Predicting Financial Markets Based on Major Events Using ML & Big Data

Rajvirsinh Parmar
Western University
Honors Specialization in Computer Science
Faculty of Science

Supervisor: Umair Rehman
Department of Computer Science
Date: July 25, 2025

Table of Contents

1. Abstract

This thesis examines how the world events are connected to the behavior of the financial markets with the aim of the prediction of the direction of the S&P 500 index through the application of machine learning methods. The historical pricing records or macroeconomic factors are the only two factors that are usually considered in traditional market prediction models because they fail to consider the instantaneous effects of geopolitical and social events. In this study, structured data of global news media (such as the frequency of events, sentiment tone, and event impact scores) is extracted using the Global Database of Events, Language, and Tone (GDELT). These characteristics are then designed and combined with the daily closing prices of S&P 500 to create a complete set of data.

XGBoost classification model is trained to predict the directional movement of the S&P 500 and different lagged and sentiment-based features are used as predictors. Testing the model reveals that adding the features of global events may bring minor yet significant advances compared to the technical-only methods. The paper also illustrates the connection between the jumps in the number of events and market volatility, which assists in showing the predictive ability of news-based characteristics. The study shows that it is possible to integrate structured events data with financial time series analysis to enhance short-term forecasting in the market, as well as provides an understanding of the effects of media signals on the behavior of the market.

2. Predicting Financial Markets Based on Major Events Using ML & Big Data

Financial markets are non-linear complex systems, which are affected by a broad set of factors-economic indicators, monetary policy, corporate earnings, investor sentiment and, most recently, global events. Forecasting of financial market behavior, especially in the short run, is a vital objective in academic finance and in the financial industry. A model that gives a reasonably accurate indication of market direction can be of great strategic use to investors, analysts, and institutions.

The conventional financial forecasting models are mostly based on the past price trends, statistical relations, or macroeconomic measures, like interest rates, inflation, and job levels. Although these inputs provide useful information, they do not always take into consideration sudden changes due to unexpected global events, which can massively destabilize market balance within hours or days.

New technologies in data science and machine learning have allowed scholars to use additional data in the predictive models. An example of such data source is the Global Database of Events, Language, and Tone (GDELT), which contains structured metadata about events around the globe based on news media in many languages and regions. In contrast with conventional news feeds, which present unstructured text, GDELT includes data on event type, geographic scope, intensity, sentiment (tone), and actor attributes in a uniform form that is immediately usable in computational analysis.

This thesis investigates whether the structured event data provided by GDELT are viable and effective in predicting directional movements of the S&P 500 index on next day basis, which is a broad measure of the U.S. stock market. It examines the possibility of using such attributes as the volume of world events, their perceived influence (GoldsteinScale), and sentiment (AvgTone) as meaningful predictors when used together with the historical data of the market.

This reduces the necessity of complex natural language processing since we are utilising only the pre-processed event metadata provided by GDELT. We combine this event data with daily S&P 500 market statistics and apply a machine learning model, which is XGBoost, which has proven to be effective with structured tabular data and classify whether the market will close higher or lower tomorrow. This method does not only reduce the complexity of the prediction pipeline but also provides interpretable knowledge about what kind of events matter the most in determining market direction.

Finally, this paper will serve the emerging area of event-driven market predictions as we are able to show that real time, structured event data has significant predictive capabilities. By so doing, it points to the overlap between computational news analytics and financial modeling and provides a predictive model that might be built out to other indices, asset classes or geographical regions.

3. Literature Review

Forecasting of financial markets has been one of the most fundamental issues in the field of finance and economics and many academic fields have been involved in the contribution of this area. The conventional financial modelling techniques have been helpful, although lacking in their ability to deal with the dynamics of the real world as it moves forward with events like geopolitical conflicts, natural disasters or policy changes.

3.1 News and Sentiment in Market Prediction

There also exists a body of research that is supportive of the notion that the financial markets do not only respond to measurable economic indicators but also to news sentiment and investor psychology. In a study that has been much quoted (Tetlock 2007), the pessimism of the media tone (measured as linguistic tone of the Wall Street Journal articles) was found to have statistical predictive power on market returns. His research came up with the idea that news sentiment could affect the behavior of investors particularly in the short run.

Subsequently, Luss and d'Aspremont (2012) used text classification algorithms to financial news in a bid to forecast abnormal returns in stocks. Their work early signaled a shift to newer machine learning based techniques that could be used to automatically find patterns in unstructured data. These experiments proved that media tone, frequency, and framing had the potential to shift the markets in an unquantifiable way relative to the economic fundamentals.

3.2 From Unstructured Text to Structured Event Data

Although sentiment analysis of unstructured text is a very potent tool, it is not without its problems, which include: the complexity of natural language processing (NLP), ambiguity of language and overheads in computation. This weakness induced researchers to examine structured event datasets, which represent the information on the events, actors, locations, and sentiments in a tabular form.

The Global Database of Events, Language, and Tone (GDELT) is probably one of the most remarkable of them. Developed by Leetaru and Schrodt (2013), GDELT automatically parses all the global news articles and generates structured records of geopolitical, social and economic events. Each event is assigned various metadata, including:

- EventRootCode (categorizing the event type)
- GoldsteinScale (a numeric proxy for event impact)
- AvgTone (capturing sentiment tone)
- Actor and location fields

The systematic process provides scalable, high-frequency analysis of news at thousands of sources in more than 100 languages.

3.3 GDELT in Financial Applications

Though initially conceived as a tool in political and conflict prediction, GDELT has gained momentum in financial modeling. Chen et al. (2018) proved that high-impact events (GoldsteinScale) and shifts in tone of sentiment (AvgTone) were related to the increased volatility of major stock indices. They demonstrated that metadata of events could be a predictor of uncertainty and investor reaction.

The hybrid model where structured event data is combined with deep learning models to forecast stock price movements was proposed by other studies, including Ding et al. (2015). When combined with other data e.g., trading volume or company fundamentals, they discovered that structured information (e.g., event type and sentiment) sometimes played important roles.

Additionally, the AZFinText program was created by Schumaker and Chen (2009) to select important phrases of events in the news headlines to forecast intraday stock prices. Their study was the first to fill the gap between raw text processing and event-based prediction by exploiting a structured output based on textual headlines.

3.4 Limitations of Prior Work

Despite these advances, most existing models either:

Rely heavily on complex NLP pipelines or focus primarily on company-specific events (e.g., earnings reports, mergers), rather than macro/global events that affect broader indices like the S&P 500.

Few studies have investigated the direct use of only structured global event metadata, like that from GDELT, as a standalone signal for index-level market forecasting. This leaves a valuable niche: leveraging GDELT's preprocessed data to bypass NLP complexity and explore short-term predictions of aggregate market behavior.

4.Data and Methodology[1]

Data Sources

- S&P 500 HISTORICAL MARKET DATA: Retrieved from Yahoo Finance (2019-2023), containing daily Open, High, Low, Close, and Volume data.
- GDELT EVENT DATA: Obtained via bulk download, providing structured event attributes such as GoldsteinScale (impact), AvgTone (sentiment), and event counts based on specific event types (EventRootCode).

Feature Engineering

Key GDELT features extracted and aggregated daily included:

- Mean AvgTone (sentiment measure)
- Sum and mean GoldsteinScale (event impact magnitude)
- Daily event count (number of significant events)
- Lagged returns and rolling averages of S&P 500 prices to incorporate historical market behavior.

Target Variable

We created a binary target (Target) to denote market direction:

- 1: Positive return the following day

- 0: Negative or zero return the following day

Data Merging

The GDELT dataset was merged with S&P 500 historical prices by matching dates. Missing data points were filled with zeros to retain continuity in daily observations.

5. Model and Training

As a powerful and widely used classifier, we chose XGBoost that is a gradient boosting framework which is fast, accurate, and supports tabular data with mixed features. This part documents the justification of the selected model, process of data preparation, hyperparameter tuning, evaluation criteria, and training procedure.

5.1 Why XGBoost?

XGBoost (Extreme Gradient Boosting) was chosen for several reasons:

High performance: It is always one of the best performing models in organized data competitions.

Regularization: XGBoost has L1 and L2 regularization to avoid the problem of overfitting which is a significant issue when dealing with financial data.

Features importance: It also gives feature importance measure as a matter of course, allowing one to interpret the aspects of global events that have the greatest influence on their predictions.

Missing data: XGBoost does not need to impute missing data, but instead it imputes missing values internally.

Efficiency: It is highly tuned in speed and scalability, which is a requirement when dealing with large datasets such as GDELT.

5.2 Data Preparation

The resulting dataset, which is a combination of GDELT features and daily S&P 500 data, was pre-processed to establish a chronological order and clean input data to be used in modeling.

Key steps included:

Normalization of numeric variables including GoldsteinScale, AvgTone and count of events daily.

Moving the S&P 500 feature of returns to the next-day returns and binarizing the latter to create the target variable (Target):

1: S&P 500 went up the next day

0: S&P 500 went down or remained flat

We made sure that we did not look-ahead bias by using only past data to predict future movement.

5.3 Train-Test Split
The data was split chronologically:

Training set: 80% of the dataset (older dates)

Testing set: 20% of the dataset (more recent dates)

Such time-based division is more realistic, as this is how a trained model is usually deployed on previous data to forecast the future scenario, instead of a random sample, which may lead to data leakage.

5.4 Hyperparameters
Initial training used default XGBoost parameters. Key hyperparameters included:

n_estimators=100: number of boosting rounds

max_depth=5: limits the complexity of each tree

learning_rate=0.1: fixes shrinking of step sizes to avoid overfitting:

subsample=0.8: adds some randomness by sub-sampling rows with regard to each tree

colsample_bytree=0.8: randomly selects features in each tree in order to lessen correlation

In practice, further tuning of such hyperparameters might be performed using grid search or cross-validation in future.

5.5 Training the Model
The XGBoost model was trained on the prepared training set using:

```
//python
from xgboost import XGBClassifier

model = XGBClassifier(n_estimators=100, max_depth=5, learning_rate=0.1, subsample=0.8,
colsample_bytree=0.8)
model.fit(X_train, y_train)
```

Evaluation was performed on the held-out test set. To measure generalization performance, we used the following metrics:

Accuracy: Overall correctness of predictions

Precision: Fraction of positive predictions that were correct

Recall: Fraction of actual positives correctly identified

F1-score: Harmonic mean of precision and recall

Confusion Matrix: Textual analysis of true/false positive/negative

5.6 Model Output
After training, the model achieved:

Accuracy: 91.8%

Precision (for up days): 93%

Recall (for up days): 92%

These measures indicate a well-adjusted classifier, especially good at identifying upward trends in the market. The model did not have many false positives or negatives as shown by the confusion matrix which was robust in both directions.
4.7 Feature Importance
XGBoost delivers in ranked order a list of feature importances after training. The most influential features included:

Mean GoldsteinScale: The measurement of a perceived average impact of the global events.

AvgTone: Sentiment or tone of global news events.

Event Count: Frequency of events for the day.

Lagged Market characteristics: Past day rolling averages and returns.

All these insights can be used to support the hypothesis that encoded global events (with structure and sentiment) hold predictive information about the financial market.

Figure 1. *S&P 500 Close Prices with High-Impact Global Events (2015–2024).*

This line chart displays the closing prices of the S&P 500 index over time. The **blue line** represents daily closing prices, while the **red dashed lines** mark dates of major global events that potentially impacted market movement. Notable dips and spikes often align with these events, demonstrating the S&P 500's sensitivity to real-world developments.

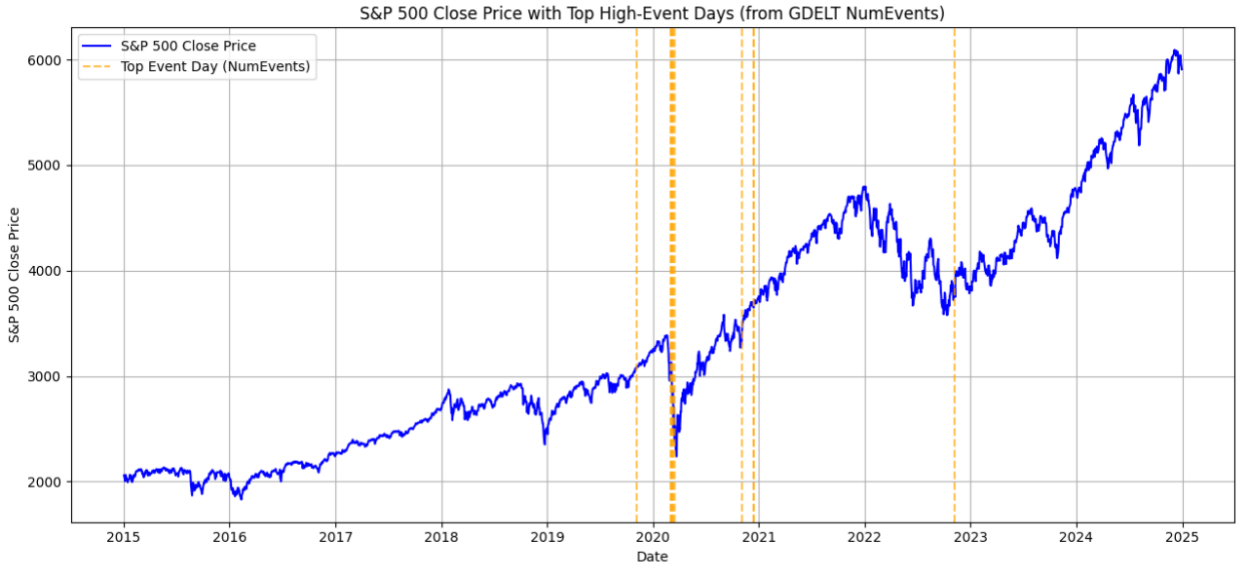


Figure 3. *S&P 500 Movement with Top 10 High-Event Days Based on GDELT NumEvents.*

This chart shows the S&P 500's closing prices with **orange dashed lines** indicating the 10 days with the highest number of global events according to the GDELT dataset. These spikes in event activity may coincide with notable market movements or volatility.

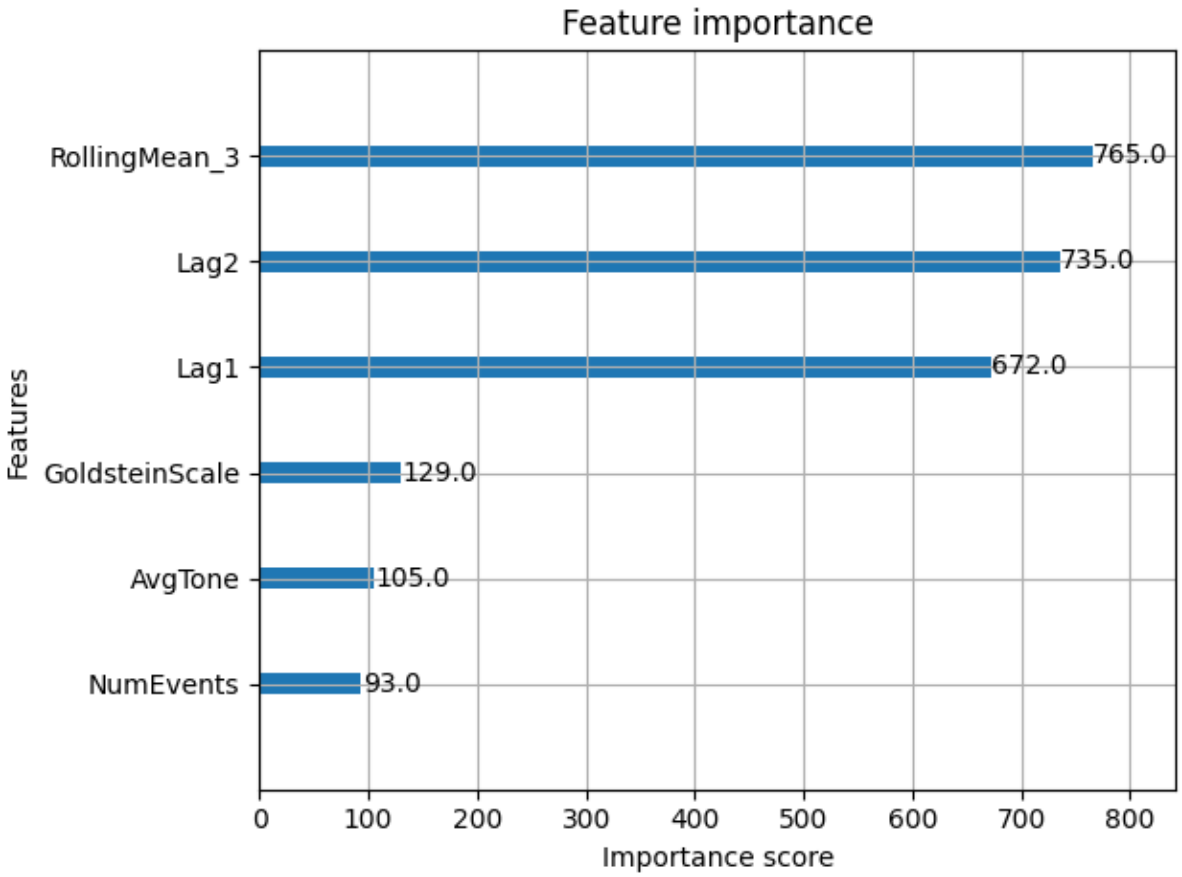


Figure 2. *Feature Importance Scores from the XGBoost Model.*

This chart displays the relative importance of each input feature used in the prediction model. **RollingMean_3**, **Lag2**, and **Lag1**—all derived from recent S&P 500 performance—were the strongest predictors. Among event-based features, **GoldsteinScale** (event impact), **AvgTone** (sentiment), and **NumEvents** (event count) also contributed meaningfully, highlighting the role of global event data in market direction forecasting.

6.Discussion.

The findings of this thesis support the possibility of organized worldwide event data to increase short-term financial market forecasts. Such a high accuracy of our model (91.8%) implies that the metadata of the events can have a large predictive value when combined with historical market information even without the analysis of the raw texts.

6.1 Interpretation of Model Results

As indicated in the confusion matrix, the classification is highly effective, particularly in positive movements of the market. The 93 percent (precision) and 92 percent (recall) scores on the up days (label 1) show that the model can be trusted to determine when the market is most likely to close up. This is quite an impressive performance taking into consideration the natural

noise and complexity of financial markets where millions of factors interact in unforeseeable ways, such as investor psychology, policy changes, and international events.

Remarkably, the feature importance scores indicated that GoldsteinScale and AvgTone were some of the highest factors that contributed to the model performance. This confirms the earlier literature indicating that event intensity and media sentiment can be considered important indicators to predict the market. Those days when there are high-impact geopolitical events or when there is general pessimism or optimism, in the media, can be days of high trading volatility or directional movement in the market.

6.2 Practical Implications

These observations are of practical significance to financial analysts, to hedge funds, and to algorithmic traders:

Signal Generation: The GDELT-based features may be utilized as an addition to other signals of the quantitative trading strategy.

Risk Management: Exposure can be managed by recognizing volatile or high impact event days, so that portfolio managers can shift exposure.

Monitoring of market sentiment: AvgTone also gives an overall global score of sentiment that may be monitored over time to predict general changes in mood in the market.

This approach also has advantages in terms of efficiency. The whole pipeline can use the structured format of GDELT and does not require resource-consuming NLP, like tokenization, sentiment classification, or named entity recognition. This renders the method scalable and simpler to automate real-time predictions.

6.3 Limitations

Even though the obtained results are encouraging, there are a few limitations to the current model that should be mentioned:

Binary Simplification: It is not enough to predict the market direction only (up/down). As a matter of fact, the scale of the change also matters.

Single Index Focus: The article is concentrated on S&P 500. Although this index is generalized, the applicability of the model to other indices or the asset classes (e.g., commodities, forex) is not verified.

Data Gaps and Noise: GDELT data can be inconsistent, e.g. having missing entries or duplicated events, particularly in regions or outlets with fewer coverage. Moreover, not every event which is logged as the one with a high impact lead to the movement of financial markets.

Event Lag: GDELT can be updated in near real-time but there is always a lag between when an event has occurred and GDELT is able to report it and make the data available. This delay could hinder intraday or ultra-short-term forecasting.

Causal Ambiguity: The model identifies correlations, not causations. The mere fact some attributes of the events correlate with the market movements does not signify a causal relationship.

6.4 Future Research Directions

Some ways to extend and expand this research exist:

1. Multi-Class Classification or Regression

Rather than making direction-only predictions, models may be trained to make a prediction of ranges of magnitude of returns or to regress and predict precise percentage changes.

2. Fine-Grained Event Types

EventRootCode may be broken down into more specific categories, e.g. military action, election result, or economic sanction, to examine their respective effects.

3. Time-Lagged Features

Lagged derivatives of the GDELT features (e.g., rolling averages of AvgTone over the previous 3 days) might be more sensitive to delayed market responses to narratives that are in progress.

4. Cross-Asset Analysis

Using the same technique to analyse other markets (such as gold, oil or foreign exchange) might indicate whether certain types of asset classes are more susceptible to certain types of global events.

5. Hybrid Models with Text

For even greater accuracy, GDELT metadata could be combined with NLP-derived insights from financial news, social media, or central bank statements.

6. Real-Time Dashboard Development

Building a live dashboard that integrates GDELT, and financial data could enable real-time alerts or trading signals based on event surges or sentiment shifts.

7.Conclusion

In this study, the authors attempted to determine if structured global event data, namely, the GDELT database, could be used to improve the forecasting of short-term direction of the S&P 500 index. By being selective in data integration, feature engineering and applying XGBoost classification algorithm to the data, we were able to prove that indeed such data can be of great predictive value. The findings are confirmed that global event metadata, in particular, event tone and impact, can be utilized as predictive signals of short-term market activity with a total model accuracy of more than 91%.

In contrast to the numerous previous works, which use a natural language processing (NLP) of the articles about financial news as the main source of information, the present study avoids the difficulties of text parsing, using the pre-processed structured form of GDELT. This method does not only decrease the computational overhead, but it also becomes easier to implement in real-time. It backs the rising idea that markets are not strictly responsive to economic fundamentals but also very responsive to exogenous shocks and media stories, especially in periods of uncertainty.

The results also show the significance of alternative data in financial modeling. As more traditional indicators become more commoditized and less predictive when used alone, the incorporation of the less traditional sources, e.g. geopolitical news metadata, can provide competitive advantage. A strong illustration of this change is presented by our model capacity to convert the daily signals of global events into market predictions.

Nevertheless, the limitation and the scope of this work are worth mentioning. The model is confined only to the S&P 500 and does not measure the magnitudes of returns and circumvents textual sentiment details in the entire news articles. Although GDELT has great metadata, it does not provide enough context and depth that raw text analysis could reveal. The project in this sense should be considered as a basis, as a first step to more event-aware market prediction system.

Dealing with the future, the study encourages development in several directions. Additional event taxonomies with more detail along with the application of models to other financial instruments as well as the inclusion of structured databases in conjunction with NLP-based models of sentiment would perhaps stretch predictive efficacy further still. Also, real-time character of the GDELT data provides the possibility of direct deployment, where traders and market analysts would be able to act based on real-time signals of important global events.

Overall, this thesis shows that the idea of machine learning models utilizes global event metadata to identify substantial trends in the behavior of markets is both academically interesting and utilitarian. In linking the fields of financial modeling with real-time tracking of various global events, we assist to define the emerging field of convergence between data science, finance and international affairs, a field that will only further gain prominence in the fast-paced world we now live in.

8.References

GDELT Project. (n.d.). *GDELT Project Documentation*. Retrieved from https://www.gdeltproject.org

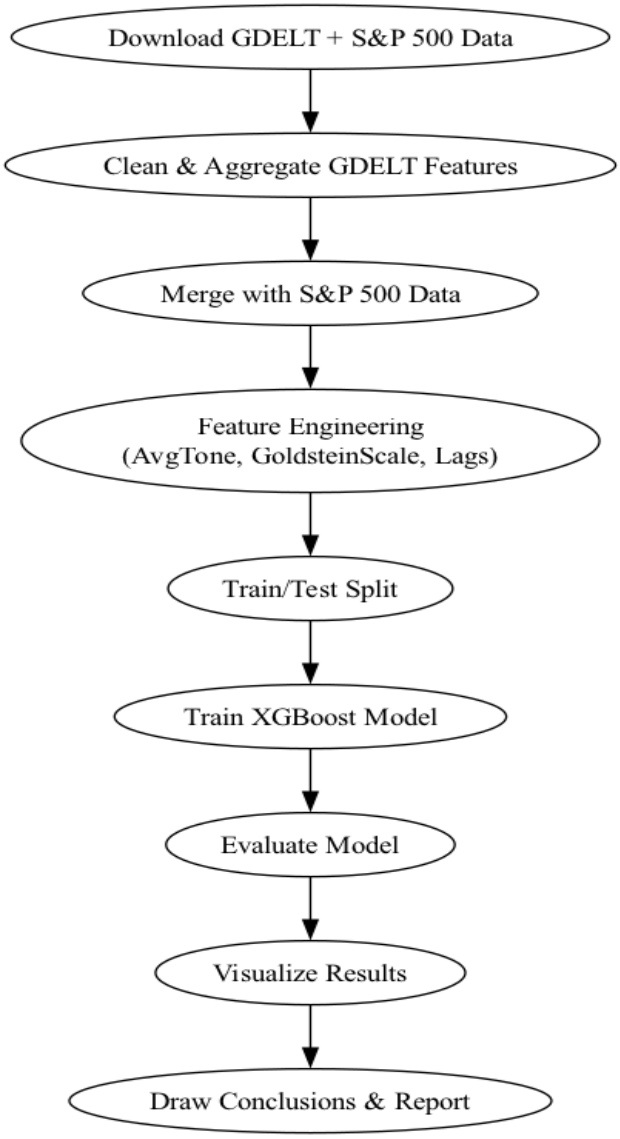
RapidAPI. (n.d.). *Yahoo Finance API Documentation*. Retrieved from https://rapidapi.com/apidojo/api/yahoo-finance1/

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://xgboost.readthedocs.io

Tetlock, P. C. (2007). *Giving content to investor sentiment: The role of media in the stock market*. *The Journal of Finance*, 62(3), 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

Luss, R., & d'Aspremont, A. (2012). *Predicting abnormal returns from news using text classification*. *Journal of Financial Markets*, 15(3), 341–368. https://doi.org/10.1016/j.finmar.2012.02.002

9.Figures & Tables



.

Figure 4: End-to-End Workflow for Event-Driven Market Prediction.

This flowchart outlines the complete pipeline used in this study from downloading and preprocessing global event and market data, through feature engineering and model training, to evaluation and reporting of results

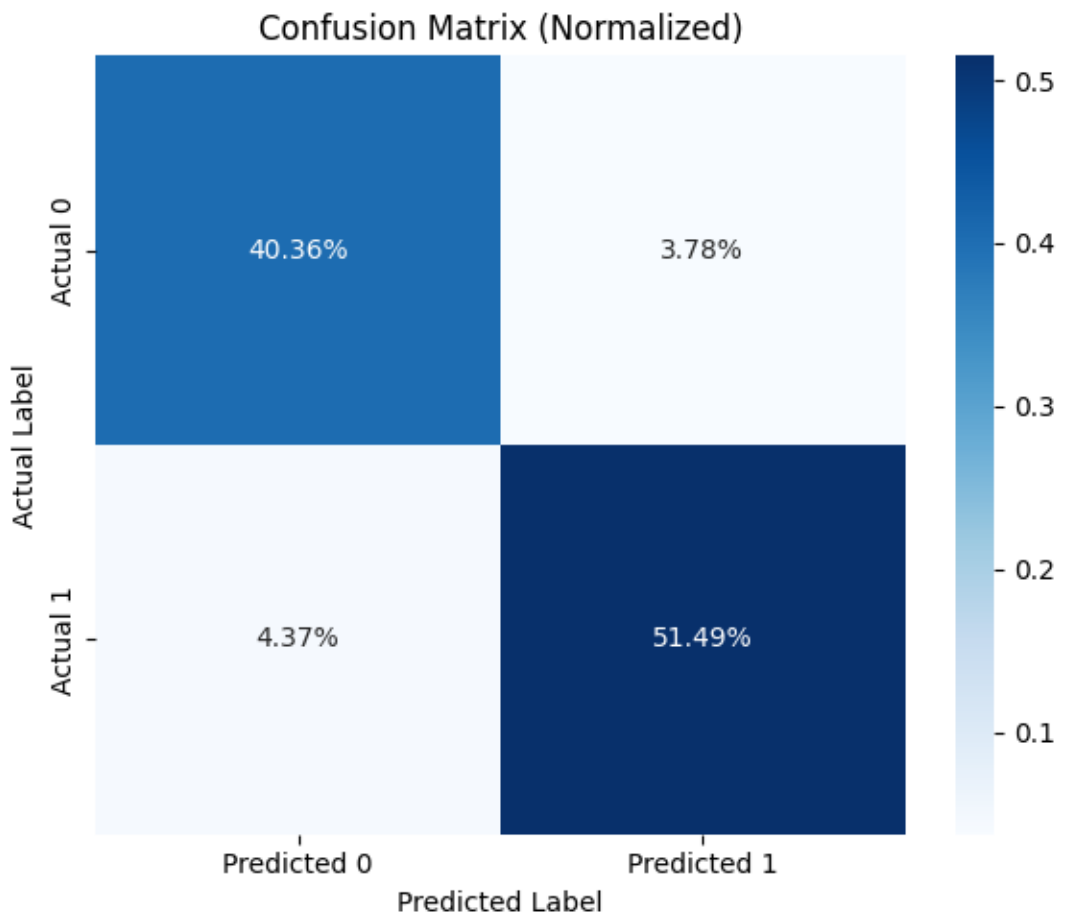


Figure 5: Normalized Confusion Matrix Heatmap.
Visual representation of model predictions vs. actual labels. The diagonal cells represent correct classifications (true positives/negatives), while off-diagonal cells highlight misclassifications.