



# Social sentiment and impact in US equity market: an automated approach

José Antonio Nuñez-Mora<sup>1</sup> · Román A. Mendoza-Urdiales<sup>2</sup>

Received: 18 July 2023 / Revised: 11 August 2023 / Accepted: 15 August 2023 / Published online: 2 September 2023  
© The Author(s) 2023

## Abstract

In this study, a database of approximately 50 million tweets was used for the estimation of the positive and negative sentiment factors for 2557 companies operating in US stock market. For each company, the sentiment factors were calculated through the mean equations on GARCH models of different orders. Our findings show that, for 503 companies the negative factor effect has a greater impact than the positive factor effect. The period analyzed was from October 2022 to January 2023, using hourly observations. Results provide evidence to support that there is an asymmetric effect from the factors traveling to the stock market and it takes at least an hour the signal to travel. The investors and regulatory agents can find useful the results given that news has been demonstrated a source of influence in the market. Therefore, news impact can be modeled into portfolio theory using GARCH which is easy to implement and to interpret. Given the exposure of prices and volatility to news, it can be considered that these findings provide evidence to support efficient market hypothesis. Modeling returns and volatility for the assets through GARCH family is a widely known tool. Including the news sentiment on social media is dually a novelty: the empirical demonstration of the effects of social comments on the stock performance and volatility, in addition to the use of a large data set of social network comments in an hourly frequency.

**Keywords** Computational Thinking · Natural Language Processing · Sentiment Analysis · High Frequency Trading · Computational Complexity · Data Mining

## 1 Introduction

A recent research topic is defined through the impact of the textual analysis on the prices of stock prices of companies which receive comments in social media. Twitters provoke an important influence on stock prices given that many persons are participating with different opinions and all participants are informed of the rest of the opinions. The analysis of text and sentiment has been applied in different contexts, health in Corti et al. (2022), presidential elections Ali et al. (2022), forecasting short-term prices of stock in Shen and Shafiq (2020), event detection in Kolajo et al.

(2022) and consumer sentiment in Kaur and Sharma (2023). In the financial field, the efficient market hypothesis established by Fama in 1965 recognizes the influence on prices of the information accessible to investors (Fama 1965). In this sense the mood of the investor has an impact of time series of prices and so we can say the psychological factor affect the investments of financial agents (Shiller 2003). The impact of news dissemination has two components: the size and the direction. In general, positive news has a positive impact on returns and negative news has a negative impact on returns, see (Shiller 2003; DeGennaro and Shrieves 1997; Atkins et al. 2018; Audrino et al. 2020; Ren et al. 2020; Li et al. 2014). Sentiment not only has this impact on prices, but there also exists an asymmetric effect, i.e., negative news has a stronger impact than positive news (Mendoza-Urdiales et al. 2022). Moreover, an important issue is the prediction of prices using social sentiment.

A challenge emerges to investors and risk managers given that the information available and necessary to include in decision making process has been increasing exponentially in volume and the decision time reduced. In particular,

✉ Román A. Mendoza-Urdiales  
roman.alejandro.mendoza.ur@gmail.com  
José Antonio Nuñez-Mora  
janm@tec.mx

<sup>1</sup> Monterrey Institute of Technology and Higher Education, Monterrey, Mexico

<sup>2</sup> Institute for the Future of Education, Monterrey Institute of Technology and Higher Education, Monterrey, Mexico

Twitter for many of the financial markets' actors is one of the most popular and accepted methods to deliver and gather relevant information to the general public.

There are documented cases in which a fake news affects the stock price of companies. One recent relevant example is from last November 12, 2022, in which a fake Eli Lilly (Major insulin provider) Twitter account claimed that insulin was free. The announcement was not well received by investors; consequently, the stock price of the company fell 4.37%.<sup>1</sup> Going further, we can observe that even when a "positive" announcement was given, with initially mostly positive responses and fewer negative comments, the negative impact in the stock price supports the hypothesis in which the negative effects are greater than positive ones, an effect we have analyzed and presented in this study.

Additionally, the impact on the volatility of stock returns can be considered through the family of generalized autoregressive conditional heteroskedasticity (GARCH) models. As a contribution of this paper, we analyzed several models of this family with different lags for 2,557 companies in stocks exchanges of the US economy. Most of these companies have showed that the returns are affected in an asymmetric way using models like EGARCH and APARCH. Over 50 million tweets were extracted, analyzed, and categorized for a period of 4 months (850 hourly observations) to create negative and positive sentiment factors, which were used as independent variables in the mean equation of the GARCH models.

The objective of this study is to use all public available information mentioning the companies' tickers and measure the sentiment. With it, calculate the total exposure of each company performance to its corresponding sentiment factors (positive and negative) independently from the rest of the companies. Finally, the questions to answer are:

1. Is it possible to separate positive and negative sentiment for each company and calculate the exposure to each factor using GARCH models?
2. Is there an asymmetric effect between negative and positive factor?
3. Increasing the dataset to all the available information allows to increase the frequency of the observations? If so, does the signal takes less than an hour to travel from social media to the stock market?

The rest of this paper is structured as follows. "Related work" section introduces related works. "Problem definition" and "Method" sections represent the proposed methodology. "Results and discussion" section shows experimental and evaluation of the proposed method. Finally,

<sup>1</sup> Fake Eli Lilly Twitter Account Claims Insulin Is Free, Stock Falls 4.37%. Source: <https://www.forbes.com/sites/brucelee/2022/11/12/fake-eli-lilly-twitter-account-claims-insulin-is-free-stock-falls-43/?sh=6e45d51a41a3>.

"Conclusion" section concludes the paper and gives some future research directions.

## 2 Related work

Before the emergence of social networks, the question of the effect of news on the stock prices has been studied in the financial literature. For example, in Chan (2003) bad news is followed by a strong drift return using headlines about individual companies. Antweiler et al. (2004) studied the influence of messages in Yahoo! Finance and Raging Bull on 45 Dow Jones Industrial Average stock prices. The main finding is the existence of influence on market return and volatility but without predictive capability. In Dougal et al. (2012) it is shown that returns of Dow Jones Industrial Average (DJIA) can be predicted using a column from Wall Street Journal (Abreast of the Market).

Tetlock (2007) developed an analysis of the relationship between stock market and media (Wall Street Journal news) using the vector autoregressive technique. Similarly, consequences of negative words in Wall Street Journal and Dow Jones News Service on S&P500 earnings and returns are studied in Tetlock et al. (2008) using quantification of language.

More recently has been recognized the influence of social media on the prices. For example, a sentiment analysis by topic is applied in Nguyen et al. (2015) for prediction of 18 stocks using support vector machine. In Daniel et al. (2017) the impact of popular events tweets on thirty companies of the Dow Jones Average is developed using sentiment analysis and four algorithms, after filtering the noisy tweets.

In a theoretical model, Yang and Li (2013) showed that investors sentiment has impact on the asset prices and in Bollen et al. (2011) the mood in large scale set of twitters is used for prediction of the Dow Jones Industrial Average (DJIA) and some dimensions of sentiment cause in the Granger sense the DJIA. Applying a textual analysis which works in a phrase level, Boudoukh et al. (2012) showed a strong relation of this information with the change of prices.

Heston and Sinha (2016) used 900 thousand news stories to demonstrate that returns are predicted by this news and the predictability depends on the aggregation level of news. Daily news is useful for one or two days ahead. In a study of 181 altcoins prediction, Steinert et al. (2018) used sentiment analysis to show the prediction of such altcoins prices.

In Derakhshan and Beigy (2019) support vector machine algorithm is used to demonstrate the level of prediction of stocks using sentiment and in Yang and Wu (2021) an asset pricing model is developed showing that public information has an impact on the sentiment of investors and the sentiment investors has an impact on asset price.

For the period of COVID Outbreak, Das et al. (2022) showed for the case of Nifty-50 data that the level of

accuracy is better when mood is combined with news headlines together stock data. Using snapshots of information and sentiment time series Anbaee Farimani et al. (2022) studied the impact on price prediction for EUR/USD, GBP/USD, USD/JPY32 and BTC/USD.

In Figà-Talamanca and Patacca (2022) a study of the impact of textual content of newspapers or social media postings on the return and variance of components of S&P500 using a GARCH family approach is studied. In Jiang et al. (2021) a GARCH model including Baidu index is used to forecast the returns of Chinese stock during the COVID-19 period. The Baidu index affects the sentiment of the investors and therefore has an influence on the stocks returns.

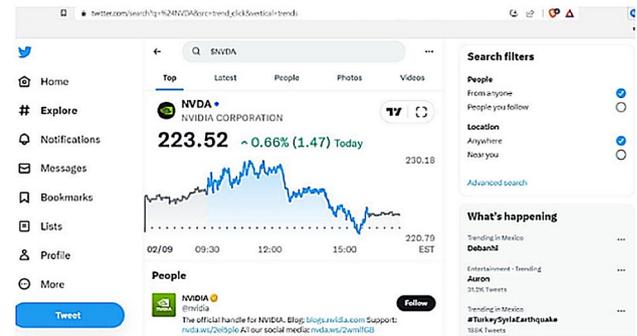
In our paper we applied GARCH-type models to take into account the variance equation together the inclusion of the positive and negative indexes in the equation of the mean similar to the sentiment impact on variance developed in Olson and Nowak (2020) for Dow Jones Average using a GARCH (1,1) model and Mendoza et al. (2022) for a sample of 24 companies over a 10-year period.

## 2.1 Problem definition

Our previous example in which the stock company was affected by a fake account could be an indicator that the public has been increasingly relying in social networks for collecting information. The reviewed literature presented studies that contribute empirical evidence and some models that measure the impact of the social sentiment in the stock market behavior. Some of these studies do so in small periods of time and the ones with large time frames use relatively small tweets/companies' samples (i.e., Mendoza et al. 2021, 2022). The initial problem to solve in this study was, if the total universe of comments mentioning a large number of companies would allow clearer results regarding the sentiment factor impact in the stock returns. As a secondary hiatus was, the larger dataset would allow to measure in greater frequency the impact of social sentiment factor in the stock returns and calculate the asymmetric effect. In Fig. 1, it is presented financial information of company Nvidia consulted in Twitter.

## 2.2 Methodology

Before the emergence of social networks, there have been papers relating sentiment expressed in texts from financial news like Chan (2003), Antweiler et al.(2004) and Dougal et al. (2012) with prices and predictions of these prices. In a very outstanding way Tetlock (2007) and Tetlock et al. (2008) applied an econometric with the same objective. Similar to our paper, these papers studied language to translate it into a variable which can be related to the stock market



**Fig. 1** The use of the Social Network Twitter for stock market monitoring has increased in the last years. Source: [https://twitter.com/search?q=%24NVDA&src=typeahead\\_click](https://twitter.com/search?q=%24NVDA&src=typeahead_click). Accessed on Feb 9, 2023

prices. However there have been big changes and challenges in the diffusion of news with the emergence of social media like twitter. Analogously to our paper, Boudoukh et al. (2012), Heston and Sinha (2016), Steinert and Herff (2018), Derakhshan and Beigy (2019) and Das et al. (2022) among others used different econometric and machine learning techniques to show a link between news and stock prices in the financial markets. Moreover, in this paper an hourly frequency of twitter is applied to find an asymmetric effect of negative news respect to the positive news.

In Mendoza et al. (2022), it has been established a causal relationship between the social sentiment factor and the returns for a small sample of companies over a long period of time. Extracting the top 1% of the comments mentioning 23 companies, in this study, the approach will be to an increased number of companies using the total universe of comments in English language, allowing us to measure the sentiment factor impact in hourly observations, instead of the original daily observations. In Table 1, we summarize the difference between both studies.

The framework utilized in this study is presented in Fig. 2; it is the automated process designed for the extraction of the data, sentiment calculation, time series construction of quantitative and qualitative data and GARCH simulations.

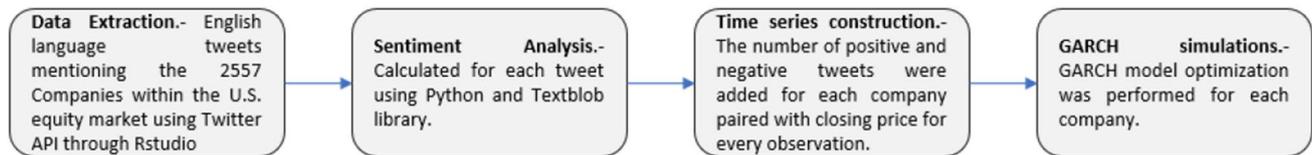
## 2.3 Variable selection

It has been defined in the previous section the asymmetric effect of behavioral finance we are aiming to capture and measure for the equity market. In this section we will focus on the experiment design.

Approximately 60% of the total capital being traded in the world is concentrated in USA, for which we will focus on companies based there. We have selected 2,557 companies that represent approximately 85% of US equity markets and extract all the English language comments that mention each company ticker in Twitter (\$ + Ticker). Another way

**Table 1** Evolution of data sampling and comparison between approaches

	Previous Study	Current Approach
Number of companies	24	2,557
Time period	10 years	120 days
Frequency of observations	Closing daily price	Closing hourly price
Extraction criteria	Top 1% of comments	Total number of comments
Companies	83%	93%

**Fig. 2** Framework used to model sentiment factor for each company

the tickers are identified in the social network is *cashtag*, a variation from hashtag. With this selection criteria, we are covering approximately 50% of the global equity markets.

## 2.4 Data extraction

Twitter API tool was used to extract the data through Academic Research Access, and RStudio was the software used for this purpose. The *rtweet* package version 0.7.0<sup>2</sup> was used to perform the requests of the 2,557 companies analyzed in the present study. The requests need to be configured accordingly and will be explained next:

- Type of request. – The *rtweet* package has different types of requests according to the level of access granted from Twitter. The most common type of accessibility is the “*search\_tweets*” requests which covers up to 7 days before the extraction with a maximum volume of 500,000 tweets.
- The search term or terms. – We are looking for all the tweets mentioning each company ticker(*cashtag*) which is constructed “\$” + ‘Ticker’.
- Number of tweets (n). –The maximum number of tweets that the requests can extract, and the limit desired for the consultation, we always requested the maximum allowed which is 500,000 (n = 500,000).
- Tweet language (lang). – We are focusing on the English language tweets (lang = ‘en’).
- Tweet mode (*tweet\_mode*).—Originally the tweets were capped at 140 characters but by now it has been

extended to 280. By requesting the extended version of each tweet (*tweet\_mode* = ‘extended’), we are ensuring that the full text is extracted for each tweet.

Additionally, the *retryonratelimit* = TRUE option was selected to persist the request to ensure that the full dataset was extracted. This request was performed for each company ticker systematically each week at the same hour, to make sure that all the comments were captured. In Fig. 3 we present the code used to extract the tweets using RStudio.

For the companies’ financial data extraction, we used Refinitiv Eikon API<sup>3</sup> (library: *refinitiv-data*) in Python. It is important to mention that there exists a difference between a company’s Ticker and RIC.<sup>4</sup> Refinitiv software requires to the request to be in RIC form. The full table with the companies’ names, tickers and RICs in in the repository. The requests need to be configured accordingly and will be explained next:

- Type of request. – For eikon the request type for time series is *get\_timeseries*.
- The search term or terms. – The company RIC, which is the American stock market Ticker.
- Interval. – Frequency of observations. The elected interval is hour.
- start\_date*. – Starting date of data extraction.
- end\_date*. – Ending date of data extraction.

The period selected for extracting the data was from October 1, 2022, to Jan 1, 2023, in which both the tweets

<sup>2</sup> *rtweet* is an R package of free Access to extract Twitter data via its API. For further information please visit: <https://cran.r-project.org/web/packages/rtweet/rtweet.pdf>

<sup>3</sup> Refinitiv <https://developers.refinitiv.com/en/api-catalog/eikon/eikon-data-api>

<sup>4</sup> RIC.

```
search_tweets('$Ticker', n = 500000, retryonratelimit = TRUE, lang = "en", tweet_mode='extended')
```

**Fig. 3** Request form for Twitter API using rtweet package in RStudio. This was performed on a weekly basis for each of the 2,557 companies. Source: <https://cran.r-project.org/web/packages/rtweet/rtweet.pdf>. Accessed on Feb 9, 2023

```
1 for stock in stocks:
2     try:
3         df = ek.get_timeseries(stock, interval='hour', start_date="2021-12-31", end_date="2023-02-01")
4
5         df.to_csv(r'C:/Users/Downloads/rics/'+stock+'.csv')
6
7     except ek.EikonError as e:
8         print(stock, ' not available')
9
```

**Fig. 4** Request form for Refinitiv API using eikon library in Python

were extracted from Twitter API and the hourly observations of the closing stock prices were extracted from Refinitiv API. This concludes the extraction section, dealing with the rounding of the observation's hours and time zones transformation will be addressed in the Time Series Construction section. In Fig. 4, we present the code used for extracting the stocks financial information from Refinitiv Eikon using Python.

## 2.5 Sentiment analysis

Having obtained the data samples for our dependent (stock prices) and independent variables (tweets), in order to run some statistical analysis, data transformation from qualitative form into quantitative using automated processes was needed. Python was the language in which a natural language processing algorithm was coded, and the library used for said task was TextBlob<sup>5</sup>. This library calculates the sentiment by breaking each text analyzed individually into the words that compose the text. Single letter words are ignored and for the rest of the text, a numeric value is given for each word that is already assigned inside the library, a value for polarity and subjectivity. When composed expressions are used (i.e., 'very1 great2'), the library recognizes the emphasizing word 'very' that precedes 'great', for which polarity is ignored and multiplies the intensity for the following words polarity.

This library was selected given that it includes interpretation of popular abbreviations and emojis, which are heavily found in tweets. Additionally, comparing the results with other libraries such as Vader, the results show Textblob has

more sensibility to the common language found in social media.

In addition to the existing algorithm, helping the software in cleaning each tweet phrase, we improved the technique changing abbreviations to the full extent of the words (i.e., 'ive', to 'I have', 'im' to 'I am', etc.); this step was very needed since the abbreviation of words is very common in twitter given the limited space for each tweet (280 characters). Given that for words that the library does not detect or identifies, the resulting assigned value is zero. By cleaning each tweet, we reduced considerably the margin error. Example, we have taken a real tweet from November 10, 2022 (Fig. 5):

“We are excited to announce insulin is free now.”

We can understand that the user is stating that the company Eli Lilly would start giving away insulin. This is the example mentioned in the introduction section in which a fake message was delivered affecting the stock price of Eli Lilly. The first step is to clean the tweet from characters other than letters and from abbreviations. The algorithm returned the clean sentence:

“we are excited to announce insulin is free now.”

The second step the algorithm breaks the tweet in sentences and words:

```
[Sentence ('we are excited to announce insulin is free now')]
WordList(['we', 'are', 'excited', 'to', 'announce', 'insulin', 'is', 'free', 'now'])
```

Finally, it analyzes the polarity and subjectivity for the sentence adding the individual score for each word. The individual score is also prerecorded in the library.

Sentiment(polarity = 0.3875, subjectivity = 0.775)

<sup>5</sup> Textblob is a Python library that is used for sentiment analysis, which calculates sentiment polarity and subjectivity. Source: <https://textblob.readthedocs.io/en/dev/>. Accessed on Feb 9, 2023.



Fig. 5 Tweet from a fake account impersonating Eli Lilly, major insulin provider

The result for this example was a polarity ( $P$ )=0.3875 meaning that it has a 38.75% of positivity according to our algorithm. The sentiment analysis process was applied for each tweet in our extracted data.

### 2.6 Time series construction (pairing qualitative and quantitative data)

Having extracted the data from Twitter and Refinitiv and calculated the Sentiment for each tweet, pairing qualitative and quantitative data is performed in 4 steps:

- (1) *Matching time zones across all datasets* According to the API configuration of Twitter API, the date format is configured to UTC time zone. It was necessary to transform it to US/EST time, which is the time zone in which NYSE operates. Additionally, the date format of Refinitiv API is GMT, which was also needed to be transformed into US/EST time zone. With the previous date and hour homologation it is ensured that the matching of sentiment and financial data is correctly done.
- (2) *Rounding hours in continuous distribution data* Another consideration before pairing the sentiment and financial data is rounding to the hour the sentiment data. The raw form in which the data is extracted from the API is to the millisecond, but which rounding criteria is a fair question to be answered. The usual way would be to the nearest hour, but this presents a problem given that if you have a comment at 15:17, rounding to the closest hour (15:00) would pair the observation to a stock return prior to the comment occurring. By rounding to the ceiling (16:00), the comment occurrence is paired to the following stock return and would ensure that the causality from the sentiment is measured toward the company and not the other way around. In Table 2 we present the differences of resulting rounded hours between rounding to the nearest and the following hour (ceiling).

Table 2 Apple test grouping difference between nearest and ceiling methods. This ensures that the sentiment observation is paired with the closing price return that affects

Hour	Nearest	Ceiling
15:48	16:00	16:00
15:48	16:00	16:00
15:48	16:00	16:00
08:16	08:00	09:00
13:38	14:00	14:00
11:42	12:00	12:00
11:30	12:00	12:00
10:15	10:00	11:00
12:15	12:00	13:00
17:18	17:00	18:00

- (3) *Grouping of sentiment data by date and company* Having transformed the data time to US/EST time zone and rounded the tweets hours to the next hour the time series creation was performed. If the tweet polarity is  $< 0$ , it is labeled negative  $N$ , if its polarity is  $> 0$ , it is labeled positive  $P$ . After labeling all the tweets, they are aggregated by date  $t$ , creating the time series  $N$  for negative tweets and  $P$  for positive tweets. The stock price return  $R$  for date  $t$  is calculated by subtracting the closing price  $S$  of  $t-1$  from the date  $t$ . The same notation of (Mendoza et al. 2022) is used as follows:

$$N_t = \sum_{n=1}^i x_n \tag{1}$$

$$ZNeg_t = \frac{N_t - \mu_N}{\sigma_N} \tag{2}$$

$$Neg = [ZNeg_t, ZNeg_{t+1}, ZNeg_{t+2}, \dots] \tag{3}$$

$$P_t = \sum_{n=1}^i y_n \tag{4}$$

$$ZPos_t = \frac{P_t - \mu_P}{\sigma_P} \tag{5}$$

$$Pos = [ZPos_t, ZPos_{t+1}, ZPos_{t+2}, \dots] \tag{6}$$

$$R_t = S_t - S_{t-1} \tag{7}$$

$$Z_{R_t} = \frac{R_t - \mu_R}{\sigma_R} \tag{8}$$

$$R = [Z_{R_t}, Z_{R_{t+1}}, Z_{R_{t+2}}, \dots] \tag{9}$$

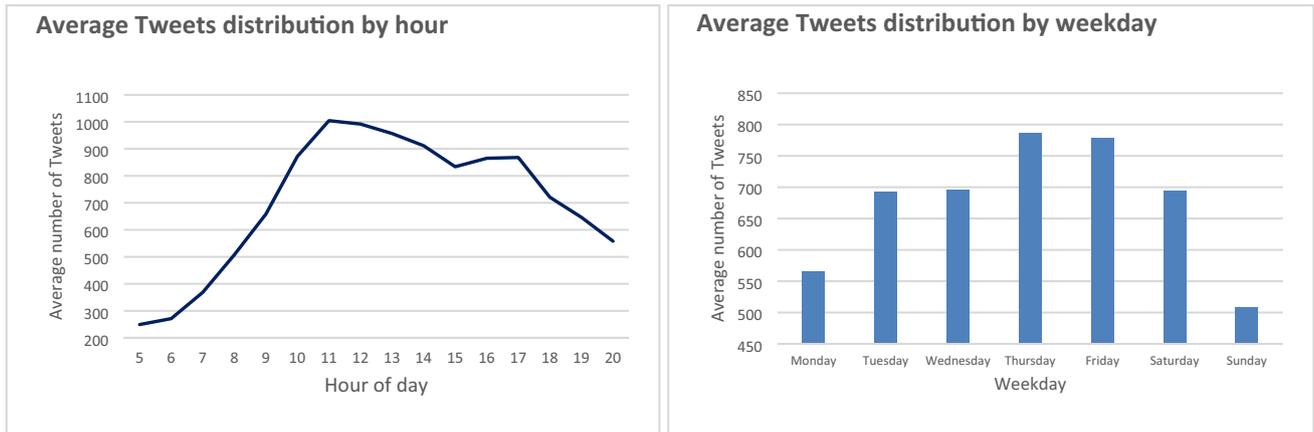


Fig. 6 Average number of tweets distribution by hour and weekday for Tesla. Period (Oct 2022, Jan 2023)

$$T_t = N_t + P_t \tag{10}$$

$$ZT_t = \frac{T_t - \mu_T}{\sigma_T} \tag{11}$$

$$T = [Z_{T_t}, Z_{T_{t+1}}, Z_{T_{t+2}}, \dots] \tag{12}$$

$$ACWI_t = ACWI_t - ACWI_{t-1} \tag{13}$$

$$Z_{ACWI_t} = \frac{ACWI_t - \mu_{ACWI}}{\sigma_{ACWI}} \tag{14}$$

$$ACWI = [Z_{ACWI_t}, Z_{ACWI_{t+1}}, Z_{ACWI_{t+2}}, \dots] \tag{15}$$

- (4) **Pairing Qualitative and Quantitative data.** Neg and Pos, calculated in Eqs. (3) and (6), are what we have introduced as negative and positive factors. These factors then are paired with their respective hourly closing price return ( $R$ , developed in Eq. 9) of each company. Here is where we have matched qualitative (sentiment factors) with quantitative (stock returns).
- (5) **Substituting observations with no comments.** After constructing the time series sentiment factors (Pos and Neg) and  $R$  for each company and paired by US/EST date (YYY-mm-dd hh:mm), the companies were categorized by percentage of hours with no comments. Even when for every company all the comments were extracted, and the index was constructed by using hourly observations, there would be a percentage of companies with periods of time with no comments due the varying levels of attention that companies get. Meaning, that a controversial company such as Apple or Tesla would have a greater volume of tweets than

a less controversial company, such as Starbucks. For that, the companies were grouped in 3 clusters, the first cluster with no comments in (d):  $0\% > d \leq 10\%$  of the observations. The second cluster with  $10\% > d \leq 20\%$  and the third cluster with  $20\% > d \leq 30\%$ . All the companies with no comments and thus no sentiment calculated in greater than 30% of the observations were discarded, keeping 556 of the original 2557.

This modification was introduced given that in our previous studies when there were zero comments on an observation, the sentiment was considered to be zero, which introduces a limitation because in the sentiment definition, zero means that there is indeed a comment with sentiment zero, meaning that the commentary is neutral. This could have introduced unnecessary noise to the results in our previous studies. By ensuring that in every observation there is a number of comments, the noise is at mitigated, or reduced.

In Fig. 6, it is presented the average distribution of tweets for Tesla during the analyzed period in which the volume of tweets follows a cyclical behavior. This trend was benchmarked to filling periods of time with no comments, meaning that for each company the average number of tweets, negative and positive factors were used to fill periods of time with no comments. As an example, if for a Monday at 14:00 there were not negative and positive factor observations for Starbucks, because there were not observations, the empty spaces were filled using the average negative and positive values of all the Monday 14:00 observations of Starbucks, this process was replicated for the 556 companies.

### 2.7 GARCH simulations

To present a proposal for measuring the direction and percentage of stock returns that is affected by movement in

**Table 3** Summary of results according to missing data clustering criteria. It can be observed that the fewer missing data, the higher is the success rate

Cluster	Missing data ( <i>d</i> )	Companies	Feasible solution	S.F. condition met <sup>a</sup>	Success rate (%)
1	10% ≥ <i>d</i>	216	209	199	95.22
2	20% ≥ <i>d</i> > 10%	168	162	151	93.21
3	30% ≥ <i>d</i> > 20%	172	168	153	91.07
Total		556	539	503	93.32

<sup>a</sup>The aim of this study is to fit a model that satisfies that the negative factor negatively influences the stock performance in a greater way than positives influence from the positive factor

negative and positive factors, we fitted a GARCH model for each company (*c*) using their corresponding sentiment factors in addition to the number of tweets and returns of ACWI (ACWI is an index that represents the performance of the world global stock market, named All Country World Index created by MSCI). The mean equation can be written as:

$$R_c = \beta_0 + \beta_1 T_c + \beta_2 Neg_c + \beta_3 Pos_c + \beta_4 ACWI_c + u_c \quad u_c \sim N(0, \sigma_t) \quad (16)$$

Several GARCH models were created for each company by using all possible combinations of the orders and distributions until there was found a feasible solution, additionally, the companies that solution met the asymmetric effect, in which negative factor coefficient (having negative sign) was in absolute terms greater than the Positive Factor coefficient (having positive sign). Three different specifications were found for the companies' data. The GARCH, EGARCH and APARCH (*p, q*).

Variance equation for each case is:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^q \alpha_i u_{t-i}^2 + \sum_{j=1}^p \beta_j \sigma_{t-j}^2 \quad (17)$$

Variance equation for EGARCH (*p, q*)

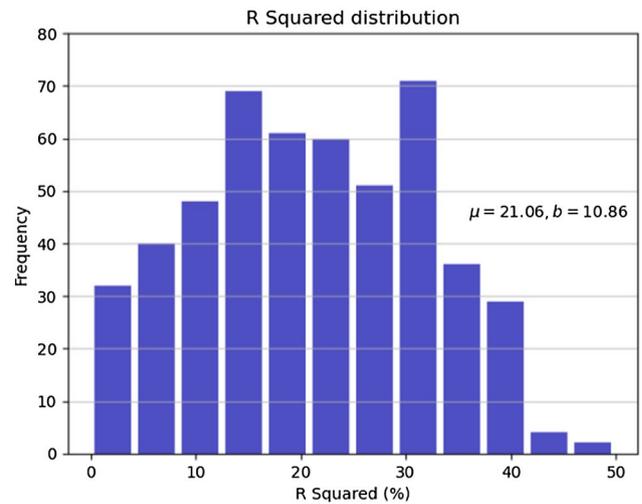
$$\log \sigma_t = \frac{\omega}{2} + \sum_{i=1}^p a_i \log \sigma_{t-i} + \sum_{i=1}^q b_i \delta_{t-i} |\theta_{t-i}| \quad (18)$$

Variance equation APARCH (*p, q*)

$$\sigma_t^\delta = \omega + \sum_{i=1}^q \alpha_i (|u_{t-i}| - \gamma_i u_{t-i})^\delta + \sum_{j=1}^p \beta_j \sigma_{t-j}^\delta \quad (19)$$

### 3 Results

Of the 2,557 analyzed companies and calculated sentiment factors paired with the hourly closing stock price, the companies with less than 30% of missing observations— independently of the reason for this—were selected for the

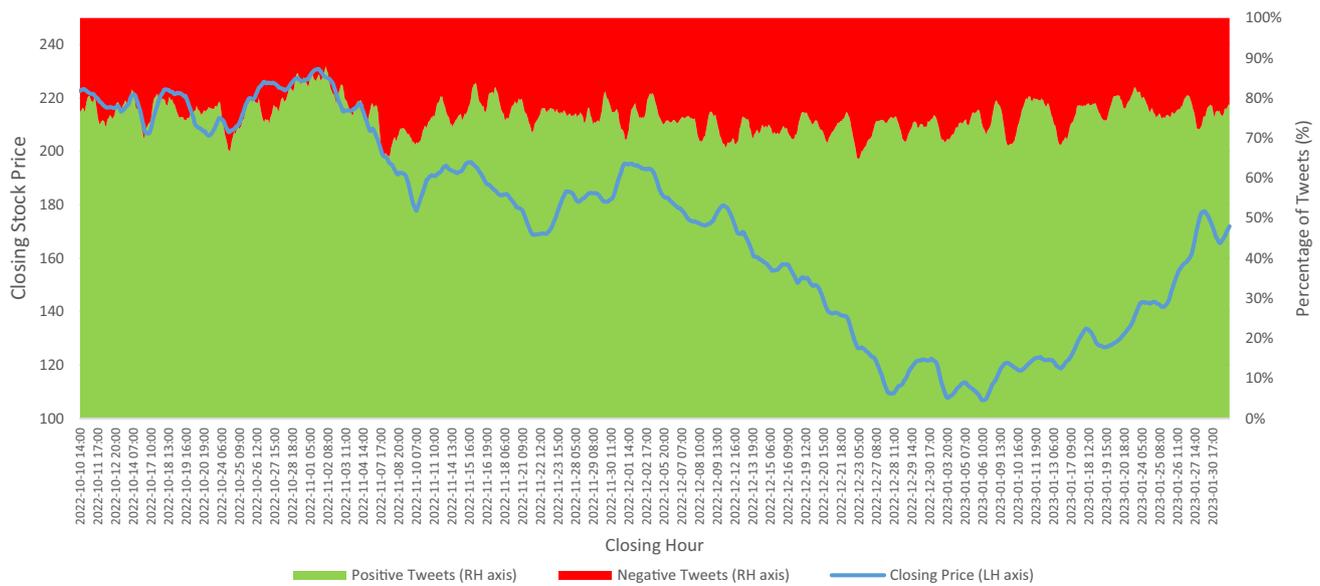


**Fig. 7** R Squared distribution for the 503 companies that GARCH model found a feasible solution and the sentiment factor conditions met, in which the negative factor (Neg -with negative sign-) had an absolute value greater than positive factor (Pos -with positive sign-)

GARCH modeling. Finally keeping 556 companies divided in three clusters presented in Table 3:

Of the 556 companies selected for the GARCH modeling, there were feasible solutions for 539 cases, in which 93.32% of the cases the asymmetric condition in which the negative factor coefficient (with negative sign) absolute value was greater than the positive factor coefficient (with positive sign). An interesting characteristic of the results is that the success rate is higher for the clusters with fewer missing observations. In Fig. 7, the R squared for the 503 companies is presented as a distribution chart; it can be appreciated that the solutions fitted models ranges up to 50%, but the majority are concentrated between 10 and 30%.

In Table 3, we present the solution for each company; it can be observed that the models orders and distributions were included, in which the solutions volatility equations varied between GARCH, EGARCH and APARCH. An Important remark is that in previous works, the sentiment factors were calculated using daily observations. In this case, we have been able to measure the asymmetric influence in hourly observations, meaning that the velocity in which the



**Fig. 8** Percentage of positive and negative tweets and closing price of Tesla (Cashtag \$TSLA). Period: Oct 10, 2022, 08:00–Jan 31, 2023, 17:00 EST

sentiment travels from the social network to the stock market occurs under an hour.

Another remark is that in all our studies, the sample data from social networks the volume of positive comments is greater than the volume of positive comments. And even with this disbalance between the data, the models have captured the asymmetric effect. In Fig. 8, there is presented the percentage of positive tweets versus negative tweets for Tesla (Ticker TSLA) in which the mean equation for an EGARCH model that satisfied the sentiment factor asymmetry resulted:

$$R_{TSLA} = -0.056 + 0.018T_{TSLA} - .0007Neg_{TSLA} + .00008Pos_{TSLA} + 0.48ACWI_{TSLA} + u_{TSLA}$$

In which the negative factor has an impact nearly 10 times as greater than positive factor. This effect has been present in our two previous studies. Let us recall that we are working with hourly observations in this study and that the volume of tweets is somewhat large at this point.

### 4 Discussion

There is still debate in academia regarding how to measure the signal from social networks to stock market performance. In this study, consistent with our two previous works, we have been able once again to measure the impact of negative and positive factor sentiment in stock market for a wide number of companies. Compared to Olson and Nowak (2020), in which

the authors measure the negative and positive impact based on negative and positive word count for panel data for the Dow Jones average performance. In this case we have been able to breakdown the effect to each company analyzed and create an individual sentiment factor.

In this study it is presented a methodology to measure the level of overreaction to negative news through the Negative sentiment Factor and that the signal travels as fast as an hour. The econometric model used is around GARCH family showing the asymmetric impact of news on the returns of stock prices. These effects have been recognized to exist showing that negative news tend to have a greater impact on the dependent variable than positive news. For the case of the firms reported in our paper, this stylized fact is found in the modeling of the variable of sentiment divided into positive and negative mood related to the prices (returns) of the stocks.

The present study is useful for the investors and regulators given the insights for the impact on portfolios which involve stocks and want to investigate for possible sources of returns and risk in the financial markets. As limitation here we used a member of the GARCH family. However, other models are more general like Markov switching GARCH, i.e., change of regime models to study the changes in the dynamic of princes/returns.

### 5 Conclusion

In this paper we have gone through the full framework that we have developed. To clarify, having coded the full process, it was a matter of automatizing the routines to have the

extraction, transformation, analysis, and summarization on a constant basis without major intervention.

Starting from the premise that there exists influence of the social networks in the stock markets, the natural next step would be to measure negative and positive impact aiming to capture the asymmetric effect. This research concludes that there is influence from the social network into the stock market and reaffirms our proposed factor calculation methodology in a previous work (Mendoza et al. 2022, 2021), the data sampling from top tweets to all available was increased and the stock companies from 24 to 2,557. Additionally, the frequency was increased from daily to hourly observations.

Another remark is that even when the volume of positive comments is on average 4 times as larger as the volume of negative comments, the impact of negative comments is as greater as 10 times than the positive comments. Additionally, if there exists persistence of negative comments, the affectation in the stock price keeps constant. Meaning that a single volume of very negative comments won't hurt a company price more than a constant not so negative flow of comments over a longer period of time.

Finally, we conclude with the success rate increase from 83% to 93% versus our previous work, meaning that for 503 of the 539 companies, it was possible to capture the asymmetric effect of negative and positive factor. At last, but not least, the speed at which the sentiment travels from the social network toward the stock market can be measured in under an hour.

Limitations of this work are the complication to extract and structure the enormous amount of information in time to make managerial decisions. Additionally, the access to the data given the acknowledgement of its importance by the managers of social media has been restricted over the last years. Policy implications are multiple, starting from including the analysis of a company popularity to make informed decisions, as to how can managers of the companies being mentioned in social media control their image to the general population. Future studies can focus in optimizing the sentiment and text trends extraction to capture further hidden data important for decision makers. Additionally, optimizing the data processing allowing managers to make decisions faster with more accurate information.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s13278-023-01116-6>.

**Acknowledgements** Tweets extracted using Twitter API through Academic Project "A comprehensive analysis of behavioural economics applied to social media."

**Authors contributions** JA conceptualized the study, performed the literature search, and wrote the original draft of the manuscript. RA designed the experiment, synthesized the data and modeled interpretation, provided supervision and resources, and assisted in manuscript

writing, review and editing. All authors approved the final manuscript version.

**Funding** The authors would like to thank the financial support from Tecnológico de Monterrey through the "Challenge-Based Research Funding Program 2022". Project ID # I001 - IFE001 - C1-T1 - E. Also, academic support from Writing Lab, Institute for the Future of Education, Tecnológico de Monterrey, México.

**Data availability** Data available at request.

## Declarations

**Conflict of interest** The authors have declared that no competing interests exist.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Ali RH, Pinto G, Lawrie E, Linstead EJA (2022) A large-scale sentiment analysis of tweets pertaining to the 2020 US presidential election. *J Big Data*. <https://doi.org/10.1186/s40537-022-00633-z>
- Anbae Farimani S, Vafaei Jahan M, Milani Fard A, Tabbakh SRK (2022) Investigating the informativeness of technical indicators and news sentiment in financial market price prediction. *Knowl Based Syst*. <https://doi.org/10.1016/j.knsys.2022.108742>
- Antweiler W, Frank MZ (2004) American finance association is all that talk just noise? The information content of internet stock message boards. *J Financ* 59(3):1259–1294
- Atkins A, Niranjana M, Gerding E (2018) Financial news predicts stock market volatility better than close price. *J Financ Data Sci* 4:120–137
- Audrino F, Sigrist F, Ballarín D (2020) The impact of sentiment and attention measures on stock market volatility. *Int J Forecast* 36:334–357
- Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. *J Comput Sci* 2(1):1–8. <https://doi.org/10.1016/j.jocs.2010.12.007>
- Boudoukh J, Feldman R, Kogan S, Richardson M. (2012) Nber working paper series which news moves stock prices? A textual analysis. <http://www.nber.org/papers/w18725>
- Chan WC (2003) Stock price reaction to news and no-news: Drift and reversal after headlines. *J Financ Econ* 70(2):223–260. [https://doi.org/10.1016/S0304-405X\(03\)00146-6](https://doi.org/10.1016/S0304-405X(03)00146-6)
- Corti L, Zanetti M, Tricella G, Bonati M (2022) Social media analysis of Twitter tweets related to ASD in 2019–2020, with particular attention to COVID-19: topic modelling and sentiment analysis. *J Big Data*. <https://doi.org/10.1186/s40537-022-00666-4>
- Daniel M, Neves RF, Horta N (2017) Company event popularity for financial markets using Twitter and sentiment analysis. *Expert*

- Syst Appl 71:111–124. <https://doi.org/10.1016/j.eswa.2016.11.022>
- Das N, Sadhukhan B, Chatterjee T, Chakrabarti S (2022) Effect of public sentiment on stock market movement prediction during the COVID-19 outbreak. *Soc Netw Anal Min*. <https://doi.org/10.1007/s13278-022-00919-3>
- DeGennaro RP, Shrieves RE (1997) Public information releases, private information arrival and volatility in the foreign exchange market. *J Empir Financ* 4:295–315. [https://doi.org/10.1016/S0927-5398\(97\)00012-1](https://doi.org/10.1016/S0927-5398(97)00012-1)
- Derakhshan A, Beigy H (2019) Sentiment analysis on stock social media for stock price movement prediction. *Eng Appl Artif Intell* 85:569–578. <https://doi.org/10.1016/j.engappai.2019.07.002>
- Dougal C, Engelberg J, García D, Parsons CA (2012) Journalists and the stock market. *Rev Financ Stud* 25(3):640–679. <https://doi.org/10.1093/rfs/hhr133>
- Fama EF (1965) The behavior of stock-market prices. *J Bus* 38:34–105
- Figà-Talamanca G, Patacca M (2022) An explorative analysis of sentiment impact on S&P 500 components returns, volatility and downside risk. *Ann Oper Res*. <https://doi.org/10.1007/s10479-022-05129>
- Heston SL, Sinha NR (2016) News versus sentiment: predicting stock returns from news stories. *Financ Econ Discuss Ser* 2016(048):1–35. <https://doi.org/10.17016/feds.2016.048>
- Jiang B, Zhu H, Zhang J, Yan C, Shen R (2021) Investor sentiment and stock returns during the COVID-19 pandemic. *Front Psychol*. <https://doi.org/10.3389/fpsyg.2021.708537>
- Kaur G, Sharma A (2023) A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis. *J Big Data*. <https://doi.org/10.1186/s40537-022-00680-6>
- Kolajo T, Daramola O, Adebisi AA (2022) Real-time event detection in social media streams through semantic analysis of noisy terms. *J Big Data*. <https://doi.org/10.1186/s40537-022-00642-y>
- Li X, Chen L, Wang J (2014) Deng X (2014) News impact on stock price return via sentiment analysis. *Knowl-Based Syst* 69:14–23
- Mendoza Urdiales RA, García-Medina A, Nuñez Mora JA (2021) Measuring information flux between social media and stock prices with Transfer Entropy. *PLoS ONE* 16(9):e0257686. <https://doi.org/10.1371/journal.pone.0257686>
- Mendoza-Urdiales RA, Núñez-Mora JA, Santillán-Salgado RJ, Valencia-Herrera H (2022) Twitter sentiment analysis and influence on stock performance using transfer entropy and EGARCH methods. *Entropy*. <https://doi.org/10.3390/e24070874>
- Nguyen TH, Shirai K, Velcin J (2015) Sentiment analysis on social media for stock movement prediction. *Expert Syst Appl* 42(24):9603–9611. <https://doi.org/10.1016/j.eswa.2015.07.052>
- Olson E, Nowak A (2020) Sentiment's effect on the variance of stock returns. *Appl Econ Lett* 27(18):1469–1473. <https://doi.org/10.1080/13504851.2019.1690123>
- Ren Y, Liao F, Gong Y (2020) Impact of news on the trend of stock price change: an analysis based on the deep bidirectional LSTM. *Procedia Comput Sci* 174:128–140
- Shen J, Shafiq MO (2020) Short-term stock market price trend prediction using a comprehensive deep learning system. *J Big Data*. <https://doi.org/10.1186/s40537-020-00333-6>
- Shiller RJ (2003) From efficient markets theory to behavioral finance. *J Econ Perspect* 17:83–104
- Steinert L, Herff C (2018) Predicting altcoin returns using social media. *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0208119>
- Tetlock PC (2007) Giving content to investor sentiment: the role of media in the stock market. *J Financ* 62(3):1139–1168
- Tetlock PC, Saar-Tsechansky M, Macskassy S (2008) More than words: quantifying language to measure firms' fundamentals. *J Financ* 63(3):1437–1467
- Yang C, Li J (2013) Investor sentiment, information and asset pricing model. *Econ Model* 35:436–442. <https://doi.org/10.1016/j.econmod.2013.07.015>
- Yang C, Wu H (2021) Investor sentiment with information shock in the stock market. *Emerg Mark Financ Trade* 57(2):510–524. <https://doi.org/10.1080/1540496X.2019.1593136>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.