Methodology

This study will adopt a quantitative, econometric approach to investigate the impact of media sentiment related to macroeconomic events on the volatility of financial and crypto markets. Drawing on the principles of behavioral finance by Kapoor & Prosad, (2017) and Steve Y. Yang & Sheugn Yin Kevin Mo (2016), the research that will be conducted integrates sentiment analysis with volatility modeling to assess whether the sentiment of media exerts a significant effect on volatility within financial and crypto markets. The analysis will focus on three major equity indices, namely S&P500, NASDAQ100, and Dow Jones and two alternative investments options Gold and Crypto10 index, this ensures we capture both traditional and more unconventional asset markets.

We will be employing a deductive approach, where existing theories on media influence on investment sentiment as shown by Fang & Peress (2009) & Ren et al., (2021) are repurposed for our research goals and are tested using empirical data. This involves extracting media coverage of macroeconomic news and obtaining a sentiment score from each macroeconomic category. This score is obtained using a Natural Language Processing tool and will be incorporated into a Generalized Autoregressive Conditional Heteroskedastic framework to measure it's volatility. The aim of the research is to evaluate how external media sentiment impacts volatility patterns over 20 years.

Data Collection

All the data used in the study were sourced from the Bloomberg Terminal at the University of Pretoria, Economics Department. The Bloomberg terminal provides my research with institutional-grade access to global financial markets and news media data over multiple periods of time.

The time frame in question spans from January 1994 to December 2024, this is chosen as to account for a time frame before economic shocks in the early 2000's as the dot.com bubble and Global financial crisis leading to covid-19 and the recovery after. The choice to use monthly data is to reflect the periodic nature of macroeconomic news and event to reduce sentiment "noise" that may incorrectly reflect actual market reactions. This would give me a 360-month news data to create a comprehensive volatility guide for the start of the 21st century

The financial data for the S&P500, NASDAQ100, Dow Jones, Gold, and Crypto10 indices were extracted in the OHLC format (Open, High, Low, Close). By using the Bloomberg news feed, we can collect news media from Bloomberg News, New York Times, Financial Times, the Economist and more.

A keyword-matching method based on seven macroeconomic categories taken from academic literature was used to screen articles. Major economic factors that impact investor behaviour and market mood are represented by each category. Seven separate searches were conducted to gather articles, and duplicates were eliminated by comparing the outlet metadata, timestamp, and headline.

1. Monetary Policy and Interest Rates
   "*Interest rate changes, Federal Reserve, monetary tightening, monetary easing, central bank policy, inflation targeting, discount rate, yield curve, quantitative easing, quantitative tightening , liquidity injection, open market operations, rate hikes, rate cuts*"
2. Inflation and PPP
   "*Consumer Price Index, Producer Price Index, inflation expectations, stagflation, hyperinflation, real wages, cost-push inflation, demand-pull inflation, inflation volatility*"
3. Economic Growth and Business Cycles
   "*GDP growth, economic slowdown, recession, expansionary phase, output gap, business cycle turning points, economic contraction, investment cycles, leading economic indicators*"
4. Fiscal Policy and Government Expenditure
   "*fiscal deficit*, *budget surplus*, *public debt*, *government spending*, *tax policy*, *stimulus package*, *austerity measures*, *debt ceiling*, *sovereign default*"
5. Exchange Rates and Capital Flows
   "*currency depreciation*, *currency appreciation*, *foreign exchange reserves*, *exchange rate volatility*, *capital flight*, *hot money flows*, *carry trade*, *balance of payments crisis*"
6. Trade and Globalization
   "*trade deficit*, *tariff hikes*, *protectionism*, *trade war*, *export growth*, *import restrictions*, *supply chain disruptions*, *foreign direct investment*"
7. Commodity Prices and Supply side shocks
   "*oil price shocks*, *commodity price inflation*, *supply chain constraints*, *agriculture price index*, *input cost volatility*, *gold prices*, *energy price spikes*"

Based on the frequency of keywords and the contextual significance of the headline and metadata, each news article was categorized into one of the seven macroeconomic event categories. To prevent multi-collinearity in subsequent analysis, articles that made reference to several macroeconomic categories were categorized based on the prevailing theme.

This study has a limitation, despite its goal of offering a thorough sentiment analysis throughout the course of a whole time series of macroeconomic news. Because these sources have a greater chance of influencing investor behaviour, the data gathering procedure purposefully prioritizes stories from significant, well-known media outlets (such as Bloomberg, Reuters, and the Wall Street Journal). This strategy is based on research by Fang & Peress (2009), who contend that institutional media coverage acts as a signal to markets and that equities with less media exposure frequently generate larger returns as a result of investor underreaction.

Sentiment Analysis Process

Based on the fundamental BERT (Bidirectional Encoder Representations from Transformers) architecture, FinBERT is a transformer-based model that has been specifically adjusted for financial language. In contrast to general-purpose models, FinBERT can more accurately capture specialized syntax and semantic patterns because it has been trained on a sizable corpus of financial texts, including analyst reports, earnings calls, and market news. Tokenization using the Bert Tokenizer from Hugging Face's Transformers library is the first step in the implementation. Word Piece tokens, which keep sub word information for phrases that are not in the lexicon (e.g., "unemployment" → ["un", "##employment"), are created from each input sentence or headline. After being transformed into input IDs, these tokens are fed into the model with an attention mask that separates relevant text from padding tokens.

The fundamental BERT encoder stack, which has 12 layers, 12 self-attention heads, and 768-dimensional hidden states, is what FinBERT employs internally. Each input sequence begins with the [CLS] token, which serves as an aggregate representation of the sentence. A classification head, a straightforward feed-forward neural network, receives the token's final-layer hidden state and produces raw logits for the positive, neutral, and negative sentiment classes.

```
text = "Rising interest rates suggest that the Fed is confident in economic growth."

inputs = tokenizer(text, return_tensors="pt", truncation=True, padding=True)

outputs = model(**inputs)

logits = outputs.logits

probs = torch.nn.functional.softmax(logits, dim=-1).detach().numpy()


labels = ['positive', 'neutral', 'negative']

sentiment = labels[np.argmax(probs)]

confidence = np.max(probs)
```

The softmax function is used to convert the logits into probability scores. These show the estimated probability that the input falls into each sentiment class according to the model. Next, we use the class probabilities as weights to calculate an expected sentiment score. A continuous sentiment score ranging from -1 to +1 is the outcome of our function for each expected sentiment score.

$$E[s_i] = P_{positive} * (1) + P_{neutral} * (0) + P_{negative} * (-1)$$

Because it avoids information loss from hard categorization and preserves probabilistic nuance, this method is better for time-series volatility modelling.

Following the FinBERT calculation of each article's unique sentiment score, these values are combined monthly for each predetermined macroeconomic category. The expected sentiment scores, which range from -1 to +1, are averaged across all articles that belong to the same category and month in order to perform the aggregation using a simple arithmetic mean.

$$S_{Y;M} = \frac{1}{n}\sum_{i=1}^{n} s_i$$

With this method, the overall tone of news coverage of a particular macroeconomic topic throughout time is reflected in a continuous sentiment index. The study guarantees temporal alignment with the monthly financial market data by collecting sentiment in this way, which also makes it possible to compare sentiment directly to observed market volatility. This methodology preserves the granularity of article-level analysis while generating a standardized input suitable for time-series econometric modelling.

FinBERT is especially well-suited for sentiment analysis in financial contexts due to its many benefits. First, it can successfully read complex language, including compound sentences and tonal modifiers, thanks to its context sensitivity. For instance, lexicon-based algorithms frequently misclassify statements like "barely recovered" or "shockingly positive," which are appropriately understood in context. Second, FinBERT demonstrates domain adaptation, which enables it to identify and accurately categorize syntactic structures and industry-specific vocabulary because it was trained exclusively on financial documents. Misclassification errors, which are typical in general-purpose sentiment models, are significantly decreased by this. Last but not least, FinBERT has excellent accuracy, particularly when used with lengthier texts or multi-clause headlines that contain nuanced sentiment signals that are difficult for rule-based methods to pick up.

Nevertheless, there are costs associated with these advantages. FinBERT is computationally costly, requiring a huge amount of memory, computing power, and inference time, especially when studying large textual batches. For timely execution, GPU acceleration is frequently required. Additionally, like the majority of deep learning models, FinBERT's internal reasoning is opaque, making it challenging to understand its categorization choices without the use of explainability techniques like attention heatmaps or SHAP values. FinBERT's propensity to over-predict neutral emotion, especially in brief or syntactically confusing headlines, presents another difficulty. This may necessitate further calibration or rebalancing.

Volatility Modelling with GARCH

To measure the time changing nature of financial and commodity market volatility in response to macroeconomic media sentiment, this study will employ an Exponential Generalized Autoregressive Conditional Heteroskedasticity or EGARCH for short. Bollerslev (1986) introduced the GARCH as an extension on Engle's ARCH model by incorporating lagged conditional variables into the conditional variance equation, the EGARCH was introduced by Nelson (1991) and is a non-linear extension of the GARCH model. Its framework allows for more asymmetric volatility effects or the leverage effect. In financial markets negative news or returns will typically induce higher volatility than positive effects, this is also shown by Umar et al. (2021) where he found that heightened media coverage amplifies returns and volatility.

The standard GARCH model the was introduced by Bollerslev (1986) might struggle to differentiate between positive and negative news. The EGARCH model overcomes this limitation by modelling the logarithm of the conditional variance, thus ensuring that the conditional variance is always greater than 0 without requiring constraints within the model. The conditional Variance can be expressed as:

$$ln\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i(|\epsilon_{t-1}| + \gamma_i\epsilon_{t-1}) + \sum_{j=1}^{q} \beta_j * ln\sigma_{t-j}^2$$

$$x_t = \mu + a_t$$

$$a_t = \sigma_t * \epsilon_t$$

$$\epsilon_t \sim P_v(0; 1)$$

Where x_t is the time series value at time t, μ is the mean of the EGARCH model, a_t is the residual at time t for the EGARCH.

Within the EGARCH model the γ_i or gamma, is the asymmetric response parameter to shocks within the time series. A leverage effect is implied by a statistically significant and negative γ_i, which means that negative shocks raise volatility more than positive ones. In the context of macroeconomic sentiment, this is especially relevant because negative sentiment, such as worries about a recession, monetary tightening, or inflation spikes, is likely to cause disproportionately higher market volatility than comparable positive sentiment.

Because of the nature of the macro sentiment that is going to be developed we have to create an exogenous parameter within the conditional heteroskedastic variance $ln\sigma_t^2$. We want to estimate it without violating the model's logic and have an individual marginal effect via a δ coefficient parameter.

We are combining both the absolute sentiment level of macroeconomic events per month and the change in sentiment level month on month. This allows us to capture two distinct effects on sentiment and volatility. 1) we can observe the level effect of how sentiment influences overall volatility and 2) we can observe how sudden shocks in sentiment triggers spikes in volatility due to investor behaviour. The first effect will be the aggregated effect $S_{Y;M}$ and $\Delta S_{Y;M} = S_{Y;M} - S_{Y;M-1}$.

In order to capture both persistent and surprise-driven volatility responses, this research will use an EGARCH process to simulate conditional volatility. This process incorporates both the absolute level of macroeconomic sentiment and the first-difference of the mentioned sentiment. The definition of the extended conditional variance equation is:

$$ln\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i(|\epsilon_{t-1}| + \gamma_i\epsilon_{t-1}) + \sum_{j=1}^{q} \beta_j * ln\sigma_{t-j}^2 + \delta_1 S_{Y;M} + \delta_2 \Delta S_{Y;M}$$

Within this model the δ_1 will capture our baseline sentiment effect on volatility and δ_2 will capture our sentiment change effect on volatility. These additions ensure that we can both capture any changes in volatility either planned or surprised.

Although our planned model is expressed in logarithmic form, it remains essential to use logarithmic returns for all our price data OHLC to ensure that the variance satisfies stationarity requirements and ensures compatibility with econometric assumptions. The log return will be computed as:

$$r_{asset;t} = \ln\left(\frac{P_{asset;t}}{P_{asset;t-1}}\right)$$

Where $r_{asset;t}$ is the log return for the asset in question at time t. $P_{asset;t}$ is the asset in question adjusted closing price at time t. This is especially important when estimating the monthly movements of cryptocurrency and equities indexes.

The EGARCH model has a number of drawbacks even if it offers a strong foundation for simulating volatility asymmetries in response to macroeconomic sentiment. The model is risky in the situation of limited monthly data since it is very sophisticated and requires the estimation of more parameters than normal GARCH. Second, there is a greater chance of colinearity problems, especially when the conditional variance equation incorporates exogenous variables like sentiment and its first difference. Third, the inclusion of both sentiment level and sentiment change at the same time increases the likelihood of multicollinearity.

Econometrical Diagnostic

To ensure the time-series inputs meet the requirements of conditional heteroskedasticity, stationarity testing is done before the EGARCH model estimation. Both the standardized sentiment scores and the log return series are subjected to the Augmented Dickey-Fuller (ADF) test. All return series are stationary and mean-reverting, according to the results of the ADF test, which is in line with actual data from the cryptocurrency and equity markets.

Despite being more variable, the sentiment data will also pass the ADF test, proving that it should be included in the conditional variance equation. The Variance Inflation Factor (VIF) is calculated to address multicollinearity concerns, particularly between absolute sentiment and sentiment change $S_{Y;M}$ and $\Delta S_{Y;M}$. Significant collinearity risk would be indicated by a VIF greater than 5 for either variable, which could inflate standard errors and destabilize coefficient estimates. Only one sentiment term will be kept if such hazards are noted.

Maximum Likelihood Estimation (MLE) is used to estimate model parameters under the presumption of conditional normality. Because the EGARCH framework is nonlinear and complicated, robust inference is achieved by using robust (heteroskedasticity-consistent) standard errors, which account for minor sample abnormalities and potential model misspecification. This method increases the sentiment coefficients' hypothesis test reliability. Because the study focuses on in-sample volatility dynamics rather than prediction accuracy, bootstrapping is not used here, despite the fact that it can be useful for creating confidence intervals in forecasting scenarios.

In-sample fit and statistical adequacy are the main criteria for evaluating a model. The Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC), which weigh goodness-of-fit against model complexity, serve as a guide for choosing amongst competing models. Each estimated model's log-likelihood is also provided to assess fit without regard to penalization. The Ljung-Box Q-statistic on squared and scaled residuals is used in diagnostic evaluation to check for unmodeled structure and serial correlation. The Jarque–Bera test is used to evaluate the normality of standardized residuals, and the ARCH-LM test is performed after estimation to verify that residual heteroskedasticity is absent. These diagnostics make sure that there are no econometric pathologies and that the volatility models are properly specified.

References

Bollerslev, T. (1986). GENERALIZED AUTOREGRESSIVE CONDITIONAL HETEROSKEDASTICITY. In *Journal of Econometrics* (Vol. 31).

Fang, L., & Peress, J. (2009). Media Coverage and the Cross-section of Stock Returns. In *THE JOURNAL OF FINANCE •: Vol. LXIV* (Issue 5).

Kapoor, S., & Prosad, J. M. (2017). Behavioural Finance: A Review. *Procedia Computer Science*, *122*, 50–54. https://doi.org/10.1016/j.procs.2017.11.340

Nelson, D. B. (1991). *Conditional Heteroskedasticity in Asset Returns: A New Approach* (Vol. 59, Issue 2).

Ren, J., Dong, H., Padmanabhan, B., & Nickerson, J. V. (2021). How does social media sentiment impact mass media sentiment? A study of news in the financial markets. *Journal of the Association for Information Science and Technology*, *72*(9), 1183–1197. https://doi.org/10.1002/asi.24477

Steve Y. Yang, & Sheugn Yin Kevin Mo. (2016). *Social media and news sentiment analysis for advanced investment strategies* (W. Pedrycz & S.-M. Chen, Eds.; Vol. 639). Springer International Publishing. https://doi.org/10.1007/978-3-319-30319-2

Umar, Z., Adekoya, O. B., Oliyide, J. A., & Gubareva, M. (2021). Media sentiment and short stocks performance during a systemic crisis. *International Review of Financial Analysis*, *78*. https://doi.org/10.1016/j.irfa.2021.101896