A Report on

A Comparative Study of Sentiment Analysis for Stock Price and Volatility Prediction in Indian Markets

Prepared under the guidance of Dr. Ritika Jaiswal
Department of Economics and Finance

By

Madhav Bajaj	2019B3A70256G
Himanshu Daga	2019B3A70320G
Hruthvik M	2019B3AA0535G

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

ज्ञानं परमं बलम्

1 Abstract

Capturing investor sentiment through news and social media data analysis has been an emergent trend in the finance domain, and much of recent research has concentrated on utilizing news-derived information to forecast stock price movement. The same degree of effort hasn't been committed to assessing the effect of sentiment analysis on other vital financial variables. This paper conducts correlation analysis and the granger causality test to demonstrate the greater efficacy of sentiment analysis in forecasting volatility than price or returns. Three metrics corresponding to aggregate, positive, and negative sentiment have been assessed against daily returns and multiple proxies to represent volatility. Textual data to construct the sentiment score has been extracted from Google news feed and Twitter. The paper has been further extended to an LDA and logistic regression-based procedure to produce a model with an enhanced ability to forecast volatility movement the next day.

2 Introduction

The emergence of artificial intelligence and machine learning has renewed interest in several fields, and the financial industry is no exception. The study of financial modeling for forecasting various variables through time series models is well documented, but as the multiple subdomains within A.I. and ML developed, research on their applicability was well endorsed by academia and industry proponents alike. Natural Language Processing (NLP) is among the most prominent of these subdomains and can potentially augment prediction strategies for financial statistics. Sentiment analysis of various news feeds regarding sectoral and macroeconomic conditions can effectively capture the effect of the behavioral element of the investors on the financial markets and can be utilized to bolster portfolio optimization and risk management and even produce enhanced trading strategies. But a disproportionate amount of research has been dedicated to forecasting stock returns as opposed to other statistics, such as volatility which is significant for financial modeling. This paper aims to produce a comparative analysis of the effectiveness of sentiment analysis for the prediction of stock returns and price volatility.

3 Literature Review

A rising number of research articles employ natural language processing (NLP) approaches to investigate how the sentiment of firm-specific news, financial data, or social media affects stock market returns. Most sentiment analysis work appears to be focused on forecasting market prices or directional change. Several examples of text mining are used on stock market news data, emphasizing market price prediction. However, few research studies investigate how financial news affects stock market volatility.

Tetlock (2007) investigated potential linkages between the media and the stock market using information from the Wall Street Journal and discovered that strong pessimism creates downward pressure on market prices. Afterward, to determine if corporate financial news can forecast a company's accounting profitability and stock returns, Tetlock et al. (2007) used a bag-of-words model. The findings showed that negative news about a given business predicts low firm profitability, although market prices frequently underreact to the news with a solid negative slant. Similarly, Bollen et al. (2011) investigated whether sentiment obtained from Twitter feeds connected to the Dow Jones Industrial Average Index value (DJIA). They used the opinion-tracking programs OpinionFinder and Google-Profile of Mode States (GPOMS), which evaluate mood in six dimensions (Alert, Vital, Sure, Happy, Calm, and Kind). The findings led them to conclude that adding particular public mood characteristics, but not others, can significantly increase the accuracy of DJIA predictions.

Similarly, Loughran et al. (2011) analyze sentiment in U.S. 10-K filings. According to the authors, over three-quarters of the negative word counts in 10-K files based on the Harvard lexicon are not usually unfavorable in an economic context. In order to do this, they created an alternate lexicon that better represents sentiment in financial content. Kogan et al. (2009) used a Support Vector Machine (SVM) to forecast the volatility of stock market returns. According to the findings, "text regression model forecasts correspond with true volatility almost as well as historical volatility, with a combination model doing even better." Mao et al. (2011) forecast financial market prices using various news data and sentiment tracking techniques. The authors discover that Twitter sentiment is a major predictor of daily market returns, but if all other mood indicators, including VIX, are controlled for, sentiment indicators are no longer statistically insignificant. Similarly, Groß-Klußmann et al. (2010) discovered that disseminating highly relevant news increases return volatility, with negative news having a more significant influence than positive news.

Glasserman et al. (2019) employ an n-gram model to construct an approach that predicts volatility at both the company-specific and aggregate levels using unexpected negative and positive news. According to the authors, an increase in the "unusualness" of news with negative emotion predicts an increase in stock market volatility. Similarly, unexpectedly favorable news expectations reduce volatility. According to study findings, news has a slower impact on volatility at the aggregate level than at the company-specific level, which is consistent with the effect of diversification. Even Calomiris et al. (2020) used news items to construct a model for predicting risk and return in established and emerging market stock markets. Their findings suggest that news material's topic-specific sentiment, frequency, and unusualness can forecast future returns, volatility, and drawdowns. Similarly, Caporin et al. (2017) discovered that news-related factors could help forecast volatility. Certain news subjects, such as earnings announcements and upgrades/downgrades, are more critical in predicting market volatility than others.

In recent work, Atkins et al. (2018) forecast stock market volatility fluctuations using LDA and a straightforward Naive Bayes classifier. The authors discover that news item data can predict market volatility more accurately than price movement direction. They had a 56% success rate in forecasting the direction of stock market volatility in response to the release of new data. Also, Mahajan et al. (2009) utilized LDA to identify financial news subjects and subsequently predict stock market movements based on the topics derived from financial news. Their created classifier predicted market direction with 60% accuracy. According to Jiao et al. (2016), substantial social media activity surrounding a given firm predicts a considerable rise in return volatility, but attention from essential news publications, such as the Wall Street Journal, indicates the opposite: a decrease in return volatility.

There are a few research papers in which the authors have predicted volatility in the Indian stock market. There are a few papers that used GARCH models using time series data to predict volatility without using any machine learning models. Babu et al. (2015) used a hybrid of Autoregressive Integrated Moving Average (ARIMA) and Generalized Autoregressive Conditional Heteroscedastic (GARCH) models to preserve data trends and a non-linear ANN model to preserve the precision. They applied it to selected NSE Indian stock market data to predict their volatility. Deb et al. (2003), in their work, modeled the monthly volatility of Indian capital market indexes (Sensex and S&PCNX-Nifty) using eight different univariate models whose performances were examined using several symmetric and asymmetric loss functions. They found that the ARCH (9) model is better than the other models for investors who are more concerned with under-forecasts than over-predictions. Sen et al. (2021) presented several volatility models based on the generalized autoregressive conditional heteroscedasticity (GARCH) framework for estimating the volatility of ten equities listed on India's national stock market (NSE). Their results revealed that asymmetric GARCH models produce more accurate projections of stock volatility in the future.

Some Indian scholars have also approached volatility modeling through prevalent machine-learning models. Dixit et al. (2013) employed an artificial neural network (ANN) modeling approach used to estimate the following trading day's volatility using India VIX (a volatility index based on NIFTY Index

Option prices) based indicators. The study's findings show that ANN models can be quite valuable for anticipating VIX's downward trajectory. Choudhury et al. (2014) employed a SOM-based hybrid clustering approach merged with support vector regression for portfolio selection and accurate price and volatility forecasts, which is the foundation for the portfolio's specific trading strategy. The study looks at the top 102 stocks on the NSE stock exchange in India to determine a set of best portfolios that an investor may keep for risk reduction and high profitability.

Very few research papers have been published in the context of volatility prediction using sentiment analysis of the Indian stock market. Kumari et al. (2015), in order to create a sentiment index relevant to emerging Indian stock markets, used ten aggregate market-related sentiment proxies. They used the GARCH model in which the sentiments were incorporated and showed a significant effect of an investor's sentiment on the stock market volatility. They also used the VAR-GARCH model to capture the impact of lagged sentiment on stock market volatility and returns spread. Paramanik et al. (2020) used the asymmetric GARCH model of conditional volatility for the Indian Stock Market (SENSEX) by incorporating positive and negative sentiments. Their empirical findings indicate that negative market sentiment has a more significant influence than positive market sentiment. They also give evidence of noisy trading in the financially underdeveloped Indian stock market.

3.1 Contribution to the Literature Gap

This report provides a novel contribution to volatility forecasting research in the Indian context through a comparative study of the effectiveness of sentiment analysis on stock returns and volatility prediction. This paper has utilized two sufficiently differentiated methodologies, with the former involving correlation analysis and Granger's causality test and the latter employing LDA and a logistic regression model for prediction

4 Methodology

4.1 Data description

Two primary data sources, news articles and Twitter tweets, were employed to construct the separate sentiment scores. Python was utilized extensively during the research to conduct data extraction and preprocessing due to the vast availability of convenient libraries. Google news feed articles for all stocks featured in the BSE Sensex index for the period from 1st January 2021 and 30th June 2021 were extracted through the pygooglenews library. A total of 15,362 articles were sampled across six months. Specialized libraries such as beautiful soup and urllib were used to fetch news text data, while the daily stock data for the index was obtained from the BSE historical data website.

Almost 90,000 tweets were considered in the study during the period over which news feed data was considered. Because of limitations faced with the official Twitter data extraction API Tweepy, the snscrape python library has been utilized to exact tweets data. Snscrape is a social networking service data scraper used to extract data such as user profiles, hashtags, or searches and returns the discovered items, for example, any relevant posts based on the search parameter. It is an open-source python library and was run locally. Any related tweets containing a keyword belonging to a predetermined corpus of words (refer to Appendix A) generally associated with the targeted tweets have been extracted. Here the corpus has been defined as the hashtags most associated with the 'SENSEX' keyword during the time period considered. The keyword choice was made keeping in mind the relevance of tweets showing up for the associated keywords.

4.2 Construction of sentiment and volatility indices

The sentiment index movement is compared against five metrics constructed from the SENSEX index daily close price data: daily returns (DR_t), 3-period daily volatility ($DV3_t$), lagged 3-period daily volatility ($DVL3_t$), 7-period daily volatility ($DV7_t$), and lagged 7-period daily volatility ($DVL7_t$). The standard deviation of the closing prices over 3-day and 7-day periods, including the day for which the metric is being calculated, has been considered to represent daily volatility.

Stock Indices Formulae

$$DR_t = log\left(\frac{P_t}{P_{t-1}}\right)$$

$$DV3_t \;=\; \sqrt{\frac{\sum_{i=0}^{2}\left(P_{t-i} - \left(\frac{\sum_{k=0}^{2} P_{t-k}}{3}\right)\right)^2}{3}}$$

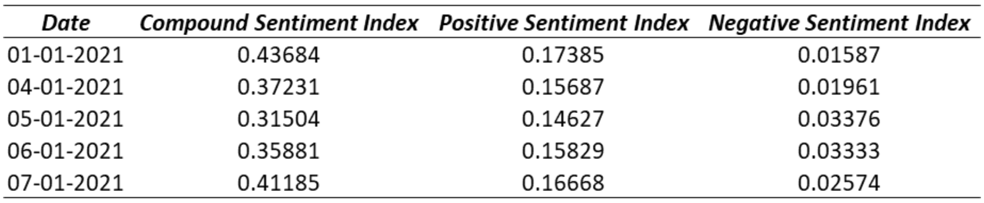
$$DV7_t \;=\; \sqrt{\frac{\sum_{i=0}^{6}\left(P_{t-i} - \left(\frac{\sum_{k=0}^{6} P_{t-k}}{7}\right)\right)^2}{7}}$$

Where P_t denotes the closing price of the SENSEX index on day t.

VADER Sentiment Intensity Analyzer has been used to develop the sentiment index. Introduced by Hutto and Gilbert (2014), VADER is a lexicon and rule-based sentiment analysis tool and has been well-received by academia engaged in natural language processing research. VADER has gone through several improvements since its introduction and has been utilized extensively to model and forecast financial metrics (refer Sohangir et al. (2018), Agarwal (2020), Nousi and Tjortjis (2021), Ekaputri and Akbar (2022) and Long et al. (2022)). VADER sentiment analyzer for a given text returns the positive, negative, neutral, and compound (aggregate) sentiment score where each statistic is normalized to a range of $[-1, 1]$.

Three sentiment metrics, compound sentiment index, positive sentiment index, and negative sentiment index, were constructed from the news feed data. Initially, google news articles for each stock in the SENSEX index were analyzed, and three metrics corresponding to the positive, negative, and compound sentiment were calculated and averaged, grouped by the date. A weighted average of individual stocks' figures has been considered to determine the sentiment metrics for the overall index, and the weight assigned to a stock is equal to the stock's contribution to the calculation of the index price. The same metrics were also considered for the tweets data set.

Table 1: Twitter Sentiment Score Sample



Date	Compound Sentiment Index	Positive Sentiment Index	Negative Sentiment Index
01-01-2021	0.43684	0.17385	0.01587
04-01-2021	0.37231	0.15687	0.01961
05-01-2021	0.31504	0.14627	0.03376
06-01-2021	0.35881	0.15829	0.03333
07-01-2021	0.41185	0.16668	0.02574

4.3 Correlation and Granger causality analysis

Pearson's correlation has been implemented to assess how significantly the changes in multiple sentiment indices affect daily returns and volatility. The correlation (r) obtained between variables ranges from -1, indicating complete linear anticorrelation, to +1, indicating complete linear correlation.

Furthermore, bidirectional Granger causality tests have been conducted for each combination of a sentiment and stock market metric. Granger causality determines whether one time series variable has some predictive power over the other time series variable. The Augment Dickey-Fuller (ADF) test has been conducted for all indices to confirm stationarity and the applicability of the Granger causality tests. All volatility measures, sentiment indices, and daily returns time are stationary. A 5% alpha level has been considered to ensure the statistical significance of the correlation coefficients and Granger causality test statistics. Both analyses were conducted separately for the news feed data and Twitter data set.

4.4 Latent Dirichlet Allocation and Logistic Regression

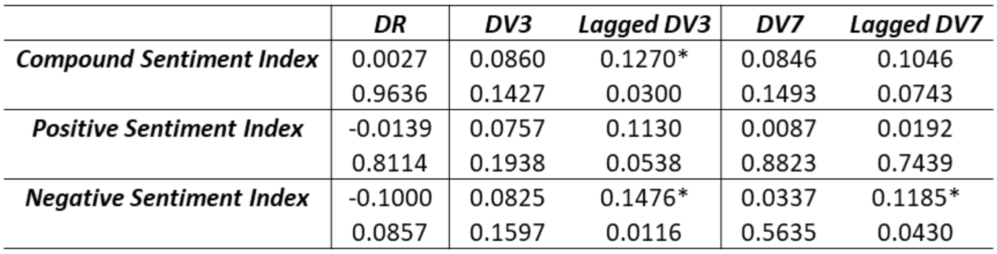
Latent Dirichlet allocation is the approach that is frequently used for topic modeling or topic finding from a large number of documents. LDA, introduced by Blei, D.M. et al. (2003), is widely seen as a powerful way of extracting topics across a range of text data. The generative topic model creates combinations of latent topics from a set of documents, and each combination of topics yields words from the lexicon of the set with a given probability. LDA execution is broken down into multiple phases. A topic is further selected based on the distribution of topics first sampled from a Dirichlet distribution. A subject is represented as a distribution over words, and each document is modeled as a distribution over themes. Each word inside a document is chosen from one of a multitude of corpus-wide subjects. The generative model is constructed using topic modeling, which on a high level, extrapolates backward from a series of documents to infer the themes that might have generated them. LDA produces a limited number of topics, reducing the data's dimensionality, but it is computationally intensive with a high complexity polynomial time $O(n^k)$, refer Porteous et al. (2008), Sontag and Roy (2011), Xiao and Stibor (2010).

LDA, along with a logistic regression exercise, has been used to develop and train a model to determine the movement of the stock market volatility the next day ('UP' or 'DOWN'). An LDA model is supplied with titles of all news feed articles published on a particular day, and 15 distributions of topics are received as output. Kelechava (2021) methodology is then used to develop a 15-dimension feature vector from the LDA output, which is further fed to a classification problem. A feature vector here refers to a topic-count sparse vector that represents the frequency with which each topic appears in specified textual data during the specified interval. The above process is repeated for all days in the considered time period to construct a data set which is then merged with daily market volatility data with the binary movement direction labels ('UP' or 'DOWN') depending on the value of the volatility metric the previous day. Thus each feature vector is paired with a movement label to create the calculated target vector. The trained model is then asked to create a feature vector from the unseen test data, and logistic regression is used to determine the movement of the volatility metric the next day. An identical exercise has been conducted with the Twitter data set.

5 Empirical Analysis

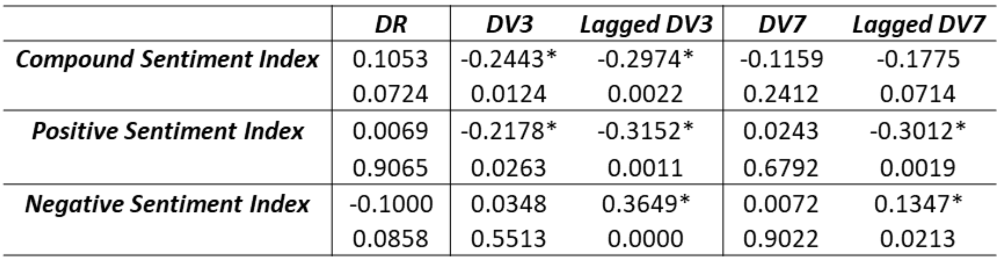
5.1 Results from the Correlation coefficient analysis and the Granger causality tests

Table 2: Correlation coefficients analysis summary for news feed data set



	DR	*DV3*	*Lagged DV3*	*DV7*	*Lagged DV7*
Compound Sentiment Index	0.0027	0.0860	0.1270*	0.0846	0.1046
	0.9636	0.1427	0.0300	0.1493	0.0743
Positive Sentiment Index	-0.0139	0.0757	0.1130	0.0087	0.0192
	0.8114	0.1938	0.0538	0.8823	0.7439
Negative Sentiment Index	-0.1000	0.0825	0.1476*	0.0337	0.1185*
	0.0857	0.1597	0.0116	0.5635	0.0430

Table 3: Correlation coefficients analysis summary for the Twitter data set



	DR	*DV3*	*Lagged DV3*	*DV7*	*Lagged DV7*
Compound Sentiment Index	0.1053	-0.2443*	-0.2974*	-0.1159	-0.1775
	0.0724	0.0124	0.0022	0.2412	0.0714
Positive Sentiment Index	0.0069	-0.2178*	-0.3152*	0.0243	-0.3012*
	0.9065	0.0263	0.0011	0.6792	0.0019
Negative Sentiment Index	-0.1000	0.0348	0.3649*	0.0072	0.1347*
	0.0858	0.5513	0.0000	0.9022	0.0213

** statistically significant at α = 5%, p-values are present below the coefficients (r)*

It can be observed from the analysis of the correlation coefficients between the compound sentiment index, and the volatility metrics that across both the data sets, 3-period volatility measures (normal and lagged) are generally statistically significant, while correlation coefficients for the 7-period volatility measures are all statistically insignificant. This aligns with the expectation that investor sentiment is continuously and near-instantaneously reflected in the market prices. The 3-period metrics capture these effects with greater efficacy than their 7-period counterparts.

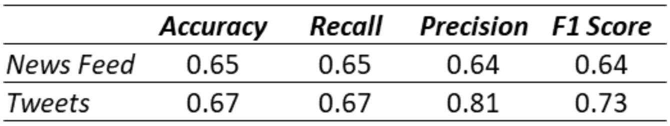
Another striking observation is the negative correlation between the positive sentiment index and volatility metrics. Some of the correlation coefficients corresponding to these combinations are also statistically significant. This implies that positive investor sentiment results in calmer less-erratic markets.

Another primary observation is the correlation coefficients for the Twitter data set are more prominent compared to the news feed data results. This could be explained by the dynamic nature of Twitter and the outsized effect of social media platforms on market participants' sentiments. Twitter and similar platforms allow investors, financial institutions and experts, corporations, and regulatory authorities to convey information and their opinions freely, while the process of publishing articles is typically longer. The most integral remark regarding these results is while the sentiment indices do express a certain degree of statistically significant correlation with volatility measures, the same is not valid for daily returns or price movement.

Since the ADF test has confirmed the stationarity for all the time series variables involved in this paper, the application of the granger causality test is valid. Analysis of the results of bidirectional granger causality tests (Refer to Appendix A) between all sentiment indices and market metrics reinforces the characteristics observed earlier. Compound sentiment and positive sentiment indices failed to reject the null hypothesis for any volatility measure and daily returns, while the negative sentiment granger causes lagged $DV7_t$ and $D3_t$ Indices.

5.2 Results from the LDA and logistic regression procedure

Table 4: LDA and logistic regression procedure results



	Accuracy	*Recall*	*Precision*	*F1 Score*
News Feed	0.65	0.65	0.64	0.64
Tweets	0.67	0.67	0.81	0.73

Results from LDA and logistic regression confirm the observation made earlier. The sentiment index constructed from the Twitter data performs considerably better than indices developed from the news feed data. Both data sets yielded results that were somewhat similar to prior studies such as Atkins et al. (2018) and Mahajan et al. (2009).

6 Conclusion

This study empirically demonstrated that information derived from textual news sources could be utilized to forecast directional fluctuations in market volatility, emphasizing that changes in volatility are better anticipated than changes in the returns of an asset or portfolio of investments. It also highlighted the role social media has taken in our lives and shows how it is more influential than traditional news sources. The findings indicate that information in the news, impacting markets through sentiment-driven behavior, has a significant effect on the financial system's second-order statistics.

7 References

Agarwal, A. (2020). Sentiment Analysis of Financial News. *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*. https://doi.org/10.1109/cicn49253.2020.9242579

Atkins, A., Niranjan, M., & Gerding, E. (2018). Financial news predicts stock market volatility better than close price. *The Journal of Finance and Data Science*, *4*(2), 120–137. https://doi.org/10.1016/j.jfds.2018.02.002

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

Calomiris, C. W., & Mamaysky, H. (2017). How News and Its Context Drive Risk and Returns around the World. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2944826

Caporin, M., & Poli, F. (2017). Building News Measures from Textual Data and an Application to Volatility Forecasting. *Econometrics*, *5*(3), 35. https://doi.org/10.3390/econometrics5030035

Clayton J. Hutto, & Eric Gilbert. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *International Conference on Weblogs and Social Media*, *8*(1), 216–225.

David M. Blei, Andrew Y. Ng, & Michael I. Jordan. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.5555/944919.944937

David Sontag, & Dan Roy. (2011). Complexity of Inference in Latent Dirichlet Allocation. *Neural Information Processing Systems*, *24*, 1008–1016.

Ekaputri, A. P., & Akbar, S. (2022). Financial News Sentiment Analysis using Modified VADER for Stock Price Prediction. *2022 9th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*. https://doi.org/10.1109/icaicta56449.2022.9932925

Glasserman, P., & Mamaysky, H. (2019). Does Unusual News Forecast Market Stress? *Journal of Financial and Quantitative Analysis*, *54*(5), 1937–1974. https://doi.org/10.1017/s0022109019000127

Groß-Klußmann, A., & Hautsch, N. (2011). When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions. *Journal of Empirical Finance*, *18*(2), 321–340. https://doi.org/10.1016/j.jempfin.2010.11.009

Han Xiao, & Thomas Stibor. (2010). Efficient Collapsed Gibbs Sampling For Latent Dirichlet Allocation. *Asian Conference on Machine Learning*, 63–78.

Jiao, P., Veiga, A., & Walther, A. (2020). Social media, news media and the stock market. *Journal of Economic Behavior &Amp; Organization*, *176*, 63–90. https://doi.org/10.1016/j.jebo.2020.03.002

Kelechava, M. (2021, December 8). *Using LDA Topic Models as a Classification Model Input*. Medium. https://towardsdatascience.com/unsupervised-nlp-topic-models-as-a-supervised-learning-input-cf8ee9e5cf28

Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., & Smith, N. A. (2009). Predicting risk from financial reports with regression. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics On - NAACL '09*. https://doi.org/10.3115/1620754.1620794

Kumar, S., Yadava, M., & Roy, P. P. (2019). Fusion of EEG response and sentiment analysis of products review to predict customer satisfaction. *Information Fusion*, *52*, 41–52. https://doi.org/10.1016/j.inffus.2018.11.001

Kumari, J., & Mahakud, J. (2015). Does investor sentiment predict the asset volatility? Evidence from emerging stock market India. *Journal of Behavioral and Experimental Finance*, *8*, 25–39. https://doi.org/10.1016/j.jbef.2015.10.001

Long, S., Lucey, B., Xie, Y., & Yarovaya, L. (2022). "I Just Like the Stock": The Role of Reddit Sentiment in the GameStop Share Rally. *Financial Review*. https://doi.org/10.1111/fire.12328

LOUGHRAN, T., & MCDONALD, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Mahajan, A., Dey, L., & Haque, S. M. (2008). Mining Financial News for Major Events and Their Impacts on the Market. *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. https://doi.org/10.1109/wiiat.2008.309

Mao, H. (2011, December 5). *Predicting Financial Markets: Comparing Survey, News, Twitter and. . .* arXiv.org. https://arxiv.org/abs/1112.1051

Nousi, C., & Tjortjis, C. (2021). A Methodology for Stock Movement Prediction Using Sentiment Analysis on Twitter and StockTwits Data. *2021 6th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*. https://doi.org/10.1109/seeda-cecnsm53056.2021.9566242

Paramanik, R. N., & Singhal, V. (2020). Sentiment Analysis of Indian Stock Market Volatility. *Procedia Computer Science*, *176*, 330–338. https://doi.org/10.1016/j.procs.2020.08.035

Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., & Welling, M. (2008). Fast collapsed gibbs sampling for latent dirichlet allocation. *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 08*. https://doi.org/10.1145/1401890.1401960

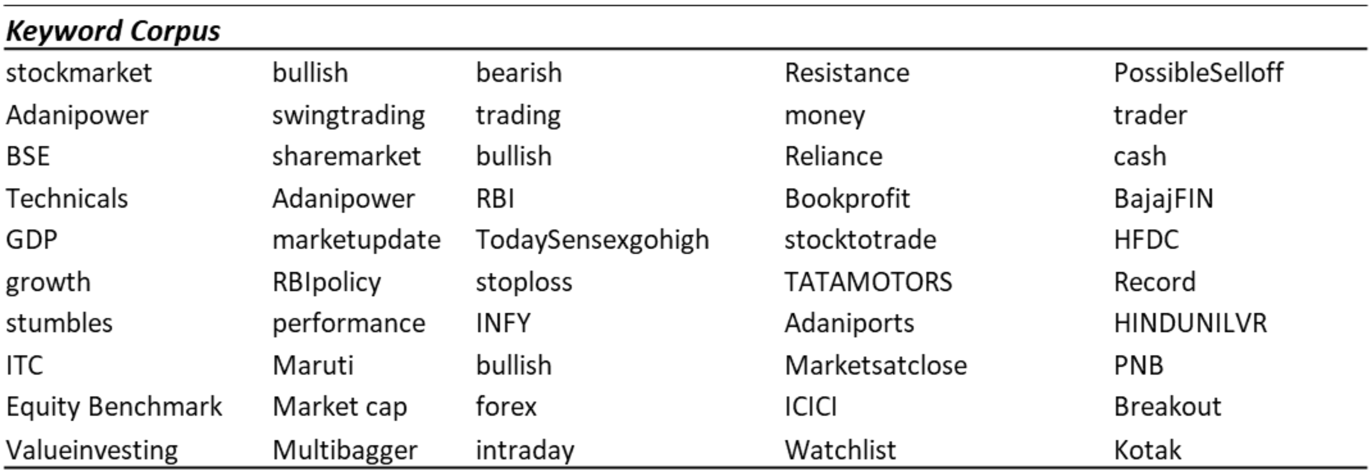
Sohangir, S., Petty, N., & Wang, D. (2018). Financial Sentiment Lexicon Analysis. *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*. https://doi.org/10.1109/icsc.2018.00052

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

TETLOCK, P. C., SAAR-TSECHANSKY, M., & MACSKASSY, S. (2008). More Than Words: Quantifying Language to Measure Firms' Fundamentals. *The Journal of Finance*, *63*(3), 1437–1467. https://doi.org/10.1111/j.1540-6261.2008.01362.x

8 Appendix

8.1 Appendix A



Keyword Corpus				
stockmarket	bullish	bearish	Resistance	PossibleSelloff
Adanipower	swingtrading	trading	money	trader
BSE	sharemarket	bullish	Reliance	cash
Technicals	Adanipower	RBI	Bookprofit	BajajFIN
GDP	marketupdate	TodaySensexgohigh	stocktotrade	HFDC
growth	RBIpolicy	stoploss	TATAMOTORS	Record
stumbles	performance	INFY	Adaniports	HINDUNILVR
ITC	Maruti	bullish	Marketsatclose	PNB
Equity Benchmark	Market cap	forex	ICICI	Breakout
Valueinvesting	Multibagger	intraday	Watchlist	Kotak

8.2 Appendix B

ADF Test:

ADF test checks for all four types of non-stationarity present within variables, namely Deterministic Trend ($y_t = \alpha + \beta t + \varepsilon_t$) and Random Walk with Drift, Deterministic Trend ($y_t = \alpha + \beta t + y_{t-1} + \varepsilon_t$), Pure Random Walk ($y_t = y_{t-1} + \varepsilon_t$) and Random Walk with Drift ($y_t = \alpha + y_{t-1} + \varepsilon_t$),
The testing procedure for the ADF test is applied to the following model:

$$\Delta \boldsymbol{y_t} = \boldsymbol{\alpha} + \boldsymbol{\beta_1 t} + \boldsymbol{\beta_2 t^2} + \boldsymbol{\gamma y_{t-1}} + \boldsymbol{\Phi_1} \Delta \boldsymbol{y_{t-1}} + \cdots + \boldsymbol{\Phi_{p-1}} \Delta \boldsymbol{y_{t-p+1}} + \boldsymbol{\varepsilon_t}$$

where:

- Δ is the first different operator
- α is a constant
- β_1 is the coefficient on a time trend
- β_2 is the coefficient on a squared time trend

This model can be estimated, and testing for a unit root is equivalent to testing that γ=0.

In sum, the ADF test hypothesis is as follows:

$$\boldsymbol{H_0 : \gamma = 0}$$
$$\boldsymbol{H_1 : \gamma < 0}$$

where:

- H_o is the null hypothesis (i.e., y_t has a unit root)
- H_1 is the alternate hypothesis (i.e., y_t does not have a unit root)

The test statistics (τ) value is calculated as follows:

$$\tau = \frac{\boldsymbol{\hat{\gamma}}}{\boldsymbol{\sigma_{\hat{\gamma}}}}$$

where:

- $\hat{\gamma}$ is the estimated coefficient
- $\sigma_{\hat{\gamma}}$ is the standard error in the coefficient estimate

The test statistics value (τ) is compared to the Dickey-Fuller Test's relevant critical value. If the test statistic is less than the critical value, we reject the null hypothesis and conclude that no unit root is present. The number of non-missing values in the input time series must be at least 10. The ADF test does not directly test for stationarity but indirectly through the unit root's existence (or absence). Furthermore, ADF incorporates a deterministic trend (and trend squared), allowing a trend stationary process to occur.

8.3 Appendix C

Granger Causality Test:

Granger causality is a statistical concept of causality that is based on prediction. According to Granger causality, if a signal X1 "Granger-causes" (or "G-causes") a signal X2, then past values of X1 should contain information that helps predict X2 above and beyond the information contained in past values of X2 alone. Its mathematical formulation is based on linear regression modelling of stochastic processes (Granger 1969). More complex extensions to nonlinear cases exist, however these extensions are often more difficult to apply in practice.

G-causality is normally tested in the context of linear regression models. For illustration, consider a bivariate linear autoregressive model of two variables X1 and X2:

$$X_1(t) = \sum_{j=1}^{p} A_{11,j} X_1(t-j) + \sum_{j=1}^{p} A_{12,j} X_2(t-j) + E_1(t)$$

$$X_2(t) = \sum_{j=1}^{p} A_{21,j} X_1(t-j) + \sum_{j=1}^{p} A_{22,j} X_2(t-j) + E_2(t)$$

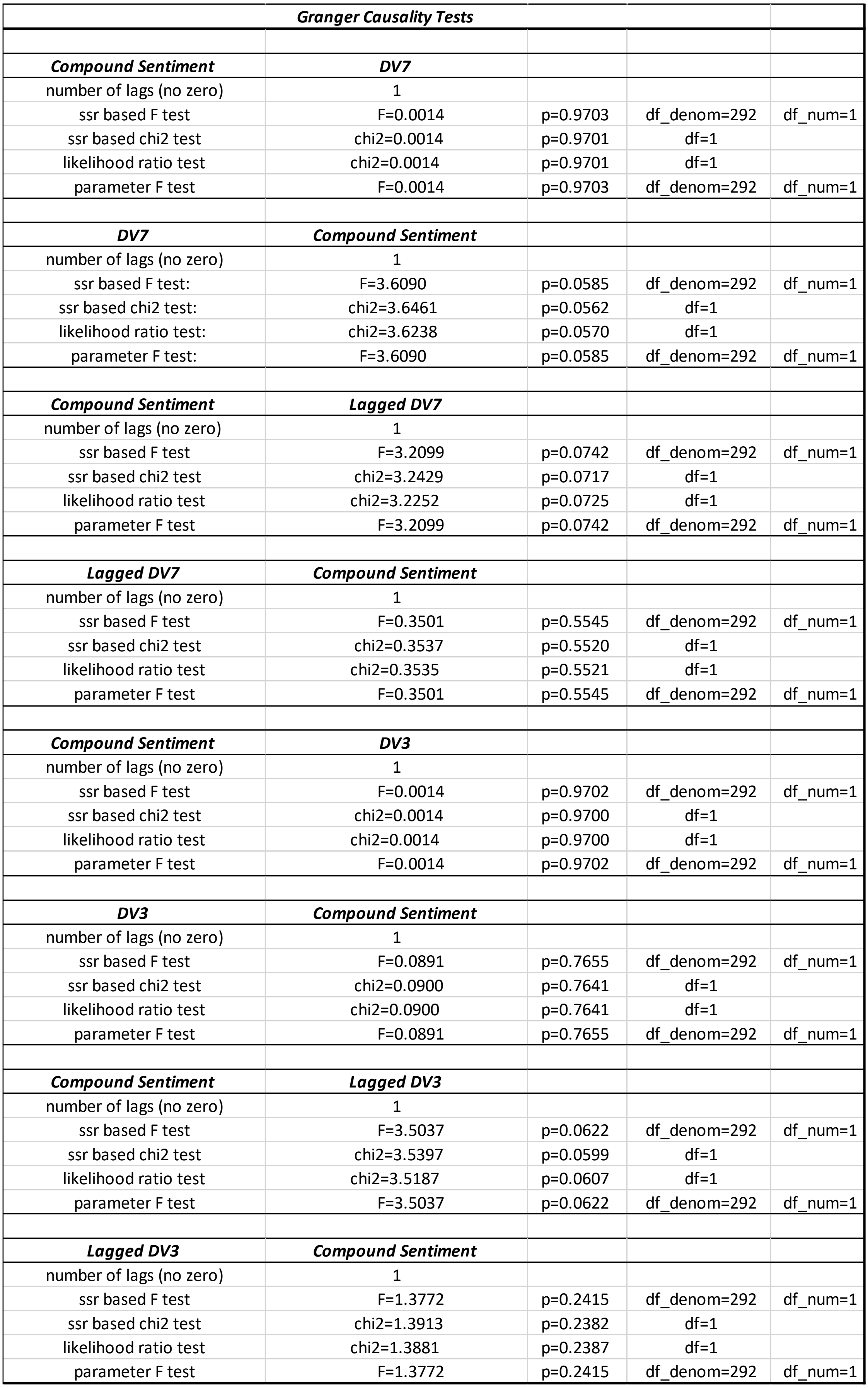
where p is the maximum number of lagged observations included in the model (the model order), the matrix A contains the coefficients of the model (i.e., the contributions of each lagged observation to the predicted values of X1(t) and X2(t), and E1 and E2 are residuals (prediction errors) for each time series. If the variance of E1 (or E2) is reduced by the inclusion of the X2 (or X1) terms in the first (or second) equation, then it is said that X2 (or X1) Granger-(G)-causes X1 (or X2). In other words, X2 G-causes X1 if the coefficients in A12 are jointly significantly different from zero. This can be tested by performing an F-test of the null hypothesis that A12 = 0, given assumptions of covariance stationarity on X1 and X2. The magnitude of a G-causality interaction can be estimated by the logarithm of the corresponding F-statistic (Geweke 1982). Note that model selection criteria, such as the Bayesian Information Criterion (BIC, (Schwartz 1978)) or the Akaike Information Criterion (AIC, (Akaike 1974)), can be used to determine the appropriate model order p.

As mentioned in the previous section, G-causality can be readily extended to the n variable case, where n>2, by estimating an n variable autoregressive model. In this case, X2 G-causes X1 if lagged observations of X2 help predict X1 when lagged observations of all other variables X3…XN are also taken into account. (Here, 'X3…XN correspond to the variables in the set W in the previous section; see also Boudjellaba et al. (1992) for an interpretation using autoregressive moving average (ARMA) models.) This multivariate extension, sometimes referred to as' conditional' G-causality (Ding et al. 2006), is extremely useful because repeated pairwise analyses among multiple variables can sometimes give misleading results. By contrast, a conditional/multivariate analysis would infer a causal connection from X to Y only if past information in X helped predict future Y above and beyond those signals mediated by Z. Another instance in which conditional G-causality is valuable is when a single source drives two outputs with different time delays. A bivariate analysis, but not a multivariate analysis, would falsely infer a causal connection from the output with the shorter delay to the output with the longer delay.

Application of the above formulation of G-causality makes two important assumptions about the data: (i) that it is covariance stationary (i.e., the mean and variance of each time series do not change over time), and (ii) that it can be adequately described by a linear model.

8.4 Appendix D

Granger Causality Test Results



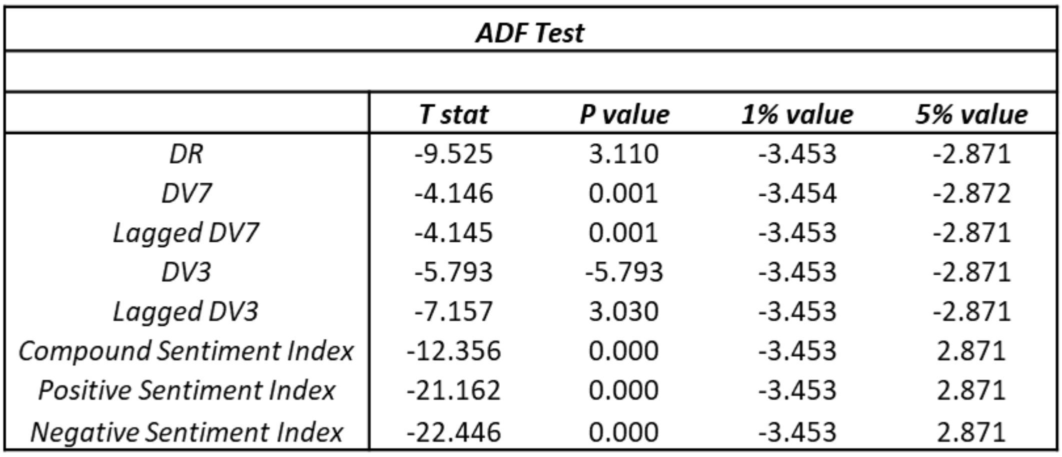
	Granger Causality Tests			
Compound Sentiment	***DV7***			
number of lags (no zero)	1			
ssr based F test	F=0.0014	p=0.9703	df_denom=292	df_num=1
ssr based chi2 test	chi2=0.0014	p=0.9701	df=1	
likelihood ratio test	chi2=0.0014	p=0.9701	df=1	
parameter F test	F=0.0014	p=0.9703	df_denom=292	df_num=1
DV7	***Compound Sentiment***			
number of lags (no zero)	1			
ssr based F test:	F=3.6090	p=0.0585	df_denom=292	df_num=1
ssr based chi2 test:	chi2=3.6461	p=0.0562	df=1	
likelihood ratio test:	chi2=3.6238	p=0.0570	df=1	
parameter F test:	F=3.6090	p=0.0585	df_denom=292	df_num=1
Compound Sentiment	***Lagged DV7***			
number of lags (no zero)	1			
ssr based F test	F=3.2099	p=0.0742	df_denom=292	df_num=1
ssr based chi2 test	chi2=3.2429	p=0.0717	df=1	
likelihood ratio test	chi2=3.2252	p=0.0725	df=1	
parameter F test	F=3.2099	p=0.0742	df_denom=292	df_num=1
Lagged DV7	***Compound Sentiment***			
number of lags (no zero)	1			
ssr based F test	F=0.3501	p=0.5545	df_denom=292	df_num=1
ssr based chi2 test	chi2=0.3537	p=0.5520	df=1	
likelihood ratio test	chi2=0.3535	p=0.5521	df=1	
parameter F test	F=0.3501	p=0.5545	df_denom=292	df_num=1
Compound Sentiment	***DV3***			
number of lags (no zero)	1			
ssr based F test	F=0.0014	p=0.9702	df_denom=292	df_num=1
ssr based chi2 test	chi2=0.0014	p=0.9700	df=1	
likelihood ratio test	chi2=0.0014	p=0.9700	df=1	
parameter F test	F=0.0014	p=0.9702	df_denom=292	df_num=1
DV3	***Compound Sentiment***			
number of lags (no zero)	1			
ssr based F test	F=0.0891	p=0.7655	df_denom=292	df_num=1
ssr based chi2 test	chi2=0.0900	p=0.7641	df=1	
likelihood ratio test	chi2=0.0900	p=0.7641	df=1	
parameter F test	F=0.0891	p=0.7655	df_denom=292	df_num=1
Compound Sentiment	***Lagged DV3***			
number of lags (no zero)	1			
ssr based F test	F=3.5037	p=0.0622	df_denom=292	df_num=1
ssr based chi2 test	chi2=3.5397	p=0.0599	df=1	
likelihood ratio test	chi2=3.5187	p=0.0607	df=1	
parameter F test	F=3.5037	p=0.0622	df_denom=292	df_num=1
Lagged DV3	***Compound Sentiment***			
number of lags (no zero)	1			
ssr based F test	F=1.3772	p=0.2415	df_denom=292	df_num=1
ssr based chi2 test	chi2=1.3913	p=0.2382	df=1	
likelihood ratio test	chi2=1.3881	p=0.2387	df=1	
parameter F test	F=1.3772	p=0.2415	df_denom=292	df_num=1

	Granger Causality Tests			
Positive Sentiment	***DV7***			
number of lags (no zero)	1			
ssr based F test	F=0.9223	p=0.3377	df_denom=292	df_num=1
ssr based chi2 test	chi2=0.9318	p=0.3344	df=1	
likelihood ratio test	chi2=0.9303	p=0.3348	df=1	
parameter F test	F=0.9223	p=0.3377	df_denom=292	df_num=1
DV7	***Positive Sentiment***			
number of lags (no zero)	1			
ssr based F test	F=1.6549	p=0.1993	df_denom=292	df_num=1
ssr based chi2 test	chi2=1.6719	p=0.1960	df=1	
likelihood ratio test	chi2=1.6671	p=0.1966	df=1	
parameter F test	F=1.6549	p=0.1993	df_denom=292	df_num=1
Positive Sentiment	***Lagged DV7***			
number of lags (no zero)	1			
ssr based F test	F=0.0001	p=0.9905	df_denom=292	df_num=1
ssr based chi2 test	chi2=0.0001	p=0.9905	df=1	
likelihood ratio test	chi2=0.0001	p=0.9905	df=1	
parameter F test	F=0.0001	p=0.9905	df_denom=292	df_num=1
DV7	***Positive Sentiment***			
number of lags (no zero)	1			
ssr based F test	F=1.3734	p=0.2422	df_denom=292	df_num=1
ssr based chi2 test	chi2=1.3875	p=0.2388	df=1	
likelihood ratio test	chi2=1.3843	p=0.2394	df=1	
parameter F test	F=1.3734	p=0.2422	df_denom=292	df_num=1
Positive Sentiment	***DV3***			
number of lags (no zero)	1			
ssr based F test	F=2.3531	p=0.1261	df_denom=292	df_num=1
ssr based chi2 test	chi2=2.3773	p=0.1231	df=1	
likelihood ratio test	chi2=2.3678	p=0.1239	df=1	
parameter F test	F=2.3531	p=0.1261	df_denom=292	df_num=1
DV3	***Positive Sentiment***			
number of lags (no zero)	1			
ssr based F test	F=0.0086	p=0.9262	df_denom=292	df_num=1
ssr based chi2 test	chi2=0.0087	p=0.9257	df=1	
likelihood ratio test	chi2=0.0087	p=0.9257	df=1	
parameter F test	F=0.0086	p=0.9262	df_denom=292	df_num=1
Positive Sentiment	***Lagged DV3***			
number of lags (no zero)	1			
ssr based F test	F=0.4183	p=0.5183	df_denom=292	df_num=1
ssr based chi2 test	chi2=0.4226	p=0.5156	df=1	
likelihood ratio test	chi2=0.4223	p=0.5158	df=1	
parameter F test	F=0.4183	p=0.5183	df_denom=292	df_num=1
Lagged DV3	***Positive Sentiment***			
number of lags (no zero)	1			
ssr based F test	F=0.1832	p=0.6689	df_denom=292	df_num=1
ssr based chi2 test	chi2=0.1851	p=0.6670	df=1	
likelihood ratio test	chi2=0.1850	p=0.6671	df=1	
parameter F test	F=0.1832	p=0.6689	df_denom=292	df_num=1

	Granger Causality Tests			
Negative Sentiment	***DV7***			
number of lags (no zero)	1			
ssr based F test	F=0.1684	p=0.6818	df_denom=292	df_num=1
ssr based chi2 test	chi2=0.1701	p=0.6800	df=1	
likelihood ratio test	chi2=0.1701	p=0.6800	df=1	
parameter F test	F=0.1684	p=0.6818	df_denom=292	df_num=1
DV7	***Negative Sentiment***			
number of lags (no zero)	1			
ssr based F test	F=1.1871	p=0.2768	df_denom=292	df_num=1
ssr based chi2 test	chi2=1.1992	p=0.2735	df=1	
likelihood ratio test	chi2=1.1968	p=0.2740	df=1	
parameter F test	F=1.1871	p=0.2768	df_denom=292	df_num=1
Negative Sentiment	***Lagged DV7***			
number of lags (no zero)	1			
ssr based F test	F=4.8306	p=0.0287	df_denom=292	df_num=1
ssr based chi2 test	chi2=4.8802	p=0.0272	df=1	
likelihood ratio test	chi2=4.8403	p=0.0278	df=1	
parameter F test	F=4.8306	p=0.0287	df_denom=292	df_num=1
DV7	***Negative Sentiment***			
number of lags (no zero)	1			
ssr based F test	F=0.3785	p=0.5389	df_denom=292	df_num=1
ssr based chi2 test	chi2=0.3824	p=0.5363	df=1	
likelihood ratio test	chi2=0.3821	p=0.5365	df=1	
parameter F test	F=0.3785	p=0.5389	df_denom=292	df_num=1
Negative Sentiment	***DV3***			
number of lags (no zero)	1			
ssr based F test	F=4.2980	p=0.0390	df_denom=292	df_num=1
ssr based chi2 test	chi2=4.3422	p=0.0372	df=1	
likelihood ratio test	chi2=4.3105	p=0.0379	df=1	
parameter F test	F=4.2980	p=0.0390	df_denom=292	df_num=1
DV3	***Negative Sentiment***			
number of lags (no zero)	1			
ssr based F test	F=0.0019	p=0.9651	df_denom=292	df_num=1
ssr based chi2 test	chi2=0.0019	p=0.9649	df=1	
likelihood ratio test	chi2=0.0019	p=0.9649	df=1	
parameter F test	F=0.0019	p=0.9651	df_denom=292	df_num=1
Negative Sentiment	***Lagged DV3***			
number of lags (no zero)	1			
ssr based F test	F=2.2471	p=0.1349	df_denom=292	df_num=1
ssr based chi2 test	chi2=2.2702	p=0.1319	df=1	
likelihood ratio test	chi2=2.2615	p=0.1326	df=1	
parameter F test	F=2.2471	p=0.1349	df_denom=292	df_num=1
Lagged DV3	***Negative Sentiment***			
number of lags (no zero) 1				
ssr based F test	F=4.2567	p=0.0400	df_denom=292	df_num=1
ssr based chi2 test	chi2=4.3004	p=0.0381	df=1	
likelihood ratio test	chi2=4.2693	p=0.0388	df=1	
parameter F test	F=4.2567	p=0.0400	df_denom=292	df_num=1

8.5 Appendix E

ADF test results



ADF Test

	T stat	***P value***	***1% value***	***5% value***
DR	-9.525	3.110	-3.453	-2.871
DV7	-4.146	0.001	-3.454	-2.872
Lagged DV7	-4.145	0.001	-3.453	-2.871
DV3	-5.793	-5.793	-3.453	-2.871
Lagged DV3	-7.157	3.030	-3.453	-2.871
Compound Sentiment Index	-12.356	0.000	-3.453	2.871
Positive Sentiment Index	-21.162	0.000	-3.453	2.871
Negative Sentiment Index	-22.446	0.000	-3.453	2.871