Sector-specific forecasting of volatility from Twitter sentiment

Shiruo Liang, Joe Wee Sham, Yiming Song (equal contribution)

Columbia Summer Undergraduate Research Experiences in Mathematical Modeling

1. Introduction

Advances in natural language processing (NLP) have allowed for financial market sentiments to be extracted automatically from textual information from sources like Twitter [1]. We explore the impact of such data on specific market sectors, specifically, how changes in public sentiment can affect the volatility of different stock market sectors. Volatility measures the magnitude and frequency of price movements and is a good analog of risk, allowing us to inform risk management strategies. We train and compare four machine learning models to determine the most reliable model to use in forecasting volatility.

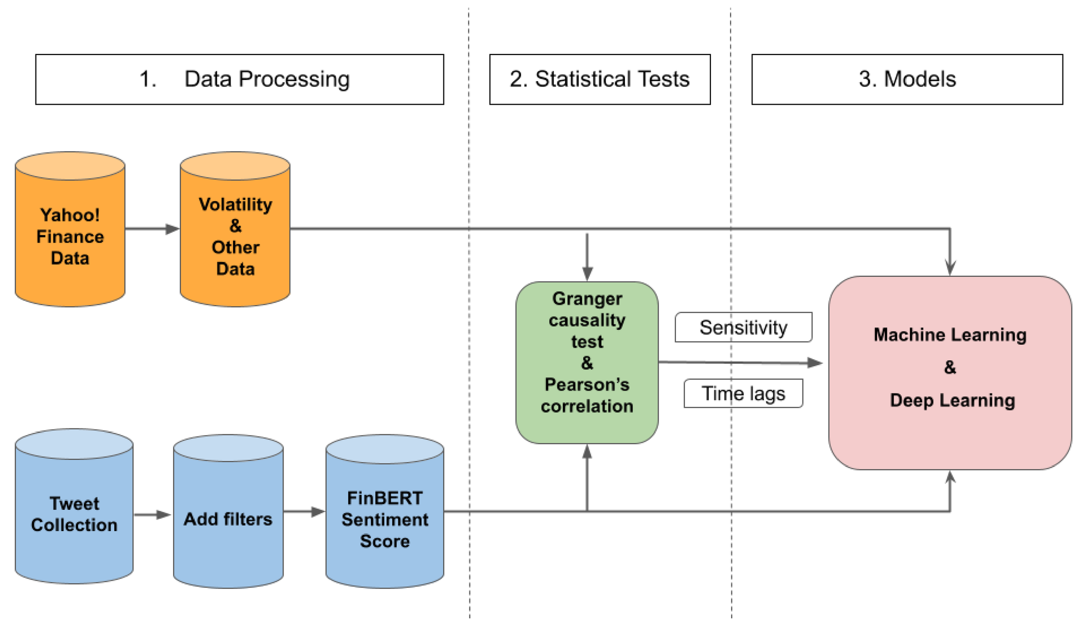

Figure 1: Research project structure.

2. Data Acquisition and Processing

We gathered 130 finance-related Twitter users, including individuals and financial institutions. We then filtered for tweets containing a list of market and emotion-related terminology, with a minimum of 20 likes and 5 retweets. We ended up with 80,000 tweets from January 2022 to June 2023.

To extract sentiment data from the Tweets, we used FinBERT, a deep-learning based language classification model that is pre-trained on financial data. Given a text input, it returns a single score between -1 and 1, corresponding to negative and positive sentiment. [2].

We used exchange-traded fund (ETF) data across the same time period downloaded from Yahoo! Finance. We looked at 11 different ETFs from Vanguard representing 11 sectors. To calculate daily volatility, we used the Garman-Klass, Rogers-Satchell, and Parkinson estimators [3] based on the open, close, high, and low prices each day.

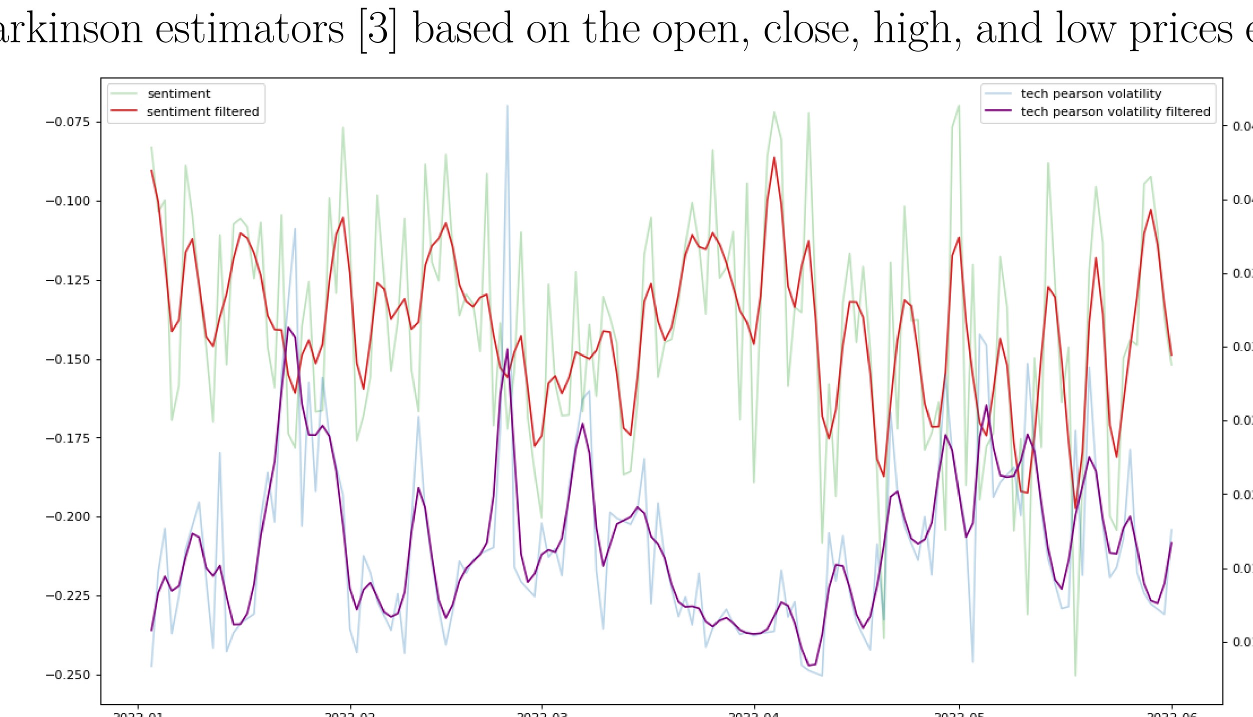

Figure 2: A section of sentiment and Parkinson's volatility for the technology sector.

3. Methods: Causality & Correlation

The Granger causality test is commonly used in econometrics to analyze unidirectional or bidirectional causation. Given two time series X and Y, we say that X Granger-causes Y if predictions of Y based on its own past values *and* past values of X are better than predictions of Y solely based on its own past values.

To do so, we first perform a regression with the values of Y only:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + ... + \beta_m y_{t-m} + err_t \tag{1}$$

Next, we include the values of X, time-lagged by a certain period:

$$y_t = b_0 + b_1 y_{t-1} + b_2 y_{t-2} + ... + b_m y_{t-m} + c_p x_{t-p} + ... + c_q x_{t-q} + err_t \tag{2}$$

We then perform a t-test on each of the individual values of X, keeping the ones that are significant, with the condition that they add explanatory power to the regression (under an F-test). Then, the null hypothesis that X does not Granger-cause Y is only accepted if no values of X are included.

To quantitatively determine the correlations, we calculated the Pearson correlation between of public sentiment and various values across the 11 market sectors (closing price, percentage change, and volatility calculated using the three different estimators). It is important to consider that the Pearson correlation test only gives information about same-day correlation; thus, we cannot infer from it whether one of the variables causes the other.

4. Methods: Machine Learning

To forecast the volatility of various market sectors, we tested four different machine learning models: Long Short-Term Memory (LSTM), Linear Regression (LinR), Decision Tree (DT) and XGBoost (XGB). We use mean absolute prediction error (MAPE) to evaluate our models, and compare model performance with/without sentiment data input:

$$MAPE = \frac{1}{n}\Sigma_{t=1}^{n}\frac{|y_{true} - y_{pred}|}{y_{true}} \tag{3}$$

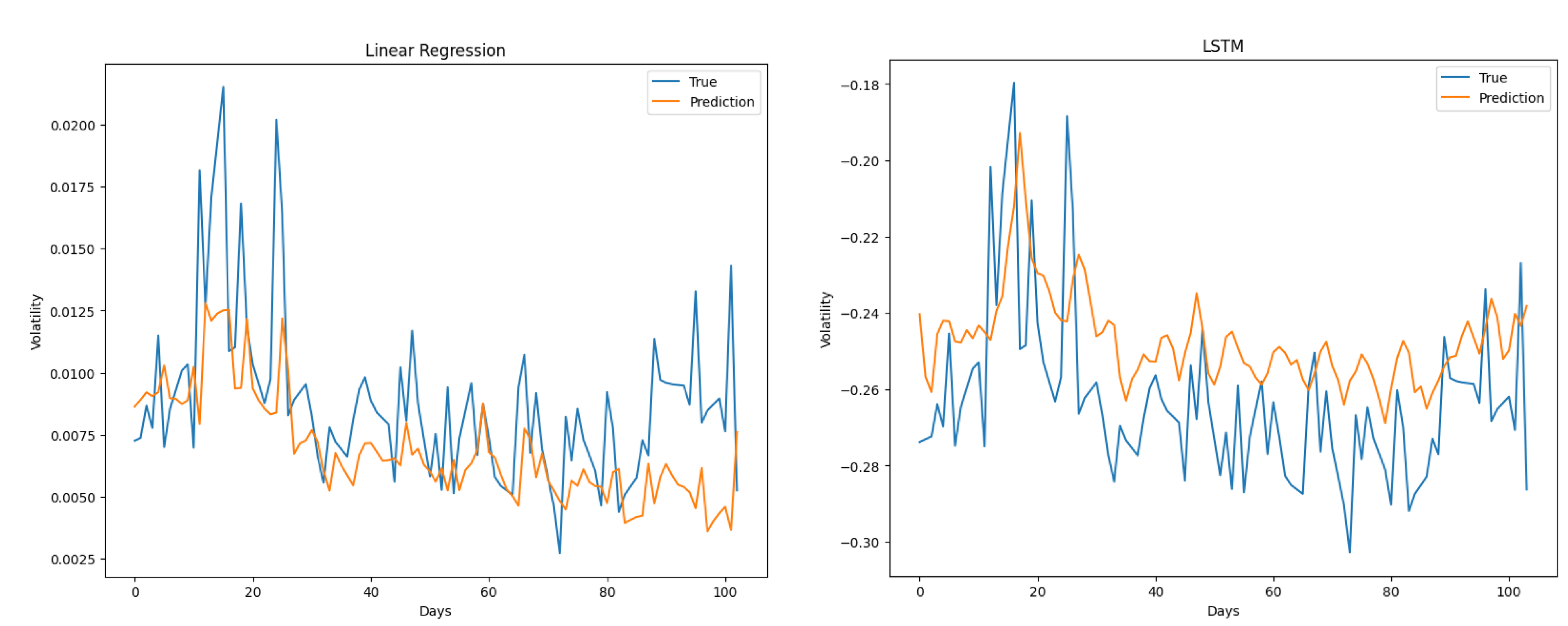

Figure 3: Comparison of LinR (left) and LSTM (right) models, predicting volatility for the technology sector. All models were trained on a section of our data and tested on a future section to ensure valid results.

5. Results

1. Pearson correlation: We find significant ($p < 0.05$) positive correlations between sentiment and market returns, and significant negative correlations between sentiment and volatility, in most sectors.
2. Granger causality: We find significant Granger-causation for sectors such as technology and consumer discretionary at certain time lags from 1 to 14 days, while the tests for other sectors like real estate and energy are insignificant.
3. Machine learning: We find that only one model, LSTM, significantly improves upon the inclusion of sentiment data, as shown in the table below.

	LSTM		LinR		DT		XGB	
	s.	n.s.	s.	n.s.	s.	n.s.	s.	n.s.
Tech.	0.072	0.446	0.294	0.294	0.967	0.795	0.650	0.513
...								
C.D.	0.099	0.41	0.241	0.242	0.656	0.902	0.759	0.765
Mean	0.090	0.323	0.202	0.201	0.437	0.434	0.57	0.348
p	0.000		0.687		0.934		0.618	

Table 1:MAPE of various machine learning models.

6. Conclusion and Discussion

We find significant relationships between sentiment and market volatility in certain sectors, and utilize an LSTM model to take advantage of this, in order to make better predictions of volatility after including additional sentiment data obtained from Twitter.

Our models can definitely be furthered optimized and tuned to produce more accurate predictions of volatility. Our training data, spanning a time period of 521 days, limited our ability to build stronger models. Also, our volatility estimators are also not perfect, as they are designed for estimation over longer time periods than one day. We also had to fill in missing days such as weekends, to ensure our sentiment data matched up. There are also features that could be included into a machine learning model other than sentiment that we did not consider, such as extra market statistics, and more, as covered in [4].

Acknowledgements

We would like to thank Columbia University's Mathematics Department, Center for Career Education, and Barnard College's Summer Research Institute. We also appreciate guidance from Professor George Dragomir and our graduate mentors Xintong Zhou, Casey Ivanovich, Tao Long, and Hanyang Zhao.

References

[1] Yang Gao and Chengjie Zhao. Investor sentiment and stock price jumps: A network analysis based on china's carbon-neutral sectors. *The North American Journal of Economics and Finance*, 68:101954, 2023.

[2] Yi Yang, Mark Christopher Siy UY, and Allen Huang. Finbert: A pretrained language model for financial communications, 2020.

[3] Jinghong Shu and Jin E Zhang. Testing range estimators of historical volatility. *Journal of Futures Markets: Futures, Options, and Other Derivative Products*, 26(3):297–313, 2006.

[4] Nan Jing, Zhao Wu, and Hefei Wang. A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. *Expert Systems with Applications*, 178:115019, 2021.