The correlation between sentiment of tweets and stock price returns

Livanis Nikolaos

A dissertation submitted in partial fulfillment of the requirements for the degree of Master of Science in Applied Economics & Data Analysis

School of Business Administration
Department of Economics

Master of Science
“Applied Economics and Data Analysis”

August 2020
Three-member Dissertation Committee

Research Supervisor: Emmanouhl Tzagarakis  Assistant Professor
Dissertation Committee Member: Dimitrios Tzelepis  Associate Professor
Dissertation Committee Member: Eleftherios Goulas  Assistant Professor

The present dissertation entitled

«The correlation between of sentiment of tweets and stock price returns »

was submitted by Livanis Nikolaos, SID 1018730, in partial fulfillment of the requirements for the degree of Master of Science in «Applied Economics & Data Analysis» at the University of Patras and was approved by the Dissertation Committee Members.
I would like to dedicate my dissertation to all the people who helped and supported me during all these years of academic education and especially to my parents who sacrificed many things for it.
Acknowledgments

I would like to express my sincere gratitude for this research, which is the result of hard work and sacrifice on my part, to my postgraduate professors and especially to my supervisor Tzagarakis Manolis where with their way, but also with the their special knowledge combined with the ability to impart their knowledge to their students, they helped me to learn new ways of research and analysis but also to develop my judgment and knowledge of economics and data analysis. Also a big thank you is a little for them as they were also my professors in my bachelor degree which means that they were the people who gave the foundations to understand economics.
## Contents

1 Introduction 7  
  1.1 Background .................................................. 7  
  1.2 Purpose ................................................... 8  

2 Related Work 9  

3 Methodology 11  
  3.1 Time Period ................................................... 13  
  3.2 Data ........................................................ 13  
    3.2.1 Text Data .............................................. 13  
    3.2.2 Market data ........................................... 14  
  3.3 Keywords Evaluation ........................................ 24  
  3.4 Dataset Explanation and Cleaning .......................... 27  
  3.5 Tweet Preprocessing ....................................... 29  
    3.5.1 Lowercase .............................................. 29  
    3.5.2 Tokenization .......................................... 29  
    3.5.3 Stopwords Removal .................................... 30  
    3.5.4 Punctuation Removal .................................. 30  
    3.5.5 Stemming and Lemmatization .......................... 31  
  3.6 Classification ............................................... 31  
    3.6.1 Feature ................................................ 32  
    3.6.2 Classifier ............................................ 34  
    3.6.3 Training Set .......................................... 36  
    3.6.4 Classifier Evaluation .................................. 38  
  3.7 Polarity .................................................... 41  
  3.8 Correlation .................................................. 41  

4 Results 42  
  4.1 Sentiment Results .......................................... 42  
  4.2 Polarity Results ............................................ 47  
  4.3 Correlation Results ........................................ 50  

5 Conclusion 56
**List of Figures**

<table>
<thead>
<tr>
<th>Figure Number</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Samsung Timeline</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>Xiaomi Timeline</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>Nvidia Timeline</td>
<td>18</td>
</tr>
<tr>
<td>4</td>
<td>AMD Timeline</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>Huawei Timeline</td>
<td>21</td>
</tr>
<tr>
<td>6</td>
<td>Tesla Timeline</td>
<td>23</td>
</tr>
<tr>
<td>7</td>
<td>Dell Timeline</td>
<td>24</td>
</tr>
<tr>
<td>8</td>
<td>Accuracy by Criterion</td>
<td>25</td>
</tr>
<tr>
<td>9</td>
<td>Accuracy by Position</td>
<td>26</td>
</tr>
<tr>
<td>10</td>
<td>Accuracy by Group of 10</td>
<td>27</td>
</tr>
<tr>
<td>11</td>
<td>Dataset Explanation</td>
<td>28</td>
</tr>
<tr>
<td>12</td>
<td>Training Set 1</td>
<td>37</td>
</tr>
<tr>
<td>13</td>
<td>Training Set 2</td>
<td>38</td>
</tr>
<tr>
<td>14</td>
<td>Classifiers Evaluation</td>
<td>40</td>
</tr>
<tr>
<td>15</td>
<td>Distribution of daily Sentiment of Tweets (Classifier 1)</td>
<td>43</td>
</tr>
<tr>
<td>16</td>
<td>The mood of each Company (Classifier 1)</td>
<td>44</td>
</tr>
<tr>
<td>17</td>
<td>Distribution of daily positive/negative tweets (Classifier 2)</td>
<td>45</td>
</tr>
<tr>
<td>18</td>
<td>Distribution of daily neutral tweets (Classifier 2)</td>
<td>46</td>
</tr>
<tr>
<td>19</td>
<td>The mood of each Company (Classifier 2)</td>
<td>47</td>
</tr>
<tr>
<td>20</td>
<td>Polarity values for Amd, Dell and Huawei</td>
<td>49</td>
</tr>
<tr>
<td>21</td>
<td>Polarity values for Nvidia, Samsung, Tesla and Xiaomi</td>
<td>50</td>
</tr>
<tr>
<td>22</td>
<td>Pearson’s Correlation Coefficients</td>
<td>52</td>
</tr>
<tr>
<td>23</td>
<td>Pearson’s Correlation Coefficients (Focused Period)</td>
<td>54</td>
</tr>
</tbody>
</table>
1 Introduction

1.1 Background

As opposed to computers, it is hard for people to consistently settle on objective choices and not let sentiments and feelings influence them. It has been noticed that there are a ton of irregularities in budgetary markets that would not show up if people were levelheaded and if stock prices were based on all accessible data about an company and nothing more. That is what the efficient market hypothesis by Fama (1970) recommends. The reality is clearly more complicated than that and to be able to predict future stock returns, human emotions need to be considered.

The ongoing technological revolution with boundless nearness of computers and Web has made a remarkable circumstance of information storm, changing significantly the way in which we take a gander at social and financial sciences. The continually expanding utilization of the Internet as a wellspring of data, for example, business or political news, set off a closely resembling expanding on the web action. The collaboration with innovative frameworks is creating gigantic datasets that archive aggregate conduct in a formerly inconceivable manner [1, 2]. At last, in this tremendous store of Web movement we can discover the interests, concerns, and aims of the worldwide populace as for different monetary, political, and social wonders.

We accept that social angles as estimated by informal organizations are especially helpful to comprehend monetary turnovers. In fact, budgetary virus and, eventually, emergencies, are regularly begun by aggregate wonders, for example, grouping among investors (or, in outrageous cases, alarm) which signal the inherent unpredictability of the financial system [3]. In this way, the possibility to anticipate anomalous conduct of investors is of incredible enthusiasm to policy makers [4–6] in light of the fact that it might consider a progressively speedy mediation, when proper.

How to gauge human feelings? A potential arrangement is to go to web based life and microblogs, for example, Twitter. Throughout a day, a lot of content are transmitted online through an assortment of internet based life channels. Inside these writings, important data about for all intents and purposes every theme exists. Through Twitter alone, more than 400 million tweets are sent for every day with more than 300 million dynamic month to month clients. With a limit of 140 characters, that is more than 56 billion characters created every day, just on Twitter. Despite the fact that each tweet may not appear to be incredibly significant, it has been contended that
collections of a lot of tweets can give important knowledge about open state of mind and slant on specific subjects [10].

Content mining includes a corpus of archives. In this unique circumstance, we for the most part center around tweets as a result of their attributes: they are short and speak to live circumstances better. What’s more, it is simpler to slither them than archives from other social networks. An online Application Programming Interface (API) is made accessible for any engineer ready to utilize Twitter information. Besides, tweets as a rule contain hashtags and images that encourage the quest for important posts. When a huge example has been gathered, creators apply diverse machine learning methods, which they can approve through strategies like cross approval and test with exactness measurements, for example F1-measure. Regardless of whether no classifier is great, they presently accomplish fulfilling results considering safe further investigation.

1.2 Purpose

The reason for this proposition is to analyze the connection between’s Twitter notion and stock returns with the relationship between’s Twitter feelings and the adjustments in stock price for a few companys.
2 Related Work

In [1] contemplated whether the day by day number of tweets predicts the S&P 500 stock pointers. A different line of research investigates the substance of tweets. In a printed investigation way to deal with Twitter information, the creators discover clear relations between the state of mind markers and Dow Jones Modern Normal (DJIA) [2–4]. In [5], the creators show that the Twitter opinion for five retail organizations has measurably critical connection with stock returns and instability. An ongoing report [6] looks at the data substance of the Twitter opinion and volume as far as their impact on future stock costs. The creators relate the intra-day Twitter and value information, at hourly goals, and show that the Twitter assumption contains altogether more lead-time data about the costs than the Twitter volume alone. They apply tough measurements which require moderately high volume of tweets over the whole time of a quarter of a year, and, as a result, just 12 money related instruments breeze through the assessment. Talks about the utilization of Twitter specically as a corpus for conclusion investigation [7]. They examine the techniques for tweet assembling and handling. The creators utilize explicit emojis to shape a preparation set for conclusion classication, a method that enormously diminishes manual tweet labeling. Their preparation set was part into positive and negative examples dependent on cheerful and pitiful emoticons. Furthermore, they break down a couple of precision improvement strategies. Essentially, [8] presents a conversation of spilling Tweet mining and slant extraction, while facilitating the conversation to incorporate sentiment mining.

As introduced by [9], one of the principal signs that there might be a connection between’s Twitter opinion and the stock market. In their work, an assumption score is associated with the DJIA and afterward took care of into a neural system to predict market movements. The creators utilize a state of mind following device named OpinionFinder to quantify mind-set in 6 measurements (Quiet, Alert, Sure, Crucial, Kind, and Cheerful). At that point, they connect the state of mind time arrangement with DJIA shutting esteems by utilizing a Self Organizing Fuzzy Neural System. Utilizing their strategies, they estimated an enhancement for DJIA expectation exactness. After distribution, this paper propelled a great part of the ebb and flow look into in the connection among twitter and market feelings.

In [10] for instance, utilized positive and negative emojis to extricate positive and negative tweets while paper title texts were utilized as instances of neutral tweets. Indeed, even [11] depended on emojis in their tweet grouping which was then used to
represent the emergency in Toyota in 2010 when the organization needed to review a huge number of vehicles because of issues with quickening agent pedals. The emoji approach could work for making a universally useful corpus of explained tweets and it accomplished work in the Toyota case, however it is faulty if tweets about traded on an open market organizations, particularly tweets with a money related nature, regularly incorporate emojis.

In [12] made their preparation corpus utilizing positive and negative emojis, a methodology depicted in the past subchapter, joined with a slant dictionary that gives a valence for each word in English. For top-16 innovation stocks as indicated by Yahoo! Finance a average accuracy of 87% was accomplished in anticipating the future movement of a stock. This is a promising outcome which makes it fascinating to attempt to discover other progressively unmistakable classifications that tweets could be ordered into that would bring about a much higher exactness. A potential inadequacy with that paper is that only technology stocks were secured. In [13] prepared their own opinion classifier which likewise utilized an estimation dictionary to give words notion scores. Looking at the prescient precision of their model for organizations in various businesses, the IT business had the most noteworthy prescient exactness when contrasted with organizations in the fields of fund, media, vitality, assembling and medication.

In [14] gathered around 21.000 tweets that incorporated a hashtag for any of the six Ekman feelings which at that point shaped the Twitter Feeling Corpus (Sleuth). This corpus has comparable issues to the one portrayed above thinking about the motivation behind this proposal. It only represents a small part of all tweets and it is unlikely that tweets about companies and their products include hashtags of Ekman feelings. Concerning assessment grouping, having a space explicit corpus to prepare the classifier on assumes an imperative job.

In [15] estimated aggregate expectation and dread on Twitter by checking tweets including mind-set words. Expectation and upbeat were utilized to discover positive tweets and dread, stress, on edge, apprehensive and upset were utilized for gathering negative tweets. Their surmising from the information was that all sort of enthusiastic upheavals, positive or negative, had a negative relationship with stock market, for example, the DJIA and S&P 500. It was estimated that individuals post increasingly emotional tweets in the time of financial vulnerability.
3 Methodology

As we have previously said, the purpose of this research is to find the impact of sentiment of tweets in the price returns. Because the source of text data will be a social media platform, we decide to focus on the technology sector, taking seven technology companies for investigation. This decision was made because the technology sector is more discussed about than other sectors in social media platforms. The companies than will investigate in the following article are Tesla, Huawei, Samsung, Xiaomi, Nvidia, Amd and Dell.

The following flow chart shows the process we followed in this research. Represents the steps we followed to collect and process our data until we reach the extraction of the final results. Clearly moving on to the next sections, all of these steps will be thoroughly analyzed. However, this diagram is a good idea to understand and get a general knowledge of the flow of steps and procedures to be followed.
3.1 Time Period
The time period is different for each Technology company. The criterion for the selection of time period is major events which take part during 2019 for every company. This could help us to extract better and bigger datasets from social media platforms. People have the tendency to write more emotional posts for companies during their major events (ex. new model presentation). Also, in this way we reduce the bias from our dataset.

3.2 Data
As we have previously said, the purpose of this research is to find the impact of sentiment of tweets in the price returns. Because the source of text data will be a social media platform, we decide to focus on the technology sector, taking seven technology companies for investigation. This decision was made because the technology sector is more discussed about than other sectors in social media platforms. The companies than will investigate in the following article are Tesla, Huawei, Samsung, Xiaomi, Nvidia, Amd and Dell.

3.2.1 Text Data
For tweet assortment, Twitter gives a somewhat hearty Programming interface. There are two potential approaches to assemble Tweets: the Streaming API or the Search API. The Streaming API permits clients to acquire constant access to tweets from an information question. The client first demands an association with a flood of tweets from the server. At that point, the server opens a gushing association and tweets are spilled in as they happen, to the client.

In any case, there are a couple of confinements of the Streaming API. In the first place, language isn’t speciable, bringing about a stream that contains Tweets all things considered, including a couple non-Latin based letters in order. Furthermore, at the free level, the gushed Tweets are just a little portion of the real Tweet body. Introductory testing with the streaming API brought about a contaminated preparing set as it demonstrated dicult to acquire an unadulterated dataset of Tweets.

In light of these issues, we chose to go with the Twitter Search Programming interface. The Search API interface is a REST Programming interface which permits clients to demand specific queries of tweets. REST, or Representational State Transfer, basically utilizes HTTP techniques (GET, POST, PUT, DELETE) to execute
various activities. The Search API permits more fine tuning of questions, including sifting dependent on language, locale, and time. There is a rate limit related with the question, however we handle it in the code. For our motivations, as far as possible has not been an issue. To really get tweets, we persistently send queries to the Search API, with a little deferral to represent as far as possible.

Both of these APIs require the client have a Programming interface key for validation. In this case we must refer to a ready-to-go Twitter Collector program which is submitted by my Professor and Supervisor Emanouhl Tzagkarakis for the purpose of this research.

The data collection have two stages. In the first stage we feed the Search API with a company name query. After the first data mining from Twitter, from the dataset of tweets which include the company name we take the 10 most frequent 2-gram keywords. In the second stage we feed the Search API with these 10 most frequent keywords. After that, we combine all the data to one big Dataset. This method is executed for each technology company which has been selected for our investigation. The benefit of this technique is that we take a large number of tweets for every company which covers most of the company’s activities. Also this technique helps us to find the trends of the specific time period which is different for each company.

3.2.2 Market data

Stock information have a few fascinating qualities to explore. The factors chose in past works mirror the assorted variety of potential applications. As a rule, a few ward factors are watched, for example, stock’s closing price or adjusted closing price, which is the closing value changed to incorporate dividends and different appropriations that happened previously. The adjusted price is the previous value, deducted of the previously mentioned profits and disseminations.

The source of our market data was Yahoo Finance. The stock of each company was active in Nasdaq Index. Unfortunately Nasdaq doesn’t work over ends of the week or occasions so there are missing values in our dataset. This would be a critical impediment of this investigation as it would impair the ability to calculate the connection between public mood and stock price movements on a given day would be ompromised. So as to sidestep this issue, a strategy for approximating the missing Nasdaq values utilizing a sunken function proposed by Mittal and Goel was received.
If a Nasdaq value on a given day is \( x \) and the following accessible is \( y \) with \( n \) days missing in between, the first missing value of \( x_{x+1} \) can be approximated using \( y(\frac{y + x}{2}) \) and repeated for other missing values. Mittal and Goel argue that this works for stock data as it usually follows a concave relationship, unless an incident of a sudden anomalous rise or fall occurs.

Returns are really regular logarithm returns \( R \) of stock prices \( S(t) \) over a period a time interval of one day. This extra activity has a few favorable circumstances, for example, standardization of the varieties.

\[
R_t = \ln(S_t) - \ln(S_{t-1})
\]

The log returns are utilized to ascertain unpredictability which is a level of variety estimated by their standard deviation over a specific timeframe. One contention of utilizing this sort of recipe instead of taking arithmetic returns is that it works better for demonstrating the stock market. To be sure, on the off chance that we expect that the closing prices are distributed log normally, we cannot accept the equivalent for arithmetic returns. Logarithm capacities preserve the typicality properties of the time series and permit processing securely insights, for example, compound returns.

As we said previously, the investigation period for each company differ and depends on a crucial event or a model launch date which took place in 2019. In this periods the stock prices for every company is more actively and tend to have an upward or downward trend. We decided to take a four month investigation period for each tech company which was enough to cover the most of their trend. In Figures 1, 2, 3, 4, 5, 6, 7 we can see the timeline for every company. Each chart shows the graphical representation of the stock prices in USD dollars and the returns prices for each of the companies we research. We observe that a new model was launched for Samsung, Tesla, Xiaomi, Nvidia and AMD during the research period which increased the stock price of the company. For Dell, an important technological event took place during the investigation period where the company had a dynamic presence. On the other hand the things aren’t so positive for huawei, which is the only company that has such a big downward trend during 2019 and more specifically during the research period because of the ban for 5g.

Samsung Galaxy S10 is a line of Android-based cell phones made, discharged and
showcased by Samsung Electronics as a feature of the Samsung Galaxy S arrangement. The Galaxy S10 arrangement is a celebratory arrangement of the tenth commemoration of the Samsung Galaxy S lead line, its top line of telephones close to the Note models. Uncovered during the "Samsung Galaxy Unpacked 2019" press occasion hung on February 20, 2019, the gadgets began delivering in specific locales. It is the tenth generation of Samsung’s Galaxy system S arrangement of cell phones.

As has been done since the Galaxy S6, Samsung disclosed flagship Galaxy S10 and Galaxy S10+ models, separated principally by screen size and an extra forward looking camera on the S10+. Likewise, Samsung additionally uncovered a littler model known as the Galaxy S10e, just as a bigger, 5G-perfect form, the Galaxy S10 5G.

As we can see in the chart of Samsung, there is no clear trend in the stock price after the launch of the new model, as well as the price of the company’s returns ranges from 0.04 to -0.04 with the highest prices being at the area of research.

![Samsung Timeline](image)

**Figure 1: Samsung Timeline**

The Xiaomi Mi CC9 Pro (known as the Mi Note 10 or Mi Note 10 Pro globally)
is an Android smartphone developed by Xiaomi which first presented in China at
the Xiaomi event on November 8, 2019 along with a series of new products of the
company. This mobile phone is the flagship of the company in the field of smart-
phones and led to significant sales and also to the consolidation of its the brand name.

Unlike Samsung from the previous chart, Xiaomi’s new model seems to have helped
a lot in the growth of its stock price. While at the beginning of the research period
the price did not have a specific trend, we observe that after the launch of the new
model there is a strong increase in the price. Its returns ranges from -0.07 to 0.07
with the highest being during the survey period.

![Xiaomi Timeline](image)

Figure 2: Xiaomi Timeline

The GeForce 20 series is a family of graphics processing units developed by
Nvidia. Serving as the successor to the GeForce 10 series, the line started shipping
on September 20, 2018, and after several editions, on August 20, 2019, the GeForce
RTX Super line of cards was announced. The 20 series marked the introduction of
Nvidia’s Turing microarchitecture, and the first generation of RTX cards, the first in the industry to implement real-time hardware ray tracing in a consumer product. In a departure from Nvidia’s usual strategy, the 20 series doesn’t have an entry level range, leaving it to the 16 series to cover this segment of the market.

Seeing how the stock price of Nvidia is activated in 2019, one can say that there is an upward trend but also some drops in its price at certain intervals. However, if we focus on Nvidia’s research period, we notice that the launch of the new graphics card helped the stock to have a clear upward trend of greater inclination than before. Also the returns of the stock ranges between -0.15 which is the highest returns we encountered and is not in the research period. On the contrary, the highest returns does not exceed 0.07.

![NVIDIA Timeline]

**Figure 3: Nvidia Timeline**

Ryzen is a brand of x86-64 chip planned and advertised by Advanced Micro Devices, Inc. (AMD) for desktop, mobile and inserted platforms dependent on the
Zen microarchitecture. It comprises of focal preparing units advertised for standard, lover and workstation fragments and quickened handling units (APUs) promoted for standard and section level portions and inserted applications. Ryzen is particularly noteworthy for AMD, since it is a totally new structure, and denotes the enterprise's arrival to the very good quality work area CPU advertise. AMD’s rival Intel had controlled this segment of the market for very nearly a long time since the arrival of AMD’s infamous Bulldozer microarchitecture in 2011. With AMD’s underlying relapse in this market section starting considerably further in 2006, when Intel discharged its earth shattering Center microarchitecture.

AMD formally declared another arrangement of processors, named "Ryzen" during its New Skyline highest point on December 13, 2016 and presented Ryzen 1000 arrangement processors in February 2017, highlighting up to 8 centers and 16 strings, which propelled on Walk 2, 2017. The second generation of Ryzen processors, the Ryzen 2000 arrangement, includes the Zen+ microarchitecture, a gradual improvement based on a 12nm procedure innovation, was discharged in April 2018 and highlighted a minor exhibition increment over Ryzen 1000 processors that previously discharged in 2017. On October 8, 2019 AMD launch the third generation of processors, the Ryzen 9 3900 which become famous for its specifications.

The same phenomenon as in Nvidia is observed in the case of Amd. We notice that at the beginning of the research period the price was quite stable, while from the day of the launch of the new processor of the company, the stock price is rising rapidly and at a steady pace. The growth rate is quite stable as the stock returns have lower average during the research period without extreme prices.
Huawei is the world’s No. 1 telecom provider and No. 2 telephone maker, however it’s an untouchable in nations like the US. For over a year, there’s been no deficiency of investigation on the Chinese telecom monster, and various nations have restricted the utilization of its systems administration gear. Its telephones are for all intents and purposes undetectable in the US regardless of its huge nearness around the globe.

In Pre-May 2019 it was declared by the US government that Google and US organizations need to change the manner in which they manage Huawei. The Chinese monster was boycotted by the US in the most recent engagement of the continuous exchange war. Google was especially vociferous that forestalling Huawei utilizing its rendition of Android might bring about national security issues through individuals utilizing a Huawei-created substitution operating system - presently uncovered to be HarmonyOS. As a general rule, it’s likely in light of the fact that non-Android telephones would hit Google benefits.
Unlike other companies, Huawei has a fairly active stock characterized by sharp changes in its prices, either negative or positive. However, for a long time that includes the research period, the stock price has a steady downward trend and this is due to the ban. We notice that during the ban period the share price fell rapidly as the returns reached its lowest price close to -0.10. After this shock the returns continued to fluctuate between 0.05 and -0.05.

The Cybertruck was uncovered at the Tesla Structure Studio in Los Angeles on 21 November 2019. During the introduction, Musk exhibited the toughness of the vehicle and its materials. Notwithstanding effective drop tests led on a sheet of the specific ‘Tesla armor glass’ and a fruitful pre-show test where a steel ball was tossed at the windows of the truck itself by head of plan Franz von Holzhausen with obviously no harm, the windows were harmed when Holzhausen rehashed the test during
the show. Musk flippantly shouted that "the ball didn’t make it through" and "we’ll fix it in post" after the sudden outcomes. He later clarified that the windows were harmed on the grounds that in a previous exhibit, the entryway was hit by a heavy hammer and that split the base of the glass.

Toward the finish of the introduction, the Tesla Cyberquad, an off-road vehicle (ATV), was driven onto the bed of the Cybertruck utilizing worked in slopes in the back end. The Cyberquad was connected to the Cybertruck’s locally available electrical plug to charge the Cyberquad batteries. The ATV will be ready to move as a discretionary bundle with the Cybertruck.

The Cybertruck unveiling event was covered heavily by traditional media and online blogs/social media. In social media, numerous reporters communicated abhorrence of the sharp forms and uncommon outside of the Cybertruck. Tesla Inc. stock was down 6% following the Cybertruck announcement. On 23 November 2019, Elon Musk tweeted that Tesla had gotten 146,000 pre-orders in the primary 1.5 days after the revealing.

With a downward trend until mid-2019 and with an upward trend until the end of the year we could characterize the stock of Tesla. However, the stock has higher returns values during the research period and this is due to the launch of Cybertruck. We notice that at that time the price fell instantly due to the negative event of the presentation but after a stabilization of the fall the price starts to increase exponentially. At the same time, during the research period and before the presentation of the new model, the price of Tesla receives the highest returns, which exceeds 0.15.

As with most of the companies we analyze, Dell seems that the launch of the new model had a positive impact on the stock price, creating a steady rise during the analysis period. However, we observe that after the research period, the stock returns have its lowest and highest price above -0.10 and close to 0.10 respectively.
CES (Consumer Electronics Show) is an annual trade show organized by the Consumer Technology Association (CTA). Held in January at the Las Vegas Convention Center in Las Vegas, Nevada, United States, the event typically hosts presentations of new products and technologies in the consumer electronics industry.

Every January, tech giants present their products at CES events. Dell is one of the companies that every year has the momentum in this event and manages to catch the eye with its new technologies and make people talk about them. In 2019, Dell once again had a dynamic presence at Event with the introduction of the new XPS and Alienware series of laptops to steal the interest of consumers.
3.3 Keywords Evaluation

Many research tried to find some correlation between sentiment of tweets and price stock taking a small number of tweets for their investigation company. Most of them, search for tweets with their investigation company name only. However they lose a big number of tweets which are relevant with their investigation company without containing its name. As we said previously, in our research we separate the mining technique of tweets into two stages. With this method we could collect more relevant tweets which means a bigger dataset for each company. Also we could find some trends of the company for the specific investigation period, a new models which launched this period and generally we could cover the most of the company activities. To succeed that, we should extract the most frequent keywords from the dataset of tweets which have already collected in the first stage of mining for every company. However not all keywords which are more frequent are also relevant with the company activity. To overtake this problem, in this research we prefer 2-gram keywords.
because are more complex and also minimize the probability to find a wrong frequent keyword.

In our research we use two kind of criteria to examine which keywords are most relevant. The Google trends criterion and the Personal Opinion of 2 people. We examine 175 2-gram keywords which were the 25 most frequent extracted keywords from stage one for each company. We use the accuracy as an evaluation measure. The google trends criterion match the extracted 2-gram keywords from first stage’s tweets with the keywords of google trends platform as the results of the firm name query at specific time period for every company. On the other hand, in personal opinion criterion the keywords evaluated for their relevance by 2 people. As we can see in Figure 8 54% of extracted keywords are relevant with their company activities and 36% of them are in the results of google trends with the same query.

![Figure 8: Accuracy by Criterion](image)

The results from Figure 8 aren’t so hopeful for our research because if we decide to adopt this technique only the 54% of extracted keywords will be relevant with
their company activities which means that we will take a big amount of irrelevant tweets in our dataset and the bias issues will be appeared. However, a better look in the accuracy for each frequency position in Figure 9 help us to understand that the bigger the frequency position the smaller the accuracy for each criterion. From 1st to 10th position seems that the accuracy is higher than the other positions with the top 5 frequency positions having the best relevance returns.

Creating a diagram with 3 groups of 10 in Figure 10, we can see clearly that the accuracy after the 10th position significantly falls. This help as to understand that the extracted 10 most frequent keywords have a probability of 85% to be relevant with their company activities. Because of the highly accuracy of the first 10 positions we can adopt this technique and take the first 10 most frequent keywords as a query in the second stage for every company.
3.4 Dataset Explanation and Cleaning

Using the mining technique separated into two stages we collected a total amount of 243,532 tweets. However after checking the dataset we found 35,688 duplicates tweets which should be removed. Also, there were 10,829 tweets created by theirs company’s profile which should be removed because the sentiment of these kind of tweets be always positive. After the dataset cleaning the total amount of tweets which we will use for our research is 197,758.

In Figure 11 we can see all the kinds of tweets which were collected for each company. The colored area are the total number of tweets for every company. Seems like tesla have the biggest dataset than any other company. The green area represent the amount of tweets created by theirs company’s profile and the red area the
number of duplicates. Removing these tweets, the remaining dataset is the blue and orange area which represent the clear dataset extracted from second stage and the clear dataset extracted by first stage respectively. AMD and NVIDIA have the lowest dataset of tweets, on the other hand Huawei and Samsung have the highest dataset. If we hadn’t adopt the two stages mining technique which means that we had extracted our tweets only with the company’s names as queries the dataset that we would have would be only the orange area. The benefit of two stages mining technique is the blue area for each company. For most of companys, this technique increased their dataset over 85%.

Figure 11: Dataset Explanation
3.5 Tweet Preprocessing

The content of each tweet contains numerous superfluous words that don’t add to its slant. Many tweets incorporate URLs, labels to different users, or images that have no significance. To precisely acquire a tweet’s opinion, we first need to channel the clamor from its unique state. To do this, we consolidate an assortment of methods.

3.5.1 Lowercase

The first step in processing of tweets is to convert all the letters to lowercase. This process is quite important as for people the same letter whether written in uppercase or lowercase has the same meaning, but this does not happen with the machine. The machine unfortunately recognizes uppercase and lowercase letters as different letters. For example, for humans, 'm' and 'M' are the same letter as opposed to the machine that recognizes these two letters as different characters that have nothing to do with each other. Therefore, in order to avoid this phenomenon that would cause us serious problems in the continuation of the research, we convert all the letters into small ones so that they have the same basis.

3.5.2 Tokenization

Tokenization is the demonstration of separating a grouping of strings into pieces, for example, words, catchphrases, expressions, images and different components called tokens. Tokens can be singular words, states or even entire sentences. During the time spent tokenization, a few characters like accentuation marks are disposed of. The tokens become the contribution for another procedure like parsing and text mining. A sentence of 10 words, at that point, would contain 10 tokens. Tokenization is language-explicit, and every language has its own tokenization necessities. English, for instance, utilizes void area and accentuation to signify tokens, and is moderately easy to tokenize. In reality, most alphabetic dialects follow generally clear shows to separate words, expressions and sentences. In this way, for most alphabetic dialects, we can depend on rules-based tokenization.

However the first step includes parting the content by spaces, shaping a rundown of individual words per text. This is additionally called a pack of words. We will later
utilize each word in the tweet as highlights to prepare our classifier.

3.5.3 Stopwords Removal
In registering, stop words will be words which are sifted through previously or in the wake of handling of characteristic language information (text). Though 'stop words' as a rule alludes to the most widely recognized words in a language, there is no single all inclusive list of stop words utilized by all common language preparing apparatuses, and without a doubt not all devices even utilize such a list. A few tools explicitly abstain from evacuating these stop words to help state search.

Any gathering of words can be picked as the stop words for a given reason. For some web crawlers, these are the absolute generally normal, short capacity words, for example, the, is, at, which, and on. For this situation, stop words can cause issues while scanning for phrases that incorporate them, especially in names, for example, 'The Who', 'The', or 'Take That'. Other web search tools expel probably the most widely recognized words—including lexical words, for example, 'need'— from an inquiry so as to improve execution.

For these reasons, we expel stopwords from the pack of words. Python’s Natural Language Toolkit library contains a stopword word reference. To expel the stopwords from every content, we essentially check each word taken care of words against the word reference. In the event that a word is a stopword, we filter it out. The rundown of stopwords contains articles, a few relational words, and different words that include no sentiment value.

3.5.4 Punctuation Removal
Numerous tweets contain additional symbols, for example, '@' or '#' just as URLs. The word promptly following a '@' symbols is consistently a username, which we filter out altogether, as they add no value to the content. Words following '#' are kept, for they may contain data about the tweet, particularly for classification. URLs are filtered out completely, as they add no conclusion significance to the content. To achieve the entirety of this, we utilize a regex that counterparts for these images. Furthermore, any non-word images taken care of words are filtered out also.
3.5.5 Stemming and Lemmatization

Stemming and Lemmatization are Text Normalization (or some of the time called Word Normalization) procedures in the field of Natural Language Processing that are utilized to plan text, words, and reports for additional preparing. Stemming and Lemmatization have been examined, and calculations have been created in Computer Science since the 1960’s. Languages we talk and compose are comprised of a few words frequently got from each other. At the point when a language contains words that are gotten from another word as their utilization in the discourse changes is called Inflected Language. The level of expression might be higher or lower in a language. As you have perused the meaning of affectation as for language structure, you can comprehend that an inflected word(s) will have a typical root structure. Stemming and Lemmatization encourages us to accomplish the root structures (in some cases called equivalent words in search setting) of inflected (inferred) words. Stemming is distinctive to Lemmatization in the methodology it uses to create root types of words and the word produced. Stemming and Lemmatization are broadly utilized in labeling frameworks, ordering, SEOs, Web search results, and data recovery. For instance, scanning for fish on Google will likewise bring about fishes, fishing as fish is the stem of the two words. Later in this instructional exercise, you will experience a portion of the huge employments of Stemming and Lemmatization in applications.

3.6 Classification

Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/labels or categories. Classification predictive modeling is the task of approximating a mapping function \( f \) from input variables \( X \) to discrete output variables \( y \). For example, spam detection in email service providers can be identified as a classification problem. This is a binary classification since there are only 2 classes as spam and not spam. A classifier utilizes some training data to understand how given input variables relate to the class. In this case, known spam and non-spam emails have to be used as the training data. When the classifier is trained accurately, it can be used to detect an unknown email. In a general system, classification is tied in with giving a content mark or a discrete worth gratitude to a segregating capacity based on information previously characterized physically or originating from past encounters. This function is then used to anticipate the class for new information. It works when we attempt to predict a discrete value for our information point. An example of binary classification stands in banking sector where it can tell whether a particular can receive a loan or not but it can also predicts multiple classes to answer different kinds of problems.
Deciding sentiment polarity of tweets isn’t a simple assignment. Financial specialists frequently differ whether a given tweet speaks to a purchase or a sell signal, and even people are not generally steady with themselves. We contend that the upper bound that any automated sentiment classification can accomplish is controlled by the degree of understanding between the human. So as to quantify the understanding between the human, a generous division of tweets must be explained by 3 unique individuals — for our situation, 1,000 tweets were commented on between 'positive', 'Negative' and 'Neutral' sentiment. When we accumulate a huge tweet corpus with 'positive' and 'negative' and 'Neutral' assessment, we can assemble and prepare a classifier.

3.6.1 Feature

A unigram is just a N-gram of size one, or a solitary word. For every one of a kind word in a tweet, a unigram highlight is made for the classifier. For instance, if a positive tweet contains 'advertise', an element for classification would be whether a tweet contains 'advertise'. Since the component originated from a positive tweet, the classifier would be bound to group different tweets containing 'advertise' as positive.

We separate bigrams and trigrams from our tweets as highlights to prepare our classifier. Bigrams are sets of back to back words, which add to the precision of the classifier by distinguishing words, for example, 'don’t care for' or 'not glad.' Likewise, Trigrams are triplets of successive words. These too add to the data inclusion of the classifier. In our N-grams, hole skipping isn’t permitted, and the words must follow one another. At the point when we include bigrams and trigrams, we increment the highlights in our list of capabilities by n-1 for bigrams and n-2 for trigrams, where n is the quantity of individual words among the entirety of the tweets in our corpora. In spite of the fact that this enormously builds preparing and classification time, our own experimental classifiers give us that bigrams and trigrams increment the precision of the classifier.

We feed in features from an external lexicon called SentiStrength, which is a list of words that are predened with a sentiment, either positive or negative. Their data is applied to 'short texts,' which is perfect for short things. The inclusion of the SentiStrength Lexicon allows for a broader coverage of words that may be missed.
by merely collecting from Tweets. We spent time looking into more freely available lexicons, and discovered that there are far too many to be able to feasibly measure each one’s effectiveness. Therefore, we decided that lexicon analysis was out of scope for this paper.

Different works endeavor to utilize Part of Speech Tagging to changing degrees of accomplishment. Part of Speech Tagging includes increasing words in a tweet with a relating grammatical form, just as its specific circumstance. With that information, certain highlights can hypothetically be wiped out or more heavily weighted. Past work has demonstrated that they didn’t find any significant classification improvement utilizing this procedure. Subsequently, while we tinkered with a framework that permitted part of speech tagging on tweets, we concluded that it was eventually not worth the extra exertion, and left it out of our structure.

We considered applying a third mark, 'neutral,' as a potential classification for our tweets. However, we ran into a couple of difficulties and discovered proof against utilizing this mark. Furthermore, work indicated that including a third 'neutral' name to a classifier in actuality decreased its accuracy and was not beneficial for their framework. In the spirit of simplicity, we chose to stick to this same pattern and not utilize an 'neutral' name in our classifier yet we made a classifier with three labels to check if the accuracy will fall. Of course, there are drawbacks to just arranging each tweet as 'positive' or 'negative.'

Statistical methodologies, for example, machine learning and deep learning admirably with numerical information. In any case, common language comprises of words and sentences. In this manner, before you can manufacture a sentiment analysis model, you have to change over content to numbers. A few methodologies have been produced for changing text to numbers. Bag of Words, N-grams, and Word2Vec model are some of them. A good approach of Bag of Words model is TF-IDF scheme, so as to change text to numbers. Python’s Sklearn library comes with built-in functionalities to implement TF-IDF approach. Here we will provide a brief insight into the TF-IDF approach.

In a basic bag of words, each word is given equivalent significance. The thought behind TF-IDF is that the words that occur more frequently in one document and less much of the time in different reports should be given more significance as they are increasingly valuable for classification. TF-IDF is a result of two terms: TF and IDF.
TF = (Frequency of a word in the document) / (Total words in the document)

IDF = Log[(Total number of docs) / (Number of docs containing the word)]

We calculated the TF-IDF score for each word, using each tweet as a separate document.

With the strategy portrayed over, the list of capabilities becomes bigger and bigger as the dataset increments. After a specific point, notwithstanding, it gets troublesome and superfluous to utilize each and every unigram, bigram, and trigram as a component to prepare our classifier. To cure this circumstance, we chose to utilize just the n most significant highlights for preparing. To find the best combination of parameters which maximize the accuracy of our classifier we use the GridSearchCV. GridSearchCV check every possible combination of parameters and gives us the best one.

### 3.6.2 Classifier

Precise classification is as yet an intriguing issue with regards to machine learning and data mining. Commonly, we need to manufacture a classifier with a lot of preparing information and marks. For our situation, we need to develop a classifier that is prepared on our "positive","negative" and neutral marked tweet corpus. From this, the classifier will have the option to name future tweets as either "positive" ,"negative" or "neutral" in view of the tweet’s properties or highlights. In our research we use Naive Bayes model because is better classifier in text analysis.

A Naive Bayes classifier is probabilistic classifier dependent on Bayes Rule, and the least difficult type of a Bayesian system. The classifier is easy to execute and generally utilized in numerous applications. The classifier works on a hidden "naive" suspicion of restrictive freedom about each component in its element set. The classifi
er is an utilization of Bayes Rule:

\[ P(c \mid F) = \frac{P(F\mid c)P(c)}{P(F)} \]

In our specific situation, we are searching for a class, c, so we locate the most likely class given highlights, F. We can treat the denominator, P(F) as a consistent, for it doesn’t rely upon c and the qualities are given. Along these lines, we should concentrate on tackling the numerator. To do this, we have to decide the estimation of P(F \mid c). Here is the place the autonomy presumption comes in. We expect that, given a class, \( c_j \), the highlights are restrictively free of one another, hence:

\[ P(f_1, f_2...f_n \mid c_j) = \prod_i P(f_i \mid c_j) \]

From this, we can classify a tweet with a label \( \hat{c} \) with a maximum posterior decision rule taking the most probable label from all labels C.

\[ \hat{c} = \arg\max_{c_j \in C} P(c_j) \prod_i P(f_i \mid c_j) \]

The Naive Bayes classifier is incredibly straightforward, and its restrictive freedom assumptions are not sensible in reality. Be that as it may, utilizations of Naive Bayes classifiers have performed well, better than at first envisioned. Past work from [16] examines the amazing presentation of the Credulous Bayes classifier and recommends that the conveyance of conditions among all qualities over a class prompts an ideal classification. In this paper, we utilized Python’s Natural Language Toolkit’s NLTK library.
3.6.3 Training Set

In machine learning, a typical assignment is the investigation and development of calculations that can gain from and make predictions on information. Such calculations work by settling on information driven forecasts or choices, through structure a numerical model from input data.

The information used to construct the last model as a rule originates from various datasets. Specifically, three datasets are regularly utilized in various phases of the formation of the model. The model is at first fit on a training dataset, which is a lot of models used to fit the boundaries of the model. The model (a neural net or a naive Bayes classifier) is prepared on the the training dataset utilizing an administered learning method. In practice, the training dataset frequently comprises of sets of an information vector (or scalar) and the relating yield vector (or scalar), where the appropriate response key is ordinarily meant as the target. The current model is run with the training dataset and produces an outcome, which is then contrasted and the objective, for each information vector in the training dataset. In light of the aftereffect of the examination and the specific learning algorithm being utilized, the boundaries of the model are balanced. The model fitting can incorporate both variable determination and parameter estimation.

Progressively, the fitted model is utilized to anticipate the reactions for the perceptions in a second dataset called the validation dataset. The validation dataset gives an unprejudiced assessment of a model fit on the training dataset while tuning the model’s hyperparameters. Validation datasets can be utilized for regularization by early. This straightforward system is entangled by and by the way that the validation dataset’s blunder may vacillate during preparing, delivering numerous neighborhood minima. This difficulty has prompted the making of some specially appointed principles for choosing when overfitting has really started.

Finally, the test dataset is a dataset used to give an unbiased evaluation of a final model fit on the training dataset. If the information in the test dataset has never been utilized in preparing (for instance in cross-approval), the test dataset is additionally called a holdout dataset.

In our research we took random tweets from our whole dataset. We collected 780 tweets which were classified by 2 people into 'Positive', 'Negative' (and Neutral). We created two training set, one of them have only 'Positive' and 'Negative' label and the other one have one more label, the 'Neutral' label.
In Figure 12 and 13, we can see the distribution of two training sets. The first training set has 410 observations and the second has 312 observations. Also, we can notice that each label in every training set has the same number of observations with the other labels in the same dataset. This situation will help us to use the accuracy as our Classifier Evaluation metric in the next steps.

Figure 12: Training Set 1
3.6.4 Classifier Evaluation

Passing judgment on a classification model feels like it should to be a simpler task than making a decision about a regression. All things considered, your expectation from a classification model can just either be correct or wrong, while a forecast from a regression model can be pretty much off-base, can have any degree of mistake, high or low. However, making a decision about a grouping isn’t as basic as it might appear. There’s more than one path for an order to be correct or to not be right, and numerous approaches to consolidate the various approaches to be good and bad into a bound together measurement. Obviously, all these various measurements have extraordinary, regularly unintuitive names — precision, recall, F1, accuracy.

By what method should you measure the performance of your classifier? The clear answer is to utilize accuracy: the number of examples it classifies correctly. You have
a classifier that steps through examination models and theorizes classes for each. On each test model, its estimate is either right or wrong. You essentially measure the quantity of right choices your classifier makes, separate by the all out number of test models, and the outcome is the exactness of your classifier. It’s that basic. By far most of examination results report accuracy, and numerous commonsense activities do as well. It’s the default metric. What’s going on with this methodology? Exactness is a shortsighted measure that is misdirecting on some genuine issues. Actually, the most ideal approach to get an excruciating encounter is to utilize exactness as an assessment metric.

\[
\text{Accuracy} = \frac{(\text{TruePositives} + \text{TrueNegatives})}{(\text{TruePositives} + \text{FalsePositives} + \text{TrueNegatives} + \text{FalseNegatives})}
\]

Accuracy essentially treats all models the equivalent and reports a level of right reactions. Exactness might be fine when you’re managing adjusted (or around adjusted) datasets. The further you get from 50/50, the more precision deceives.

Moreover, we estimated accuracy and review esteems for every one of the marks, 'positive' and 'negative'. Precision (additionally called positive prescient worth) is the division of pertinent examples among the recovered cases, while recall (otherwise called affectability) is the part of the aggregate sum of important occurrences that were really recovered. Both exactness and review are along these lines dependent on a comprehension and proportion of relevance. Precision here is estimated as the quantity of genuine right outcomes over the all the positive outcomes, or:

\[
\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}}
\]

\[
\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}}
\]

In Figure 14 we can see the results of our classifiers. We trained two classifiers, the first one classify the extracted tweets into 'Positive' and 'Negative' tweets and the second one into 'Positive', 'Negative' and 'Neutral' with the previous training sets.
respectively. The reason we trained two models was because we wanted to see the effect of inserting a neutral label in accuracy of our classifier, as previous studies have shown that it significantly reduces the accuracy. However in Figure 14 the accuracy of classifier 2 which has 3 labels is 52.4%, not so hopeful for our investigation. On the other hand when we removed the "Neutral" label and trained again our classifier with the two labels training set we enjoy an accuracy of 68.9%. That means that the Neutral label has a negative impact of -16.5% into classifier accuracy.

![Figure 14: Classifiers Evaluation](image)

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Label</th>
<th>precision</th>
<th>recall</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Negative</td>
<td></td>
<td>0.7020</td>
<td>0.6350</td>
<td>0.6990</td>
</tr>
<tr>
<td>1 Positive</td>
<td></td>
<td>0.6780</td>
<td>0.7410</td>
<td>0.6990</td>
</tr>
<tr>
<td>2 Negative</td>
<td></td>
<td>0.5220</td>
<td>0.6000</td>
<td>0.5240</td>
</tr>
<tr>
<td>2 Neutral</td>
<td></td>
<td>0.4780</td>
<td>0.5790</td>
<td>0.5240</td>
</tr>
<tr>
<td>2 Positive</td>
<td></td>
<td>0.5880</td>
<td>0.4170</td>
<td>0.5240</td>
</tr>
</tbody>
</table>

Figure 14: Classifiers Evaluation
3.7 Polarity

After we have classified all tweets to 'positive', 'negative'(and "neutral"), the next step is to calculate the polarity for each day or the mood of the day for every single company. Sentiment polarity $P_d$ is the difference between the number of positive and negative tweets as a fraction of non-neutral tweets

$$P_d = \frac{\text{PositiveTweets} - \text{NegativeTweets}}{\text{PositiveTweets} + \text{NegativeTweets}}$$

If there were no negative sentiments, the ratio would be 1. Conversely, if there was less positive sentiment, the ratio would be closer to 0.

3.8 Correlation

Correlation analyzes two factors, stock price returns and tweet sentiment polarity for each company, and is counterbalanced by a specific lag value. This implies we postponed costs by a specific time, for every connection count to represent day distinction required for twitter sentiment polarity to be reflected in stock price returns. In this research we use Pearson correlation coefficient.

Pearson’s correlation coefficient is the test insights that quantifies the factual relationship, or relationship, between two continuous variables. It is known as the best strategy for estimating the relationship between variables of intrigue since it depends on the technique for covariance. It gives data about the size of the affiliation, or connection, just as the heading of the relationship. Coefficient values can range from +1 to -1, where +1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and a 0 indicates no relationship exists. The condition for connection we utilized was:

$$r_{xy} = \frac{\sum_{i=0}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=0}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2}}$$
4 Results

4.1 Sentiment Results

At this stage we will discuss and present the results of our analysis. First we will analyze the results of the 2 classifiers that we constructed, the polarity we calculated based on these results and also the correlation of this polarity with the stock price returns for every company which is the main topic of our work.

After collecting tweets from twitter and stock price data of the companies for each research period which as we have previously mentioned and shown differs for each company, the next stage of our analysis was the construction and training of our classifiers. We made 2 classifiers where the only difference between them was the one extra label (Neutral) and we trained them with the training dataset which we analyzed previously where they had an equal number of observations for each label. After the training of our classifiers, we are ready to classify all the other tweets we have already collected into negative, positive (and neutral) labels for each company. The following graphs show the results of this categorization for every company.

Figure 15 shows the distribution of the percentage of positive and negative tweets for each day during the research period for each company for classifier 1 (2 labels). It seems that 50 to 60 percent of daily tweets are positive for Amd and Dell and the other 40 to 50 percent are negative. Samsung, Xiaomi and Nvidia have the highest percentage of positive daily tweets as result of the lowest percentage of daily negative tweets. For Tesla company, the percentage of two labels are very close to 50/50. On the other hand, as the results of the downward trend in stock price of Huawei, this company also have one of the highest and one of the lowest percentage of negative and positive daily tweets with most of them are between 75 to 85 and 15 to 25 percent respectively.
As an auxiliary diagram, Figure 16 shows the total number of negative and positive tweets for each company during its research period. As it is logical Huawei has an extremely high number of 35K negative tweets in relation to the number of its positive comments. Also it has the highest number of negative tweets than any other company. On the other hand, Samsung seems to have the exact opposite phenomenon. However all the other companies have more positive tweets than negative.
Figure 16: The mood of each Company (Classifier 1)

Figure 17 shows the same results with the Figure 16 with the architecture of the classifier being the only difference. In this graph we use the classifier 2 (3 labels) to classify our tweets. We observe the distribution of the percentage of positive and negative tweets for each day like the previous graph. In this case, the huawei have again the highest and the lowest percentage of negative and positive tweets respectively. Samsung, Nvidia, AMD and Xiaomi have a very low percentage of negative daily tweets but a not so high daily percentage of positive tweets which is lower than 45%. These numbers suggest us that these companies seem to have a fairly high number of neutral tweets.
As we understood from the previous chart that many companies seem to have a significant number of neutral tweets, it remains to be seen the distribution of the percentage of neutral daily tweets for each company. Figure 18 represent this kind of distribution. It is clear that, Huawei and Tesla have the lowest percentage of daily neutral tweets which is lower than 10%. The percentage for Samsung and Dell is between 15% and 25% while AMD ,Nvidia and Xiaomi have the highest percentage with over 45% of daily neutral tweets.
Like Figure 18, Figure 19 shows the total number of negative and positive tweets for each company during its research period as a result of classifier 2. Xiaomi seems to have the highest number of neutral tweets, with the Huawei and Samsung having again the highest number of negative and positive tweets respectively. The only thing that changed in this case is that the number of negative tweets for Dell is higher than its number of positive tweets.
4.2 Polarity Results

Once we have categorized all the data we have extracted from twitter into negative, positive (and neutral) the next step of our research is to use the polarity formula to extract an emotion value for each day for each company during the period research. The range of this value is from 1 (completely positive) to -1 (completely negative) while the value 0 indicates neutrality. The research period for each company is 4 months or 122 days. However, it is worth mentioning that for the continuation of the research we will use only the classifier 1 (2 labels) and the results of this classifier for two reasons. The first reason is because the accuracy increases by 16.5% which is a serious advantage as it increases the reliability of the results. The second reason is because the nature of the polarity formula uses only positive and negative emotion to calculate it. So using data that has neutral emotions will automatically force us
to greatly reduce the dataset which is not so positive in combination with the lower accuracy of classifier 2.

Therefore according to the above and the results of the classifier 1 (2 labels) we found the polarity values for each company. In the diagrams 20 and 21 we have the graphic representation of these prices during the research period for each company. The results have been divided into two figures for a better interpretation of them. As the only common feature that each research period has with the others is the number of days otherwise 122 days, in the horizontal axis we have values from 1 to 122 that represent the days within the analysis period. On the vertical axis we have the polarity values for each of these days.

Figure 20 shows graphically the polarity results for Amd, Dell and Huawei companies. Starting from Huawei, we observe that throughout the analysis period the polarity is at levels below zero except in some cases. The minimum polarity value for this company is -0.756 with an average of -0.472 and its maximum value does not exceed 0.159. From these results we conclude that there is a negative mood in tweets of this company during its research period which is logical because as we mentioned at the beginning of the research at that time the company was banned by large companies in the industry. In contrast, for Amd and Dell the average polarity is 0.282 and 0.363 respectively with the highest price reaching just over 0.6 for both companies. A difference is present in the minimum price of polarity for these 2 companies as for Amd the price reached -0.5 while for Dell it never fell below zero.
Figure 21 shows graphically the polarity results for Nvidia, Samsung, Tesla and Xiaomi companies. What is worth noting in Figure 21 is that for Nvidia, Samsung and Xiaomi the price of polarity never falls below zero in contrast to Tesla where in some places the price is below zero with the lowest reaching 0.372 while the largest is 0.288. For the other companies in the chart (Nvidia, Samsung, Xiaomi) the average polarity ranges from 0.45 to 0.65 with the highest values reaching close to 0.89. These prices help us to understand that for these companies there is a fairly positive mood during the research period. Tesla, on the other hand, has an average close to zero which leads us to the conclusion that there is primarily a neutrality in the companies’ tweets during this period.
4.3 Correlation Results

Once we have classified and found the polarity for each tweet of each day for every company which we investigate, the next stage of our research is to see if there is any correlation between the polarity of the company and its stock returns during the research period. So we used Pearson’s Correlation Coefficient formula to find the correlation of these variables. However, it is not clear that what you are writing today about a company will definitely affect its share price today. There are indications that what is being written today may affect the stock price after a few days, which means that there is some kind of delay that may be due to the not so fast information of financial market players. Also if this case is true, then there is strong evidence to disprove the Efficient Market Hypothesis that speaks to the full representation of all available information on the stock price and the fact that people are effective and act effectively without any direction from any emotion.

Therefore, in order to see all the above, we correlated the polarity values of each company with returns values of the respective stock price. Initially we started with
the same day prices and then we continued to correlate the polarity prices with the stock returns the very next day until we got to correlate the polarity with returns after 10 days. The results of these correlations are shown in Figure 22 which shows all the correlations of each company during its investigation period for any lag in returns of the stock. We have highlighted with a specific color any kind of 2 tailed statistical significance (0.01, 0.05, 0.1).

Let’s start by saying that most of the correlation values are not statistically significant while a small part has some statistical significance. In several companies such as Tesla, Samsung and Nvidia we did not find any statistically significant correlation in any time lag of returns. We also notice that Amd, Dell, Huawei and Xiaomi have correlations which are statistically significant with Huawei and Xiaomi having more than one value. Also something else that we could observe is that most of the correlation prices are negative while in no case did we find a large correlation between the polarity and the stock returns as all prices range from -0.25 to 0.15. Starting with the company Amd, we found a correlation -0.20 between the polarity of the company and the price of the returns with a time lag of 1 day which is statistically significant with a significance level of 5 and is the largest correlation for this company. The same scenario happens in the company Dell where here we have only one correlation which is negative and statistically significant at a significance level of 1 %. The only difference in this case is that the statistical significance was found in the time lag of 3 days and is also less than that of Amd by 0.3 points which means that it is quite low and close to zero. For Huawei things are different as we found correlations with a statistical significance of 5 % in three different time lags. The first was found after 4 days which is the largest negative correlation for this company, the second after 5 days which is the smallest statistically significant correlation we observe for this company. The latter was found after 10 days which is the last time lag of our analysis. It is also the only company that has a statistically significant correlation in this time lag. Respectively at Xiaomi we see three different time lags in a row having negative correlations with the polarity of the company. This phenomenon is observed after 4 days, after 5 days and after 6 days, with the time lag of 5 days having the highest correlation and the lowest level of statistical significance of the table.
Looking at the figure above we notice that the three of the seven companies do not have a correlation that is statistically significant and therefore we can analyze it. There are many reasons why we could not find any significant correlation for these companies and one of them is probably the potential noise that may be present in the dataset of these companies during the investigation period. To see if the above hypothesis is valid, we reduced the investigation period for each of the companies we analyze in such a way that each investigation period of each company starts from the day of the launch of its model (Amd, Dell, Nvidia, Samsung, Tesla, Xiaomi) or from the day an important event for the company took place (Huawei).

This will reduce the number of comments for each company but at the same time the tweets that were written before the important event or the launch of a new model.
will go away as there is a high probability that these tweets are not important and may not help extracting the correct results. This way the potential noise that may be in each company’s dataset from tweets that are not important to it and were probably written before the event will be reduced.

Therefore, for our analysis, we called this reduced research period the "focused period" for which we rediscovered the polarity correlations of each company for each time lag of its stock returns up to 10 days ahead. These results are shown in detail in Figure 23 where the statistically significant correlations have been highlighted in three different colors.

We observe that even in this case most of the correlations are negative. We also have a larger number of statistically significant correlations and at the same time for almost all companies we have at least one correlation that is statistically significant. Therefore, as a first point of view, we observe that the reason why we did this further research by reducing the research period for each company was successful as only one company in this case does not have a statistically significant correlation in contrast to Figure 22 where the number of these companies was 3. Also in this case we have a greater range in the values of the correlations with the smallest being just above -0.41 and the largest just below 0.20. Starting from Xiaomi, as in Figure 22, so here it has three negative correlations in the time lags of 4, 5 and 6 days which are statistically significant. What is worth noting in this company is that it has the largest negative correlation of the table and at the same time the lowest level of statistical significance (0.01). Continuing with Tesla we observe that there is a correlation between the polarity of the company and the returns of the share with a time lag of 8 days and a level of statistical significance of 1%. In the same scenario is Samsung, where it has only one correlation in the zero time lag which is statistically significant at a significance level of 1% with the only difference that the correlation is positive and the largest in the table. For Nvidia, things remain the same as those of the previous figure, as even in this case we did not find any correlation that has any statistical significance so that we can highlight it. For Huawei and Dell we notice that they have only one correlation in the time lag of 10 and 3 days respectively which are both negative with the largest being Huawei. Finally, AMD has two negative correlations close to -0.20 at a level of significance of 1% in the time lag of 1 and 4 days.
Comparing the two diagrams, Figure 22 and Figure 23 there are several similarities and differences that we can observe. Initially we see that the reason why we entered the process to reduce the research period for each company so that each period starts from the day that each important event of each company took place, was successful and gave us the conclusions we expected. This conclusion comes from the fact that in Figure 23 with the focused period we have found for all companies except Nvidia at least one correlation which is statistically significant in contrast to Figure 22 where we had 3 companies that had no statistically significant correlation. Therefore we conclude that the tweets before the event were not so important for the company and confused the dataset more. We also observe that the correlations that are statistically significant are on average larger in Figure 23 than in Figure 22. In other words in Figure 23 in addition to the larger number of companies with statisti-
cally significant correlations, we observe that the correlations are also comparatively larger with the Figure 22, with the smallest reaching -0.41 and -0.25 and the largest 0.20 and 0.15 in chart 23 and 22 respectively.
5 Conclusion

This section will discuss the results and conclusions of the research. Starting from the data collection, it appeared that the two-stage method initially helped to collect a larger amount of data covering most of the activity of each company, which in most cases increased the initial data size of each company by 90%. This increase in the number of data has helped us to get more reliable results as several studies that have been done with a relevant title report some problems in the results that may have been caused by the not so large volume of data they had.

The collection of a larger volume of data helped a lot in the machine learning stage of this research, as in this way our classifier was given more unique words for learning in order to increase its accuracy. We also found that adding an extra category to our classifier, in this case neutral tweets, would cost us a lot in classifier efficiency which would lead to the possibility of getting wrong results. This led us to the conclusion that our classifier has a lot of difficulty in identifying neutral tweets, which are more confusing than beneficial to the machine.

Regarding the final results of our research that have to do with the correlation of the emotion of the tweets and returns of the share of each company, they helped us to draw several conclusions that will be analyzed below. Initially, it is worth noting that each company showed a statistically significant statistical significance in the correlation between polarity and stock performance in a different time lag. But if we had to mention those hysteria times that showed most of the time some statistically significant correlation we would say that the time lags of 3 to 5 days are some of them. Also another important conclusion that we should mention is that all the statistically significant correlations are negative except for one in the case of Samsung, but at the same time close to zero which means that it is quite close to neutrality. In other words, it seems that what is written on twitter has little negative effect on returns and therefore the share price after 3 to 5 days after the publication.

We also noticed that by reducing the search period for each company, so that each one starts from the day that each important event of each company mentioned above took place, we managed to reduce the noise that seemed to be present in the companies’ tweets published before the date of the event. As a result, we have stronger statistically significant correlations between polarity and stock performance, but also an increase in the number of companies that had at least one correlation or which was statistically significant.
References


