Linköping University | Department of Computer and Information Science
Bachelor's thesis, 16 ECTS | Computer Science
2021 | LIU-IDA/LITH-EX-G--2021/071--SE

Tools and Methods for Analysis of Stock Market Manipulation using Social Media

– A Longitudinal Characterization of Time Series Dynamics

Verktyg och Metoder för Analys av Aktiemarknadsmanipulation med Sociala Medier

Carl Terve
Mattias Erlingsson

Supervisor : Niklas Carlsson
Examiner : Marcus Bendtsen

LINKÖPING UNIVERSITY

Linköpings universitet
SE–581 83 Linköping
+46 13 28 10 00 , www.liu.se

Upphovsrätt

Copyright

Abstract

Social media has proven to affect the dynamics of the stock market directly. The potential influence of social media makes it an excellent tool for stock market manipulation, especially in the era of online misinformation. Therefore, the work in this thesis aims to provide a better understanding of company discussion on social media. More precisely, collecting a dataset of company-specific discussion from Twitter, Reddit, Seeking Alpha, and Citron Research enabled us to perform time series analysis using smoothing and clustering, where associated events were identified, categorized, and summarized. A selected companies' time series was also more closely and qualitatively analyzed to build intuition and understanding of the dynamics. The results show that the dynamics of company discussion on social media can be evaluated using the discussion intensities. They also show the possibility of measuring weather a specific social media is in general reactive or proactive to abnormal events in the stock market. Moreover, external events seem to trigger discussion activity which propagates through all social media. Furthermore, the results seem to depend on the choice of bandwidth used for smoothing and the number of clusters given to the clustering algorithm, which requires further refinement as a continuation of this thesis.

Acknowledgments

We want to express our deepest gratitude to our thesis supervisor, Niklas Carlsson, who trusted our capabilities and gave us the freedom of exploration that this study required. Your support proved paramount to the progress of this thesis. A special thanks to Alireza Mohammadinodooshan, who was available for questions throughout the term and assisted with the data collection. We would also like to thank our project group members who have been giving advice and sharing ideas along the way. We also recognize our classmates who have reviewed our report and provided feedback on matters that otherwise would have been overlooked; thank you.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

The media and news landscape has seen a drastic shift in the past decade, moving more to the online paradigm. With social media taking the *pole position* in information dissemination, it has become a natural environment for discussion and information consumption. Today, fake news on social media is perhaps most frequently associated with politics and elections. But lately, it has become clear that it also applies to the financial market. Private investors partaking in online discussions are fed information, potentially resulting in poor investments and loss of capital as more attention is focused on social media for advice on potential *bullish* or *bearish* securities. Recent events have shed light on the power of social media as a tool for stock market manipulation, which may conclude in economic consequences for companies and private investors and mistrust in the use of financial instruments.

Another aspect highlighted in behavioural finance research [1] is the importance of social interaction for human decision making. In particular, emotion seems to play an essential role in how people make decisions when there is risk involved. The pessimism and optimism on topics transmit partly through the interactions on social media, affecting how people act on information on issues such as finance. Therefore, the public mood propagates to investors and the stock market, implying that the stock market can be used as a measure of social mood. Considering online activity has become a substantial part of how the stock market behaves, it is in the public's interest to better comprehend the social media dynamics and actual effects on the financial sphere.

1.2 Aim

As a first step towards understanding the effects of social media activity on the stock market, we must first examine what data can be found on such activity. Therefore, we aim to produce a comprehensive dataset with posts and comments from Reddit, Twitter, articles from Seeking Alpha, short reports from Citron Research, and metadata about the retrieved data. The collection targets a carefully selected set of companies. With this dataset, we hope to conclude what one may learn from it and how it can be helpful in the study of market manipulation. We will then perform a preliminary analysis on the dataset where we present possible strategies for a quantitative comparison on the time series using smoothing and clustering. While

plenty of related research focuses on social media predictors for stock price, this analysis will primarily seek to compare and contrast how the different social media behave before, during, and after financial events. More precisely, we ask questions regarding how social media activity correlates to these events or between themselves, if social media activity is generally reactive or proactive, and if a particular social media platform seem to act or react faster than other platforms. The answers to this preliminary analysis will lay the foundation for discussion on how the methodology performs and how the dataset may be used in future studies.

1.3 Research questions

1. What can we learn from publicly available APIs on social media, and how can they help understand social media and stock market dynamics?
2. Is social media reactive or proactive to financial events?
3. For a cluster of associated social media and stock events, what is the internal form, order, and timing?

1.4 Contributions

The main contribution of this thesis is building intuition and a foundation for the analysis of the social media landscape concerning the stock market. More precisely, the data collection highlight how stock related social media volumes have grown and how it allows for more far-reaching manipulation schemes. Another important contribution is how significant events, based on social media activity, can be identified, categorized and analyzed using clustering. Highlighting the weaknesses as well as the strengths of this approach lowers the threshold for potential future improvements. Processing the social media data in this way could prove helpful in navigating its vastness in search of valuable information.

1.5 Delimitations

Several delimitations had to be made in the scope of this work. The amount of data collected and the time it would take to download it was reduced. This was done by excluding the data between 2011 and 2019 for a large set of companies. Instead, by focusing the collection on two time periods, 2009-01-01 to 2011-04-20 and 2019-01-01 to 2021-04-20. These two periods each represent a cross-section for two different eras in the social media landscape. Another reason for the limitation was the lack of short-interest data between 2013 and 2019. We have also had to limit the number of companies analyzed to decrease the size of the dataset and the time it would take to retrieve it. Another necessary delimitation comes from not having full access to the Twitter API, which led to us only collecting Twitter data for a few companies, for which one was used in the analysis.

2 Background

In this section, we will provide a brief introduction to the different social media and online platforms. This is to better understand the origin of where the dataset has been collected from and why. They are also put in the context of market manipulation with concrete examples, which seeks to highlight their importance for further analysis. Understanding the basics of the financial market is vital for future interpretation of the results and the discussion and is therefore also explained here.

2.1 Social media discussing stocks

There are many social media platforms where stock market discussion occurs, such as Facebook, Twitter, Reddit, Seeking Alpha, Instagram, etc. However, some social media platforms provide a structure for identifying stock market discussions, e.g. Reddit with finance-specific subreddits, Twitter with the cashtag symbol ($) for identifying company tickers in tweets, Seeking Alpha, and Citron Research with its clear focus on stock market analysis articles. These social media platforms also have existing APIs for collecting historical data, which makes them especially interesting for the research of this paper.

Reddit

Reddit is a pseudonymous online discussion forum structured with subreddits. There are many subreddits available covering a wide range of topics. With its 430 million active monthly users across its subreddits, some having millions of subscribers, Reddit is considered to be one of the most favoured social networks in the world [2]. Reddit also has a unique structure for discussions. A user may create a post on a specific subreddit, comment on other posts, comment on other Redditors' comments, and upvote or downvote posts and comments. The structure allows for easy access to subreddit posts, discussion threads- and sub-threads made on a subreddit [3].

Twitter

Twitter is a global social media platform with approximately 330 million active monthly users [4]. It is one of the largest micro-blogs, especially prominent in the USA [5].

The platform allows its users to create tweets where they can share their thoughts. The users may also interact with other tweets with likes, comments, and retweets. Twitter is a some-to-many social network meaning that a small portion of the users creates most tweets.

Seeking Alpha

Seeking Alpha is an online investment community platform where people connect to share ideas, news, analysis, and more about the stock market. According to Seeking Alpha themselves, 20 million use the service monthly. Each month, 10 000 investing ideas are published from investors among the 7 000 active contributors. These ideas are reviewed before publication to ensure high quality, according to their website [6]. Compared to a tweet or Reddit post these reports are more qualitative by nature: they are usually longer, they directly comment on a stock, and the target group is other investors.

Citron Research

Citron Research is an online stock commentary website founded by Andrew Left [7]. The site aims to reveal fraud and fraudulent business models and publishes reports and investigations on companies. They have published over 150 reports since 2001, and according to their website, more than 50 companies reported have seen regulatory interventions since then.

2.2 Stock manipulation in social media

Market manipulation is the act of intentionally defrauding investors by controlling the price of a security or disrupting the natural forces of supply and demand. These acts undermine the honesty of a market and affect its integrity. One such scheme is *Pump and Dump* where adversaries obtain ownership of a significant amount of shares. Then they promote the stock to attract investors, causing the price to rise. Social media is a great place to achieve this pump phase. When the manipulators think the price is high enough, they sell and collect the profit. This usually drives the price downwards again, and investors acting on this false information lose their money. There are also schemes where trading was used to create specific patterns in the financial data, designed to fool investors [5].

As mentioned in the introduction, social media has proved to have a significant effect on the political domain, especially when it comes to the spread of fake news [8], [9]. There is also evidence showing how false information receives greater exposure and virality than true information on social media, no matter the category [10]. The fake news problem is prominent in the financial market. There are multiple examples of how social media has been used as a tool to influence the stock market directly, either utilizing fraudulent claims or by coordinated Pump and Dump schemes.

Investigations led by the U.S. Securities and Exchange Commission (SEC) in 2017 uncovered multiple scenarios where public companies hired communication firms and individuals to attract attention to their stocks on the investment site Seeking Alpha. This investigation led to 27 charges being made based on the fraudulent promotion of stocks [11].

Another example is GameStop and Reddit, where Wallstreetbets collectively drove the GameStop stock price up. In this case, the strategy was utterly transparent instead of more normal market manipulation schemes (i.e. Pump and Dump). While some bought shares to pick on hedge funds being short in the stock, others saw the opportunity to make a quick profit. However, when the former SEC Chief of the Office of Internet Enforcement, John Reed Stark, was asked in an interview in Forbes Magazine [12] when these activities would raise a legal risk for the individuals involved, he said it is complex and that this does not fall into the category of normal stock market manipulation, for which they usually charge. He mentions that the charges pressed by SEC typically fall into the category of the previously mentioned

Pump and Dump scheme or that of trading techniques to generate specific patterns in volume. With GameStop, the trading does not seem to be based on any fraud, and whether the activity is illegal is a matter of intent, he mentions [12]. Social media seems to be opening up for new schemes, legal or illicit, to manipulate the stock market.

It has also been clear that Twitter can be used to influence the market as well. For instance, influential people can move the market with a single tweet. There are multiple examples where this has been the case [13], [14]. There are also online stock commentary websites like Citron Research [7] that investigates companies and publishes reports to disclose potential fraud or misbehaviour. These reports have proven to affect the market as well. One example that gained public attention in 2018 was when they published a report on Twitter, which made the stock price drop 11 % on the same day.

These examples illustrate the potential power of social media and online forums for stock market manipulation, either intentionally or unintentionally. The market sentiment advertised by groups of people or by individuals on social media, therefore, seems to have a direct impact on the financial market.

2.3 The stock market

The financial market is broad, and it includes any marketplace where the trading of securities takes place. One of these marketplaces is the stock market, where investors may buy or sell shares in a company. An investor's stake in a company correlates to ownership of that share in the company, which they believe will eventually increase value. Stocks are mainly sold and bought at stock exchanges such as the variations of the New York Stock Exchange (NYSE) and NASDAQ [15]. Most of the trading today occurs on digital stock exchanges and have historically been one of the main contributors to economic growth as well as crashes [16].

The value of a share in a company is tightly connected to the company's assets, profits, and community opinion. Therefore owning shares in a company comes with significant risk as well as opportunity. There are opportunities to profit from either the value of a company dropping or increasing where fundamental and technical analysis of stock are two common approaches. The fundamental analysis mainly focuses on making a better prediction on a company's future financial statements than the rest of the market investors, and technical analysis focuses on historical parameters such as prior stock price, short interest, and sale volumes to detect patterns and make an investment [16].

Morningstar

For this thesis, we retrieved a collection of financial data. We retrieved the data from Morningstar's financial management and services, which is an American-based company. It allows for the collection of any security on the marketplace with an abundance of financial parameters such as the daily Open, High, Low, Close and Volume (OHLCV) and short interest and many more. In particular, the short-interest provided by morningstar is presented as a percent of shares float. This value represents the total number of shares shorted for a company divided by the total number of shares open for trade. The available shares that can be traded are often referred to as the float [17].

3 Related research

3.1 Characterization of social media dynamics

There have been numerous previous work trying to understand activity and information spread on social media. Wu et al. [18] conducted a study using Twitter to examine its information spread. Their paper shows that most of the consumed content on Twitter derives from a small subset of the user base and that the information spread is commonly homophily within classifications of users. Similarly to [18], Glenski et al. [19] examines information spread except on the social media platform Reddit. They create tree structures of comments to posts on three cryptocurrency subreddits to analyse the discussion cascades. Their results show that the discussion intensity varies between the three identified subreddits and that subjective posts on the subreddits result in a more intense discussion.

The work in this thesis differs from those previously mentioned by on a high-level examining the relations between different social media platforms' activity intensity and their correlations to each other and the stock market. More related to our work would be the paper by Kryvasheyeu et al. [20]. They use the social media platform Twitter to correlate subject-specific discussion intensity to external events, more precisely disasters geographically adjacent and close in time. They show through their study that large social media platforms can be an effective tool for rapidly evaluating disaster consequences. Here, the main differences with our work are that we correlate financial spikes to the discussion intensity from several social media platforms close in time and evaluate social media platforms as a tool for understanding external financial events.

3.2 Social media for prediction

Much previous and current research explores social media in the context of predicting future events. The wide use of social media for online discourse, together with its accessibility, speed, and reach, make it an excellent place for trendsetting and for influencing public agendas in a wide range of topics. Some view social media as a collective wisdom, and if enough people act upon this wisdom then one should be able to predict real-world outcomes. This was the assumption of Asur and Huberman [21], and their research shows how the rate of social media chatter on Twitter can be utilized to predict box-office dividends of movies. Furthermore, they conclude that there seems to be a strong correlation between topic attention

and its future popularity. Likewise, work by Gruhl et al. [22] show that online chatter volumes can be used as an early indicator of real-time events. Their work studied the influence of online blog posts as a predictor for future purchase decisions on Amazon.com. Using specific queries, they captured discussion on particular products and verified that the volume of these discussions acted as an indicator for future sales using time series analysis.

The compelling use cases for predicting stock returns make it a well-studied subject. Early work on stock predictions [23] is based on the so-called Efficient Market Hypothesis (EMH) [24]. This implies that the stock price follows a random walk and therefore cannot effectively be predicted. However, this theory has been criticized [25], and multiple studies have been done on social media sentiment as a predictor. Chen et al. [26] extract company sentiment from articles published on Seeking Alpha and find that these opinions strongly predict future stock returns. This points to the usefulness of seeking financial advice from such sources. Bollen et al. [27] analyses Twitter mood as a predictor for the value of the Dow Jones Industrial Average and finds an accuracy of 86.7 % in predicting daily up and down changes.

The work done by Asur and Huberman [21] and Gruhl et al. [22] correlates online chatter-volume to external events, which is similar to what is done in this thesis, although the approaches taken differ substantially. Since our work is not intended to predict the stock market, the work by Chen et al. [26], and Bollen et al. [27] is perhaps not directly related. However, their work further highlights the influence of social media on the financial market, particularly the dissemination of opinion through social media, such as Twitter and Seeking Alpha, etc., on stock-related topics.

3.3 Our work in relation to others

On a high level, our work differs from others in the field in that it seeks to analyze a broad dataset that spans multiple social media. In many cases a single social media is used, like Twitter in these [9], [18], [20], [21], [27], Reddit in [19] or Seeking Alpha in [26]. Most research that studies social media and the stock market together do so to predict stock market prices [26], [27] and not to evaluate the consequences of market manipulation. Nonetheless, their research all require substantial processing and characterization of large sets of social media data, and a lot can be learned by their approaches.

4 Method

4.1 Company selection strategy

In this thesis, we have focused our research on a selection of 41 companies from a variety of market sectors and market capitalization listed on the NYSE or NASDAQ exchanges. Mainly, the selection falls into two categories across two different periods. The first category is based on the amount of social media activity from Reddit, and the other on short interest. Companies were selected for both of these categories for the periods 2009-01-01 to 2011-04-20 and 2019-01-01 to 2021-04-20. These time periods were chosen to limit the amount of data-collection needed and because they represent different stages in the evolvement of social media. Using similar selection measures on different time periods also lets us compare what companies end up in the final dataset now and then. Another category of companies is chosen based on short reports, and the collection for these are done over a more extended period, 2009-01-01 to 2021-04-20.

Reddit activity

To collect companies with a high social media activity, Reddit is used as a baseline to identify a set of companies. Reddit can represent other similar social media platforms such as Twitter when it comes to which companies see many mentions. By collecting metadata using PushShifts API for Reddit [28] over companies listed on NYSE or NASDAQ [15] with a market capitalization above $1B$ dollar that has the most mentions, a top-list was created. More precisely, a query based on the companies stock symbol and name for the periods 2009-01-01 to 2011-04-20 and 2019-01-01 to 2021-04-20 resulted in a set of companies from each period which can be seen in Table 4.1. More details regarding the data collection from Reddit and the social media platforms will follow in the data collection section.

Short interest

The category of companies selected based on short interest is done differently from the other categories of companies for the two time periods due to different availability in short data. For 2019-2021, the daily short sale volume and total volume from the NASDAQ and NYSE exchanges were collected from FINRA [29] for all companies currently listed on these ex-

changes. The daily short volume represents all reported shares being sold short together with short sale exempt trades, whereas the total volume is the volume of all shares traded on that day [30]. The average short volume ratio was calculated for each company and selected the ten companies with the highest ratio. This selection disregarded companies with a market capitalization of less than $1B$ dollar. Moreover, we filtered away all companies with an IPO younger than 2019.

The short volume data from the Financial Industry Regulatory Authority (FINRA) was collected from a third-party API, called Quandl [31]. FINRA themselves [29] comment their data by stating that there might be some offset in the short sale volume due to the exclusion of any trading activity that is not publicly disclosed. This might reflect in the short ratio appearing to be higher than it is but should not corrupt the results significantly.

Since we found no short volume for the time interval 2009-01-01 to 2011-04-20, this selection was instead based on a list of the most shorted stocks in the S&P 500, published by CNBC [32].

Short reports

For the last category of companies, the selection was based on the occurrence of short reports released on Citron Research [7] within the period 2009-01-01 to 2021-04-20. The short reports were found through the Citron Research archive of articles [33] where all of their articles are stored and recorded. The short reports selected from the archive are based on the criteria that the mentioned company have to be listed on NYSE or NASDAQ [15], have a market capitalization of over $1B$ dollar, and have a reasonable volume of social media mentions for future work and analysis. Similarly to previous categories of companies, a total of ten companies were selected, which can be seen in Table 4.1.

Company list summary

The stock symbols for the companies chosen for each category: social media activity, short interest, and short reports can be seen in Table 4.1. For a full list of company names together with market capitalization and other useful information we refer to Appendix A.1.

Table 4.1: Company list

2009-2011		2009-2021	2019-2021	
Social media	Short interest	Short reports	Social media	Short interest
AMP	FSLR	IBOC	GME	WRLD
AIG	KODK	WRLD	TSLA	ORA
GM	GME	SHOP	AMD	IROQ
AAPL	AN	LYFT	AAPL	RCL
GE	AIG	MSI	MSFT	CACC
CIT	X	SNAP	AMZN	CABO
EBAY	CMS	TWTR	FB	CELC
UPS	-	GPRO	GNUS	JCOM
GS	-	NFLX	UPS	FDS
ADP	-	NVDA	-	WDFC

Figure 4.1 show the 2009-2011 companies where the two groups are separated by colour and shape, as can be seen. The figure displays the inter-company relations regarding social media activity and short interest as a percentage of shares float. In Figure 4.2, the same figure for the 2019-2021 companies can be seen but with the short volume ratio on the x-axis. In Figure 4.3 the relation between the short interest as a percentage of shares float to the short

volume ratio can be seen for the companies in 2019-2021. To see the exact distributions of data volumes for each company, look in Appendix A.1.

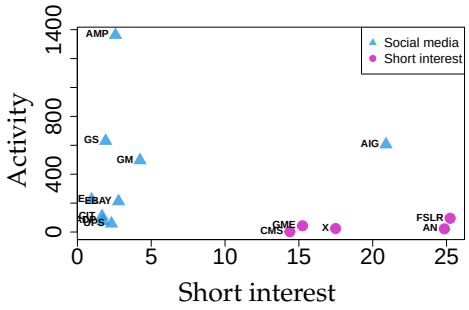


Figure 4.1: Companies collected for 2009-2011 showing total number of Reddit posts and Seeking Alpha articles as activity and short interest.

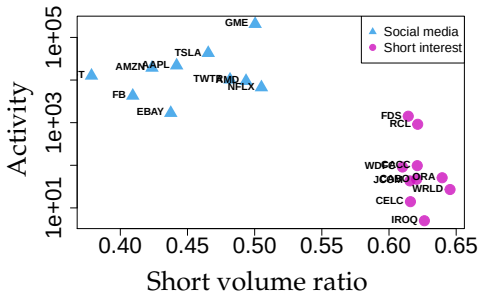


Figure 4.2: Companies collected for 2019-2021 showing total number of posts and Seeking Alpha articles as activity and short volume ratio.

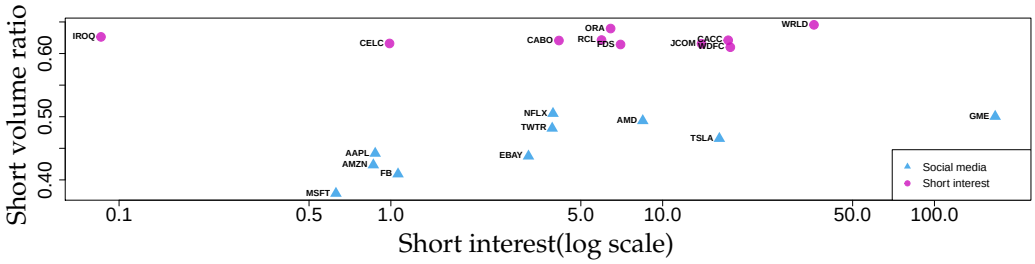


Figure 4.3: Companies collected for 2019-2021 showing short volume ratio and short interest.

4.2 Data collection

The analysis will be based on the companies from Table 4.1 and each stocks' financial data is gathered from Morningstar. The data collection focuses on social media posts from Reddit and Twitter and stock analysis articles and comments from the investment site Seeking Alpha. We mainly chose Reddit for its high popularity (19th most popular domain globally [34]) and especially for its influence surrounding events such as with GameStop and its accessible datasets [35]. The second most overlapping site to Reddit looking at the user base (calculated from analysis of familiar visitors and search keywords) is Twitter [36]. Twitter allows for the collection of large sets of historical data; it is highly popular and captures the *influencer*-side of social media. Seeking Alpha is the final social media platform used in this paper; it is also a highly popular site for investors. The articles published there capture different elements of online investment content.

Reddit

To evaluate posts and comments from Reddit, historical data was collected. In this thesis, the PushShift API is used as a foundation to retrieve data and metadata from Reddit [28]. More precisely, the API wrapper PMAW for Pushshifts API was used [37]. The API wrapper allows for collecting comments and submissions from the Pushshift database, which contains

all historical Reddit data. PMAW also allows for running requests on multiple cores to speed up the data collection process. It also allows for queries of specific keywords and phrases, for instance, a query with a company's ticker and name that returns all occurrences in either posts or comments between two specific historical dates across all or a set of subreddits. For this thesis, we used a set of subreddits related to finance, stocks, and investing was used.

Due to the queries for the data collection being based on a company's ticker and name, companies with names such as Apple have the risk of false positives in their dataset. To decrease the risk of keywords with many meanings, such as *Apple*, being associated with anything not related to finance, we used a set of relevant subreddits. The choice of subreddits is based on Reddit's search functionality on their homepage [3], where keywords can be mapped to subreddits based on relevancy and most comments. By using this functionality on Reddit we could find the subreddits most relevant or commented based on the keywords *Finance*, *Stock*, and *Invest*. Example of subreddits that appeared for these keywords are *investing*, *wallstreetbets*, and *stocks* etc. All subreddits used in this collection can be viewed in Appendix A.3. An example showing how a query using Tesla's ticker and company name and what type of data it returns can be seen below and in Table 4.2.

Query: '"TSLA | "Tesla Inc" | "Tesla"'

Table 4.2: Example of a Reddit entry in our dataset

author	id	title	selftext	comments	created	score	upvote ratio	subreddit

Twitter

Twitter provides a well-developed API for collecting tweet data from their database. For this thesis, the full-archive API endpoint of the Twitter API was used [38]. It allows for collecting any historical tweet since Twitter was created in March 2006 with many optional filters for the query. The API fetches every tweet matching the specified query and returns the tweet in a JSON format which can be converted to a suitable CSV-file format. To collect relevant tweets to a specific company, the cashtag functionality in conjunction with the ticker of each company was used. It proved to be an effective way of fetching tweets relevant to finance, stocks, and investing. Through the API, the query was also specified to only retrieve English tweets and exclude retweets to avoid any data overlaps. The API also allows for specifying which fields regarding data and metadata to retrieve from the query. An example showing how a query using Tesla's ticker and what type of data it returns can be seen below and in Table 4.3.

Query: "$TSLA" + " -is:retweet lang:en"

Table 4.3: Example of a Twitter entry in our dataset

author id	id	text	created	retweets	replies	likes	quotes

Seeking Alpha

From Seeking Alpha, a total of 7430 stock analysis articles were collected for the set of companies. This corresponds to all the articles on each company's respective page, published within the appropriate time intervals. Furthermore, the data was gathered using the Seeking Alpha

API found on Rapidapi.com [39]. More precisely, the endpoints *analysis/list* and *analysis/get-details* were used. To query a company, you provide the stock symbol and page through the results. In Table 4.4 we can see an example entry in our Seeking Alpha dataset.

Table 4.4: Example of a Seeking Alpha entry in our dataset

date	id	title	content	likesCount	commentCount	authorId	primary tickers

Morningstar

For all 41 companies, the daily open, high, low, close, and volume values together with the adjusted volume were gathered. This data was also complemented with company fundamentals such as Short Interest as a percentage of shares float, Price/Cash and Price/Sales. The 10-day volatility for each company was also included. These parameters help analyze a company's financial and economic position. In Table 4.5 we can see an example entry in our finance dataset.

Table 4.5: Example of an entry in our finance dataset

Date	Open	Close	High	Low	Volume	Adjusted Volume	SInterest	P/S	P/C	Volatility10

4.3 Longitudinal characterization of time series dynamics

The social media data collected from Reddit and Seeking Alpha, together with financial data, allows for a quantitative, data-driven, longitudinal characterization of the social media activity of the selected companies concerning their financial data. To use a data-driven approach regarding social media activity, the companies in question must have a sufficient volume of posts or articles for the social media platforms in our dataset. Therefore, some of the companies from Table 4.1 did not make the cut for the quantitative analysis. During the data collection, it was obvious which companies did not have enough data, and a visualization of this can be seen in Appendix A.1 Figures A.1, A.2, and A.3. Hence, companies with less than 200 posts for Reddit and Seeking Alpha combined will not be used for this part of the methodology. Furthermore, AMP in the 2009-2011 social media list is removed due to the lack of Seeking Alpha articles. In this section, we first describe how social media activity is defined for each of the platforms. For the remaining companies, we present a quantitative approach used to contrast them. This approach is separated into three main steps where we first *smooth* the data to capture trends in the data, then *events* are detected, and lastly, these events are segmented into *clusters*.

Defining social media activity

To compare and contrast the activity for each social media and for each company, we first need to define what we mean by activity. The goal is to generate comparable time series for each social media that can be used together with the stock price time series. For Reddit, we do so by letting the activity be defined as the number of posts mentioning a specific company each day. We neglect the attention given to a post in the form of comments and up-votes or down-votes. This is done to simplify the measure and to get a more similar process across all platforms. For Twitter, the activity is measured by the number of tweets mentioning a company each day. Similarly to Reddit, the number of retweets and likes are ignored for the same reason. Since Seeking Alpha is a different kind of social media platform, the activity is defined differently. The activity is instead defined as the attention given to the articles

published. More precisely, we say that the activity each day is defined by the number of comments on articles published on that day, plus the number of articles themselves. This is because Seeking Alpha has fewer but more qualitative posts and that an article published with zero comments would result in zero activity. However, we want the article itself to count toward the activity and therefore count it as one.

Kernel smoothing of time series

In this thesis, we are not interested in capturing the day-to-day fluctuation of time series. Instead, we aim to determine underlying trends in the data. A common way of doing so is to use variations of moving averages. They seek to create a time series by averaging over subsets of the initial time series and eliminating noise in the data. Another approach is to use kernel smoothing, which refers to the estimation of functions without relying on the assumption that the data belongs to any probability distribution. This thesis's social media and finance time series present us with the need to do scatter-plot smoothing because of the volatile nature of the data. Kernel smoothing achieves this by estimating the regression function with kernel density estimation and without introducing a parametric model [40]. Since the goal is to find essential structures in the data, it is necessary not to use a density estimator that presumes this structure.

We chose to go for kernel smoothing and, more precisely, the univariate kernel density estimator, which is the most straightforward approach. We assume our time series to be a random sample $X_1, X_2, ..., X_n$ taken from a uni-variate, continuous density f. From [40] we then have that the following formula can express the kernel density estimator

$$\hat{f}(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K\{\left(\frac{x - X_i}{h}\right)\},$$

where K is the smoothing kernel deciding how neighbouring points are weighted, satisfying $\int K(x)dx = 1$, and h is the smoothing bandwidth. When a kernel K has been chosen, h is the parameter that can be modified to effect the smoothness of the estimation.

The kernel smoothing in this study is done in R using the kernel regression smoother package, *ksmooth* [41]. The *ksmooth* function takes the kernel K as well as the bandwidth h as arguments. In this thesis, we use the default kernel, which in the package is called *normal* and refers to the Gaussian density function, also called the normal distribution. This kernel is usually chosen because it is symmetric and satisfies the $\int K(x)dx = 1$ condition. Using this kernel for estimation gives more weight to the closest neighbouring observations, whereas a rectangular kernel would estimate the average over the adjacent observations.

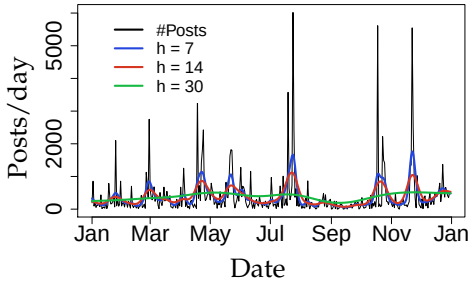


Figure 4.4: Smoothing of Tesla Reddit activity in 2019 using normal kernel for different bandwidths.

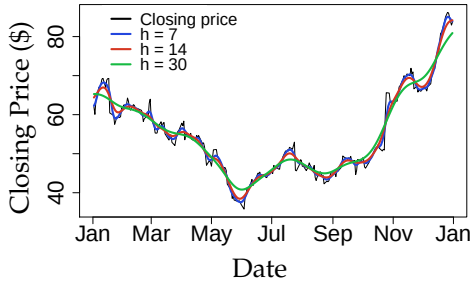


Figure 4.5: Smoothing of Tesla closing prices in 2019 using normal kernel for different bandwidths.

In Figure 4.4 the kernel regression smoothing for the Tesla Reddit activity can be seen for three different values of the bandwidth parameter h. In Figure 4.5 the same smoothing is shown for the Reddit stock price during the same period. The Reddit time series is noisier, and underlying trends are more difficult to distinguish. Already at bandwidth $h = 7$ (blue curve), the smoothing reveals trends in the Reddit data, making it more comparable to the financial time series for the same bandwidth. On the other hand, a bandwidth of $h = 30$ (green curve) reduces trends by excessively smoothing the data, which might not be desirable. As can be seen, the bandwidth has a great effect on the estimated values, and choosing a suitable bandwidth will be crucial for the next step in the method.

Defining events

Smoothing was done to capture trends in the time series and to distinguish the noise from the information we call useful in this study. Another result from the smoothing is that *events* can be easily defined as the timing of the local maxima and minima within the time series. Not only do these points define peaks and bottoms in the level of social media activity or stock price, they also represent a trend change, which makes them inviting for further analysis.

In R these local maxima and minima were retrieved from the smoothed time series using the *local.min.max()* function found in the *spatialEco* package. In Figure 4.6 and 4.7 we see an example of how the bandwidth in the previous step affects the number of events for a given period.

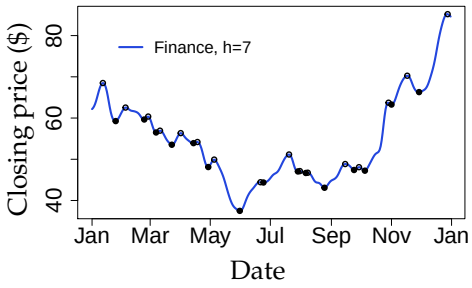


Figure 4.6: Smoothing using bandwidth 7 of Tesla closing price in 2019 with identified events.

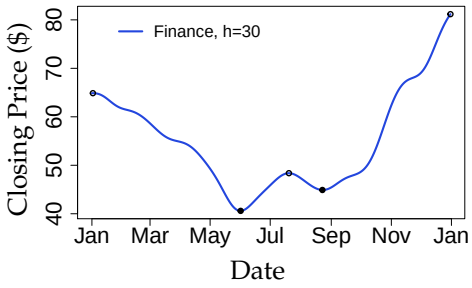


Figure 4.7: Smoothing using bandwidth 30 of Tesla closing price in 2019 with identified events.

Optimal clustering of events in 1-D data

To compare and contrast the different social media platforms and the financial data using the events, we need to identify associated events. Given a company, we extract the events from each social media time series and the company's stock price time series and add them to a mutual list. In this stage, we're only interested in the timing of the events, and therefore we represent the data as a series of one-dimensional, sorted dates. The aim is then to partition this data into k clusters such that the sum of squared distances from each cluster element to its cluster mean is minimized. This is called the k-means problem. Wang and Song [42] present an optimal solution to this problem, using dynamic programming that guarantees optimality of clustering in 1-D. Their R package *Ckmeans.1d.dp* is used to perform the clustering, which returns the cluster indexes for each data point. It is up to us to decide on an appropriate value for the parameter k. We decided to choose k to be the number of events in a company's stock price time series after smoothing and add 20 % to whatever it is. This is because we optimally would want each financial event to be in its separate cluster, together with its associated social

media events. There might be events from each time series that are not associated with any other events that the extra 20 % could isolate into separate clusters and therefore be excluded from the analysis.

For simplicity, we only analyze the clusters of size three that have precisely one financial event (F), one Seeking Alpha event (S), and one Reddit event (R). We can then determine the form, order, and timing of events within each of these clusters. This approach lets us perform a quantitative analysis on a large set of time series.

In Figures 4.8 and 4.9 the smoothed time series for Reddit, Seeking Alpha and Stock price, their events and the clustering of events are visualized for the compan Tesla (TSLA). The figures show the period 2019-01-01 to 2019-12-31 and are visualized for two different choices of the smoothing bandwidth, seven and 30. The y-axis has been min-max normalized to present the three datasets in the same graph.

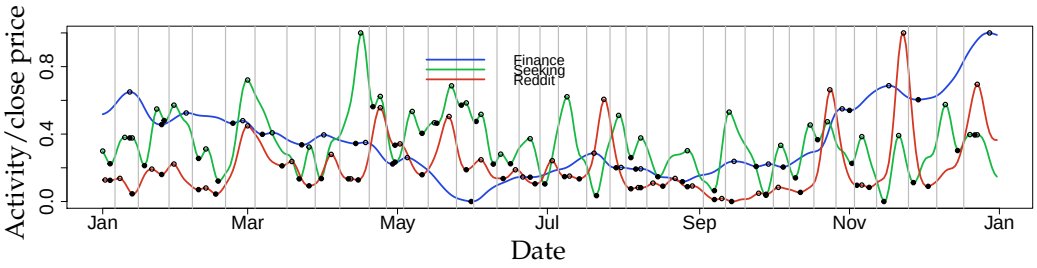


Figure 4.8: Normalized and smoothed time series (bandwidth 7) for Tesla data in 2019 where the events and clusters are annotated.

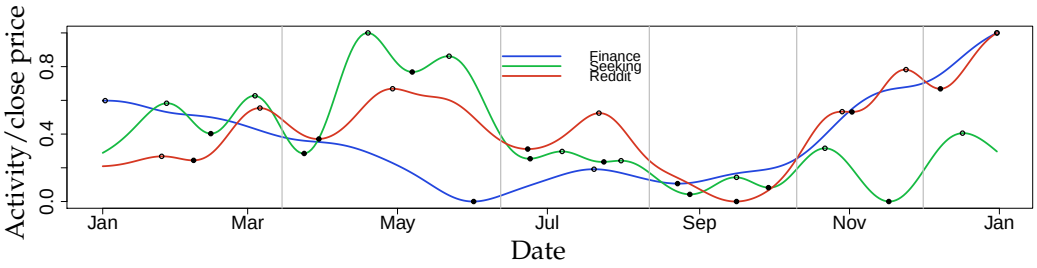


Figure 4.9: Normalized and smoothed time series (bandwidth 30) for Tesla data in 2019 where the events and clusters are annotated.

4.4 Short report method

The companies selected with short reports from Citron Research were analyzed in a slightly different way. The time series of these companies will be represented in a single graph for visualization and to manually be able to detect trends and patterns in the data. All companies from the short report category in the data collection that satisfied the activity threshold of a minimum of 200 posts or articles had figures such as Figure 4.9 made for them, but without the events and cluster lines for readability.

In comparison to the other companies, these also had activity time series from Twitter. A single company, GoPro, was chosen to be presented and analyzed more thoroughly in this thesis. The graph was annotated with the release of short reports as well as some company-specific events. The graph is also annotated with some of the most commented or liked

posts from the dataset. Additional charts will be created where certain events are given extra attention. These graphs will focus on the events from seven days before and seven days after the event, with no smoothing, and further highlight the time series dynamic.

The methodology explained in Section 4.3 was tested for the short report company as well, but with some variety. Firstly, the Twitter time series was added and the methodology performed as usual. In the second approach, the social media time series was isolated and the financial time series (F) removed. This was done to capture the inter social media dynamics without assuming the correlation to stock price events. Once again, only the clusters containing three events and exactly one Reddit event (R), one Seeking Alpha event (S), and one Twitter event (T) was accounted for. For this, the smoothing bandwidth 14 was used, but the number of clusters k was set to be the number of events detected in the smoothed Twitter time series. Then we summarized the form, order and timing of events for the two companies.

4.5 Methodology summary

In the methodology of this thesis, the first step was to identify certain companies of interest based on a specific criterion. Then, an extensive dataset for these companies was collected from multiple social media and online platforms. For the companies that met a threshold in a collected volume, a quantitative comparison of time series dynamics was done. The procedure consist of three main steps: *smoothing*, detecting *events*, and *clustering* of these events. Lastly, only the clusters with one unique occurrence of each time series event was chosen for further analysis. The statistics of the intra-cluster form, order and timing of events was then collected and summarized.

The second part of the methodology looked at one company from a different perspective. Longitudinal graphs of the smoothed time series for this company were created and annotated with specific events such as the release of a short report. Visualizing the data in this way enabled us to manually analyze trends and patterns in the data and contrast them to the results from the quantitative analysis.

5 Results

5.1 Dataset summary

The data collection was done for 41 unique companies in Table 4.1 which resulted in a dataset for each of the three time periods where some companies overlap the defined time periods. A total of 20 companies for the time period 2009-01-01 to 2011-04-20 with the categories social media mentions and short interest, 17 companies for the time period 2019-01-01 to 2021-04-20 with the same categories, and finally ten companies for the time period 2009-01-01 to 2021-04-20 with the category short reports.

The result was 334796 Reddit posts and 3391 Seeking Alpha articles for the time period 2019-01-01 to 2021-04-20. Of these, 2626 Reddit posts and 84 Seeking Alpha articles came from the group of companies with a high short volume ratio. For the 2009-01-01 to 2011-04-20 time period, a total of 2266 Reddit posts and 2717 Seeking Alpha articles were collected. Among these, 10 Reddit posts and 207 Seeking Alpha articles came from the group of companies with high short interest. In the 2019-2021 period, the Seeking Alpha articles make up approximately 1 % of the total volume, whereas in the 2009-2011 period they make up 55 % of the volume. Barplots showing the per-company volume for each time period can be seen in Appendix A.1 Figures A.1, A.2, and A.3. The companies with short reports in the period 2009-01-01 to 2021-04-20 had 54118 Reddit posts, 242845 Twitter posts(GPRO and WRLD), and 5734 Seeking Alpha articles. Across the three datasets, the companies with the most data per social media platform were GME (2019-2021) with its 206071 Reddit posts, NFLX (2009-2021) with 810 Seeking Alpha articles, and GPRO (2009-2021) with 228769 Twitter posts.

For the 2019-2021 companies chosen by activity, the average short volume ratio is 0.45, and for the companies chosen based on the short ratio, it is 0.62. The short interest as a percent of shares float for the same categories are 20.64 % and 10.97 % respectively. GameStop (GME) has the highest short interest as a percent of shares float with an average of 167.20 %, which if removed makes the average to be 4.35 % instead. For 2009-2011, only the short interest of shares float are available, and the average is 4.34 % for the companies with more social media volume, and 19.44 % for the companies from the short-list.

5.2 Preliminary results from clustering of events

This part of the result shows a preliminary attempt at identifying the social media and finance dynamics for a few metrics using the datasets from 2009-2011 and 2019-2021 and the methodology described in Section 4.3. The result for the dataset for the companies with short reports can be seen in Section 5.3. Firstly, statistics on how the method performs for different bandwidths are displayed. This is followed by three subsections that each consist of a separate metric. The first is the form of intra-cluster events, the second is the order of intra-cluster events, and the last is the timing of intra-cluster events.

Bandwidth accuracy

In Table 5.1, the total number of clusters that satisfy the criteria of containing three unique events are shown for different bandwidths for each category. The proportion that the events in these clusters make up are also shown in its separate column for the same bandwidths. The bandwidths tested were 7, 14, 30, 40, 50, and 60. The companies that were accounted for in Table 5.1 and their respective result can be seen in Appendix A.2 Table A.4. Both tables are divided into the remaining categories after removing companies that did not meet the volume criteria, excluding the short report category.

From Table 5.1 we can see that lower bandwidths generally give the most clusters and has a similar proportion to larger bandwidths. It also shows that few clusters were identified for the period 2009-2011, and the proportions are low.

Table 5.1: Total number of clusters that satisfies the criteria for each category for each bandwidth and the average proportion of events these clusters make up of the total number of events found.

-	Clusters						Proportion of events					
Bandwidth / Category	7	14	30	40	50	60	7	14	30	40	50	60
2019-2021 Social media	92	42	27	14	17	17	0.17	0.091	0.14	0.092	0.15	0.16
2019-2021 Short interest	2	3	0	2	2	2	0.0032	0.008	0.00	0.16	0.0078	0.023
2009-2011 Social media	6	3	4	3	4	6	0.0033	0.002	0.009	0.015	0.036	0.083

Form of intra-cluster events

The metric *form* of events shows the intra-cluster form of events without any consideration for their internal order. For instance, a cluster for a companies' time series with a finance event at a minimum, Reddit at a maximum, and Seeking Alpha at a minimum would result in an increment in the column *Min-max-min* for the company in question in Appendix A.2 Table A.5. In Table A.5 we can see the results for each company, while in Table 5.2 the summary of totals for each category is shown. These are the same categories as in 5.1.

Table 5.2 show that the few companies in the period 2009-2011 had very few clusters and, therefore, few events to extract the form from in comparison to the period 2019-2021. Looking at the 2019-2021 social media category, two combinations of form make up more than 50 % of the total. These are *Min-min-min* and *Max-max-max*, meaning most often, the events are either all minimum or all maximum within a cluster.

Table 5.2: Form of intra-cluster events.

Companies \ Form(F-R-S)	*Min-min-min*	*Min-min-max*	*Min-max-min*	*Min-max-max*	*Max-min-min*	*Max-min-max*	*Max-max-min*	*Max-max-max*
2019-2021 Social media	12	7	3	3	3	0	4	10
2019-2021 Short interest	0	1	0	1	0	1	0	0
2009-2011 Social media	0	0	0	1	0	0	0	2

Order of intra-cluster events

The metric *order* of intra-cluster events, unlike the previous section, takes into account the internal order of the different types of events. In Table 5.3 we can see all possible combinations of intra-cluster order of events for the set of companies. The table is similar to the previous segments table, divided into the remaining categories and time periods. The first segment being the period 2019-2021 for the category social media mentions, the second the time period 2019-2021 for the category short interest, and the third segment the time period 2009-2011 for the category social media mentions. The missing companies did not make the volume criteria for the methodology. Each part ends with a total showing the most common order of intra-cluster events.

In Table 5.3 we can observe that the time period 2009-2011 similar to in Table 5.2, had few clusters compared to the time period 2019-2021. Though we can from the first segment of Table 5.3 observe that there are two orders of intra-cluster events that stand out. The most common order for the time period 2019-2021 is a finance event followed by a Seeking Alpha event and a Reddit event, pointing to both of them being reactive. The second most common order is a Reddit event followed by a finance event and a Seeking Alpha event. The other combinations have a similar distribution between them.

Table 5.3: Order of intra-cluster events

Companies \ Order	**F-R-S**	**F-S-R**	**R-F-S**	**R-S-F**	**S-F-R**	**S-R-F**
2019-2021 Social media	5	13	10	5	4	5
2019-2021 Short interest	1	1	0	0	0	1
2009-2011 Social media	2	0	0	0	1	0

Timing of intra-cluster events

The metric *timing* is the difference (in days) from a social media event to its associated financial event within the same cluster. This is because a negative timing implies the event happened before and a positive that it happened after the financial event. The results presented here are the overall timing for Reddit and Seeking Alpha, respectively, for the companies in the 2019-2021 category selected on social media activity.

When using bandwidth $h = 7$, the Reddit event happened on average 0.054 days before the financial event with a standard deviation of 3.34 days. For Seeking Alpha and the same bandwidth, the event occurred on average 0.065 days before the financial event with a standard deviation of 3.016. Using a smoothing bandwidth of $h = 14$, the Reddit event happened on average 0.52 days after the financial event with a standard deviation of 6.79 days. While

for Seeking Alpha, the Reddit event happened 0.95 days after the financial event with a standard deviation of 5.77 days.

In Figures 5.1 and 5.2, the exact distributions of the observed timings for Reddit and Seeking Alpha can be seen for the two different bandwidths. Figures 5.3 and 5.4 shows the cumulative distribution function (CDF) for the timings observed.

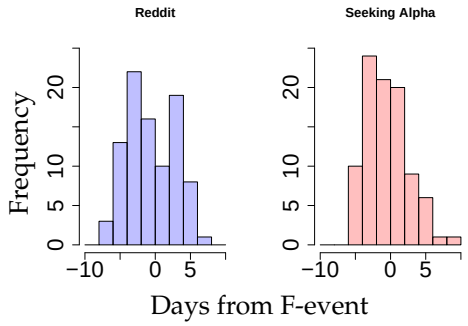


Figure 5.1: Bandwidth 7: Distribution of Reddit and Seeking Alpha events in relation to the financial event measured in days

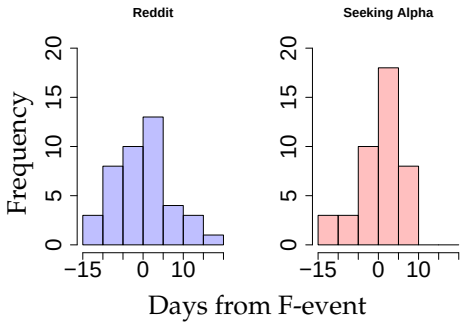


Figure 5.2: Bandwidth 14: Distribution of Reddit and Seeking Alpha events in relation to the financial event measured in days.

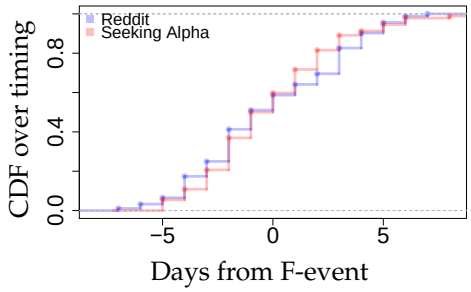


Figure 5.3: Bandwidth 7: Cumulative density function

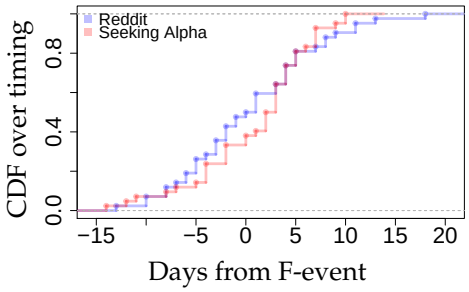


Figure 5.4: Bandwidth 14: Cumulative density function

5.3 Short reports

For the data collection, this paper identified three categories of companies, one of these categories being companies that have short reports written about them on Citron Research in the period 2009-01-01 to 2021-04-20, which can be seen in Table 4.1. This part of the results in addition to Reddit, Seeking Alpha, and Finance close price, also add Twitter activity and Citron Research short reports. To illustrate where in time short reports appear in relation to social media, financial, and external events, they can be annotated in the corresponding companies dataset plot.

GoPro time-series events

The company GoPro Inc. ($GPRO) is one of the identified companies in the period 2009-01-01 to 2021-04-20 that has had short reports written about them, as can be seen in Table 4.1. This part shows a time series for GoPro Inc. with annotated events since their IPO in 2014-06-27.

In Figure 5.5 we can see some events annotated in the figure in proximity to spikes in the social media activity and financial data. The events annotated in the figure are described in more detail below and cover popular posts on the social media platforms identified through our dataset, the Citron Research GoPro short report, and significant company actions.

- A few days after the stock market launch of GoPro, a Tweet on Twitter with 916 retweets states, "Wow trading $gpro what a rush ha ha".
- On the 4th of November, not long after the stock market launch of GoPro and when the stock price was at an all-time high of close to $87, Citron Research released a short report stating that GoPro would drop to $30 within 12 months.
- On the 19th of November 2015, the most commented Seeking Alpha post in the companies time series states, "GoPro needs to change its business model".
- On the 13th of January 2016 plans to cut 7 % of the workforce in response to poor fourth-quarter results was announced.
- On the 13th of November, plans to cut another 200 workers together with the GoPro president Tony Bates stepping down was announced.
- On the 28th of September 2018, the GoPro camera product HERO6 Black was released.
- On the 8th of January 2018, plans to cut 20 % of the workforce and their exit of the drone industry was announced.
- On the 31st of May 2019, the Reddit post with the highest score discussing GoPro was published.

Generally, we can see in Figure 5.5 that the three social media platforms follow similar trends in activity for the time series.

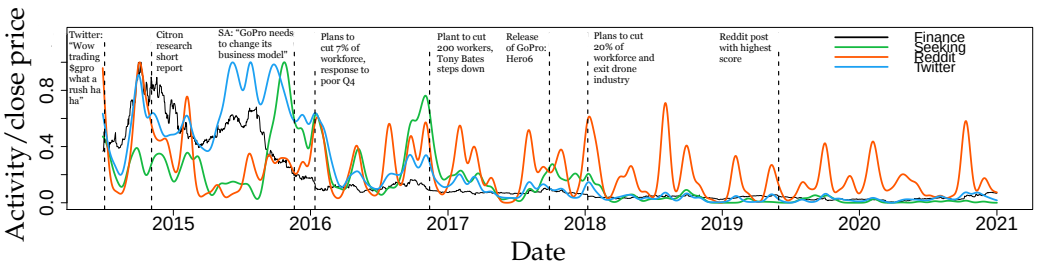


Figure 5.5: GPRO time series with annotated external events and social media time series with bandwidth 30 and unsmoothend stock price.

The social media activity and stock value behaviour of GoPro surrounding a selection of the above events can be more closely viewed in Figures 5.6 and 5.7. In Figure 5.6 we can see the non-smoothened graphs for the week prior to and after the Citron Research short report publication, which is dotted in the figure. In Figure 5.7 we can see the two adjacent events at the turn of the year 2016, the first dotted line is the day that the most commented Seeking Alpha article discussing GoPro was released, and the second dotted line is GoPro announcing their plans to cut 7 % of their workforce after poor fourth-quarter results.

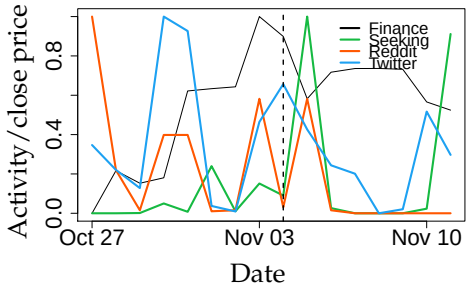


Figure 5.6: Citron research short report article publication isolated in the GPRO time series.

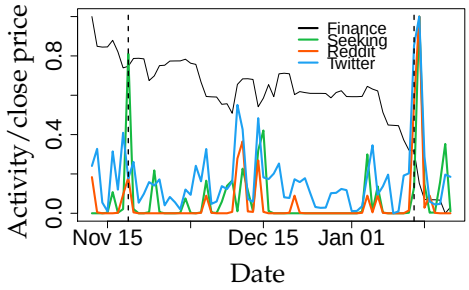


Figure 5.7: Seeking Alpha most commented post and workforce cut events isolated in the GPRO time series.

Applying the main methodology for form and order for the GoPro time series in Figure 5.5. The results for the time series can be seen in Table 5.4 for which we note that Twitter (T) is now added.

Table 5.4: Form and order of GPRO intra-cluster events for F-R-S-T with bandwidth 14. Excluding form and order combinations with zero results.

Min-min-min-min	Min-max-min-max	Min-max-max-max	Max-min-min-min	Max-max-max-max	F-R-T-S	F-S-R-T	F-T-R-S	R-T-S-F	S-R-F-T	S-R-T-F	T-R-S-F
2	1	1	1	8	2	1	1	5	1	2	1

Applying the primary methodology for form and order for the GoPro time series in Figure 5.5 excluding finance close data. The results for the time series can be seen in Table 5.5.

Table 5.5: Form and order of GPRO intra-cluster events for R-S-T with bandwidth 14.

Min-min-min	Min-min-max	Min-max-min	Min-max-max	Max-min-min	Max-min-max	Max-max-min	Max-max-max	R-S-T	R-T-S	S-R-T	S-T-R	T-R-S	T-S-R
16	0	4	2	0	3	1	31	5	26	11	2	8	5

6 Discussion

6.1 Result

In this section, we discuss the resulting dataset as well as the results from Sections 5.2 and 5.3. We highlight and criticize interesting observations when necessary, all related to the research questions from section 1.3.

Dataset

When comparing the dataset volumes for the two time periods, 2009-2011 and 2019-2021, it is clear that social media is more widely used to discuss stock-related topics today than it was ten years ago. The natural differences in the Seeking Alpha and Reddit data characteristics make it useless to compare their volumes for a single period. However, when comparing how these volumes differ from 2009-2011 to 2019-2021, it is noticeable how Seeking Alpha articles constituted a significantly larger share of the volume in 2009-2011 than they did in 2019-2021. This likely has multiple explanations, but one could be that more non-experts regarding investments comment on stocks and partake in online discussion today and does so on social media such as Reddit instead of Seeking Alpha. As more people seek advice from social media, the effects of events such as that with GameStop [12] or with the false promotion of stock as in [11] will be even more extensive, especially since misinformation tend to spread more widely [10].

Looking at data volumes for the companies in the 2019-2021 time period with lots of social media activity, they are all companies within technology or services that heavily rely on technology and among these, the top two are GameStop ($GME) and Tesla ($TSLA) which can be seen both in Figure 4.2 and in Appendix A.1 Figure A.2. In Figure 4.3, GameStop and Tesla also have the highest short interest as a percentage of shares float. Thus, it seems a company should be in the tech business and highly shorted to trend on Reddit. For the 2009-2011 companies, a similar observation can be made. The second most talked-about company on Reddit seen in Appendix A.1 Figure A.1 is the American International Group ($AIG), it is also the most highly shorted by far as can be seen in Figure 4.1. When it comes to Seeking Alpha articles in 2009-2011, Apple is the company with the most articles among all the companies but is the second least shorted. In 2019-2021, Apple remains in the top three in the number of articles, still with one of the lowest short interests. However, the first place for the number

of articles among these companies is now taken by Tesla, which on the other hand is one of the most shorted of the companies with high data volume. This may imply that investment sites, in general, write articles about "stable" companies with low short interest but also have started giving highly shorted companies attention, just like Reddit. It could also be that Tesla is just an exception and that no such correlation exists.

In general, one can learn a lot about the discussions regarding these companies on various social media platforms, as can be seen in Tables 4.2, 4.3, and 4.4. Looking beyond the frequency of posts and articles, different measures of sentiment can be extracted, such as likes, comments and retweets. These all play central roles in social media interaction and can be used in data analysis to understand individual posts' impact better. This is especially true since emotion seems to have an important role in human decision-making [1]. Still, this was not done within the scope of this thesis but is proposed as future work.

Preliminary analysis

Research questions two and three from Section 1.3 ask whether social media is proactive or reactive to what we in this report call financial events and if we can say something about how associated events behave surrounding one of these events.

The first that can be said about the methodology described in Section 4.3 is that it streets ahead for companies with sufficient, longitudinal social media volumes. With that said, given the parameters used for the bandwidth and number of clusters, Table 5.2 suggests that there is some correlation between the events within the identified clusters and that they most often coincide in all being locally maximum or all local minimum. It can also be derived that Reddit and Seeking alpha both take the same form in approximately 67 % of the cases, at least for the 2019-2021 social media companies. On the other hand, the least common combination is that the stock price reaches a local maximum together with Seeking Alpha, while Reddit is at a local minimum. What this says is difficult to interpret. More easily interpreted is the combination with all local minimum within a cluster. This suggests that discussion starts building up on both social media platforms in reaction to a price trend change, or the opposite, that the price moves toward a positive trend as a reaction to the discussion taking form on social media. To evaluate this better, one would need to know in what order these events occurred and potentially use tools for sentiment during the trend change in activity.

In Table 5.3 the first thing to notice is that two combinations (F-S-R and R-F-S) stand out while the others have an almost equal distribution, at least for the category of companies with the most social media mentions 2019-2021. Looking closer at these, F-S-R is at the top with 13 occurrences which suggest that in more than 30 % of the clusters, both the events for Seeking Alpha and Reddit happen after the financial event. This goes together with [20], where it is showed that social media activity intensity could be used to evaluate external events. This implies that social media reacts to the external event and that the intensity builds up, where the external event in our case would be a top or dip in stock price. However, this is contrary to what is said in [22], where it is showed that online chatter volumes can be used as an early indicator of an external event. For us, this would mean that social media events happen before the external financial event. There is no distinct trend confirming this either. However, the second most common combination is (R-F-S) with ten occurrences (24 %) where the Reddit event happens before the financial event suggesting that it is proactive while Seeking Alpha is reactive. But in total, Reddit is proactive in less than (50 %) of the clusters, which tells us that no apparent trend exists. However, the number of observations is few, making it difficult to confirm or deny such trends.

In Figures 5.1 and 5.2 the distribution of when the social media events happened in relation to the financial event can be seen. One could say that this implies that both Reddit and Seeking Alpha events, in general, occur after the financial events. Another observation is that the choice of bandwidth has a significant effect on the result. This may be obvious since more smoothing results in fewer events in each time series, but since the time standard deviation

for the distribution grows more prominent as the bandwidth is increased, this could mean that un-associated events are "forced" into the clusters. A better result would be if the distribution were similar for both bandwidths. This could mean that we immediately identify the "true" events, cluster them together, and that these largely stay the same as we further smooth the data. The fact that the means and standard deviations are similar for both Reddit and Seeking Alpha also raises some warning flags, especially since they seem to follow each other when the bandwidth increases. This can be more clearly seen in Figures 5.3 and 5.4, showing the cumulative density function over timing for two bandwidths. Nonetheless, the choice of these parameters both resulted in Reddit and Seeking Alpha being reactive to financial events, with less than a day on average.

Short reports

The most longitudinal analysis of the social media time series and stock price data was done partly manually. In Figure 5.5 the complete graph for the GoPro time series is displayed, which is the company chosen from the list of companies with short reports. We first notice that that the release of Citron Research's short report happened on November 4th, 2014. The report predicted that the GoPro stock price would go from the then-current 82$, down to 30$ within 12 months. Looking at the stock price on November 4th, 2015, exactly a year after, it is traded at 25$. In this example, Citron was right. We look closer at what happened on the way down. Firstly, in Figure 5.6, a snapshot of the price and social media dynamics, un-smoothened, can be seen a week prior and a week after the release of the report. It shows that all social media platforms in the prior week saw a local maximum in activity, perhaps in response to the stock price working its way up. Looking at these in Figure 5.5 they seem to correlate to the all-time high in stock price. On the day of the release, it is clear that Twitter responded since it reached a maximum. However, the activity took form already the day before the release day. There is some lag for Reddit and Seeking Alpha, but they both catch up on activity the next day, probably in response to the report or that the report potentially triggered a price drop which can be seen as well. In the following months, the company saw a significant decrease in valuation. However, it is difficult to say if this was due to the report's release or that the report pinpointed specific faults in the business model that would draw the company down eventually.

Following the time series in Figure 5.5 as it passes by the annotated events makes it clear that certain events catch all the social media attention. The most commented article for GoPro on Seeking Alpha discusses its business model, following a far-reaching drop in stock price. Prior to this, both Twitter and Seeking Alpha saw an all-time high as the price dropped in the second half of 2015. On the other hand, Reddit had its all-time high in late 2014 and coincides with one of Twitters more significant peaks, together with the rise in stock price. These observations point to social media activity taking form either as the price climbs or falls fast. Another important observation comes from the fact that the external events annotated in the graph all seem to coincide with increased social media activity. This is clearly displayed in Figure 5.7, where the two events near the turn of the year 2016 are zoomed in on and presented un-smoothened but normalized. Both of these events trigger distinct peaks in activity.

The last thing done with the GoPro data was applying the clustering method to it in two ways. Firstly, since there was now Twitter data available for the company, we performed the analysis the same way as before, only adding a fourth time series for Twitter which resulted in Table 5.4. In the table, it is once again clear that the most common form of events are either all being maximum or all being minimum. The analysis also shows how the combination R-T-S-F is the most common for order, which suggests that most social media are often proactive to the financial event. In Figure 5.5, this could be when all social media peak in activity reacting to the company releasing news, and the stock price taking longer to reach its minimum in response to the news. Secondly, the clustering was applied to only Reddit, Seeking Alpha

and Twitter, disregarding the stock price time series. The result for this can be seen in Table 5.5. Using the same bandwidth, the number of clusters found that satisfied the criteria more than quadrupled. The volume of clusters also exposed more distinct trends. This could mean that the social media is internally much more correlated than they are to events in the stock price data or that the stock price events see some delay due to infrequent updates, which unables them from being included in the "correct" cluster. Looking closer at the results, with conviction, the most frequent combination of form observed is *Max-max-max* and *Min-min-min*. Among the 56 identified clusters, almost half had the order *R-T-S*. In 64 % of the clusters, the Reddit event happens before the Twitter event. The second most common form is *S-R-T*, which shows the same internal order between Reddit and Twitter but now with Seeking Alpha taking the lead. The overall results from this analysis coincide more with what can be visually seen in Figure 5.5, indicating that the method can be used to single out noteworthy events in time series data quantitatively. These could then be analyzed more closely with more qualitative measures to learn more about stock events' social media dynamics.

6.2 Method

In this section, we discuss the methodology sections 4.1, 4.2, 4.3, and 4.4. We highlight weaknesses and convey what steps would benefit from improvements.

Dataset

The risk of false positives in the data collection varies between the three social media platforms due to their respective structures and APIs for the data collection. Twitter's API allows special characters such as the cashtag ($) in its queries, while PushShifts API for Reddit does not. This essentially means that the likelihood of the Twitter API correctly identifying a ticker is greater than for Reddit's API. Hence, Reddits API poses the risk of wrongly identifying words as tickers or company names. To increase the quality of the Reddit data collection, a selection of finance-related subreddits were specified, as can be seen in Appendix A.3 Table A.7, which resulted in a cleaner dataset except with a lower data volume. The same subreddits, dates, and queries would have to be used for reliable data collection to replicate the dataset. On the other hand, Seeking Alpha has a structure specifically focused on finance and an API that allows for proper data queries, limiting possible non-related articles to be retrieved with the API.

One of the categories in Table 4.1 was based on top social media mentions for the periods 2009-01-01 to 2011-04-20 and 2019-01-01 to 2021-04-20. Due to rate limits for Seeking Alpha, the category had to be based on the most-mentioned companies on Reddit, which essentially implies that Reddit would be representative across the two social media platforms used in the category. For the period 2019-2021, we saw that Reddit made up for 99 % of the data volume and hence, Seeking Alpha only 1 %, which could either indicate that the most-mentioned companies on the two platforms do not overlap or that this is the relative difference in volume between the two platforms. For the period 2009-2011, the relative data volume was instead that Seeking Alpha stood for 55 % and would perhaps, if the rate limits allowed for it, be better suited as a source for identifying the most-mentioned companies on the social media platforms for the period.

Another of the categories in Table 4.1 is short interest. The short volume ratio proved to be a less optimal category for choosing companies in the data collection and analysis. Although shorted companies are attractive for this thesis to examine, most of the identified companies had very low data volumes for the time periods. Essentially, making some of the companies impossible to analyze using the methodology and include in the results. However, as mentioned in Section 6.1 the most shorted company was also the most mentioned company for the time period 2019-2021, which fits the purpose of this thesis very well.

Preliminary analysis

The preliminary analysis with clusters and events quickly turned out that the companies with larger volumes of data for the time periods provided more reliable results. For instance, some promising companies with high short interest and many short reports such as $WRLD had weeks and sometimes months with zero posts or articles found through the data collection, which can be seen in Appendix A.2 Figure A.4. Essentially, the results depend heavily on which company the methodology is applied on in terms of the quantity of clusters matching the criteria for selection. Due to the varying number of data points for each company, the best fitting choice of bandwidth also varied. For companies with more significant fluctuations in their social media activity graphs, usually those with lower data volumes, a higher bandwidth would be prefered to smoothen the spikes in the data, and for companies with high data volumes, instead, a lower bandwidth works. This also made the data more easily processed. However, this posed a risk for introducing bias into the data. Further testing on how the bandwidth parameter affects the steps in the methodology would be of interest. One could also contrast and compare how the kernel density estimation performs compared to other means of smoothing. This is an important step in the method because it affects the number of events given for a specific time series. However, it does not necessarily skew the timing of the events, it only eliminates those events that are less characteristic.

In addition to the bandwidth, the choice of number of clusters is another parameter that varied greatly depending on the data volumes. Using a method where one has to provide the number of clusters beforehand also poses a risk of introducing bias. Even though the method used for this finds the optimal clusters given a particular value, it does not necessarily mean that intra-cluster events are associated. Therefore, it is important to perform the analysis on a large number of clusters for social media to find *hidden* patterns and to make the substandard clusters matter less. As of now, the number of clusters chosen are based on intuition. It would therefore be useful to find a metric to use for the evaluation of this choice. Another aspect of clustering, especially the selection of clusters, is that some events are disregarded entirely. This was done to more easily be able to compare between the different social media and to be able to automate the process of evaluating frequencies of trends where all types of events were present. However, this could also indicate flaws in the validity of the result since the choice of clusters essentially becomes weighted towards those with one of each event. Of course the number of clusters also play an important role in capturing the cluster of events that are indeed associated, which again highlights its importance to the method. Choosing this number properly could possibly reduce the risk of ignoring events that should belong to a specific cluster.

The two parameters, bandwidth and number of clusters, are somewhat co-dependent. Picking a low bandwidth results in more maximum and minimum events being identified, increasing the total number of clusters. Essentially, choosing one value for the bandwidth and number of clusters for all companies is not optimal and proved to be challenging to determine. Therefore a bandwidth of 14 and a cluster number equal to a factor 1.2 of the number of finance events were used to produce the results in this thesis as stated in Section 4.3, which is paramount to replicate the results. How to choose these also depend on what kind of trends one wishes to analyse within the data. More smoothing implies that more general trends are analysed while little smoothing makes the data more volatile which could be desired to capture day to day trends etc.

Short reports

In Section 5.3 covering the short report from Citron Research and a selection of company actions and events for GoPro, the same flaw with choice of bandwidth and number of clusters apply to the part using the main methodology with clustering. However, the visual results provided by the annotated Figures 5.5, 5.6, and 5.7 capture high-level trends in social media

dynamics in correlation to financial events but no extensive conclusions can be drawn from visually analyzing a single companies' time series and leaves much for the observer to reflect on. However, adding the Twitter time series seemed to give promising results as its volume was significant.

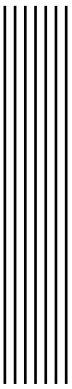
6.3 The work in a wider context

Even if social media do not widely manipulate the stock market, the effects of the forum discussions may still linger for its participants. Private investors may be influenced into taking uneducated investment actions resulting in economic loss. With the increase in data volumes over the past decade for the social media platforms studied in this thesis, the means for stock manipulation are becoming increasingly significant and more far-reaching in its effects. Manipulation and other impactful events to the stock market as illustrated in [11], [5], [13], [14], and [7] has the risk of becoming more frequent if the trend of increasing prominence continues with social media platforms.

7 Conclusion

Using a comprehensive dataset of stock-related content collected from various social media platforms, this thesis analyzed the social media dynamics on identified financial events quantitatively and manually. We found that the dataset allowed for an understanding of how company-specific discussion evolved and propagated between different social media over time. It is too early to conclude if social media are proactive or reactive to the stock market. However, the preliminary methodology showed indications of trends and correlations between the time series of social media and stock price. Most prominent being that discussion intensity coincided more frequently when excluding the stock price time series in the scope of the methodology. In general, the dynamics such as form, order and timing can be quantitatively evaluated using the methodology in this thesis, but the challenge of validating the result remains.

Our future efforts in the research would be performing the clustering on a more extensive set of companies with additional Twitter data focusing solely on the social media time series to identify and validate correlations. By identifying and extracting significant events, sentimental indicators such as comments, likes, and retweets can be used to go beyond the work in this thesis and dwell deeper into the actual characteristics of the discussion. Furthermore, the use of machine learning and natural language processing opens up for answering how social media behaves and what they discuss. This could finally be used to identify manipulation and evaluate the long-term effects on the stock market.



Bibliography

[1] J. R. Nofsinger, "Social mood and financial economics," en, *J. Behav. Financ.*, vol. 6, no. 3, pp. 144–160, 2005.

[2] *Reddit - statistics & facts*, Accessed: 2021-3-18. [Online]. Available: https://www.statista.com/topics/5672/reddit/.

[3] *Reddit.com*, Accessed: 2021-3-18. [Online]. Available: http://reddit.com.

[4] *Twitter: Number of users worldwide 2019-2020*, Accessed: 2021-4-30. [Online]. Available: http://www.statista.com/statistics/303681/twitter-users-worldwide/.

[5] *Sec.gov*, Accessed: 2021-3-16. [Online]. Available: https://www.sec.gov/files/Market%5C%20Manipulations%5C%20and%5C%20Case%5C%20Studies.pdf.

[6] *About seeking alpha*, Accessed: 2021-3-18. [Online]. Available: https://seekingalpha.com/page/about_us.

[7] *Citron research - andrew left*, en, https://citronresearch.com/who-is-citron-research/, Accessed: 2021-5-14, Dec. 2015.

[8] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," en, *J. Econ. Perspect.*, vol. 31, no. 2, pp. 211–236, 2017.

[9] A. Bovet and H. A. Makse, "Influence of fake news in twitter during the 2016 US presidential election," en, *Nat. Commun.*, vol. 10, no. 1, p. 7, 2019.

[10] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," en, *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[11] *SEC.gov*, Accessed: 2021-3-16, Apr. 2017. [Online]. Available: https://www.sec.gov/news/press-release/2017-79.

[12] B. Brumberg and JD, "Reddit and GameStop lessons: Former SEC enforcement chief explains stock manipulation and how to avoid trouble," *Forbes Magazine*, Feb. 2021.

[13] C. Thompson, *Twitter trading: 8 tweets that moved markets*, Accessed: 2021-5-14, Apr. 2013. [Online]. Available: https://www.cnbc.com/2013/04/25/Twitter-Trading:-8-Tweets-That-Moved-Markets.html.

[14] S. Shead, *Elon musk's tweets are moving markets — and some investors are worried*, Accessed: 2021-5-14, Jan. 2021. [Online]. Available: https://www.cnbc.com/2021/01/29/elon-musks-tweets-are-moving-markets.html.

[15] *Nasdaq.com*, Accessed: 2021-5-18. [Online]. Available: https://www.nasdaq.com/market-activity/stocks/screener.

[16] R. W. Hafer and S. E. Hein, *The Stock Market*. Westport, CT: Greenwood Press, 2006.

[17] *Morningstar*, Accessed: 2021-4-30. [Online]. Available: https://www.morningstar.com/.

[18] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts, "Who says what to whom on twitter," in *Proceedings of the 20th international conference on World wide web - WWW '11*, New York, New York, USA: ACM Press, 2011.

[19] M. Glenski, E. Saldanha, and S. Volkova, "Characterizing speed and scale of cryptocurrency discussion spread on reddit," in *The World Wide Web Conference*, New York, NY, USA: ACM, 2019.

[20] Y. Kryvasheyeu, H. Chen, N. Obradovich, E. Moro, P. Van Hentenryck, J. Fowler, and M. Cebrian, "Rapid assessment of disaster damage using social media activity," en, *Sci. Adv.*, vol. 2, no. 3, e1500779, 2016.

[21] S. Asur and B. A. Huberman, "Predicting the future with social media," in *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, vol. 1, IEEE, 2010, pp. 492–499.

[22] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins, "The predictive power of online chatter," in *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining - KDD '05*, New York, New York, USA: ACM Press, 2005.

[23] E. F. Fama, "Efficient capital markets: II," en, *J. Finance*, vol. 46, no. 5, pp. 1575–1617, 1991.

[24] ——, "The behavior of stock-market prices," *J. Bus.*, vol. 38, no. 1, p. 34, 1965.

[25] B. G. Malkiel, "The efficient market hypothesis and its critics," *J. Econ. Perspect.*, vol. 17, no. 1, pp. 59–82, 2003.

[26] H. Chen, P. De, Y. (Hu, and B.-H. Hwang, "Wisdom of crowds: The value of stock opinions transmitted through social media," en, *Rev. Financ. Stud.*, vol. 27, no. 5, pp. 1367–1403, 2014.

[27] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," en, *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, 2011.

[28] *API documentation - pushshift.io*, en, Accessed: 2021-4-30, May 2018. [Online]. Available: https://pushshift.io/api-parameters/.

[29] *Short sale volume daily*, https://www.finra.org/finra-data/short-sale-volume-daily, Accessed: 2021-5-18.

[30] *Finra.org*, Accessed: 2021-5-26. [Online]. Available: https://www.finra.org/sites/default/files/2020-12/short-sale-volume-user-guide.pdf.

[31] *Quandl*, Accessed: 2021-5-18. [Online]. Available: https://www.quandl.com/.

[32] *International: Top news and analysis*, Accessed: 2021-4-30, Sep. 2016. [Online]. Available: https://www.cnbc.com/world/?region=world.

[33] Citron Research, *Citron reports archives - citron research*, Accessed: 2021-5-18. [Online]. Available: https://citronresearch.com/category/citron_reports/.

[34] *Alexa - top sites*, Accessed: 2021-3-18. [Online]. Available: http://www.alexa.com/topsites.

[35] K. Grant, "GameStop: What is it and why is it trending?" *BBC*, Jan. 2021.

[36] *Reddit.Com competitive analysis, marketing mix and traffic*, Accessed: 2021-3-18. [Online]. Available: https://www.alexa.com/siteinfo/reddit.com.

[37] *Pmaw*, Accessed: 2021-4-30. [Online]. Available: https://pypi.org/project/pmaw/.

[38] *Search api*, Accessed: 2021-5-27. [Online]. Available: https://developer.twitter.com/en/docs/twitter-api/premium/search-api/overview.

[39] *Seeking alpha API documentation free with API key & SDK*, Accessed: 2021-5-14. [Online]. Available: https://rapidapi.com/apidojo/api/seeking-alpha.

[40] M. P. Wand and M. C. Jones, *Kernel Smoothing*, en. Philadelphia, PA: Chapman & Hall/CRC, 1995.

[41] *Ksmooth function - RDocumentation*, Accessed: 2021-5-13. [Online]. Available: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/ksmooth.

[42] H. Wang and M. Song, "Ckmeans.1d.Dp: Optimal k-means clustering in one dimension by dynamic programming," en, *R J.*, vol. 3, no. 2, pp. 29–33, 2011.

A Appendix

A.1 Companies

Table A.1: Companies 2019-2021

Ticker	SInterest	AvgShort	Name	Market Cap	Sector	SA	Reddit
AAPL	0.88	0.44	Apple Inc.	2189185921800	Technology	717	21203
AMD	8.45	0.49	Advanced Micro Devices Inc.	90701315858	Technology	291	9187
AMZN	0.86	0.42	Amazon.com Inc.	1649335302977	Consumer Services	341	19025
CABO	4.15	0.62	Cable One Inc.	10347296906	Consumer Services	0	47
CACC	17.42	0.62	Credit Acceptance Corporation	7227768267	Finance	12	86
CELC	0.99	0.61	Celcuity Inc.	336790595	Health Care	0	14
EBAY	3.21	0.44	eBay Inc.	42027030325	Consumer Services	55	1629
FB	1.06	0.41	Facebook Inc. Class A	894475736534	Technology	322	3953
FDS	7.00	0.61	FactSet Research Systems Inc	12359869723	Technology	14	1404
GME	167.30	0.50	GameStop Corporation	12781383107	Consumer Services	146	206071
IROQ	0.09	0.62	IF Bancorp Inc	71155417	Finance	0	5
JCOM	13.91	0.61	j2 Global Inc.	5647596068	Technology	9	34
MSFT	0.63	0.38	Microsoft Corporation	1846591448414	Technology	252	12378
NFLX	3.95	0.51	Netflix Inc.	216797333740	Consumer Services	273	6448
ORA	6.44	0.64	Ormat Technologies Inc.	3658020173	Public Utilities	2	49
RCL	5.96	0.62	D/B/A Royal Caribbean Cruises Ltd.	21521403343		28	888
TSLA	16.17	0.47	Tesla Inc.	555677902320	Capital Goods	810	42180
TWTR	3.92	0.48	Twitter Inc.	41981460791	Technology	109	10096
WDFC	17.74	0.61	WD-40 Company	3331809847	Basic Industries	18	73
WRLD	36.03	0.65	World Acceptance Corporation	1029548235	Finance	1	26

Table A.2: Companies 2009-2011

Ticker	SInterest	Name	Market Cap	Sector	SA	Reddit
AAPL		Apple Inc.	2189185921800	Technology	863	129
AIG	20.90	American International Group Inc. New	44428882102	Finance	280	326
AMP	2.58	Ameriprise Financial Inc.	29804939635	Finance	4	1359
AN	24.85	AutoNation Inc.	8480961429	Consumer Services	22	0
CIT	1.67	CIT Group	5320849000	Finance	68	41
CMS	14.39	CMS Energy Corporation	18230163089	Public Utilities	0	2
EBAY	2.79	eBay Inc.	42027030325	Consumer Services	173	39
FSLR	25.25	First Solar Inc.	7606611365	Technology	88	7
GE	0.96	General Electric	115439100000	Consumer Durables	182	42
GM	4.24	General Motors	81295600000	Capital Goods	285	211
GME	15.25	GameStop Corporation	12781383107	Consumer Services	42	1
GS	1.91	Goldman Sachs Group Inc.	125457190135	Finance	580	50
KODK		Eastman Kodak Company	539318880	Miscellaneous	31	0
UPS	2.31	United Parcel Service	188287500000	Transportation	24	34
X	17.49	United States Steel Corporation	7108272659	Basic Industries	24	0

Table A.3: Companies 2009-2021

Ticker	SInterest	Name	Market Cap	Sector	SA	Reddit	Twitter
GPRO	19.68	GoPro Inc. Class A	1516685381.0	Miscellaneous	437	2216	228769
IBOC	1.66	International Bancshares Corporation	3164184117.0	Finance	9	28	
LYFT	10.71	Lyft Inc. Class A	16371895642.0	Technology	89	3225	
MSI	2.23	Motorola Solutions Inc.	34043820995.0	Technology	52	181	
NFLX	3.95	Netflix Inc.	216797333740.0	Consumer Services	2331	15228	
NVDA	1.71	NVIDIA Corporation	352655003931.0	Technology	1038	8927	
SHOP	3.43	Shopify Inc. Class A	136372474720.0		241	2815	
SNAP	7.92	Snap Inc. Class A	82131389405.0	Technology	457	6600	
TWTR	3.92	Twitter Inc.	41981460791.0	Technology	1048	14843	
WRLD	36.03	World Acceptance Corporation	1029548235.0	Finance	32	55	14066

Bar plots for total posts and articles on the social media platforms

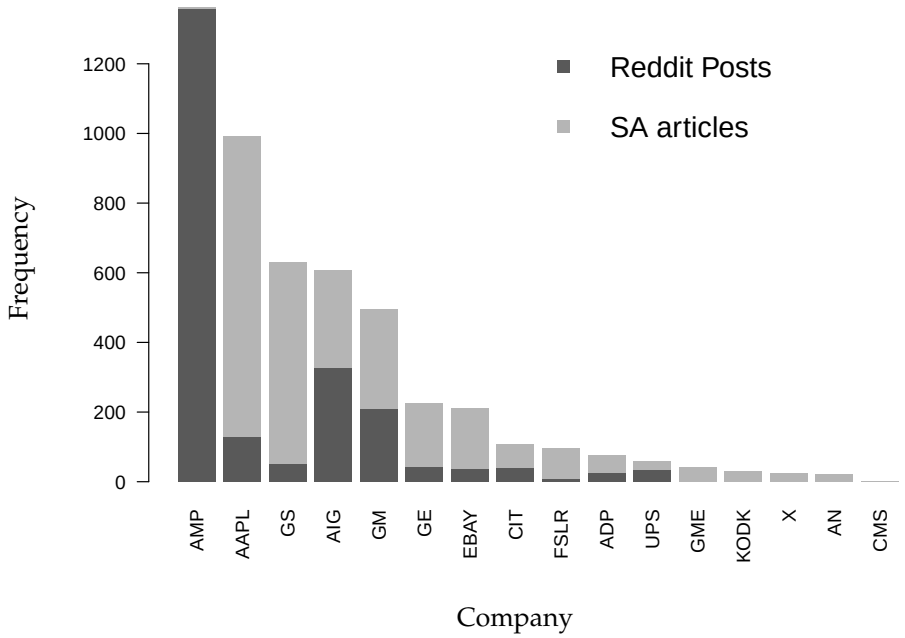


Figure A.1: Total number of Reddit posts and Seeking Alpha articles for all companies in the category social media mentions and short interest, time period 2009-01-01 to 2011-04-20.

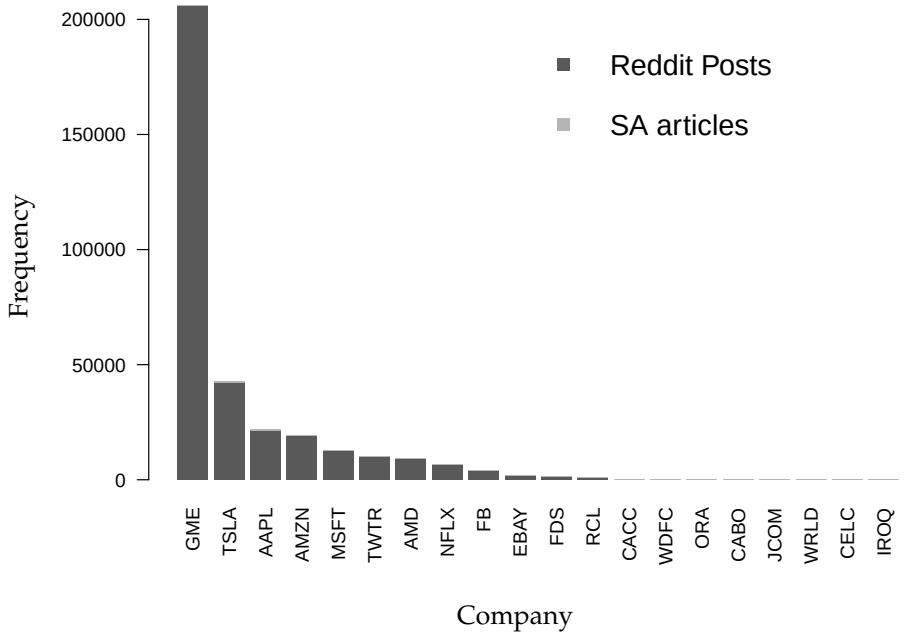


Figure A.2: Total number of Reddit posts and Seeking Alpha articles for all companies in the category social media mentions and short interest, time period 2019-01-01 to 2021-04-20.

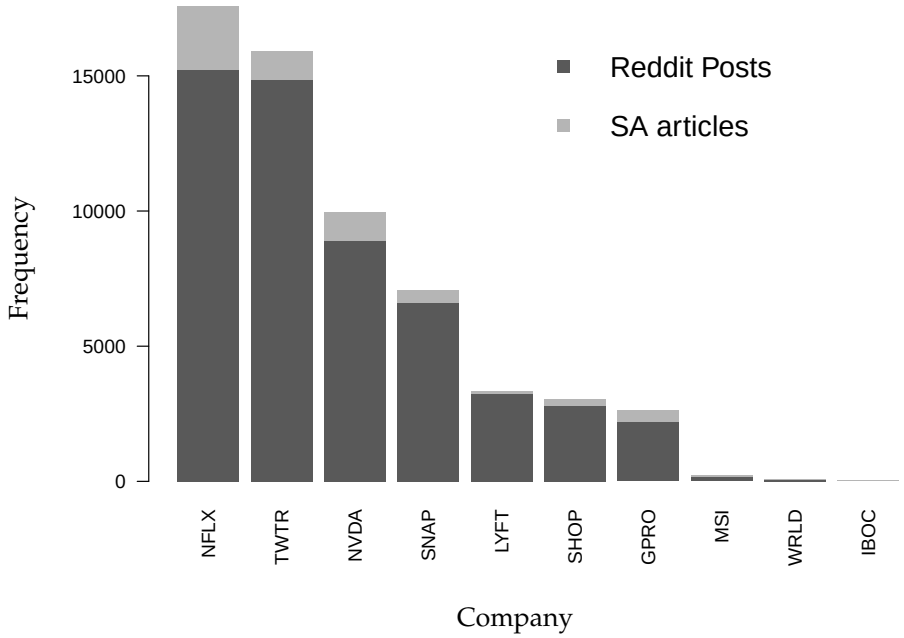


Figure A.3: Total number of Reddit posts and Seeking Alpha articles for all companies in the category short reports, time period 2009-01-01 to 2021-04-20.

A.2 Results

Preliminary analysis

Table A.4: Number of clusters found that satisfies condition, and proportion of events used to total number of events for a company. The first segment of companies are from the 2019-2021 social media list, the second is from the 2019-2021 short interest list and the last is from the 2009-2011 social media list.

-	Clusters						Proportion of events					
Company \ Bandwidth	7	14	30	40	50	60	7	14	30	40	50	60
GME	14	7	7	1	1	1	0.12	0.15	0.34	0.061	0.091	0.10
TSLA	14	1	0	1	2	3	0.13	0.022	0.00	0.070	0.16	0.26
AMD	18	12	3	2	1	0	0.17	0.23	0.13	0.11	0.077	0.00
AAPL	4	2	4	0	3	3	0.040	0.045	0.20	0.00	0.24	0.27
MSFT	9	5	0	0	0	0	0.096	0.11	0.00	0.00	0.00	0.00
AMZN	15	6	5	4	3	4	0.14	0.12	0.22	0.27	0.34	0.36
FB	12	4	2	2	3	1	0.12	0.078	0.082	0.12	0.22	0.091
UPS	5	5	5	3	3	4	0.027	0.065	0.24	0.18	0.21	0.32
GNUS	1	0	1	1	1	1	0.0025	0.00	0.012	0.017	0.023	0.035
Total | Avg	92	42	27	14	17	17	0.17	0.091	0.14	0.092	0.15	0.16
RCL	2	0	0	2	1	2	0.0063	0.00	0.00	0.031	0.019	0.046
FDS	0	3	0	0	1	0	0.00	0.016	0.00	0.00	0.073	0.00
Total | Avg	2	3	0	2	2	2	0.0032	0.008	0.00	0.16	0.0078	0.023
AIG	4	0	2	2	3	6	0.016	0.00	0.047	0.094	0.24	0.58
GM	0	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00	0.00
AAPL	1	0	0	0	0	0	0.0036	0.00	0.00	0.00	0.00	0.00
GE	0	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00	0.00
EBAY	0	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00	0.00
GS	1	3	1	1	1	0	0.0035	0.014	0.008	0.010	0.015	0.00
ADP	0	0	1	0	0	0	0.00	0.00	0.0051	0.00	0.00	0.00
Total | Avg	6	3	4	3	4	6	0.0033	0.002	0.009	0.015	0.036	0.083

Table A.5: Form of intra-cluster events with the internal order disregarded. The frequency of occurrences for each combination of form, for each company, is displayed.

Companies \ Form(F-R-S)	Min-min-min	Min-min-max	Min-max-min	Min-max-max	Max-min-min	Max-min-max	Max-max-min	Max-max-max
GME	3	0	0	1	0	0	0	3
TSLA	0	0	0	0	0	0	0	1
AMD	1	4	1	1	2	0	2	1
AAPL	1	0	0	0	0	0	0	1
MSFT	3	1	0	0	0	0	0	1
AMZN	2	0	1	1	0	0	0	2
FB	1	0	0	0	1	0	1	1
UPS	1	2	1	0	0	0	1	0
GNUS	0	0	0	0	0	0	0	0
Total (2019-2021)	12	7	3	3	3	0	4	10
RCL	0	0	0	0	0	0	0	0
FDS	0	1	0	1	0	1	0	0
Total (2019-2021)	0	1	0	1	0	1	0	0
AIG	0	0	0	0	0	0	0	0
GM	0	0	0	0	0	0	0	0
AAPL	0	0	0	0	0	0	0	0
GE	0	0	0	0	0	0	0	0
EBAY	0	0	0	0	0	0	0	0
GS	0	0	0	1	0	0	0	2
ADP	0	0	0	0	0	0	0	0
Total (2009-2011)	0	0	0	1	0	0	0	2

Table A.6: Order of intra-cluster events where the frequency of each combination for each company is displayd as well as the total for each combination.

Companies \ Order	F-R-S	F-S-R	R-F-S	R-S-F	S-F-R	S-R-F
GME	2	2	2	0	1	0
TSLA	0	0	1	0	0	0
AMD	0	4	4	2	1	1
AAPL	0	1	0	0	1	0
MSFT	1	3	1	0	0	0
AMZN	0	1	2	1	0	2
FB	0	1	0	0	1	2
UPS	2	1	0	2	0	0
GNUS	0	0	0	0	0	0
Total (2019-2021)	5	13	10	5	4	5
RCL	0	0	0	0	0	0
FDS	1	1	0	0	0	1
Total (2019-2021)	1	1	0	0	0	1
AIG	0	0	0	0	0	0
GM	0	0	0	0	0	0
AAPL	0	0	0	0	0	0
GE	0	0	0	0	0	0
EBAY	0	0	0	0	0	0
GS	2	0	0	0	1	0
ADP	0	0	0	0	0	0
Total (2009-2011)	2	0	0	0	1	0

Short report graphs

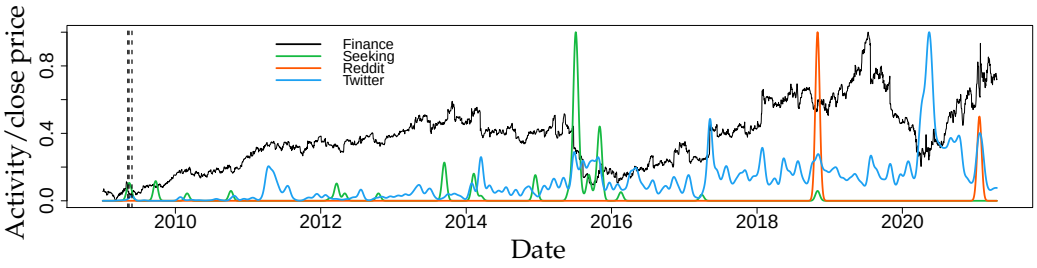


Figure A.4: WRLD time series with Citron Research short reports annotated as dotted vertical lines.

A.3 List of subreddits

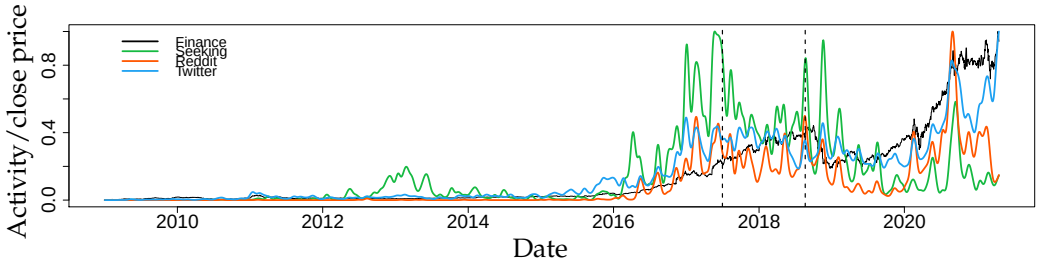


Figure A.5: NVDA time series with Citron Research short reports annotated as dotted vertical lines.

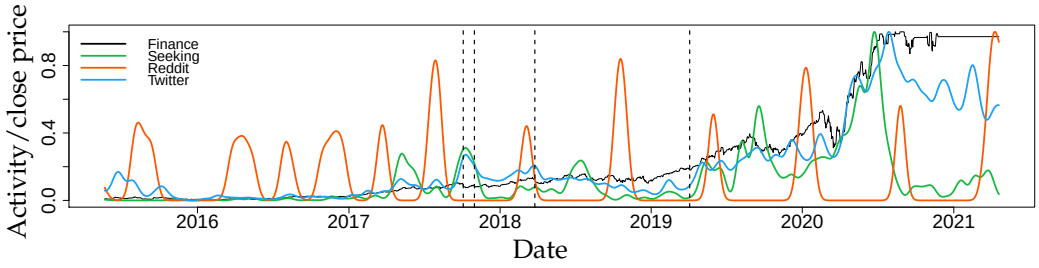


Figure A.6: SHOP time series with Citron Research short reports annotated as dotted vertical lines.

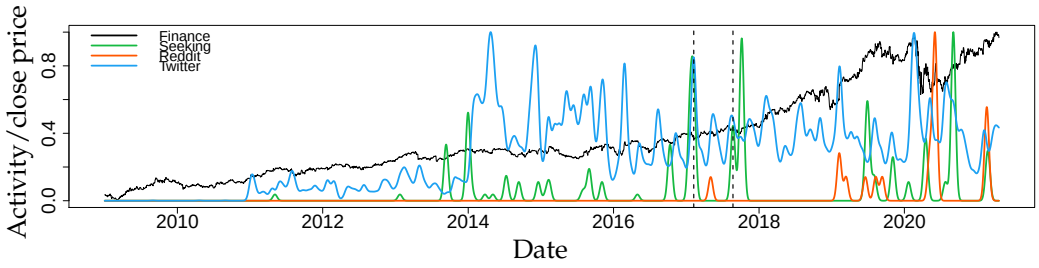


Figure A.7: MSI time series with Citron Research short reports annotated as dotted vertical lines.

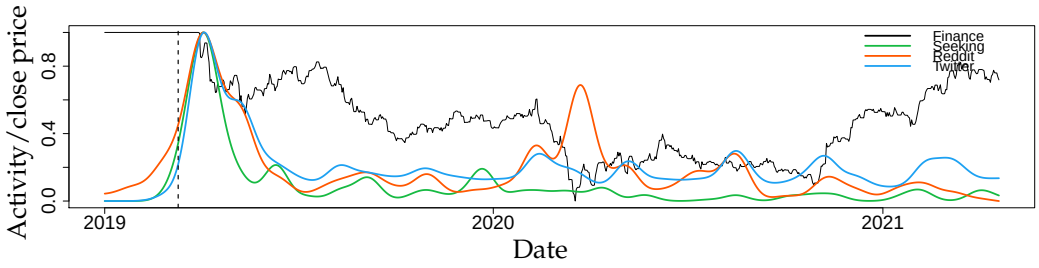


Figure A.8: LYFT time series with Citron Research short reports annotated as dotted vertical lines.

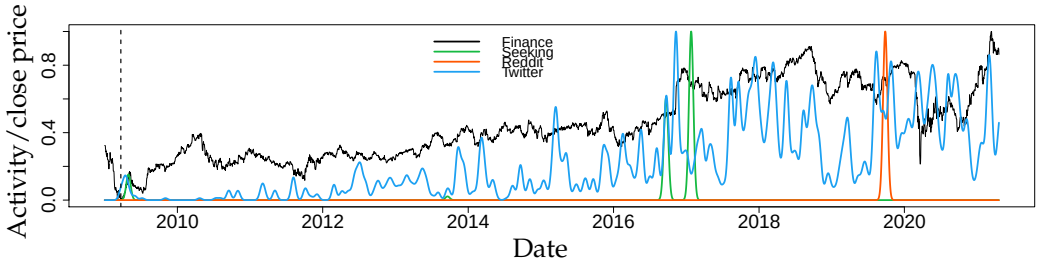


Figure A.9: IBOC time series with Citron Research short reports annotated as dotted vertical lines.

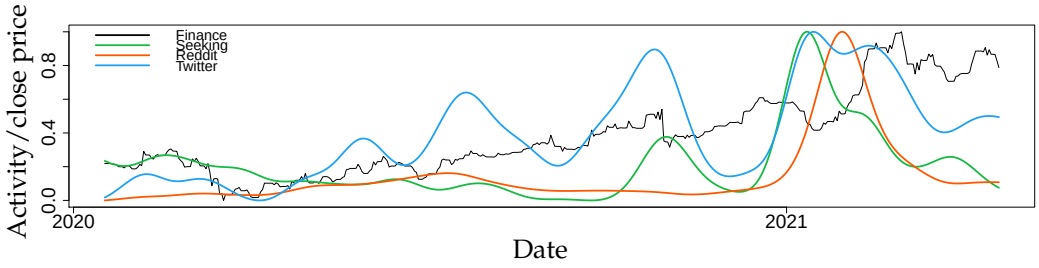


Figure A.10: TWTR part 1 time series with Citron Research short reports annotated as dotted vertical lines.

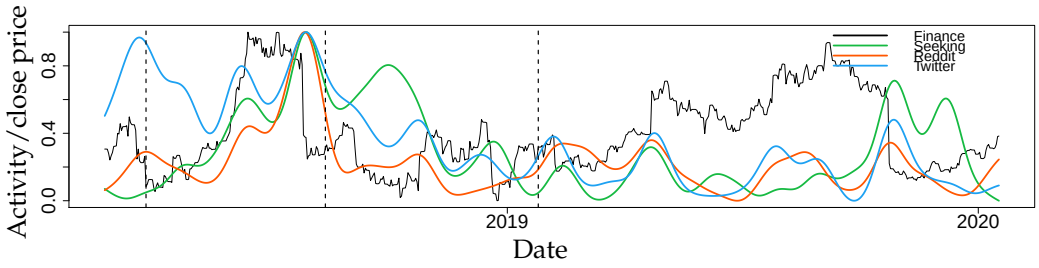


Figure A.11: TWTR part 2 time series with Citron Research short reports annotated as dotted vertical lines.

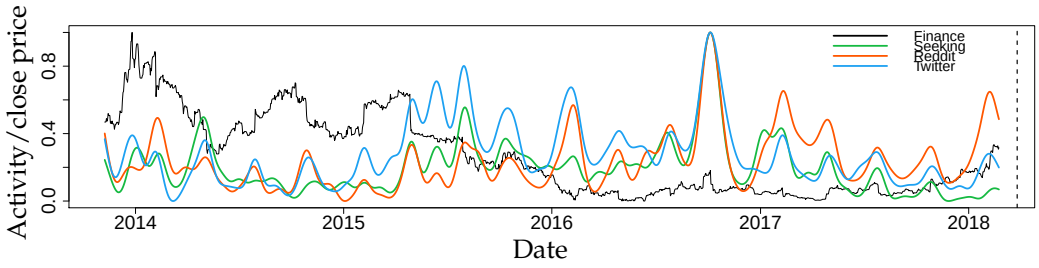


Figure A.12: TWTR part 3 time series with Citron Research short reports annotated as dotted vertical lines.

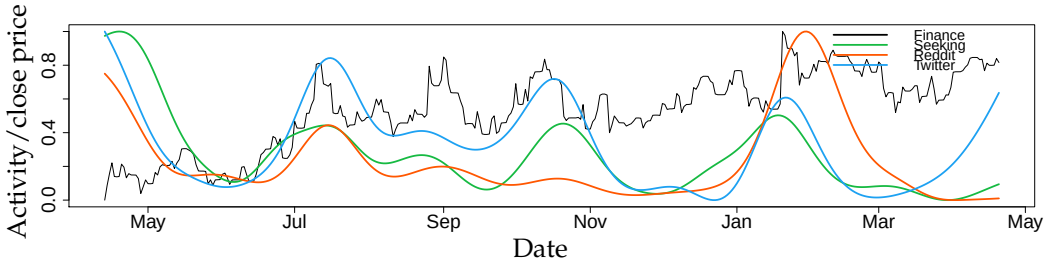


Figure A.13: NFLX part 1 time series with Citron Research short reports annotated as dotted vertical lines.

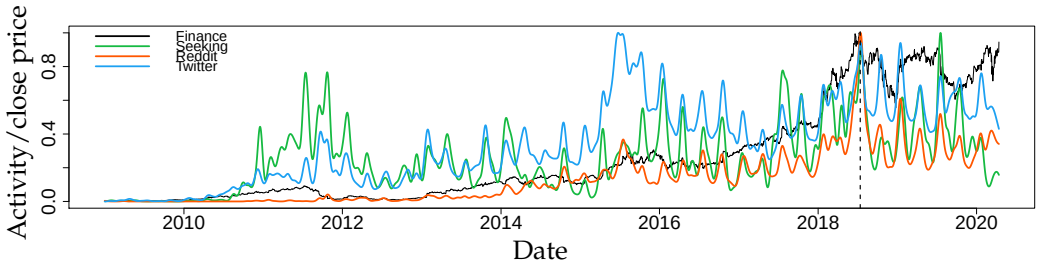


Figure A.14: NFLX part 2 time series with Citron Research short reports annotated as dotted vertical lines.

Table A.7: Subreddits

Subreddit			
stock	Stock_picks	Economics	stocks
RobinHoodPennyStocks	StockMarket	investing	SecurityAnalysis
pennystocks	RobinHood	WallStreetbetsELITE	finance
investing_discussion	Trading	Daytrading	Forex
personalfinance			