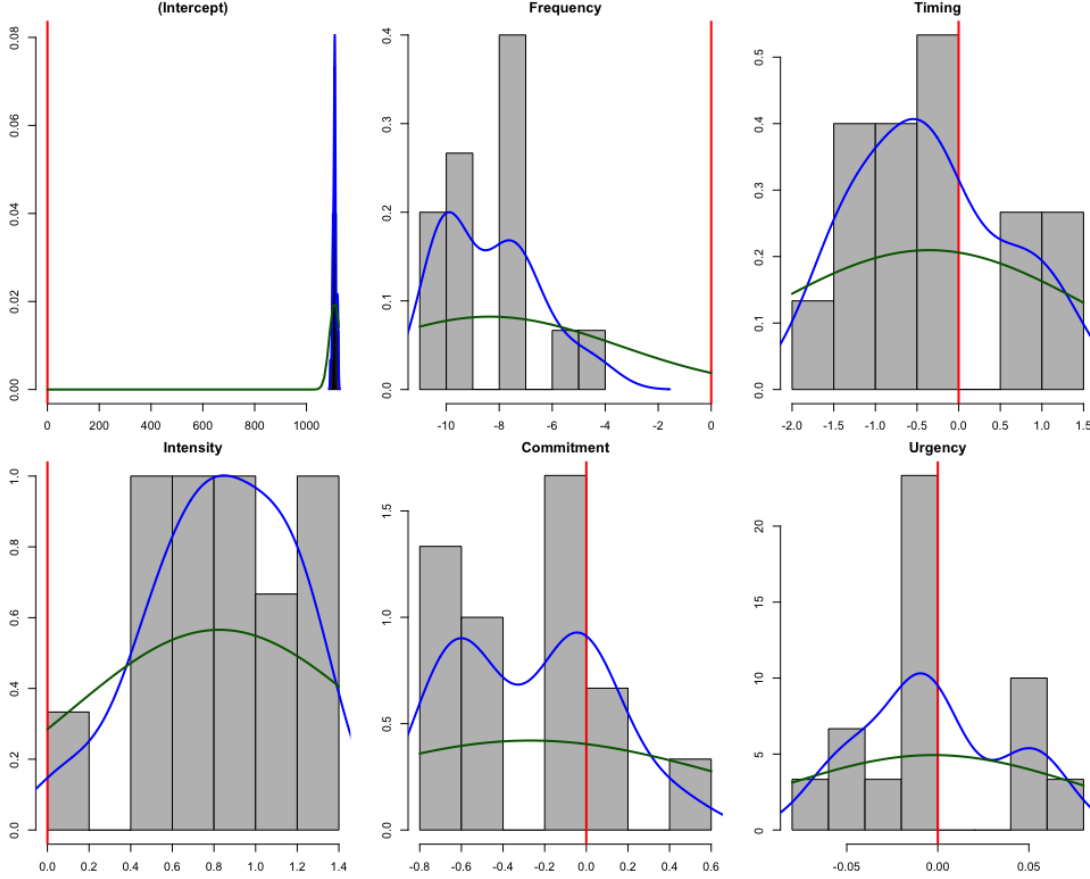
D.1.1 All Combinations of Doubtful Variables

The process and formula for EBA can be found in C.1.1. Thus, I obtained the results shown in Table D.1 and Figure D.1. They summarize the results of the EBA analysis, evaluating the robustness of each explanatory variable under both normal and generic models. Variables with beta coefficients consistently skewed to one side of zero across models are considered more robust. The accompanying histograms and density plots visually reinforce this interpretation, with greater divergence from the null line ($\beta = 0$) indicating stronger and more consistent effects (Zheng and Lucey, 2025c).

Table D.1: Sala-i-Martin's EBA(In the Absence of Political Priorities-All Combinations of Doubtful Variables)

Variable	N: CDF($\beta \leq 0$)	N: CDF($\beta > 0$)	G: CDF($\beta \leq 0$)	G: CDF($\beta > 0$)
(Intercept)	0.000	100.000	0.000	100.000
Frequency	95.677	4.323	95.076	4.924
Timing	57.267	42.733	56.136	43.864
Intensity	12.404	87.596	14.706	85.294
Commitment	61.090	38.910	59.708	40.292
Urgency	51.245	48.755	51.024	48.976

Figure D.1: In the Absence of Political Priorities-All Combinations of Doubtful Variables



My primary focus is on the statistical significance and estimation results of the coefficients of these variables within the framework of a generalized model, in order to better understand their impact on the clean energy index in a broader modeling context. To assess statistical significance, I adopt the approach proposed by Hlavac (2016), considering a variable to be reliable within the Generic model if its coefficient exhibits at least an 80% probability of being either strictly positive or strictly negative.

As shown in Table D.1 and Figure D.1, the coefficient of the variable "Frequency" has over a 95% probability of lying to the left of zero, indicating a robust negative association. Similarly, "Intensity" demonstrates considerable stability, with more than 85% of its coefficients being positive under the Generic model. In contrast, "Timing", "Commitment", and "Urgency" exhibit less robustness in both the Normal and Generic models, as their coefficient distributions are relatively symmetric

around zero, suggesting a lack of clear directional influence. These preliminary results from the Extreme Bounds Analysis (EBA), conducted without filtering central bank speeches based on political priority, indicate that the frequency of authentic expressions and the intensity of tone in governors' speeches are more significantly correlated with the clean energy market index. However, the market appears less responsive to the timing or urgency of these speeches, implying that such dimensions may be less influential than media-related factors in shaping clean energy market dynamics. The influence of media-related factors has been previously examined in Zheng and Lucey (2025c), and will be further integrated into the analysis in 4.3.2.3 to provide a more comprehensive analysis.

D.1.2 Verifying the Accuracy of Variable Classification

The process and formula for EBA can be found in C.1.2. The resulting outcomes are summarized in Figure D.2 and Table D.2.

Table D.2: Sala-i-Martin's EBA (In the Absence of Political Priorities—Verifying the Accuracy of Variable Classification)

Variable	N: $\text{CDF}(\beta \leq 0)$	N: $\text{CDF}(\beta > 0)$	G: $\text{CDF}(\beta \leq 0)$	G: $\text{CDF}(\beta > 0)$
(Intercept)	0.000	100.000	0.000	100.000
Frequency	95.677	4.323	95.076	4.924
Timing	57.267	42.733	56.136	43.864
Intensity	12.404	87.596	14.706	85.294
Commitment	61.090	38.910	59.708	40.292
Urgency	51.245	48.755	51.024	48.976

Figure D.2 and Figure D.1 present consistent results, as do Table D.2 and Table D.1, further validating the robustness of my classification strategy. This consistency suggests that my variable classification is appropriate and does not bias the results. The core explanatory variables remain robust across both unrestricted and structured EBA settings.

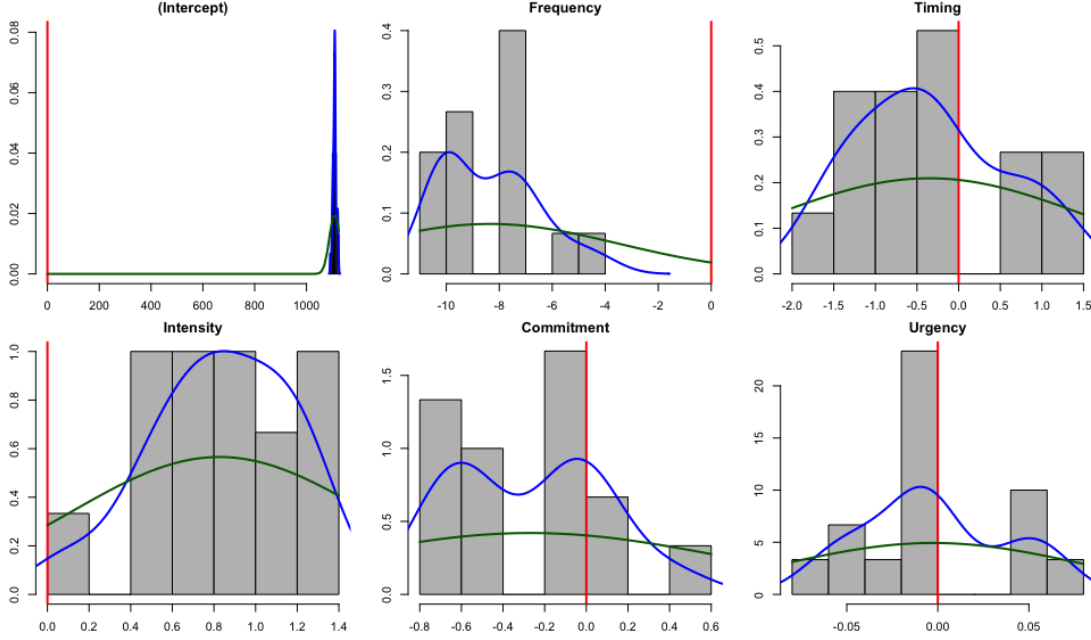
D.2 Presence

D.2.1 All Combinations of Doubtful Variables

The process and formula for EBA can be found in C.2.1.

APPENDIX D. EXTREME BOUND ANALYSIS-SOME RESULTS

Figure D.2: In the Absence of Political Priorities-Verifying the Accuracy of Variable Classification



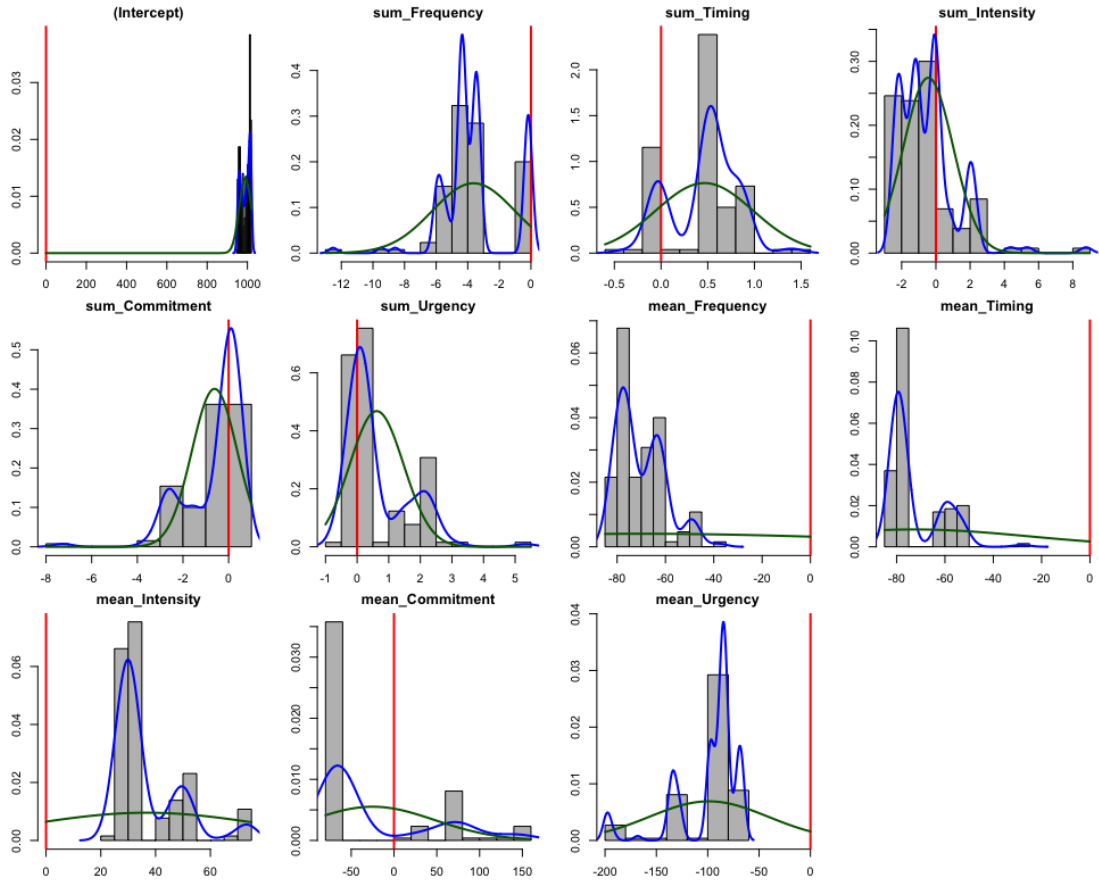
I followed a similar experimental procedure and obtained the results presented in Table D.3 and Figure D.3, where all variables are treated as doubtful variables.

Table D.3: Sala-i-Martin's EBA (In the Presence of Political Priorities-All Combinations of Doubtful Variables)

Variable	N: $\text{CDF}(\beta \leq 0)$	N: $\text{CDF}(\beta > 0)$	G: $\text{CDF}(\beta \leq 0)$	G: $\text{CDF}(\beta > 0)$
(Intercept)	0.000	100.000	0.000	100.000
sum_Frequency	88.266	11.734	91.633	8.367
sum_Timing	25.697	74.303	28.692	71.308
sum_Intensity	59.912	40.088	70.296	29.704
sum_Commitment	63.303	36.697	60.381	39.619
sum_Urgency	35.194	64.806	41.360	58.640
mean_Frequency	76.049	23.951	75.952	24.048
mean_Timing	93.771	6.229	93.209	6.791
mean_Intensity	18.913	81.087	19.878	80.122
mean_Commitment	62.882	37.118	69.292	30.708
mean_Urgency	94.885	5.115	95.210	4.790

Table D.3 and Figure D.3 demonstrate that the coefficients of "sum_Frequency", "mean_Timing", "mean_Intensity", and "mean_Urgency" have over an 80% probability of being either strictly positive or strictly negative. This suggests that, when political

Figure D.3: In the Presence of Political Priorities-All Combinations of Doubtful Variables



priorities are filtered from central bank governors' daily speeches, the frequency of truthful content is significantly negatively correlated with the clean energy index. Interestingly, the direction of this effect aligns closely with the findings presented in Table 4.2 and Figure 4.1. However, whether this variable exerts a robust influence on the clean energy market remains to be further verified through additional robustness checks—such as the inclusion of Variance Inflation Factors (VIF), heteroskedasticity-robust standard errors, and Likelihood Ratio Index (LRI).

Once political content is isolated, the variables "Timing" (speed of response) and "Intensity" (strength of response) consistently and robustly affect the clean energy market—negatively and positively, respectively. The composite indicator "Urgency" also consistently predicts market behavior. All variables represent daily averages.

D.2.2 Verifying the accuracy of variable classification

Table D.4: Sala-i-Martin's EBA (In the Presence of Political Priorities-Verifying the Accuracy of Variable Classification)

Variable	N: $\text{CDF}(\beta \leq 0)$	N: $\text{CDF}(\beta > 0)$	G: $\text{CDF}(\beta \leq 0)$	G: $\text{CDF}(\beta > 0)$
(Intercept)	0.000	100.000	0.000	100.000
sum_Frequency	88.266	11.734	91.633	8.367
mean_Frequency	76.049	23.951	75.952	24.048
sum_Timing	25.697	74.303	28.692	71.308
mean_Timing	93.771	6.229	93.209	6.791
sum_Intensity	59.912	40.088	70.296	29.704
mean_Intensity	18.913	81.087	19.878	80.122
sum_Commitment	63.303	36.697	60.381	39.619
mean_Commitment	62.882	37.118	69.292	30.708
sum_Urgency	35.194	64.806	41.360	58.640
mean_Urgency	94.885	5.115	95.210	4.790

The process and formula for EBA can be found in C.2.2. The consistency of results across Figure D.4, Table D.4, and Table D.3 with Figure D.3 lends further support to the reasonableness of my variable classification scheme.

D.3 Combining Traditional, Media-driven and Political Text-based Effects

D.3.1 All Combinations of Doubtful Variables

The process and formula for EBA can be found in C.3.1.

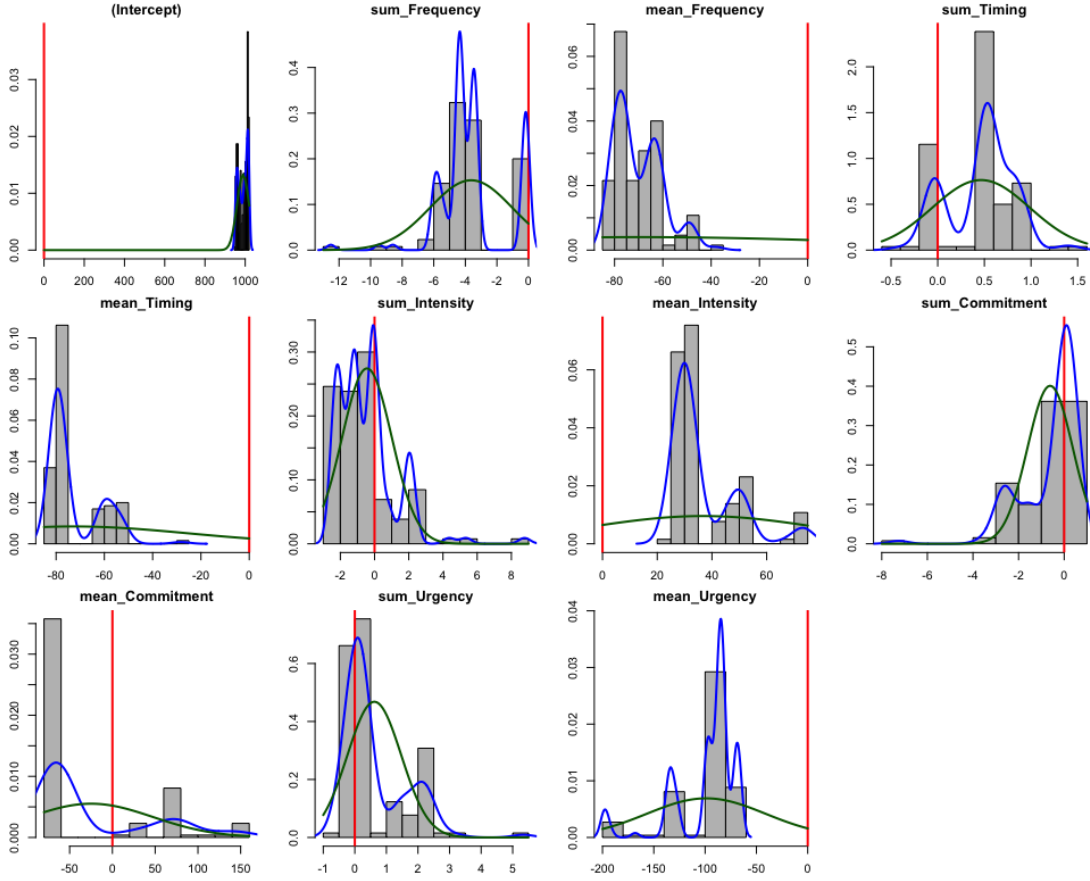
The effects of various dimensions on the clean energy market have been previously examined in Zheng and Lucey (2025c), as well as in 4.3.2.1 and 4.3.2.2. However, it should be noted that the significance tests for each coefficient may differ, as they are based on datasets with differing numbers of observations. In the following analysis, I change my focus toward assessing the influence of media-related and political text-based factors on the clean energy market, particularly in comparison to conventional determinants. Besides, I aim to evaluate the relative explanatory power of these two emerging dimensions in shaping market dynamics. To ensure the robustness of my findings, I further perform a time-aligned comparison, examining the stability of these variables across identical time windows. This approach allows us to assess whether temporal factors might represent omitted variables, potentially biasing previous estimates.

D.3. COMBINING TRADITIONAL, MEDIA-DRIVEN AND POLITICAL TEXT-BASED EFFECTS

Table D.5: Sala-i-Martin's EBA (All-All Combinations of Doubtful Variables)

Variable	N: CDF($\beta \leq 0$)	N: CDF($\beta > 0$)	G: CDF($\beta \leq 0$)	G: CDF($\beta > 0$)
(Intercept)	0.423	99.577	41.928	58.072
Frequency	29.300	70.700	28.977	71.023
Timing	22.941	77.059	26.007	73.993
Intensity	6.646	93.354	12.480	87.520
Commitment	6.450	93.550	9.356	90.644
Urgency	39.208	60.792	37.593	62.407
sum_Frequency	80.525	19.475	74.453	25.547
sum_Timing	34.424	65.576	52.981	47.019
sum_Intensity	64.729	35.271	68.600	31.400
sum_Commitment	53.789	46.211	61.555	38.445
sum_Urgency	39.998	60.002	56.290	43.710
mean_Frequency	72.037	27.963	69.206	30.794
mean_Timing	89.092	10.908	81.925	18.075
mean_Intensity	38.889	61.111	46.467	53.533
mean_Commitment	51.539	48.461	49.918	50.082
mean_Urgency	84.290	15.710	74.429	25.571
ReportCount_env	100.000	0.000	99.878	0.122
ReportCount_whole	100.000	0.000	96.121	3.879
average_goldstein_scale	99.969	0.031	91.786	8.214
average_confidence	0.000	100.000	0.842	99.158
AverageTone	0.000	100.000	0.001	99.999
ENV.number	28.219	71.781	34.281	65.719
TRI	25.032	74.968	28.916	71.084
PRI	32.667	67.333	41.003	58.997
GBI	9.597	90.403	28.873	71.127
Oil_Close	0.007	99.993	18.790	81.210
VIX_Close	0.693	99.307	15.441	84.559
GFSI	83.707	16.293	76.065	23.935
CHN_market	0.000	100.000	0.000	100.000
US_market	0.000	100.000	1.647	98.353
SP_IT	0.000	100.000	0.309	99.691
DXY	100.000	0.000	78.053	21.947
USD_CNY	100.000	0.000	99.995	0.005
MSCIEF	0.000	100.000	1.156	98.844

Figure D.4: In the Presence of Political Priorities-Verifying the Accuracy of Variable Classification



D.3.2 Verifying the Accuracy of Variable Classification

The process and formula for EBA can be found in C.3.2.

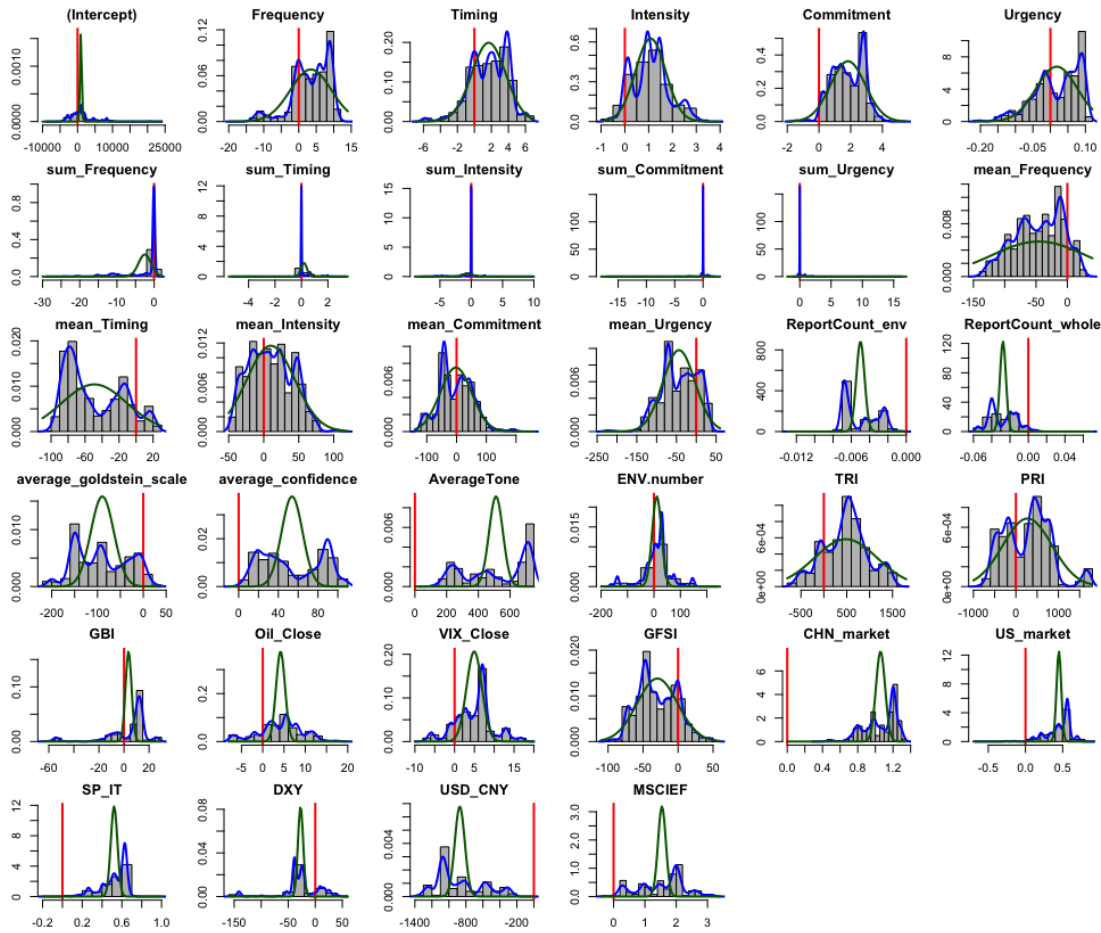
The rationality of the variable classification has been demonstrated by Figure D.6 and Table D.6.

D.3. COMBINING TRADITIONAL, MEDIA-DRIVEN AND POLITICAL
TEXT-BASED EFFECTS

Table D.6: Sala-i-Martin's EBA (All-Verifying the Accuracy of Variable
Classification)

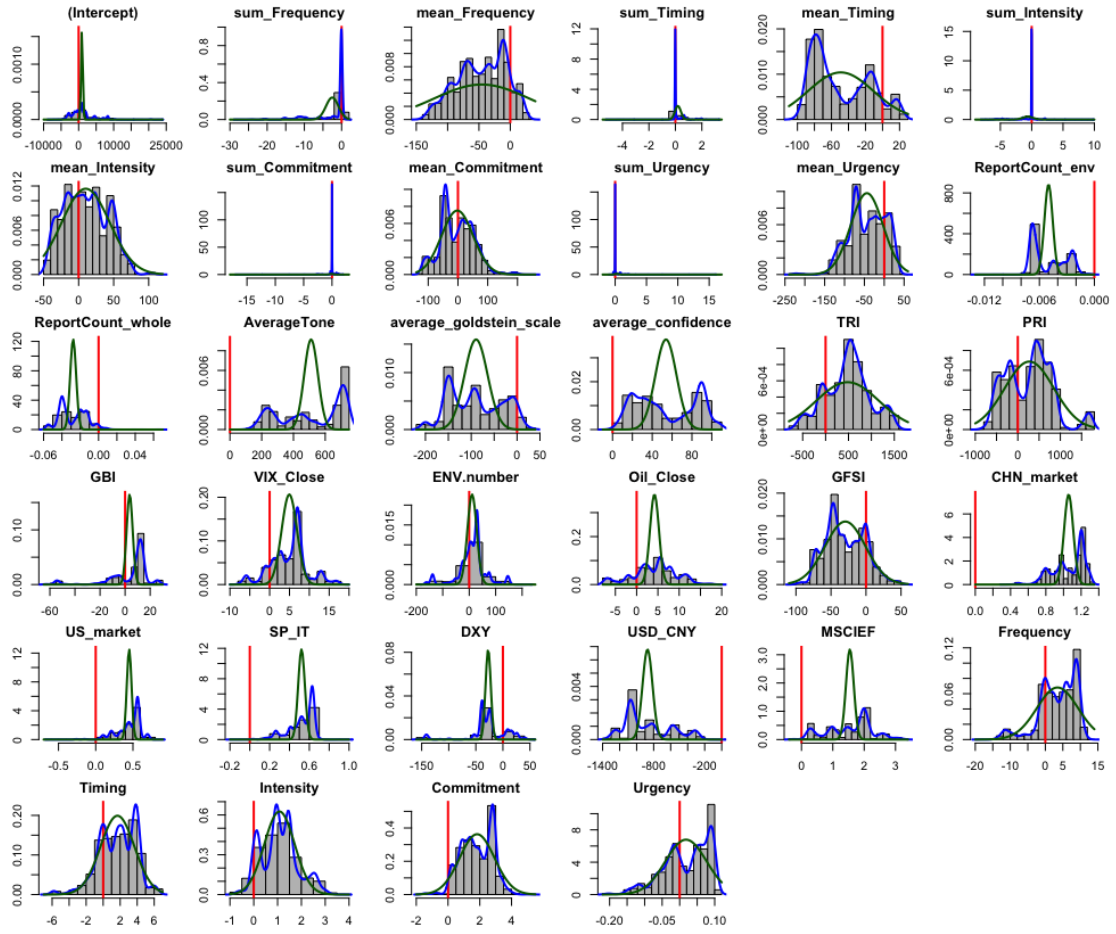
Variable	N: CDF($\beta \leq 0$)	N: CDF($\beta > 0$)	G: CDF($\beta \leq 0$)	G: CDF($\beta > 0$)
(Intercept)	0.423	99.577	41.928	58.072
sum_Frequency	80.525	19.475	74.453	25.547
mean_Frequency	72.037	27.963	69.206	30.794
sum_Timing	25.697	74.303	28.692	71.308
mean_Timing	89.092	10.908	81.925	18.075
sum_Intensity	64.729	35.271	68.600	31.400
mean_Intensity	38.889	61.111	46.467	53.533
sum_Commitment	53.789	46.211	61.555	38.445
mean_Commitment	51.539	48.461	49.918	50.082
sum_Urgency	39.998	60.002	56.290	43.710
mean_Urgency	84.290	15.710	74.429	25.571
ReportCount_env	100.000	0.000	99.878	0.122
ReportCount_whole	100.000	0.000	96.121	3.879
AverageTone	0.000	100.000	0.001	99.999
average_goldstein_scale	99.969	0.031	91.786	8.214
average_confidence	0.000	100.000	0.842	99.158
TRI	25.032	74.968	28.916	71.084
PRI	32.667	67.333	41.003	58.997
GBI	9.597	90.403	28.873	71.127
VIX_Close	0.693	99.307	15.441	84.559
ENV.number	28.219	71.781	34.281	65.719
Oil_Close	0.007	99.993	18.790	81.210
GFSI	83.707	16.293	76.065	23.935
CHN_market	0.000	100.000	0.000	100.000
US_market	0.000	100.000	1.647	98.353
SP_IT	0.000	100.000	0.309	99.691
DXY	100.000	0.000	78.053	21.947
USD_CNY	100.000	0.000	99.995	0.005
MSCIEF	0.000	100.000	1.156	98.844
Frequency	29.300	70.700	28.977	71.023
Timing	22.941	77.059	26.007	73.993
Intensity	6.646	93.354	12.480	87.520
Commitment	6.450	93.550	9.356	90.644
Urgency	39.208	60.792	37.593	62.407

Figure D.5: All-All Combinations of Doubtful Variables



D.3. COMBINING TRADITIONAL, MEDIA-DRIVEN AND POLITICAL TEXT-BASED EFFECTS

Figure D.6: All-Verifying the Accuracy of Variable Classification





SHAP VALUE-SUPPLEMENTARY EVIDENCE

SHAP (SHapley Additive exPlanations) integrates Shapley values from game theory with local explanation methods to provide consistent and interpretable feature attributions. It ensures coherence across different attribution techniques, particularly in high-dimensional settings. SHAP values offer local interpretability by measuring each feature's contribution to individual predictions, while traditional feature importance reflects a feature's overall contribution to model performance (Liu et al., 2023).

To assess the relative importance of each factor in predicting the clean energy index, I applied the SHAP value approach. The original dataset was randomly split into training and testing sets using an 80%/20% ratio. After conducting hyperparameter tuning, the resulting model achieved an R^2 of 0.978, indicating a strong fit. Therefore, the feature importance results derived from the SHAP values can be considered robust and reliable.

CONCLUSION

5.1 Summary of Results

This thesis makes four primary contributions. First, it delivers a reusable code and workflow for high-quality GDELT pre-processing and variables extraction, improving the reliability and reproducibility of media and sentiment signals for financial research. Second, I use these signals to see how they relate to the S&P Global Clean Energy Index and to compare them with traditional market factors. I also double-check the findings in several ways for their robustness. Third, it offers forecasting guidance by horizon based on a comparative evaluation of several time-series models, emphasizing that a simple baseline (ARIMA) is often competitive, complex models can overfit in small samples, and Prophet-related models are particularly well suited to the clean energy market. Fourth, it introduces and quantifies speech-based drivers linked to central bank governors' communications (e.g., "Commitment", "Intensity", "Timing", "Urgency"). Overall, the analysis focuses on correlations, not causal effects. The comparative overview is shown in Table 5.1.

5.1.1 Methods and Data Engineering

I develop custom code for data pre-processing and calculated the daily averages of Focus variables within the GDELT database. This approach lays a strong foundation for future research utilizing cloud computing with the GDELT database

Table 5.1: Comparative Overview

Element	Chapter 2	Chapter 3	Chapter 4
Dataset	GDELT + Bloomberg + Economic Policy Uncertainty	GDELT + Bloomberg + Economic Policy Uncertainty	BIS + GDELT + Bloomberg + Economic Policy Uncertainty
Methodology	EBA, Isolation Forest	SHAP, Statistical, Machine Learning, Hybrid models	EBA, SHAP, NLP metrics
Stable ables/Simulated Data	vari- Media confidence, sentiment, preferences (EBA robust) Oil price, volatility, Green Bond Index (partly stable)	Random Forest Simulated Data (more suitable)	Policy commitment, intensity, timing, urgency (speech-based), Media-related and policy commitment variables remain the most consistently correlated with clean energy index dynamics across methods
Feedback mechanism	Media–Market link	Media–Market signal response	Policy–Media–Market interaction
Research question	"What drives?"	"How to predict?"	"Who/why drives?"

and supports subsequent analyses of media topics. The incorporation of the EBA analysis, Isolation Forest, and the utilization of valuable data from the GDELT database significantly enhances the methodological originality and contribution of this study. I also propose a text–event alignment strategy that links unstructured political text fragments to their corresponding policy-event dates through an Iterative Matching approach, improving temporal precision.

5.1.2 Correlation Findings

5.1.2.1 Media, Sentiment, and Traditional Factors

Media and sentiment variables show stable and significant correlations with the clean energy index across EBA analysis. Several traditional predictors highlighted in prior studies lose robustness under my analysis. Market volatility is specification-sensitive and stabilizes only after heteroscedasticity adjustments and likelihood-ratio index weighting in Normal models. The Global Financial Stress Indicator is weaker than commonly assumed; although the Green Bond Index and U.S. Dollar Index appear robust under simpler Normal settings, they exhibit substantial uncertainty under stricter checks, suggesting that earlier conclusions should be reassessed. Other traditional controls remain relatively stable.

5.1.2.2 Speech-Based Drivers(Central Bank Communications)

Quantifying "Frequency", "Commitment", and "Intensity" shows that stronger, more emphatic statements are positively associated with investor perceptions and expectations regarding clean-energy investment, lending credibility to stated policy intentions. Analyses of "Timing" and "Urgency" indicate that shorter expected implementation lags and higher urgency correlate with clearer policy paths and faster policy responses, shaping investors' confidence and risk appetite. These patterns are particularly informative for the clean-energy markets of China and the United States.

5.1.3 Forecasting Results

5.1.3.1 Different Time Periods

5.1.3.1.1 Short term (1 week). My results suggest that ARIMA and Prophet provide reliable forecasts even without exogenous features, as the historical dynamics of the index produces tight intervals. For trend signaling, the Prophet_Features model, trained on Random Forest–simulated data, enhances performance.

5.1.3.1.2 Medium term (1 month). I find that ARIMA serves as a strong and interpretable baseline, which can be further refined with Prophet-based models. In this horizon, incorporating Random Forest–simulated features enhances the accuracy of trend forecasting.

5.1.3.1.3 Long term (6 months). The evidence indicates that, when augmented with Random Forest–simulated features, VAR provides the highest accuracy in point forecasting, while the Prophet_Features model provides the clearest forward-looking trend signals.

5.1.3.2 Model Interpretation

ARIMA should not be discounted merely because of its simplicity; in this setting it serves as a strong, competitive baseline rather than an underperformer. Although LSTM architectures can capture nonlinear dynamics, limited sample sizes make them prone to overfitting, which helps explain their comparatively lower accuracy. The Prophet model effectively captures the trend and seasonality of the clean energy index and, when enhanced with exogenous features, provides

reliable forward-looking signals. Feature engineering is consequential: Random Forest–simulated feature sets consistently improve performance across short-, medium-, and long-horizon forecasts. The accompanying summary table is intended as a concise reference for strengths, limitations, and accuracy; it does not imply causal relationships, nor does it designate a universal "best" model. My approach to tuning parameters and identifying the most influential factors works across different time periods in the clean energy markets.

5.1.4 Media Perspectives and Implications

Average Goldstein scores for the United States and China have been negative in recent years, but this does not mean that "more negative coverage improves performance". A more likely explanation is that adverse events, such as trade tensions, led both countries to put more focus on clean energy, which supported sector growth and investor confidence. Global news confidence and environmental sentiment are positively related to the index. The volume of coverage is statistically significant but small in economic terms, and the share of environmental reporting is unstable.

In practice, investors can align allocation and hedging with my horizon-specific guidance. Policymakers can adjust communication and macro-level tools to support the energy transition. Analysts can apply the text to event alignment and robustness checks to other markets.

5.2 Future Research and Outlook

Finally, this thesis gives investors a simple plan for portfolio allocation and risk management by forecast horizon (short, medium, long). It also offers policymakers a practical way to link communication, market expectations, and market reactions. I do not claim causality. However, I find that positive, trustworthy media coverage and clear policy communication are consistently linked with better clean-energy market performance.

My data pipeline and text-to-event framework can be reused. Future work can test these results in more markets and longer samples, or combine them with external policy shocks and formal identification to explore causal channels and pilot policies.

BIBLIOGRAPHY

- Afzali, E., Adegoke, A., Jin, Z., Qiu, W., and Wang, L. (2021).
Hybrid VAR-LSTM networks modeling and forecasting COVID-19 data in canada.
In Proceedings of the Statistical Society of Canada Annual Meeting. Statistical Society of Canada.
- Ahmad, W., Rahman, M. A., Shrimali, S., and Roy, P. (2024).
Tuning into climate risks: extracting innovation from television news for clean energy firms.
Technical Report 2409.08701, arXiv.
- Akaike, H. (1974).
A new look at the statistical model identification.
IEEE Transactions on Automatic Control, 19(6):716–723.
- Anand, A., Basu, S., Pathak, J., and Thampy, A. (2020).
Who moved the market? analyzing the role of central bank speeches.
Technical report, SSRN.
- Araujo, D. K. G. (2023).
gingado: A machine learning library focused on economics and finance.
- Arseneau, D. M. and Osada, M. (2023).
Central bank mandates and communication about climate change.
Technical report, Working paper.
- Avcı Azkeskin, S. and Aladağ, Z. (2025).
Evaluating regional sustainable energy potential through hierarchical clustering and machine learning.
Environmental Research Communications, 7(1):015002.
- Awad ALAfnan, M. (2020).

- COVID-19: the foreign virus: Media bias, ideology and dominance in chinese and american newspaper articles.
International Journal of Applied Linguistics and English Literature, 9(1):56.
- Baek, Y. and Kim, H. Y. (2018).
ModAugNet: A new forecasting framework for stock market index value with an overfitting prevention LSTM module and a prediction LSTM module.
Expert Systems with Applications, 113:457–480.
- Bagi, N. (2022).
Anomaly detection in financial data using isolation forest.
- Bengio, Y. (2009).
Learning deep architectures for AI.
Foundations and Trends in Machine Learning, 2(1):1–127.
- Bertoni, E., Fontana, M., Gabrielli, L., Signorelli, S., and Vespe, M., editors (2023).
Handbook of Computational Social Science for Policy.
Springer International Publishing, Cham.
- Bieganowski, B. and Slepaczuk, R. (2024).
Supervised autoencoder MLP for financial time series forecasting.
Technical Report 2404.01866, arXiv.
- Bilgen, S., Kaygusuz, K., and Sari, A. (2004).
Renewable energy for a clean and sustainable future.
Energy Sources, 26(12):1119–1129.
- Broadstock, D. C., Cao, H., and Zhang, D. (2012).
Oil shocks and their impact on energy-related stocks in china.
Energy Economics, 34(6):1888–1895.
- Bua, G., Kapp, D., Ramella, F., and Rognone, L. (2024).
Transition versus physical climate risk pricing in European financial markets: A text-based approach.
The European Journal of Finance, 30(17):2076–2110.
- Cherief, A., Sekine, T., and Stagnol, L. (2025).
A novel nature-based risk index: Application to acute risks and their financial materiality on corporate bonds.

Ecological Economics, 228:108427.

Cho, Y.-M., Pang, D., Thapa, S., Sherman, G., Ungar, L., Tay, L., and Guntuku, S. C. (2024).

Language-based valence and arousal expressions between the united states and china: A cross-cultural examination.

Technical Report 2401.05254, arXiv.

Chorzempa, M. (2021).

China, the united states, and central bank digital currencies: How important is it to be first?

China Economic Journal, 14(1):102–115.

Ciner, C., Kosedag, A., and Lucey, B. (2023).

Predictors of clean energy stock returns: an analysis with best subset regressions. *Finance Research Letters*, 55:103912.

Consultancy, D. (2025).

Unleashing the power of databricks: Azure synapse analytics and spark optimization techniques.

Dawar, I., Dutta, A., Bouri, E., and Saeed, T. (2021).

Crude oil prices and clean energy stock indices: lagged and asymmetric effects with quantile regression.

Renewable Energy, 163:288–299.

Dias, R., Horta, N., and Chambino, M. (2023).

Clean energy action index efficiency: An analysis in global uncertainty contexts. *Energies*, 16(9):3937.

Dikau, S. and Volz, U. (2021).

Central bank mandates, sustainability objectives and the promotion of green finance.

Ecological Economics, 184:107022.

Ehrmann, M. and Fratzscher, M. (2007).

Communication by central bank committee members: Different strategies, same effectiveness?

Journal of Money, Credit and Banking, 39(2–3):509–541.

- El Ghouli, S., Guedhami, O., Nash, R., and Patel, A. (2019).
New evidence on the role of the media in corporate social responsibility.
Journal of Business Ethics, 154(4):1051–1079.
- Elsayed, S., Thyssens, D., Rashed, A., Jomaa, H. S., and Schmidt-Thieme, L. (2021).
Do we really need deep learning models for time series forecasting?
Technical report, arXiv.
- Facebook (2017).
Prophet: Forecasting at scale.
- Fama, E. F. (1965).
Random walks in stock market prices.
Financial Analysts Journal, 21(5):55–59.
- Farmer, J. D. and Lo, A. W. (1999).
Frontiers of finance: Evolution and efficient markets.
Proceedings of the National Academy of Sciences, 96(18):9991–9992.
- Farouq, I. S., Umar Sambo, N., Ahmad, A. U., Jakada, A. H., and Danmaraya, I. A. (2021).
Does financial globalization uncertainty affect CO₂ emissions?
Quantitative Finance and Economics, 5(2):247–263.
- Fordahl, C. (2018).
Authenticity: The sociological dimensions of a politically consequential concept.
The American Sociologist, 49(2):299–311.
- Ghosh, I. and Jana, R. K. (2024).
Clean energy stock price forecasting and response to macroeconomic variables: A novel framework using Facebook’s Prophet, NeuralProphet and explainable AI.
Technological Forecasting and Social Change, 200:123148.
- Gómez, R. (2023).
geonamescache: A python library for accessing world geographical names and codes.
- Govindarajan, H. K. and Ganesh, L. S. (2022).
Integrating energy governance and environmental justice: Role of renewable energy.

- Renewable Energy Focus*, 43:24–36.
- Guo, T. and Lin, T. (2018).
Multi-variable LSTM neural network for autoregressive exogenous model.
Technical Report 1806.06384, arXiv.
- Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002).
Gene selection for cancer classification using support vector machines.
Machine Learning, 46(1):389–422.
- Hartvig, Á. D., Pap, Á., and Pálos, P. (2023).
EU climate change news index: forecasting EU ETS prices with online news.
Finance Research Letters, 54:103720.
- Henriques, I. and Sadorsky, P. (2008).
Oil prices and the stock prices of alternative energy companies.
Energy Economics, 30(3):998–1010.
- Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006).
A fast learning algorithm for deep belief nets.
Neural Computation, 18(7):1527–1554.
- Hlavac, M. (2016).
Extremebounds: Extreme bounds analysis in r.
Journal of Statistical Software, 72(9):1–23.
- Hollihan, T. A., Liang, S., Zou, Y., and Nguyen, H. (2021).
A comparative study of mediated public arguments on Trump’s trade war in the United States and China.
In Hample, D., editor, *Local Theories of Argument*, page 7. Routledge, New York.
- Imandoust, S. B. and Bolandraftar, M. (2013).
Application of k-nearest neighbor (KNN) approach for predicting economic events: Theoretical background.
International Journal of Engineering Research and Applications, 3(5):605–610.
- International Energy Agency (2021).
Net zero by 2050: A roadmap for the global energy sector.
Iea report, International Energy Agency, Paris.

Iomete (2025).

The ultimate guide to apache spark.

Javed Awan, M., Shafry Mohd Rahim, M., Nobanee, H., Munawar, A., Yasin, A., and Mohd Zain Azlanmz, A. (2021).

Social media and stock market prediction: A big data approach.

Computers, Materials & Continua, 67(2):2569–2583.

Jones, B. (2016).

Authenticity in political discourse.

Ethical Theory and Moral Practice, 19(2):489–504.

Kallunki, J.-P. (1997).

Handling missing prices in a thinly traded stock market: implications for the specification of event study methods.

European Journal of Operational Research, 103(1):186–197.

Kareem, M. S. (2024).

Anomaly detection in streaming data using isolation forest tree.

Master's thesis, University of Al-Qadisiyah, College of Computer Science and Information Technology.

Khalfaoui, R., Mefteh-Wali, S., Viviani, J.-L., Ben Jabeur, S., Abedin, M. Z., and Lucey, B. M. (2022).

How do climate risk and clean energy spillovers and uncertainty affect U.S. stock markets?

Technological Forecasting and Social Change, 185:122083.

Kyriazos, T. and Poga, M. (2023).

Dealing with multicollinearity in factor analysis: The problem, detections, and solutions.

Open Journal of Statistics, 13(3):404–424.

Larsson, K. and Ling, F. (2021).

Time series forecasting of the SP global clean energy index using a multivariate LSTM.

Master's thesis, KTH Royal Institute of Technology, School of Electrical Engineering and Computer Science (EECS).

- Leetaru, K. H. (2015).
Gdelt event database codebook, version 2.0.
- Leetaru, K. H. and Schrod, P. A. (2013).
Gdelt: Global data on events, location, and tone, 1979–2012.
In Proceedings of the ISA Annual Convention.
Conference paper introducing the GDELT database.
- Li, J., Song, L., Wu, D., Shui, J., and Wang, T. (2023a).
Lagging problem in financial time series forecasting.
Neural Computing and Applications, 35(28):20819–20839.
- Li, P., Pei, Y., and Li, J. (2023b).
A comprehensive survey on design and application of autoencoder in deep learning.
Applied Soft Computing, 138:110176.
- Liang, T., Chai, J., Zhang, Y.-J., and Zhang, Z. G. (2019).
Refined analysis and prediction of natural gas consumption in china.
Journal of Management Science and Engineering, 4(2):91–104.
- Lim, B. and Zohren, S. (2021).
Time-series forecasting with deep learning: a survey.
Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 379(2194):20200209.
- Liu, J., Li, C., Ouyang, P., Liu, J., and Wu, C. (2023).
Interpreting the prediction results of the tree-based gradient boosting models for financial distress prediction with an explainable machine learning approach.
Journal of Forecasting, 42(5):1112–1137.
- Lo, A. W. and Hasanahodzic, J. (2010).
The Heretics of Finance: Conversations with Leading Practitioners of Technical Analysis.
John Wiley and Sons.
- Loughran, T. and McDonald, B. (2020).
Textual analysis in finance.
Annual Review of Financial Economics, 12:357–375.

- Lyytimäki, J., Nygrén, N. A., Pulkka, A., and Rantala, S. (2018).
Energy transition looming behind the headlines? newspaper coverage of biogas production in finland.
Energy, Sustainability and Society, 8(1):15.
- Mangla, S. K., Luthra, S., Jakhar, S., Gandhi, S., Muduli, K., and Kumar, A. (2020).
A step to clean energy – sustainability in energy system management in an emerging economy context.
Journal of Cleaner Production, 242:118462.
- Matsubara, T., Akita, R., and Uehara, K. (2018).
Stock price prediction by deep neural generative model of news articles.
IEICE Transactions on Information and Systems, E101.D:901–908.
- McFadden, D. (1974).
Conditional logit analysis of qualitative choice behavior.
In Zarembka, P., editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, New York.
- Microsoft (2025a).
Microsoft azure.
- Microsoft (2025b).
Microsoft learn documentation.
- Mitchell, T. M. (1997).
Machine Learning.
McGraw-Hill, New York.
- Morrow-Howell, N. (1994).
The m word: multicollinearity in multiple regression.
Social Work Research, 18(4):247–251.
- Ning, Y., Kazemi, H., and Tahmasebi, P. (2022).
A comparative machine learning study for time series oil production forecasting: ARIMA, LSTM, and prophet.
Computers & Geosciences, 164:105126.
- Nyman, R., Ormerod, P., Smith, R., and Wiesenthal, J. (2021).

- News and narratives in financial systems: exploiting big data for systemic risk assessment.
Journal of Economic Behavior & Organization, 161:237–253.
- Obama, B. (2017).
The irreversible momentum of clean energy.
Science, 355(6321):126–129.
- Ofer, A. R. and Siegel, D. R. (1987).
Corporate financial policy, information, and market expectations: An empirical investigation of dividends.
The Journal of Finance, 42(4):889–911.
- Park, J., Lee, H. J., and Cho, S. (2023).
Hot topic detection in central bankers’ speeches.
Expert Systems with Applications, 230:120563.
- Pham, L. (2019).
Do all clean energy stocks respond homogeneously to oil price?
Energy Economics, 81:355–379.
- Polamuri, D., Srinivas, K., and Mohan, A. (2019).
Stock market prices prediction using random forest and extra tree regression.
International Journal of Recent Technology and Engineering, 8:1224–1228.
- Ren, X., Li, J., He, F., and Lucey, B. (2023).
Impact of climate policy uncertainty on traditional energy and green markets: Evidence from time-varying granger tests.
Renewable and Sustainable Energy Reviews, 173:113058.
- Roche, L., Arendt, R., Bach, V., and Finkbeiner, M. (2023).
The social impacts of resource extraction for the clean energy transition: a qualitative news media analysis.
The Extractive Industries and Society, 13:101213.
- Saeed, T., Bouri, E., and Alsulami, H. (2021).
Extreme return connectedness and its determinants between clean/green and dirty energy investments.
Energy Economics, 96:105017.

- Salmimaa, L. (2024).
Enhancing outlier detection with semantic feature extraction.
Master's thesis, Tampere University.
- Schoppa, L., Disse, M., and Bachmair, S. (2020).
Evaluating the performance of random forest for large-scale flood discharge simulation.
Journal of Hydrology, 590:125531.
- Shah, D., Campbell, W., and Zulkernine, F. H. (2018).
A comparative study of LSTM and DNN for stock market forecasting.
In *2018 IEEE International Conference on Big Data (Big Data)*, pages 4148–4155.
- Shahzad, K., Raza, H. A., and Shahbaz, M. (2024).
Environmental regulations' impact on clean energy consumption: Under the assistance of governance measures.
Renewable Energy, 237:121766.
- Shang, J., Nakamura, T., and Hamori, S. (2023).
Does the sentiment index help predict crude oil prices?
IETI Transactions on Data Analysis and Forecasting, 1(1):54–65.
- Shen, S., Xia, L., Shuai, Y., and Gao, D. (2022).
Measuring news media sentiment using big data for chinese stock markets.
Pacific-Basin Finance Journal, 74:101810.
- Siame-Namini, S., Tavakoli, N., and Siame Namin, A. (2018).
A comparison of arima and lstm in forecasting time series.
In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1394–1401.
- Ślepaczuk, R. and Zenkova, M. (2018).
Robustness of support vector machines in algorithmic trading on cryptocurrency market.
Central European Economic Journal, 5(52):186–205.
- S&P Dow Jones Indices (2024).
S&p global clean energy index: Equity report.
Equity index report.

- Sposito, H. (2021).
poldis: Tools for analyzing political discourse.
- Sposito, H. (2024).
Radiating truthiness: Authenticity performances in politics in brazil and the united states.
Political Studies, page 00323217241261229.
- Su, Y., Lucey, B. M., and Jha, A. K. (2024).
Finance research and the UN sustainable development goals: an analysis and forward look.
Research in International Business and Finance, 71:102463.
- Sutton, R. S. and Barto, A. G. (1998).
Reinforcement Learning: An Introduction.
MIT Press, Cambridge, MA.
- Taylor, S. J. and Letham, B. (2017).
Forecasting at scale.
Technical report, PeerJ Preprints.
- The GDELT Project (2022).
Gdelt.
- Toms, M. C. (2011).
The technical analysis method of moving average trading : rules that reduce the number of losing trades.
- Wang, J., Ma, F., Bouri, E., and Zhong, J. (2022).
Volatility of clean energy and natural gas, uncertainty indices, and global economic conditions.
Energy Economics, 108:105904.
- Wei, X. (2025).
Prediction and influence factors analysis of IP backbone network traffic based on prophet model and variance reduction.
Heliyon, 11(1):e41472.
- White, H. (1980).

- A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity.
Econometrica, 48(4):817–838.
- Xu, Y., Dong, H., Zhou, M., Xing, J., Li, X., and Yu, J. (2021).
Improved isolation forest algorithm for anomaly test data detection.
Journal of Computer and Communications, 9(8):48–60.
- Zhang, H., Hong, H., and Ding, S. (2023a).
The role of climate policy uncertainty on the long-term correlation between crude oil and clean energy.
Energy, 284:128529.
- Zhang, Y., Tang, S., and Yu, G. (2023b).
An interpretable hybrid predictive model of COVID-19 cases using autoregressive model and LSTM.
Scientific Reports, 13(1):6708.
- Zheng, J., Xin, D., Cheng, Q., and Tian, M. (2024).
The random forest model for analyzing and forecasting the US stock market in the context of smart finance.
Technical Report 2402.17194, arXiv.
- Zheng, Y. and Lucey, B. (2025a).
Central bankers’ political discourse as a driver of clean energy markets.
Research in International Business and Finance, 80:103126.
- Zheng, Y. and Lucey, B. (2025b).
Supervised learning models, statistical models or hybrid models? a prediction of clean energy stock based on fear and fundamental factors.
Finance Research Open, 1(2):100006.
- Zheng, Y. and Lucey, B. (2025c).
What really drives clean energy stocks – fear or fundamentals?
Energy Economics, 148:108558.
- Zhou, Y., Li, Y., and Xia, S. (2009).
An improved KNN text classification algorithm based on clustering.
Journal of Computers, 4(3):230–237.

Zivot, E. and Wang, J. (2003).

Vector autoregressive models for multivariate time series.

In *Modeling Financial Time Series with S-Plus*, pages 385–429. Springer, New York, NY.