UCLouvain
FUCaM MONS

LOUVAIN
School of Management

Louvain School of Management

# Twitter and its relationship with returns and trading volume of European stocks

Auteur : Elodie Michaux
Promoteur : Catherine D'Hondt & Christophe Desagre
Année académique 2018-2019
Master en ingénieur de gestion, majeure en corporate finance

## Acknowledgements

I would like to thank first my professors Mr Christophe Desagre and Mrs Catherine D'Hondt for answering my questions and providing me a guidance all along this final step of my studies. Our dataset of tweets was acquired with the support of Mr Hubert Naets and for that, I would like also to thank him.

Then, I'm grateful to Louis Gérard and Théophile Goffin for their help respectively in the programming part and for the second reading.

Finally, I would like to give special thanks to my parents for their support during these five years that may have sometimes seemed long.

## Abstract

In the past, decision-making was only made based on information gathered from traditional media. However, the rise of social media has changed the relation to news. Its accessibility is more easy, fast and efficient making social media an unavoidable source of information for investors. Thus, it allows them to make better and more informed decisions thanks to an exchange of information that would have never been imagined before.

This master's thesis is set in a behavioural perspective as it tries to understand the stock market reaction towards different non-fundamental information about companies. The medium of news used in this study is the social network Twitter during a period of time from 2008 to 2018. Companies examined are part of the indexes CAC 40, AEX 25 and BEL 20 giving a European orientation to this research and distinguishing it from the previous ones. What is more, this master's thesis investigates whether Twitter, through its volume of tweets and sentiments embedded, influences the stock market and then extends the existing literature by examining its link with the percentage of retail investors and market capitalization of companies.

At the end, our results assert the explanatory power of Twitter in the European market. Indeed, the tweet volume as well as the negative opinions that can be reflected in messages are linked to stock returns. Nevertheless, our results were not consistent concerning a relationship with stock trading volume. Moreover, we report a connection between the market capitalization of companies and the Twitter activity concerning their stocks.

# Table of contents

## List of figures

# List of tables

## List of appendixes

## Section I: Introduction

In recent years, the social media utilization, including Twitter, has increased due to the larger availability of Internet and of ways to connect people around the world. Unexpected outcomes have emerged from them including one useful for investors. It reveals that social media and more precisely microblogs can be used to make profit on the market and stands on the information democratisation. Indeed, social media have allowed people to access information, including the investing data, making their searches less costly and time consuming.

The idea of our study came from the drastic fall of 1,3 billion of Snapchat capitalization, namely six percent, last year. It happened after a tweet of Kylie Jenner, Kim Kardashian's stepsister, on which she pretended to not use Snapchat anymore (*appendix 4*). The stock price drop was relevant as the reality show celebrity is influential among youngers, the website target (*Clement, 2018*). However, another tweet from the starlet was published thereafter to attest that she was still loving Snap and the stock price recovered slowly. That example is not lonely. Regularly, celebrities and influential people tweet about companies and by doing so affect their stock returns. This includes Oprah Winfrey and its claim to have lost 12 kilos thanks to Weight Watchers, company of which she is an important shareholder (*Clement, 2018*). Besides TV famous characters, CEO also impact their company stock price through tweets. For instance, Elon Mush's tweet in 2015 caused an increase of four percent of Tesla market capitalization while last year he claimed that his company will quit the Stock Exchange soon, again increasing Tesla stock price. For that, he and his company get a 20 million fine (*La Libre, 2019*). This penalty came as an answer to the violation of the amical agreement between the SEC and Elon Musk preventing him from publishing tweets that could impact Tesla stock price. However, he offended again this year in 2019 with a "fake news" tweet about the production level expected for the same year. Moreover, politicians are also important in this landscape including the omnipresent Donal Trump who shudders the market at each of its tweets.

So, evidence in the news but also in the literature asserts the existence of a link between Twitter and the stock market. Unfortunately they all concern the American market and the most famous studies focused on a sample localized before 2015. Therefore, this master' thesis examines whether Twitter is also a valuable source of

information concerning companies in Europe from 2008 to 2018. The stocks examined are part of the indexes CAC 40, AEX 25 and BEL 20. Moreover, it extends the existing literature by adding a novel part reporting the impact of some corporate features on Twitter activity.

Thus, our study is articulated around three axes:

- The link between the volume of tweets and the stock market.
- The relationship between sentiments embedded in tweets and the stock market.
- The impact of corporate features on Twitter activity.

For each of them, the question will be to know whether the tweet volume, sentiments or corporate variables are good explanatory variables or not. We chose the percentages of positive and negative tweets as measures for sentiments to distinguish ourselves from other studies often having computed a global polarity feeling among tweets. Through this choice, we are able to account for differences between positive and negative emotions. What is more, the stock market movements will be reflected through the stock returns and trading volume while Twitter activity will be expressed by the tweet volume.

Consequently, the objective of this document is to answer the following research question: "What is the impact of Twitter activity and investor sentiments embedded on the stock market in Europe?".

On the top of that, our study on Twitter has an impact for both investors and managers. Through it, they will both catch the potential impact of their tweets on the European stock market but also of the others' publications. However, finding the one relevant tweet in a sea of millions of exchanges is difficult. One solution is to follow particular Twitter accounts such as the ones of CNBC, Stockwits or Wall Street Journal Markets. But another solution would be to follow an index composed of an aggregation of tweets as Bloomberg does. Since Twitter has understood the hidden value behind its tweets, it has monetized the value extraction from its contents. Nevertheless, the huge amount of information embedded in Twitter makes it difficult for investors to use it. As the role of computers in the stock market is increasing, we may imagine an algorithm that will analyse the entire pool of information from the market, social media and personal

information in real time to come to instantaneous and unbiased investment decisions. These algorithms may even replace traders and analysts and Twitter will play an important role in that change.

Besides, we will continue our topic presentation by a theoretical part divided in two sections. The first one contains all the necessary background about new technologies linked to the topic of this study: User-Generated Contents, social media, microblogs, Twitter and the information evolution related to the change of media conception. Similarly, the second part includes the financial concepts on which our study is based.

Then, a literature review is exposed. It integrates the evolution of the paperwork in reference with the theory previously mentioned: explanations about headlines are followed by User-Generated Contents and microblogs and finally by Twitter, the heart of our study.

Next, the methodology of our research is explained beginning by the hypotheses and followed by our sample and data collection techniques presentations. In the data description section, our dataset is further exposed with its processing and the variables resulting.

Finally, the regression results are developed for each of the two models. The first one and the most basic shows the links between tweet and market features while the second one includes corporate parameters in the previously realized regressions. Then, another method of sentiment analysis is performed and the divergences between results are exposed. We conclude by a summary of our results and the limits of our analyses.

We would like to end this introduction by a Benjamin Franklin's sentence that has attracted our attention during our researches and is relevant to our topic:
*"A slip of the foot you may soon recover, but a slip of the tongue you may never get over"*

## Section II: Social media theories

This section first deals with the evolution of information related to the change of media conception and then explains the different concepts linked to social media and that are unavoidable to understand the topic of our master's thesis.

As well as for products, demand and supply exist for information. The demand of information refers to the amount of data desired by consumers in order to make a purchase, sale or hold decision while the supply of information is the total amount of news available on a specific topic. The first one is reflected among others in the Google Search Intensity while the second one is expressed in the information on traditional or social media, official documents from governments or companies, etc. Specifically, this study deals with the information supply.

While information is often sporadically disclosed at some events such as press releases, earnings announcements or corporate reports, investors need to enter into an information quest in order to be up-to-date at every moment. Therefore, they have to look for information on alternative sources than financial statements and official disclosures. Particularly because firms are trying to keep negative news as more as possible hidden, it requires people to search additional opinions by their own. Without forgetting the fact that news can be unanticipated by businesses meaning that they do not have a communication plan prepared to react promptly to it. Yet, this quest is time and resource consuming, as information is a rare commodity, and so call fall into the insider trading. The latter refers to an illegal process of trading at one's own advantage based on confidential information.

Moreover, in the past, the information broadcast was only quantitative. Thus, the analysis made for the literature, the economic or the business applications were limited. However, the apparition of statistical studies on textual data changes the relation to news. With the development of computer and data science, more data can be analysed and patterns extracted from them. Now, every type of information can be studied adequately in respect of time or resources. This move from a number aspect to a textual form emphasis was strengthen by social media.

Social media in their globality are one of the solutions found by people in their information quest. They have the advantage of containing all kinds of information. It goes

from a press article posted by a press agency to rumour published in an anonymous' account. Some of these data are predictable, such as earnings announcements or CEO interview rebroadcasts, while others are not (a tweet, a consumer review, etc). Plus, the information can be presented in a structured form (for instance, official disclosures or press releases) while others can be informal, often when they are written by consumers. They can also arise from an official source (a company, a newspaper, a government) or from people in the street.

The apparition of User-Generated Contents (i.e. UGC) has modified the relation of people to news as it is more important in terms of quantity, contagiousness, accessibility, creates communities and yet is reliable. These characteristics make clear the greater frequency of UGC from both the writer and reader's perspectives and explain the development of all their forms: blogs, social media, wikis, microblogs, etc. Their principal subdivisions are shown in figure 1.

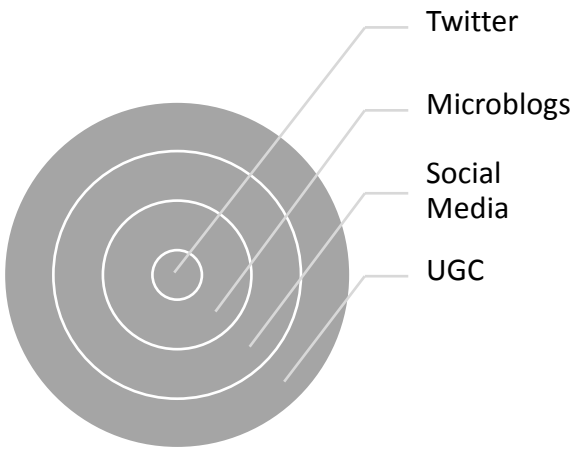



*Figure 1: Classification of UCG types*

Each layer represented in figure 1 has appeared successively. They will be defined and described in their order of occurrence hereafter. In order to better understand Twitter, the evolution of UGC will be given as well as some of the biggest elements concerning these categories. Indeed, many of Twitter characteristics come from its belonging to social media which in turn is derived from UGC.

For some years now, people are moving from their passive role of readers to an active purpose of experience sharing *(Tirunillai & Tellis, 2012)* and ultimately of influencers. This change goes even further as some sources of information are now only written by users requiring an investigation work comparable to an expert's one. For instance, wikis or blogs are only developed by non-specialists. The User-Generated

Contents (UGC) consist in all these materials voluntarily created by members of a website (*Oxford dictionaries, 2019*). They are publicly available, often free for all users and represent a new kind of word-of-mouth. In comparison to the previous form of hearsay, UGC have a deeper impact because of their instantaneity, popularity, fast-growing and ease of use (*Tirunillai & Tellis, 2012*). These characteristics will be explained for microblogs and especially for Twitter later in this section. Besides them, privacy and security are two "sine qua non" conditions for investors' willingness to share information on social media and especially for women (*Fogel & Nehmad, 2009*). On the one hand, privacy refers to the rights that everyone has on their private data for its use and disclosure. On the other hand, security means data protection. They can be jeopardized by thefts or hacker attacks (*Heather, 2013*).

What is more, UGC are mostly producted by private individuals rather than by professionals (*Tirunillai & Tellis, 2012*). Notwithstanding, consumers' opinions make these platforms valuable sources of information because they benefit from the "Wisdom of Crowds". It refers to the theory that a crowd can be better than individuals taken separately. That reliability is amplified by the number of users sharing an idea and so by its popularity. A well-known example of the Wisdom of Crowd effect is Wikipedia which has a quality comparable to encyclopaedias written by experts (*Gilles, 2006*). Aware of this phenomenon, companies are crowdsourcing tasks for problems solving or product development in many sectors: high-technology, software (e.g. Cisco), chemicals, medical devices, media (e.g. Netflix), homecare products (e.g. P&G), mining (e.g. Goldcorp), etc. (*Coeurderoy, Neysen & Paque, 2018*).

In the offline world, there are four conditions for wise crowds: knowledge, motivation, diversity and independence. We can assume that they are the same for the online world (*Nofer, 2014*). The diversity in terms of information, culture or demographic characteristics (i.e. age, education, etc) allows the population to take into account more alternatives and perspectives (*Watson, Kuman & Michaelsen, 1993*). In any case, its diversity pushes the crowd to communicate more than a homogeneous team and by doing so reduces the demographic boundaries and allows information, perspective and experience exchanges. Independence is another condition to obtain the best possible solution from the crowd. It means making one's decision freely, without being influenced by others. However, some studies have shown that the communication among members

of a crowd can improve the group performance (e.g. Miller & Steyvers, 2011). But the overall opinion remains that independence is key.

Then, following the UGC diversification, one of its branches has known a tremendous success: social media. According to the online dictionary Merriam-Webster (2019), social media are defined as "forms of electronic communication (such as websites for social networking and microblogging) through which users create online communities to share information, ideas, personal messages, and other content (such as videos)". They are composed of social networks (e.g. Facebook, Google +), blogs (e.g. Skyblogs, Tumblr), microblogs (e.g. Twitter, Reddit, Raging Bull), content communities (e.g. Youtube) and virtual worlds (e.g. World of Warcraft). On these platforms, users show their preferences and make recommendations on diverse topics.

One of the social media forms has attracted our attention: the microblogging. Its value added from other social media relies on its frequency of utilization. In opposition with usual blogs, a length criterium restricts the number of words per messages. As a consequence, the frequency of posts increases from 1/day in a regular blog to several posts/day in a microblog and by doing so reduces noise and irrelevancies in messages. That facilitates the diffusion of public investing information (e.g. earnings announcements, IPO) following the occurrence of events at a greater, if not constant, frequency. Its instantaneity goes further than for classical UGC and has made it a successful type of platform. Two categories of microblogs exist: chat rooms, which are live forums keeping no archive of the conversations, and bulletin boards, on which posts can be retrieved later by users. Examples of the second class in an investing context are Raging Bull or Reddit.

Specifically, Twitter is a microblogging website of the bulletin board type created in 2006 on which people can exchange information on various topics such as politics, sports, hobbies, events, etc. It was created after Raging Bull, Reddit or Yahoo!Finance[1] and is not dedicated to a single purpose on the contrary to these forums. A message on the platform is known as a tweet and is viewed by everyone among the publisher's connections. In order to be someone's connection, you need to "follow" this person by clicking on the button of the same name.

[1] Platforms devoted to investing discussions.

As a proof of the popularity of microblogs, Twitter reports more than five hundred million tweets per day (*Twitter, n.d.*) and a stable number of active users since 2015 (*Statista, 2019*) (*appendix 3*). A major part of Twitter success comes from the possibility to retweet messages meaning sharing someone's tweet to one's own community. Another one is the use of hashtags allowing to aggregate tweets on the same topic more easily. A "hashtag" relates a tweet to a topic (e.g. #GameOfThrones) while the "cashtag" identifies a stock ticker (e.g. $AAPL[2]) and a "at" relates it to a username (e.g. @Google) (*Forbergskog & Blom, 2013*). So, by tagging a dollar ("S") followed by a stock abbreviation at the end of their message, Twitter users refer to a company ticker. By the way, as all members of the microblogging family, Twitter has a length restriction. Indeed, a tweet is limited to 140 characters.

Based on surveys from We Are Social and Hootsuite (2019), we were able to determine that Twitter is the twelfth social media in the world in terms of active users with 326 million subscribers *(appendix 1).* In comparison, social media are used by 3.484 billion people worldwide with a nine percent increase each year (*Chaffey, 2019*) and 38 million in France. As exhibited in appendix 3, Twitter experienced a blazing growth in the first years of 2010 when it decupled its community but has known a stagnation since 2015. Most of worldwide Twitter users are between 25 and 34 years old (31%) but the distribution is quite homogeneous between 18 and 49 years old (from 21% to 31%) (*Statista, 2019*) *(appendix 2).*

Despite its ranking at the twelfth spot, Twitter is one of the most analysed social media with Google for financial outcomes. Certainly because there are appreciated by investors. Now, social networks specifically dedicated to investors exist. For example, Stockwits or Scutify. But these financial microblogs integrate Twitter which reflects the importance of this platform for equity participants. At the opposite of the traditional microblogging form which requires users to pursue information, Twitter pushes news to them through notifications for example. That change makes Twitter a distinctive microblog and allows it to attract new customers.

Its popularity can also be explained by the quantity of information available on its platform. The diversity of news available on social media, as explained in the beginning of

[2] AAPL being the ticker for Apple.

this section, is also one of the characteristics of Twitter. Especially since the agreement from the SEC in 2013 to companies to use social media for official disclosures. Thus, earning announcements, product launches or diverse statements are published on their Twitter official accounts. Their CEO's accounts can also be used as official channels if shareholders have been beforehand warned of this communication plan. In that respect, security and privacy remains prerequisites in Twitter. In 2013, the website recognized that approximately 250 000 user accounts were hacked. Similarly, Facebook faced a scandal last year related to data protection: the Cambridge Analytica Scandal. It affected its reputation and was reflected in its stock price. Therefore, privacy leakage as well as security breach influence the consumer behaviour on social media and should be taken seriously.

Consequently, User-Generated Contents have evolved throughout the years under many formats. From social media to microblogs they have changed but always kept the instantaneity, popularity, fast-growing and ease of use that made the UGC success. In all their forms, they are worthwhile data sources both in terms of quantity and quality. However, their contribution to the stock market is not yet defined. Moreover, users in their read of these data are subject to interpretations. Focusing on the behavioural finance, the following section will expose a summary of its inherent concepts. Some reminders of the Efficient Market Hypothesis will be given but this theory will not be exposed completely as it is not the core of our study. So, the point of the next part is to understand how investors' behaviour can explain the financial markets through the behavioural finance.

## Section III: Finance theories

According to Brunswick's Digital Investor Survey (*Manson, 2019*), 98 percent of investors use digital sources to search for information and 88% of them base their investment decisions (sell or buy) on the online data they found. This number is evolving throughout years as it reached only 41 percent in 2015. Therefore, insights about how investors gather and treat information is required not to mention the biases affecting publishers and readers.

Investors – fund managers, traders, institutional investors but also men in the street - evaluate financial assets based on macro- and microeconomic information from both numerical and textual sources. As explained in the previous section, UGC are one of the sources for this information. In the beginning, the literature was focusing on the association between the stock market and fundamental economic criteria and especially on how explaining the former with the latter. This emphasis on a macroeconomic aspect was followed by an analysis of micro-figures and then by an emotional computation. The emotional feature has seen the emerging of the behavioural finance. Its further development was the reaction to the statistical analysis of emotional data. It has revolutionized the market study by opposing itself to the Efficient Market Hypothesis (i.e. EMH). According to the EMH, by definition, unpredictable news creates random and unforeseeable stock price movements which are referenced as "random walk". However, the randomness is not synonym of irrationality as investors behave rationally in the case of new information. The behavioural finance changes that idea.

At all events, there are two types of investors: arbitrageurs and noise traders. The first category is not prone to sentiment while the second one is subject to emotions and non-essential information and thus tend to over- or underreact to news. Following the Efficient Market Hypothesis, all market participants do not need to be rational, but their behaviour follows a normal distribution. It also states that noise traders influence stock prices only for a short period of time as rational arbitrageurs bring prices back to the market equilibrium after exogenous shocks (*Fama, 1965*). At the opposite, the behavioural finance hypothesis argues that the power of arbitrageurs is limited due to short selling constraints, trading risks, positive feedback strategies, etc (*Bodie, Kane & Marcus, 2014; De Long, Shleifer, Summers & Waldmann.,1990*). So, with a limited arbitrage, sentiments can influence share prices by creating volatility. This is the reason why it is

difficult to outsmart the market. An example of this assumption is the market bubble of 2000 in which people were buying stocks even though fundamental analyses showed them that they were overvalued (*Bodie et al., 2014*). In other words, stock prices give a signal to investors which causes an increase or decrease of the market volume depending on the situation.

So, according to the behavioural finance, idea on which our study is based, investors even with accurate market information make their investment decisions based on their emotions, knowledges and beliefs leading to inconsistent decisions (*Bodie et al., 2014*). That concept is endorsed by behavioural biases: even with an accurate information process, people tend to make non-rational decisions. First, investors choices are affected by the frame of possibilities offered. Depending on how they are presented (in terms of losses or gains), people become risk averse or seeking. Secondly, they make different mental accounting of their investments based on their goals. For instance, an investment dedicated to children's education is considered differently than another one. Another example is that risks are more likely to be taken using the gains previously realized. Investors also avoid regrets: as they blame themselves more when an unconventional investment turns out badly, they prefer choosing traditional stocks. Similarly, they may decide to invest in companies that create a higher affect in their perception. For example, putting money in corporations socially responsible, with good working conditions or popular products.

Moreover, investors do not process information correctly and thus do not calculate right future rates of return. That idea is supported by experiments conducted by Kahneman and Tversky (*1972 and 1973*). They demonstrated that people when making forecasts give a higher weight to recent events in comparison to previous ones. Besides, the overconfidence of investors towards their beliefs and abilities impact their decision making, especially for men rather than women. In connection to this bias, investors show a conservatism tendency towards their beliefs which causes a mitigation of external signals. They tend to underreact to public information but to overreact to the private one (*Daniel, Hirshleifer and Subrahmanyam, 2002*). Plus, they tend to believe that a small sample is as representative as a bigger one leading to an extrapolation of patterns too quickly. On the contrary, a perverse effect link to the Wisdom of Crowd is the "information cascades". It means people ignoring their private information and opinions to blindly

comply with the crowd. Information cascades lead to private data loss and can occur even if the information provided by the crowd is not correct. Moreover, following the assumption that experts have more knowledges and skills to pick stocks, people follow their opinions despite the "Wisdom of Crowd" effect.

In the same idea, the "investor recognition hypothesis" (*Merton, 1987*) and "price pressure hypothesis" (*Barber & Odean, 2008*) stress the effect of awareness when it comes to information detection. The first one states that when a company becomes more famous, it increases public attention, which gives more information to investors who do not have investments in it. Therefore, some of them may buy stocks of that company (*Merton, 1987*). Then, the "price pressure hypothesis" or "attention theory" affirms that as investors do not have enough time or resources to evaluate individually each stock available, they buy stocks that attracts their attention (*Barber & Odean, 2008*).

Next, the behavioural finance and neurofinance are trying to connect investors' emotions to their trading behaviour. One of the outcomes was the fact that investors tend to value losses more heavily than gains (*Tversky & Kahneman, 1991*). Another one was the Affect Infusion Model (AIM) which affirms that people in positive mood states rely on positive signs to make decisions. This is why they associate risks positively in opposition with people in negative mood. They study mood states as a proxy for investors' sentiments.

Their analysis of mood states takes a turn with the arrival of social media. Traditionally, public's sentiments are often computed through surveys such as the AAII Sentiment Survey[3] or are based on market sentiment indexes like the Fear & Greed Index. But this type of investigation is expensive, time-consuming and can be biased because of the human dimension among surveys (*Mao, Counts & Bollen, 2011*). Otherwise, a company sentiment can be evaluated using UGC such as stock message boards (e.g. Antweiler & Frank, 2004), consumer reviews (e.g. Tirunillai & Tellis, 2012) or Twitter (e.g. Zhang, Fuehres & Gloor, 2010; Mao et al., 2011*).* That is already in application today with Bloomberg which added a sentiment analysis derived from Twitter to one of its tools[4].

[3] The results of this survey can be found here : https://www.aaii.com/sentimentsurvey/sent_results

[4] Bloomberg added a sentiment analysis derived from Twitter to its EDF (Event-Driven Feeds), a tool already including data from headlines, financial signals, global economic indicators, etc. On a daily basis, company can get an idea of mood states through this tool. The platform has started to integrate tweets in its functionalities since 2013 (*Bloomberg, 2018*).

Nowadays with social media, traders can investigate the effect of mood states on stock prices in real-time basis due to the instantaneity of messages. Plus, the impressive amount of data collected on this environment redefines the possibilities of study. What is more, the ease of use of social media has deeper impacted the researchers' work. Some years ago, they had to analyse messages manually to determine whether they were positive, negative or neutral. Now, with the social media development, stock prediction communities have grown and researchers may even not to interpret messages as they can formally include advisors' actions: the sell, buy or hold tips. Thus, the advice signification is clearly identified through whether the writer actually buys or sells the stock.

In addition to the behavioural finance, market anomalies are price distortions contradicting the EMH hypothesis as all information seems not to be reflected in stock prices (*Fama, 1970*)[5]. Investors by taking into account these anomalies can elaborate offering/bidding strategies and make profit. The inefficiencies can be classified as calendar, technical or fundamental. The first class will not be discussed in this master's thesis while the second and third ones, as they will be mentioned, need to be reminded. The second category includes the well-known Momentum effect (i.e. a short-term amplification of price movements[6]). The most famous anomalies of the last class are the Price-to-Earnings (i.e. a higher return for stocks with low price-to-earnings because of their underestimation) and the Size effect (i.e. small firms being likely to generate abnormal returns because they tend to grow faster).

Moreover, information leakage, meaning news release to a small group of people before the official announcement, can happen and interfere with the stock market. In the case of good (bad) news, the stock return will increase (decrease) before the official release[7]. To some extent, recommendations may be influenced by personal interests. For

[5] According to the EMH, markets are « informationally efficient » meaning that all information is publicly available at a current time and that in the case of price distortions, market forces will correct them.

[6] If a stock price is rising (falling), according to the momentum effect, it will keep moving up (down) during a short-term period. One explanation may be the underreaction of investors to new information.

[7] It depends on the EMH form on which we stand. As a reminder, the 3 types of Efficient Market Hypotheses are (Bodie et al., 2014):

- The weak form which states that stock prices reflect all information coming from the market (past data, trading volume, etc).
- The semi-strong form which asserts that publicly traded information such as stock prices, firm's product line, management, balance sheet or earnings forecasts is reflected instantaneously in stock prices.
- The strong form which affirms that public as well as private data are expressed in prices.

Therefore, information leakage only stands in the weak or semi-strong forms as a distinction has to be done between private and public information.

example, trying to boost the price of a stock after having bought it or posting recommendations about the company for which you work in the hope to boost its stock price. These are illegal price manipulations for which authorities are severe as in the example of Elon Musk versus SEC. Last year, Elon Musk hit the headlines for what the SEC considered as price manipulation. It is difficult to know if he is doing information leakage at a large scale or if he simply lied in order to positively impact its company stock price. In any cases, the SEC gave him a fine of 20 million dollars. Therefore, taking a careful look at online posts and their sometimes hidden messages in all circumstances is essential.

In consequence, the behavioural finance tries to understand investors' attitude towards information both in respect of the news treatment and reaction to it. The former concerns the emotions, beliefs or knowledges that bias the news analysis. Mood states are investigated for several years now and social media are bringing a new approach to this mood state assessment. But irrational traders' behaviour is difficult to understand and predict. In that respect, several assumptions exist and some of them were described in this section. As the activity of noise traders cannot be completely mitigated by arbitrageurs, excess volatility is created in the market and impact stock prices.

With clear definitions of the media used in this study and their characteristics (*section II*) and a reminder of financial theories (*section III*) in mind, we are now moving to a literature review (*section IV*). It establishes the link between User-Generated Contents and their financial aspects as well as consequences for the stock market. All of these contains a behavioural aspect referring to a sentiment or mood state analysis.

## Section IV: Literature review

In the literature history, the stock market was analysed in many aspects. First, several relationships were established between the financial market and macroeconomic variables but they are out of scope in this study. Then, with the apparition of the behavioural finance, the analysis of headlines in a cognitive point of view came out. This was followed by the emergence of User-Generated Contents and its relationship with the stock market. Different types of UGC were studied among the years. This evolution of the literature to Twitter will be briefly retraced hereafter to come to our subject.

But first, besides studies on information supply, our core topic covered after this paragraph, the demand is concisely exposed here because studies have been conducted on both demand and supply of information and have revealed that they are positively correlated *(Vlastakis & Markellos, 2012).* The demand was analysed based on both ticker symbol *(*e.g. Da, Engelberg & Gao, 2011*)* and company name (e.g. Vlastakis & Markellos, 2012; Takeda, & Wakao, 2014) searches on Google. The demand of information impacts not only individual stocks but also the overall market as it is positively correlated with volatility and trading volume in general (*Vlastakis & Markellos, 2012; Takeda & Wakao, 2014; Mao et al., 2011*). In other words, when the volatility is high, the number of searches increases. It goes up also in periods of higher returns and with higher investors' risk aversion (*Vlastakis & Markellos, 2012*). In fact, information demand is positively correlated with the level of investors' risk aversion: if they are more risk averse, they ask for further information. Consequently, according to Vlastakis and Markellos (2012), Takeda and Wakao, (2014) and other scientists, the search intensity is positively correlated with stock returns.

What is more, Mao et al. (2011) by computing the Tweet volumes of financial search terms (TV-FST) reveals a correlation between the TV-FST and the stock market. In fact, they discovered that Twitter is a better financial market predictor than the Google search volume intensity even though they are correlated.

Besides the information demand, the first stress in the supply of information after macroeconomic data was put on headlines. According to Takahashi, S., Takahashi, M., Takahashi, H. and Tsuda (2007), headlines addressing a subject positively[8] create a higher

[8] They represented 59 percent of their sample.

excess return both in terms of cumulative and simple returns. The reaction to news is the biggest on the announcement day and the influence of news is overlapped as it is repeated. Hence, Birz and Lott Jr (2011) have not used articles published after information releases since the information they provided was old. They have analysed[9] four subjects of headlines - GDP, unemployment, retail sales and durable goods – and classified them as positive, negative, neutral or mixed. They concluded that news about GDP and unemployment rate affects stock returns. Else, Daniel et al. (2002) have found that stock return movements are higher for bad rather than good news as a negative drift happen up to twelve months. In fact, prices tend to reflect bad information more slowly showing an underreaction to this type of information.

Afterwards, with the evolution of technology and information channels, researches have been carried out on multiple areas in order to detect the impact of media on the economy: demand/sales of products, stock market performance, brand awareness[10], marketing, etc. They have shown the effect of company name appearances on stock prices and specifically how it can be predicted by the former. According to an experiment by Fehle, Tsyplakov and Zdorovtsov (2005) on a sample over the 1969 to 2001 period, companies, whose ads are broadcast during the Super Bowl, see their stock prices increase. A similar consequence was demonstrated for CEO interviews on CNBC *(Kim & Meschke, 2011)*. In another theme, music sales can be forecasted using MySpace (*Dhar & Chang, 2009*) and the same stands for movie sales using the Hollywood Stock Exchange (*Skiera & Spann, 2003*) and book sales through Amazon (*Chevalier & Mayzlin, 2006*). Furthermore, the impact of professional publications, such as reviews of products *(Tellis & Johnson, 2007)* and printed news *(Tetlock, 2007)* on the stock market were demonstrated.

Then, after traditional media, the advent of social media has altered investment, trade, acquisition as well as information sharing because they represent a continuous source of information. For example, messages on Yahoo! Finance and Raging Bull were analysed among others by Antweiler and Frank (2004) as well as Tumarkin and Whitelaw (2001) and their impact on stock returns and trading volume was illustrated.

[9] Based on a headline analysis made by Lott and Hassett (2006).

[10] UGC has even become more important than the conventional *marketing (Trusov, Bucklin & Pauwels., 2009)* and therefore may be used to get information about the brand strength.

User-Generated Contents in their globality are linked to the stock market *(Antweiler & Frank, 2004).* They demonstrate a slight correlation with abnormal returns ("AR"), risk and trading volume *(Tirunillai & Tellis, 2012).* Hill and Ready-Campbell (2011) even pointed out that following advices on UGC allows to outperform the S&P500. Nevertheless, according to Tirunillai and Tellis (2012), negative UGC have an impact on AR and trading volume while the positive ones have not.
More recently, Twitter has seen studies computing its impact not only for the commodity market but also for currency rate fluctuations (*Ciftci & Ozturk, 2015*) or box office revenues (*Asur & Huberman, 2010*).

Researchers in a behavioural approach have computed mood states on microblogs in order to understand their impact on the stock market. They reveal that investors in a good mood are more willing to invest in stocks and to take more risks as they tend to be more optimistic in regard to uncertain future events. Thus, emotions impact the stock return (*Johnson & Tversky, 1983*). Consequently, risks taken by individuals depend on their mood states. For instance, the weather effect (*Hirshleifer & Shumway, 2003*), the roles of seasonal depression (*Kamtra, Kramer, & Levi, 2003*), sport events (*Edmans, Garcia & Norli, 2007*) or air pollution (*Levy & Yagil, 2011*) affect investors' mood state which in turn influences the stock market.

As stated in Tumarkin and Whitelaw (2001), investors' positive[11] opinion about Internet firms on Raging Bull is correlated with abnormal returns on the event day and is linked with prior stock prices. According to Antweiler and Frank (2004), there is a significant negative correlation between the number of postings on a day and the next day stock returns. For the trading volume, there is no link to sentiments prior to the event unlike on its day and day + 1 on which an increase in trading volume was noticed (*Tumarkin & Whitelaw, 2001*). A correlation was also established by Antweiler and Frank (2004) and Das and Chen (2007) between message board activity and volatility, trading volume and bullishness.

In addition to these computations in microblogs, Twitter is a good predictor of survey sentiment indicators showing its relevance as an alternative when it comes to mood states analysis *(Oliveira, Cortez & Areal, 2017).* This confirms our theory about the

[11] Results were non-significant for negative opinions.

evolution of the emotion investigation in phase with the development of social media explained in the third section. Thus, several researches have proven the existence of a correlation between sentiments derived from Twitter, principal market indicators (Dow Jones, S&P500, NSDAQ) and stock returns. In fact, the correlation between Twitter sentiment and returns is lower for single stocks than for indexes (Sprenger, Tumasjan, Sandner and Welpe, 2014; Ranco, Aleksovski, Caldarelli, Grčar & Mozetic, 2015). These sentiments in tweets include the feeling of fear and hope (*Zhang et al ,2010*), the percentage of bullish tweets represented by the Twitter Investor Sentiment (i.e. TIS) (Mao et al., 2011), as well as other indicators of mood states defined by Sprenger et al. (2014), Ranco et al. (2015), Smailovic, Grcar and Znidarsic (2012)[12].

Besides, TIS has a negative correlation with VIX. This is coherent since an increase in the bullish feeling is linked with a decrease in investors' fear, represented by the VIX (Mao et al., 2011). At a day level, a high volatility causes a greater message volume exchange meaning that in high volatility time, investors want to consult their peers. Plus, a high volatility induces disagreement between investors which in turn leads to an increase in the trading volume. In an intraday perspective, there is a feedback effect between volatility and sentiment as well as tweet volume (*Behrendt & Schmidt, 2018)*. Thus, message volume helps predicting trading volume and vice versa[13] (*Sprenger et al., 2014*).

However, in microblogs, message board activity (i.e. the number of messages and weighted opinions) cannot help predicting future stock price movements and trading volumes because opinion changes are not the cause of trading volume movements after the event day (*Tumarkin & Whitelaw, 2001; Antweiler & Frank, 2004; Das & Chen, 2007*). This is consistent with the EMH, as all information is supposed to be already reflected in the stock market, and contradicts the idea of stock price manipulation through online postings.

Nevertheless, the study of Tumarkin and Whitelaw (2001) can be criticized because of its sample only composed of companies from the Internet sector. This choice was made in an attempt to get the most representative model: technological companies

[12] For instance, they use financial tweets about Apple to identify events and predict Apple stock price movements.
[13] But the effect is stronger from trading volume to message volume.

were more likely to be impacted by opinions on Internet forums. However, this idea, potentially true in 2001, is not likely to be correct anymore as Internet discussions have spread and no longer concern specifically the technological sector. In contrast to their study, Wysocki (1999) found that the posting volume on Yahoo!Finance predicts changes in trading volume and returns of the next day. Similarly, in a more recent research from Oh and Sheng (2011) on Stocktwits.com[14], microblogging sentiments were proven to be predictors of simple and market adjusted returns.

Twitter sentiment is also a good predictor for market returns (Mao et al., 2011; Ranco et al., 2015; Smailovic et al., 2012; Pagolu, Nayan Reddy Challa, Panda & Majhi, 2016[15]). But according to Bing, Keith, Carol (2014), stock price prediction based on tweets depends on the type of industry. On the contrary, according to Sprenger et al. (2014), bullishness cannot be used to predict market returns. But the latter affects the former. On average even if some tweet features can be used to predict market features, the effect is stronger for market features impacting tweet features (Sprenger et al., 2014).

This link can be explained by a momentum: people buying stocks having already known an increasing performance prior to the post (*Dewally, 2003*), and thus increasing again its performance. On second thoughts, Das and Chen (2007) also found a predictive power from stock returns on sentiments. It suggests that bloggers extrapolate from prior returns when writing their messages. Previously, a momentum effect was as well found by Daniel et al. (2002) on headlines.

Furthermore, with the increasing speed of trading, Twitter has been analysed in an intraday perspective (*Behrendt & Schmidt, 2018*). Indeed, as explained in the second section, Twitter allows investigations at a higher frequency than traditional blogs. However, including Twitter sentiment and activity into intraday prediction models do not seem to improve their performance (*Behrendt & Schmidt, 2018*).

In any case, the assumed power of Twitter on the market at a daily frequency is impacted by the "mood contagion". It is referred in the study of Nofer (2014) by the number of Twitter followers and gives the idea that mood states can spread among users. Supposing someone has a higher number of followers, he will exert a larger influence on

[14] A microblogging platform with a good reputation and gathering only postings about stocks. It also includes tweets.
[15] Analysis based on Microsoft over 1 year.

the community. By including the number of followers in a study, we can see if the predictive ability of mood states is improved when considering social interaction between users (*Nofer, 2014*). The author found no prediction of stock returns by the only Twitter mood states (i.e. without the follower number) unlike the follower-weighted mood states (i.e. by integrating the number of user followers). That asserts the importance of mood contagion. So, the popularity of a user impacts stock prices as a bigger number of potential investors will follow its tips (*Nofer, 2014*). This can be put in link with the Wisdom of Crowd effect explained in the section II. As a proof, in a study from the same author, investments based on online user tips reach a return on average 0.59 percent higher annually than decisions from professionals.

On the top of that, Twitter users providing above average investment advices are retweeted more often and have more followers which amplifies their share of voice (*Sprenger et al., 2014*). Therefore, Twitter users have an incentive to share valuable information as they want to be followed, re-tweeted and increase their popularity. The popularity of a user is defined by its number of retweets, mentions and followership according to Cha, Haddadi, Benevenuto and Gummardi (2010). In link with the impact of followership, the term "sidelined investors" was created to describe investors trading a stock only because they have heard others were trading it (*Sprenger et al., 2014*). By contrast, the impact of fake news endangers this theory providing that bad quality tweets can spread rapidly. However, through the followership and its automatic improvement of the quality of content, Twitter gives a major value added in comparison to other microblogs. Following that idea of above average investment advices, Groß-Klußmann, König & Ebner (2019) separate Twitter users based on the quality[16] of their tweets and their number of followers. The ones define as "experts" are mainly responsible for the link between Twitter sentiment and the market (*Groß-Klußmann et al., 2019*). According to these authors, based on experts' sentiments from tweets and the momentum effect described earlier, profitable investment strategies may be established for trading futures.

By the way, there is a positive correlation between the market capitalization as well as the number of messages posted the previous day and the daily average message

[16] They define quality as a finance related content without noise.

postings in microblogs. It means that a bigger firm gets a higher number of postings and reveals an autocorrelation in postings. (*Wysocki, 1999; Tumarkin & Whitelaw, 2001*).

Besides that, according to Sprenger et al. (2014) and their study on tweets, the average cumulative abnormal returns ("CAR") is abnormally increasing after positive sentiment non-earning announcement events while it decreases after the negative ones. On second thoughts, according to Takahashi et al. (2007) it is possible to create excess return by publishing news. Therefore, the idea of price manipulation, as addressed in the second section with the example of Elon Musk, is meaningful. On the top of that, according to Ranco et al. (2015), differences in cumulative abnormal returns before events are not significant while the Sprenger et al. (2014)'s study shows significant CAR changes before events. By showing abnormal returns before official announcements, these authors point out potential frauds to insider trading rules. They reinforce the information leakage theory already demonstrated in the past for earnings announcements with Chambers & Penman (1984).

Consequently, a link exists between the posting volume on microblogs including on Twitter and the stock market movements but the predictive power of the former towards the latter does not achieve unanimity. The same stands for the post sentiments and stock market. Therefore, there are still divergences between studies about the direction and predictiveness of the relation behind tweets. Nevertheless, the consensus of the correlation between sentiments in tweets and returns has already created applications. A hedge fund based on the theory of Mao et al. (2011) uses Twitter to get the day mood and achieve better results than the market or other hedge funds. An algorithm inside the hedge fund takes the number of "calm"[17] but also "alert" or "happy"[18] words to invest (*Greenfield, 2011; Kelly, 2011*). Unfortunately, it closed after one month of use for obscure reasons.

Finally, even though a part of the information embedded in Twitter can be seen as irrelevant noise, the value of tweets relies in its volume of users and messages capturing the Wisdom of Crowd effect. The contagion of emotions between users amplifies this effect.

[17] A fall in the "calm" sentiment leading to a drop in the Dow Jones index.
[18] The predictive power of the words linked to « alert » or « happy » is inferior.

# Section V: Methodology

Several studies have demonstrated that Twitter is connected to lots of settings from the cinema to the currency rate, the stock market including. For some variables, it has also proven to be a good predictor. In this master's thesis, we look for answering the question "What is the impact of Twitter activity and investor's sentiments on the stock market in Europe?". Specifically, we want to examine whether the Twitter activity and sentiments embedded can be used to explain stock returns and trading volume. Some of our assumptions have already been asserted for the US market but we want to expand it to the European trade and to add new orientations by including corporate features. In the first part of this section, for each of our ideas, we define a null-hypothesis and give the reasoning from which it is derived. The alternative ones are not mentioned but consist in the negation of the corresponding null-hypothesis. Then, we present our sample and the ways we gathered the different data types.

## Hypotheses

Firstly, people may post tweets about companies on which they have traded whether to share trading advice or simply to talk about their activities. These tweets can affect the stock trading volume[19] because of a mass effect. In fact, if people hear about a good deal or a tip, they may decide to follow it. These people are referred as sidelined investors and were already defined in the literature review. Thus, the same choice taken by an important number of people at the same time lead to a mass effect and by doing so can affect the global market. Moreover, since the moment a wisdom does the same action, people, by relying on the Wisdom of Crowd effect, can believe that the action is a good decision and do the same. So, the phenomenon is again amplified. These are the reasons why we expect the Twitter activity represented by the volume of tweets to impact the trading volume.

$H_1$: Tweet volume helps explaining trading volume.

Besides Twitter activity, tweet sentiments are viewed as a proxy for the market mood state. They are correlated to stock returns as demonstrated by previous studies in the literature review but the direction of the link does not achieve unanimity. A similar information can generate different emotions depending on the reader's perception and

[19] As demonstrated by Sprenger et al. (2014).

biases as explained in the third section. In order to get the feelings derived from tweets a sentiment analysis method must be applied. It consists on understanding the writer's opinion about a topic based on diverse tools explained in the next section. These feelings will be divided into three categories: positive, negative or neutral. The investor's perception at the reading of the tweet should be different depending on the message connotation as demonstrated in the third section in the behavioural finance part. If the investor's impression is different, his reaction should be too and so likewise the impact on the stock market. Stock market movements take in stock returns as well as trading volume. Thus, two hypotheses are derived from the sentiment analysis:

$H_2$: Positive sentiment helps explaining stock market movements.

$H_3$: Negative sentiment helps explaining stock market movements.

Moreover, based on the Nofer (2014)'s study, the popularity of a tweet and its publisher impacts stock returns. In fact, if a user shares a tweet with one's own community, it means he values it. Plus, if he answers a message, he shows his interest to the subject addressed[20]. In other words, the number of replies in a tweet may reveal its popularity both in a positive or negative way, depending on people's agreement or disagreement with the topic. Therefore, the number of replies and retweets of a message are proxies for a tweet popularity. If a message is popular, that is to say retweeted or replied, its spread increases and by doing so its potential impact on people and so on the stock market is multiplied. As for the second and third hypotheses, the effects on the stock market can be seen both in terms of returns and trading volume. So, our fourth hypothesis is:

$H_4$: Number of replies and retweets impacts the stock market movements.

Our next assumptions concern the corporate point of view. On the one hand, the ownership of companies and its link with Twitter. Companies with a higher percentage of ownership by retail investors should experience a greater impact of Twitter activity on their stock movements. Individuals should exchange more information through Twitter than institutions. In fact, according to Sprenger et al. (2014), Twitter reflects more the

[20] The number of replies is not a direct measure of the quality of messages as an answer can be published to disagree with the original tweet. But it creates debate based on which the initial quality can show an improvement. However, we should interpret the derived results with caution.

activity of private traders than institutional ones. Thus, if the percentage of individual owners is higher in a company, it should see its Twitter activity increases likewise the impact of Twitter on its stocks. The fifth hypothesis focuses on the first link between retailers and Twitter:

$H_5$: The percentage of retail investors impacts Twitter activity.

On the other hand, the market capitalization of corporations is studied. Our next assumption relies on the idea that social media are platforms for news sharing and scandals spread. People may prefer to talk about companies they know well and of which they are sure others are aware. So, we suggest that the volume of tweets is higher for bigger firms as they are more famous. We measure the size of company by its market capitalization.

$H_6$: Companies' market capitalization influences Twitter activity.

As a result, our hypotheses are of two types: the link between Twitter parameters and the stock market but also the relationship between corporate features and Twitter activity. Now that our assumptions have been defined, we are moving to our dataset presentation.

## Sample

In this study, we have chosen to focus on the western European market as, so far as we are concerned, studies about it do not exist. They focus principally on S&P500 and Dow Jones and so on the American market. The merge of the Paris, Brussels and Amsterdam securities markets has created in 2000 the Euronext, the largest stock exchange in Europe now. Even if others have joined the association since that time, these three places remain important equity players in Europe. This is the reason why we have decided to focus on companies coming from the most important indexes of these markets: CAC 40, BEL 20 and AEX 25 respectively. Through corporations originated from Paris, Brussels and Amsterdam stock exchanges, we were able to get an insight within the European market and the Euronext, differentiating ourselves from previous studies.

As a reminder, CAC 40, AEX 25 and BEL 20 are all "free float market capitalization weighted index" (*Euronext, 2019*[21]*)* returning the performance of the largest shares in their respective stock exchanges in which they are the most widely used index. The CAC 40 was created in 1987 and is composed of 40 stocks (*appendix 5*) quarterly reviewed for a market capitalization of 1.627,1 billion euros in June 2019 (*Euronext, 2019*[22]) . In theory, an outgoing company is replaced by a member of the CAC Next 20 whereas twelve corporations have never left the index since its creation. It includes the well-known L'Oréal, LVMH or Michelin. Else, on the top of also quarterly interim reviews, the 25 elements composing the AEX 25 as well as the 20 components of the BEL 20 are fully revised annually in March. The French market remains the biggest one followed by the Dutch index with a global market capitalization of 671,7 billion euros in June 2019 (*Euronext, 2019*[23]) and the Belgian one with a market capitalization of 307 billion euros also in June 2019 (*Euronext, 2019*[24]). The elements of these indexes are shown in appendix 6 and 7.

[21] Based on the documents about the three indexes extracted from the Euronext website and that can be found at the links given below.

[22] The information comes from a document published by the Euronext and renewed frequently. The version on which we extracted this number is from June 2019. It can be downloaded here: https://live.euronext.com/en/product/indices/FR0003500008-XPAR/market-information

[23] The information comes from a document published by the Euronext and renewed frequently. The version on which we extracted this number is from June 2019. It can be downloaded here: https://live.euronext.com/en/product/indices/NL0000000107-XAMS/market-information

[24] The information comes from a document published by the Euronext and renewed frequently. The version on which we extracted this number is from June 2019. It can be downloaded here: https://live.euronext.com/en/product/indices/BE0389555039-XBRU/market-information

Besides its geographic location, another particularity of our sample is its large length of time. Indeed, it ranges from 2008 to 2018 for CAC 40 and AEX 25 companies and from 2009 to 2018 for BEL 20 entities, that to say ten years on average. This is a particularly extended period of time in comparison to similar studies on the US. Benefits deriving from it will be exposed in the next chapter.

As a consequence, the information included in our sample are spread in a long time period. Hence, our sample with all the restrictions explained in the data collection part hereafter contains 11.339 tweets. However, as some tweets speak about several stock tickers, they are in fact 11.169 different tweets with unique identifiers published by 4.287 distinct users.

Market data namely stock prices and volume trading for both individual stocks and indexes are exposed on each day out of the 10 years of our sample. Moreover, our information integrates each company market capitalization and percentage of retail investors for the same period as well as the industry and stock index of each corporation.

In the next part we will describe how we gathered all these information category by category.

## Data collection

The data categories we extracted include tweets, market information and corporate data as mentioned previously. For each of them we define their selection criteria as well as characteristics. We also report in this subsection the difficulties we met during the data collection.

Firstly, we began by retrieving Twitter data as it is the hearth of our subject. Twitter, as a database, is in opposition with investing platforms in respect of the quality of the information provided. These websites, as people often pay for accessing them, tend to include members with the most valuable opinions and personalities influential among the investing community. Therefore, subscribers are more willing to follow the advices on these websites. An example of this kind of platform is "Morningstar" on which the premium paying account allows you to get experts' trading tips *(Breed, n.d.).* On the contrary, Twitter being public, it integrates both precious a well as non-valuable opinions but their quantity is bigger. The platform with its huge database in terms of number of tweets and users represents a good medium to analyse the equity market. This is amplified by the advantage of Twitter to push news to users rather than to let them pursue it as explained in the second section. Nevertheless, that huge quantity of information to analyse creates a difficulty for our study. Fortunately, we succeeded in reducing the noise with one of the selection criteria of tweets. In fact, according to Da et al. (2011)[25], using the stock ticker to select information is preferable because the name of the company can be searched for other reasons than investment. In Twitter, the stock ticker is reflecting inside the cashtag already defined in the second section. Thus, filtering by company names instead of cashtags create issues. They concern the contents of messages, which may include non-investing information, the spellings of company names and the difficulty to compile all of them as well as the time constraint linked to the amount of information that would be collected. By gathering tweets with the cashtag, we are able to find the ones referring to a same stock more easily and thus to avoid noise.

After having chosen Twitter as our database and having defined a first restriction on our sample based on the cashtag, we still needed to clearly specify the type of information we were going to gather. In fact, it is difficult to know how investors interpret releases due to their biases (explained in the third section) but also which publications

[25] They used the stock ticker in their study on the Google Search Volume Intensity.

are relevant for stock price determination. According to the Efficient Market Hypothesis, only new information is considered by investors. But the difficulty remains to know which information is actually new[26]. Unfortunately, the same information can be interpreted differently by authors of articles. Consequently, different opinions can be reflected in postings: positive, negative or even neutral. Plus, the spread of an information increases as tweets are shared. Therefore, in our study we are taking into account all tweets and not only the ones integrating new information.

Now that the data type is defined, the data extraction can be processed. There are two different data types in Twitter: the historic and real time tweets. The first one is accessible through the "Twitter search API" and represents the data which has already been published. In the free version, the number of tweets extractable with a query is limited as well as the number of requests over a period of time. However, Twitter being now aware of the potential of its data, the free access to the API is restricted. The account request is checked by the website regulators and approved or not. The second tweet category can be instantaneously retrieved using "Twitter Streaming API" or "Twitter Firehose" at the moment they are published and up to seven days. The first method is free but limited in terms of tweet quantity while the second one is expensive, relying on Twitter independent providers but complete (i.e. delivering one hundred percent of the tweets needed). Our desire for this study was to perform an analysis based on a long time period and not on real-time basis tweets. Therefore, because of the restrained access to the Twitter search API in terms of time, number of tweets and cost, our database was provided by Hubert Naets[27] through one of our supervisor, Christophe Desagre.

As a result, the information collected in our Twitter database includes the tweet characteristics: text, author, time, date and location. It also integrates references such as an identification number, the "tweet_id", and the author's designation, the "user_id". They allow to trace each tweet and author specifically and individually. The language of the text is also given as well as the number of replies and retweets. At first, tweets were available in English, Spanish, Portuguese, Greek, Japanese, Turkish, Dutch and French. We have chosen to select only tweets written in English to match with the majority of text analysis

[26] Previous researches had calculated the economic surprise linked to releases. It is computed as the difference between the release and previous expectations measured by surveys. If the issue was unanticipated, the information was new.

[27] He uses an algorithm for the extraction of tweets.

methods developed for English words and to keep a guiding principle through the analysis. It avoids going from tweets in all languages in the volume study to only English tweets for the sentiment analysis. A check was performed on the database and some tweets defined as non-English were manually assigned to the English language and the unidentified languages were classified. We also retrieve the hour, day, month and year from the time and date which will be useful for further analyses.

Secondly, the Twitter database being prepared, we added the stock index and type of industry gathered from the Euronext website for each stock. The complete list of indexes and industries among them is available in appendixes 5 to 7. However, during our investigations, we noticed that some companies inside of our Twitter database are part of two indexes like Unibail or Galapagos. As the same element could not be present twice in our sample or attached to two indexes, we had to make a choice in the reference of the ticker. We based our decision on the ISIN of the stock. For instance, Unibail, available in both the CAC 40 and the AEX 25, has the ISIN "FR0013326246". The first two letters referencing to the country, we placed Unibail in the CAC 40 index. The only exception was ArcelorMittal. It is present in both the CAC 40 and the AEX 25 but its ISIN is "LU1598757687". The indicative "Luxembourg" being not helpful, we referred to the Euronext website which places ArcelorMittal in the Amsterdam Stock Exchange. Therefore, we put it into the AEX index. Finally, we ended up with thirty-eight different companies for the CAC 40, fifteen on in the AEX 25 and thirteen for the BEL 20.

Besides tickers present in several indexes, the difference between the total number of companies in our sample and their actual numbers in indexes is explained by the constraint in the data retrieval. For instance, L'Oréal (ticker "OR") was not available for the CAC 40, Barco (ticker "BAR") for the BEL 20 or ABN AMRO Group (ticker "ABN") for the AEX 25. Consequently, one company is missing from the CAC 40, eight from the AEX 25 and 5 from the BEL 20. Twitter prevented us from adding manually tweets about those companies to our database despite our attempts to collect them in- or directly. However, this obstacle guarantees the integrity of the data initially gathered.

Thirdly, the stock prices and trading volumes of each company were retrieved from Bloomberg and Yahoo! Finance on a daily basis depending on the time horizon of their corresponding indices in our sample. So, the information for CAC 40 and AEX 25 companies were retrieved from 2008 to 2018 whereas the time horizon was from 2009

to 2018 for BEL 20 corporations. These data were reprocessed to make them fit with tweet information already on hand because the ticker and date details needed to be of the same format and exact spelling for both matrices. The same was applied for the market prices and trading volumes of the third indexes because they will serve as control variables later on.

Quarterly, market capitalization and information concerning the ownership of each company were both retrieved on Orbis, based on the ISIN code of companies in order to assure the information quality. The first one was calculated by Orbis itself based on the data from financial statements at the end of each year. It is the number of shares multiplied by their value. This figure was systematically computed by Orbis from 2009 and therefore the quantity of data available for tickers in 2008 is limited. Except for that year and some other particular cases, the figures were globally complete. All the values were put in billion euros[28] since the euro was the currency in which the majority of the market capitalizations were given in Orbis[29]. As the market capitalization was only annually available, we defined the same number for each date inside a same year.

After that, the current ownership of all the companies inside our sample was collected. Unfortunately the past data was not available for downloading but a distinction was made between direct and total shares. Since indirect shares are often got via funds and do not confer the same rights as a direct ownership, we focused on the latter for our analysis. Inside it, we were able to separate all known investors into two categories on whether they were individuals or institutions. One exception remains for one type: public investors. They consist of people buying/selling shares on the stock exchange and for whom the identity is not known. So we were not able to classify them as individuals or institutions. This category represents 22,15 percent on average inside companies. By the way, individuals are defined here as employees, managers or more largely people with or without link to the company in opposition with institutions which are entities often exchanging large quantities of shares at the same type. Finally, we get the percentage of direct shares of both categories. However, as the data provided by Orbis are not fully

[28] First the data was left in million euros but the results of our analyses described later were not meaningful. Therefore we changed our initial dataset by putting it into billion euros.

[29] The changes were made based on the currency rates retrieved from the ECB website. The table comprising all the currency rates from 2008 to 2018 can be found here : https://www.ecb.europa.eu/stats/policy_and_exchange_rates/euro_reference_exchange_rates/html/index.en.html

complete, the total percentages of shares do not often reach one hundred percent but come close to it which allows us to draw conclusions anyway. For instance, the exact percentage of ownership for some groups was not exactly known and was approximated by a ">" by Orbis (e.g. ">30%"). These "bigger than" were replaced by an "equal to" in our study. So all our ownership percentages must be seen as "at least" numbers in our analysis but remain close to the reality.

At that moment, all the necessary data were on hand and the analyses could begin. They were performed using Excel, R and RStudio. The first one contains our database both for tweets, market as well as corporate data and was used to get familiar with them and make the pre-processing. While the other ones were chosen because they offer accessible statistical functions. They are free open source software already used by researchers, companies and students all around the world. Thanks to their packages[30] allowing lots of additional functionalities, their flexibility makes them the convenient tools for our work. We used them for the data processing, to get the sentiment value we were looking for, to implement all the constraints explained in the following section, to merge tweets, market and corporate information and to perform the empirical analyses.

[30] A list of the packages used is available in the appendix 18.

# Section VI: Data description

In this section, we first describe our dataset with its main distributions. Then we give the processing we realized to make our information fits our constraints and which provides us new features necessary for our further analyses and displayed in the "variables" subsection.

## Data set

As explained in the sample part, the large length of time on which our data were retrieved allows us to benefit from different phases. On the one hand, crises and non-difficult periods are included in our dataset. The crises integrated are mainly the Financial crisis of 2008 and the European Sovereign Debt crisis. On the other hand, our sample includes moments during which cashtags even if they existed were not yet widely spread and others during which they were. So, it brings an insight about the evolution of this tool in Twitter. Furthermore, it goes back two years after Twitter creation and ends up just one year ago allowing to see the evolution of the platform over the years. In consequence, this sample is well diversified in a time horizon aspect. The evolution of tweets over the years is represented in figure 1 below[31]. We see that the number of messages in our sample quintuple from the years 2010 to 2018 moving from around four hundred to 2.253 tweets. This is linked with the evolution of Twitter popularity in the same decade (*appendix 3*). The emphasize put in recent years differentiates us from the most famous studies pointing their sample around 2010-2012 (e.g. Sprenger et al., 2014, Ranco et al., 2015).

[31] The distribution of the yearly tweet volume per stock index is also given in appendix 17

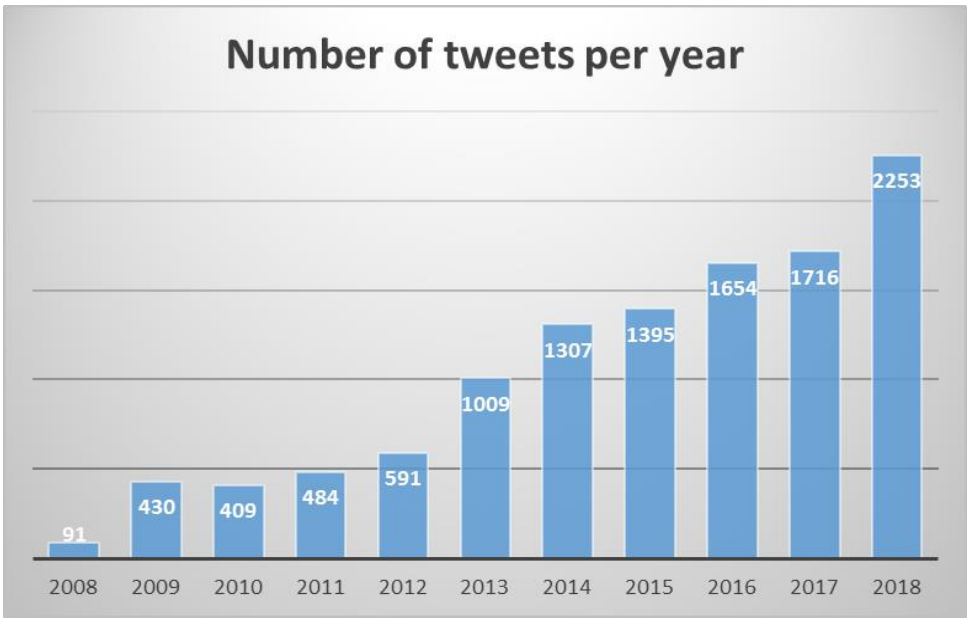



*Figure 2: Yearly tweet volume*

Then, we completed the repartitions of our database in terms of the industry, index, tweet volume and time. The results are shown from the appendixes 8 to 10 as well as below. Corporations in our sample are very well diversified. Their industrial distribution ranges between 2 to 8 percent meaning that no sector is more represented than another. These numbers were computed based on the composition of our sample and the sectorial description extracted from the Euronext website and available from appendix 5 to 7. However, in the social media world we may imagine that some areas are more impacted by consumers' opinions on Internet and therefore we will control for this effect in our further analyses. The total number of tweets per each stock is given in a table in appendix 9. However, in a wish to increase the understandability of the data, the top and bottom five stocks in terms of tweet volume are represented below in figures 3 and 4:


*Figure 3: Top 5 tweet volume stocks*


*Figure 4: Bottom 5 tweet volume stocks*

Thus, Airbus (ticker "AIR"), AXA ("CS"), Vinci ("DG"), Aegon ("AGN") and Sanofi-Aventis ("SAN") are the most talked stocks while ABInbev ("ABI"), Cofinimmo ("COFB"), Colruyt ("COLR"), Group Bruxelles Lambert ("GBLB") and Ageas ("AGS") are in the bottom five. Only the top four comes apart with at least around four hundred messages. After them, a tendency appears around 300-350 tweets followed by a decline. The bottom tickers consist of companies with less than 10 messages. It is obvious that with a so weak level these corporations could hardly individually be used to draw successful conclusions. But it is the total amount of tweets in our sample that allows us to give efficient results. By the way, four out of five of the top stocks are part of the CAC 40 whereas the five bottom ones come from the BEL 20. This shows a difference in terms of tweet quantity between indexes. However, our sample is diversified in respect of numbers of tweets per sector. This classification seems to contradict the Tumarkin and Whitelaw (2001)'s idea of some

sectors being more impacted by social media. In fact, it strengthens our initial impression that social media affect every company in the present world and that they cannot be ignored. Conversely, four out of the five most talked companies are located in the top 25 percent of market capitalization of the CAC 40, already our biggest index among the three whereas bottom stocks are from the BEL 20, a smaller index in a market cap perspective compared to the CAC 40.

Besides its diversification in terms of sectors, companies and tweet volumes, our sample should also be examined in a time perspective. As a matter of fact all the tweets in our sample are defined at the Greenwich Mean Time (i.e. "GMT") hour wherever and whenever they were published. The three stock exchanges on which our indexes are traded are placed under the Central European Time Zone. This time is usually one hour ahead of the Greenwich Mean Time (GMT+1). But in a wish to save daylight, the time is shifted forward one extra hour during the summer period leading to a new time period (GMT+2) and an extra gap toward the GMT time value. The Brussels and Paris stock exchanges close at 17h30 while the Amsterdam market closing time is ten minutes later but they both open their gates at 9h00. So, in a GMT equivalent, the trading hours start at 7h or 8h and end at 15h30[32] or 16h30[33] depending on the season[34]. By taking into account these hour changes, we notice that almost half of tweets are published during trading hours (*figure 5*). However, this number is lower than what previous studies had recorded. Moreover, the U-shape characteristic of the information exchange during trading hours is not present. Instead, we notice an increase in the number of tweets at openings, followed by a peak around the beginning of the afternoon and a stagnation after closings. Message exchanges still performing after closing hours are explained by after work discussions about already done trades. As a result, our shape of tweet volume during trading hours looks like a "S".

The divergences in respect of tweet distribution between our report and previous studies can be explained by the randomness of the tweets selected. Indeed, with the new policy of Twitter only a fraction of the whole volume of messages can be transmitted. When a user launches a request, the tweets he gets are function of their popularity (for a

[32] 15h40 for Amsterdam.
[33] 16h40 for Amsterdam.
[34] And the location of the market for the 10 additional minutes in Amsterdam.

given part) but are also randomly distributed (for a major part)[35]. Thus, on the contrary of previous studies, our sample seems to present a different repartition. In consequence, studies on Twitter that do not rely on independent and expensive tweet providers (as they are the only one to offer 100 percent coverage) are subject to the randomness of the tweets selected. This will be one of the limit of our study explained in the ninth chapter.

Furthermore, the month distribution represented in appendix 10 tends to confirm this idea of a random distribution as a peak in tweet volume is recorded in July without any explanation, the other months showing a stable number of messages.

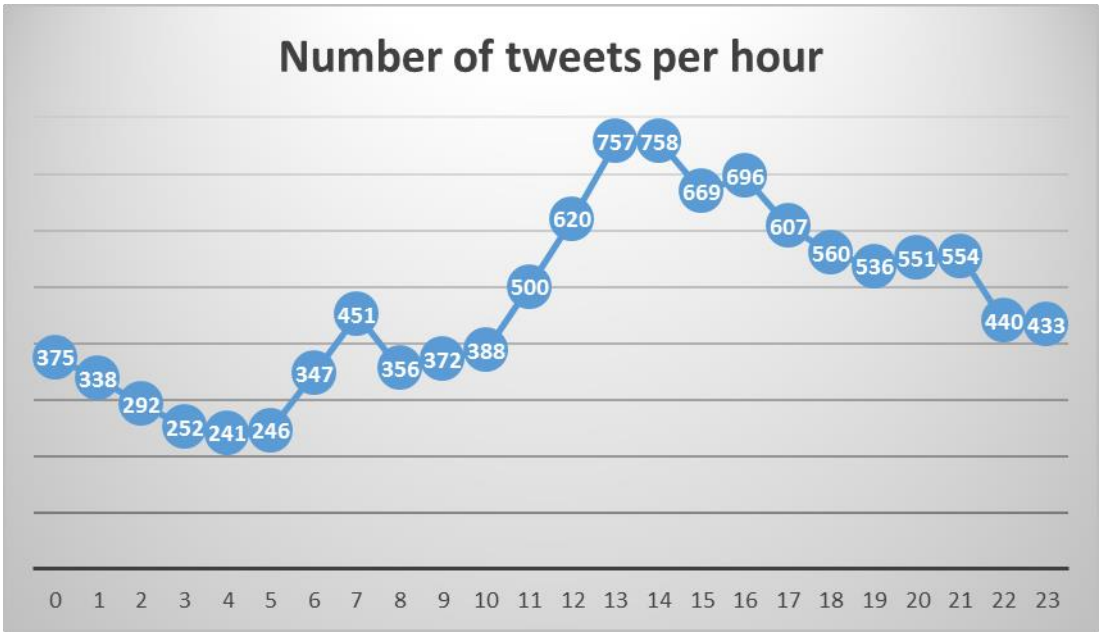



*Figure 5: Hourly tweet volume*

Conversely, the activity is higher during operating days (*figure 6*) which is in link with the findings of previous studies and the idea that people exchange more during trading time.

[35] Between two samples of tweets about indexes retrieved by Mr Naets, only 4,6 percent were identical. This reflects well that some parts of the results are given to popular tweets while the others are random.

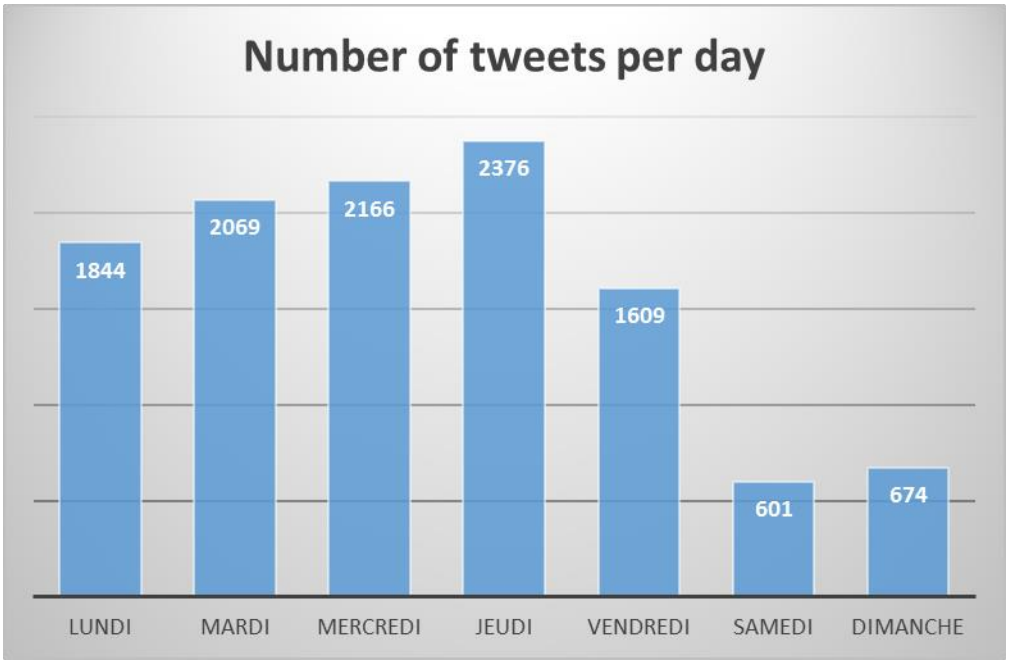



*Figure 6: Daily tweet volume*

In order to carry on further analyses, a processing of inputs was necessary. First it concerned only Twitter data with the sentiment analysis and time constraint implementation but then it included the aggregation of all information types. It will be described in the following part.

## Data processing

In addition to the first pre-processing of Twitter data done in Excel and explained in the data collection part, a new treatment was necessary in order to carry on the sentiment analysis. Therefore, all the contents of tweets was first put into lower cases. In line with the study of Ranco et al. (2015), we deleted cashtags as well as usernames and hashtags from the messages in order to make tweets independent from a company, user or topic. Moreover, the links and punctuations were also removed. Then, we pulled out the stop words. They are determiners (e.g. a, an, the), adjectives (e.g. nice, cool, bad), prepositions (e.g. to, across), etc. In a sentiment analysis context, removing adjectives would be detrimental. So the stop words removed were only neutral words carrying no useful information. In other words, texts have to be as clean as possible in order to make the sentiment analyser works. This is the reason why we deleted all useless elements such as the cashtags, punctuations, etc. To understand deeper why we did that, we need to explain the principle of sentiment analysis.

The sentiment study part was performed using an unsupervised analysis, which refers to the use of algorithms and statistics, in opposition to the supervised analysis which utilizes machine learning classifiers. More precisely, Natural Language Processing (i.e. NLP) refers to an algorithm allowing to read data and to analyse it like a human would have done. We decided to use this technique. Providing a sentiment indicator superior to the previous ones achieved in the literature is out of scope in this study. This is the reason why to process words and to give them an orientation from positive to negative, we needed to rely on an already established dictionary. In a sentiment analysis context, a dictionary contains a list of words classified according to emotions by experts. The software uses it to assign a sentiment value to each word in the target text. These given numbers vary depending on the dictionary used and the sentiment studied but are often comprised between -1 (for the most negative words) and +1 (for the most positive words). After having scaled each term with the dictionary, the NLP sums the word orientations in order to get the global sentiment of each text (here represented by tweets).

In order to compute tweet sentiments, we selected five dictionaries as well as two NLP packages on R and compared their results. Appendix 13 summarizes their outcomes.

In the first NLP attempt, the dictionary used was the Syuzhet one developed by the Nebraska Literary Lab. It is based on three existing lexicons: BING[36], AFINN[37] and NRC[38] (*Jockers, n.d.; Naldi, 2019*). Then, a second NLP was implemented through different dictionaries[39]. The first one GI refers to the psychological Harvard-IV dictionary as used in the General Inquirer software[40]. This tool was developed by the University of Harvard and assigns to words among others a sentiment in a scale going from -1 to +1 depending on their connotation: positive or negative. This well-known dictionary was already used in financial researches (e.g. Tetlock, 2007) and so has revealed to be a good instrument for sentiment indication. Then, HE is the Henry's Financial dictionary published in 2008 and among one of the first financial purpose dictionary. LM relates to the well-known Loughran-McDonald dictionary also dedicated to finance and certainly the most widely used in this area since 2011. It was based on analysis of Form 10-K, an official summary that companies have to handle to the SEC. By examining several of these documents, the authors established six lists of words[41] including the negative and positive ones. Finally in the last package QDAP concerns a set of dictionaries put together by R programmers[42].

Syuzhet and QDAP were not selected as there are global dictionaries, to the best of our knowledge, not already used in a recent financial area. So, three solutions remained, all concerning dictionaries already used in a financial context. We based our choice on the results provided by each of the possibilities and available in appendix 13. The lexicon giving the less neutral elements was the Harvard-IV dictionary and this is the reason why we relied on this one for our analyses. Nevertheless, as the Loughran-McDonald lexicon is a famous financial dictionary used in a lots of studies and the second in our test, we decided to implement also our study with it. So, the following results will be the one deriving from the sentiment analysis from the Harvard-IV dictionary. But at the end of

[36] It classifies words between two categories either positive or negative and was created by Bing Liu.
[37] It ranks words in a scale going from -5 to +5 and was made by Finn Årup Nielsen.
[38] It categories words into ten classes (positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise and trust) and was announced by Saif Mohammad and Peter Turney.
[39] Information about these dictionaries was found among others via Kim (2018).
[40] The following link allows to get the list of words used in the General Inquirer software : http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm
[41] Fin-Neg (negative words), Fin-Pos (positive words), Fin-Unc (uncertainty words), Fin-Lit (litigiousness words), MW-Strong (strong modal words) and MW-Weak (weak modal words).
[42] The list can be found here https://cran.r-project.org/web/packages/qdapDictionaries/qdapDictionaries.pdf

this essay, a complement will be given about results deducted from the Loughran-McDonald sentiment analysis.

After having decided which dictionary and NLP would be used, we computed the results and aggregated the sentiment scores to an ordinal scale of three values: “negative”, “neutral” or “positive” depending on whether the score was lower, equal or higher than zero respectively following the study of Ranco et al. (2015). The table summarizing the sentiment value statistics of our sample is available in appendix 11. It reveals that most of the tweets were assigned a positive number as the first quartile is already at 0. What is more, the mean is also positive at 0,068. In fact, 5.219 tweets were given a positive opinion in contrast to 2.014 negative. That represents respectively 46 and 17,8 percent. 35,3 percent that is to say 4.000 messages were neutral. The remaining one percent stands for the NA values which means that the NLP was not able to find words on which to assign a rating (often because the content of messages was only composed of the cashtag(s) of the company(ies)). As they only represent 106 messages out of the 11.339 from the beginning, they do not impact our work. By the way, all the dictionaries from the second NLP also returns 106 NA values (*appendix 13*).

Then, based on several studies including Takahashi et al. (2007), if the news is released in the afternoon the stock price change must be computed on the basis of the following day closing price. Following Takahashi et al. (2007)’s idea, tweets published after trading hours should be assigned to the next trading day. As explained previously, the closing hours diverge between Brussels, Paris (17h30) and Amsterdam (17h40), the stock exchanges of our indexes. Moreover, the shift between these hours and the GMT depends on the season: +1 during the winter or +2 during the summer. So we faced time and location issues. The time of each tweet in our database being set at the GMT timeline, we decided to put back the closing hours to their GMT equivalent: 15h30/40 or 16h30/40 depending on the situation. So, the closing hour for each day was defined depending on the date and the index. As a consequence, tweets published after these selected closing hours were assigned to the next trading day. Furthermore, the same idea stands for tweets written during the weekends. So, they were allocated to the following Mondays as stock exchanges are closed during weekends. Thus, tweets written after trading hours on Fridays were put in link to the next Monday stock price. Actually tweets published during

public holidays follow the same rule as after-hour news but they represented only less than one percent of our final sample. Therefore, they were not investigated.

Next to the tweet processing phase came the aggregation of the different data categories. This step was divided into five segments, some of them requiring loops to be implemented: the union of tweet information and stock market data[43], their aggregation to unique date[44], the addition of company characteristics[45] as well as index market data[46] and finally the creation of firm(or index)-specific variables. In consequence, for each stock and for each date, new variables referenced in table 1 were created. Their descriptions as well as their descriptive statistics are given hereafter.

[43] Stock prices and trading volumes retrieved from Bloomberg and Yahoo! Finance and treated previously were merged to the already on hand tweet data, creating one single entity with stock market and tweet information. No row was removed in order to keep the integrity of the data and to allow efficient analyses and conclusions.
[44] Since, at some points in time, there were several tweets about a same company at a same date, the data needed to be concatenated to a single row for each stock at each available date.
[45] The market capitalization and percentage of retail investors were joined to the previous data.
[46] Index prices and trading volumes retrieved from Bloomberg and Yahoo! Finance and treated previously were merged to the already on hand data (the tweets, stock and company information).

## Variables

Unlike other studies focusing on a shorter period of time, our decision to increase the time length due to our tweet arrangement, had a major consequence on the number of missing values. In fact, for all stocks, a tweet was not systematically displayed on a daily basis. In their study, Sprenger et al. (2014) defined the message volume, tweet bullishness and agreement among messages alternatively at zero and as missing values. Because the results were sensibly the same[47], we decided to leave NA values. Another reason for this choice is the risk of a downward bias if all the missing values were replaced by zeros. So, for all variables except the Twitter message volume, in the absence of tweet, a "NA" value is displayed. The table resuming the descriptive statistics of our variables is displayed in appendix 12. Its principal data will be given in the new variables presentation hereafter.

The first added parameters referred to the sentiments. They needed to be considered in relative terms because the total number of tweets per day varies. As tweets from the weekends are aggregated on Mondays, the number of tweets on these days is more important and cannot be compared in absolute terms to the other days. Therefore, we first computed the number of positive and negative tweets on each day and for each ticker which allowed us to calculate their respective percentages (i.e. "$POS_{i,t}$", and "$NEG_{i,t}$"). These last variables were obtained by computing the number of positive (or negative) tweets on a given day and for a given ticker divided by the total number of tweets posted that same day for that ticker. Computing a percentage on these two categories enabled us to create a distinction between emotions and so to visualize a potential difference in their impact on Twitter. This is in contrast with other studies which have computed a polarity on tweets in order to reflect their global emotional impression.

After the aggregation, 8.703 tweets with different dates and tickers were leaving. Among them, 4.329 were positive and 1.800 negative. The distribution by sentiment is thus similar to the one before the aggregation. The message volume per day varies between 0 to 35 tweets but the mean is low at 0,06 as the parameter is defined at 0 for days without any tweets. By contrast, the trading volume fluctuates from 0 to 217.503.585 shares per day for the stocks and from 0 to 531.247.600 shares per day for the indexes. The 0 is explained by some errors in the data retrieved from Yahoo!Finance but they

[47] However, their data set was quite complete with more than 80 percent coverage.

remain marginal[48]. Of course, the average trading volume for stocks is lower than for indexes: 2.971.824,74 vs 101.795.424,86 shares.

From now on, the message volume (i.e. "$MSG_{i,t}$") is set as the natural logarithm[49] of the number of tweets on a given stock per day. The same formula is applied to the trading volume (i.e. "$VOL_{i,t}$") in line with the studies of Sprenger et al. (2014) for Twitter and more generally of Antweiler & Frank (2004) for microblogs. Taking the logarithm allows to compute elasticities and to control for the non-stability of volume over time by coming closer to a normal distribution. Indeed, our sample follows the growth of Twitter throughout the years and so is not homogeneous. Plus, as it has already been explained before, Mondays are characterized by a bigger message level than the other days due to the aggregation, leading to a non-equal distribution between days.

Finally, the last variables obtained for Twitter concern the measure of tweet popularity. They are represented by the number of retweets and replies and used as proxies for the weight given to tweets. We summed all the retweets and replies for each ticker and for each day to get the variables $RTW_{i,t}$ and $RPL_{i,t}$. They both exhibit a wide range of values from 0 to more than 600 retweets/replies per day. However, extreme observations are rare since the means are low: below 5 retweets/replies.

Then, in a market perspective, we computed a logreturn for the stock (i.e. "$RET_{i,t}$") based on the closing market stock prices:

$$\mathrm{RET}_{i,t} = \ln\left(\frac{P_{i,t}}{P_{i,t-1}}\right) * 100 \qquad [1]$$

with $P_{i,t}$ being the market closing price of the stock i at time t.

For the indexes, we defined the market return (i.e. "$MRET_{z,t}$") and market trading volume (i.e. "$MVOL_{z,t}$") on the same principles as their stock equivalents but with an index-basis instead of a stock one.

The stock return $RET_{i,t}$ shows a mean close to zero which is coherent with the EMH stating that sustainable returns are unachievable. The same is observed for the index market return $MRET_{z,t}$.

[48] 20 null values for the stock trading volume which represents 0,011% of the total sample and 9.254 for the index trading volume which stands for 5,1%.
[49] Ln(1+x), x being the variable on which the natural logarithm is computed.

At the end, we added four corporate variables: the market capitalization (i.e. "$CAP_{i,t}$"), the percentage of retail ownership (i.e. "$RETAIL_i$"), the sector (i.e. "$IND_i$") and index (i.e. "$INDEX_i$") of the stocks. The retailers and institutions' ownership percentages were computed previously and explained in the data collection part. We selected the first one as a new corporate parameter.

The market capitalization has a variation from 0,48 to 183,98 billion euros which reflects the diversity of our sample in respect of company size. The second corporate parameter has a minimum percentage of zero for some corporations against 65 percent for companies with the highest ratio of retail investors. But the average is located around five percent, a low figure. Then, principal information concerning the stock index and the industry on which companies operate were already given in the methodology and data description sections. They were not displayed in the appendix 12 for this reason and because they are categorical variables.

| **Variables** | **Definitions** | **Initial data sources** |
|---|---|---|
| $POS_{i,t}$ | Percentage of positive tweets for a stock i on a day t | Twitter |
| $NEG_{i,t}$ | Percentage of negative tweets for a stock i on a day t | Twitter |
| $MSG_{i,t}$ | Twitter message volume for a stock i on a day t | Twitter |
| $RTW_{i,t}$ | Total number of retweets for a stock i on a day t | Twitter |
| $RPL_{i,t}$ | Total number of replies for a stock i on a day t | Twitter |
| $RET_{i,t}$ | Market return for a stock i on a day t | Bloomberg, Yahoo! Finance |
| $VOL_{i,t}$ | Market trading volume for a stock i on a day t | Bloomberg, Yahoo! Finance |
| $MRET_{z,t}$ | Market return for an index z on a day t | Bloomberg, Yahoo! Finance |
| $MVOL_{z,t}$ | Market trading volume for an index z on a day t | Bloomberg, Yahoo! Finance |
| $IND_i$ | The sector on which the stock i operates | Euronext |
| $INDEX_i$ | The index to which the stock i belongs | Euronext |
| $RETAIL_i$ | Percentage of current direct shares hold by retail investors for a stock i | Orbis |
| $CAP_{i,t}$ | Market capitalization of a stock i on a day t | Orbis |

*Table 1: Variable descriptions*

Henceforth, the variables will reference to these newly-computed parameters displayed on table 1. For example a mention to the volume of messages will in fact refer to the $MSG_{i,t}$ parameter and so to the natural logarithm of tweet volume. Now that all the necessary features in the three classes (Twitter, market and corporate) have been defined, we can move to the regressions.

# Section VII: Results

The results displayed in this section are divided into three parts. The first one includes the most basic model reporting for the links between Twitter and the market (that is to say stock returns and trading volume). The next subdivision shows more advanced analyses and describes the relationships between corporate and Twitter features. The final one gives the results when applying the Loughran-McDonald dictionary and account for differences between both methods.

## Relationships between tweet and market features

To begin with pairwise correlations, we chose to interpret the Spearman method instead of the classical Pearson correlation because we had, at the beginning, no certainty about the linearity inside the relationships between our variables. The Spearman correlation is also advised when logarithmic data are treated. Moreover, we compared the correlation results with both methods and revealed them in the table 2. The Spearman method retrieves the larger number of significant results. The pairwise correlations in the upper part of table 2 are related to the Pearson method whereas the lower part refers to the Spearman one.

| | POS | NEG | VOL | MSG | RET | RTW | RPL |
|---|---|---|---|---|---|---|---|
| **POS** | 1 | -0.4263*** | -0.0286*** | 0.0076 | 0.0188* | -0.0057 | 0.0197* |
| **NEG** | -0.4103*** | 1 | 0.0473*** | 0.0040 | -0.0385*** | 0.0175 | 0.0022 |
| **VOL** | -0.0318*** | 0.0598*** | 1 | 0.0591*** | -0.0119*** | 0.0661*** | 0.0254** |
| **MSG** | 0.0201* | 0.1263*** | 0.0573*** | 1 | 0.0039* | 0.2816*** | 0.1291*** |
| **RET** | 0.0169 | -0.0229** | -0.0173*** | 0.0049** | 1 | 0.0014 | -0.0051 |
| **RTW** | -0.0315*** | 0.0696*** | 0.1417*** | 0.2483*** | -0.0026 | 1 | 0.4264*** |
| **RPL** | 0.0188* | 0.0613*** | 0.0966*** | 0.2341*** | 0.0195* | 0.4173*** | 1 |

****: p < .01, **: p < .05, *: p < .1*

*Table 2: Correlation matrix between tweet and market features*

Firstly, in the links between Twitter features, a fairly strong negative correlation seems to exist between positive and negative tweets. It is logical since an increase of the percentage of positive tweets suggests a decrease in the negative tweet percentage. Then, in a bigger volume of messages, more negative tweets are likely to be seen than the positive ones as tweet volume and negative emotion are more correlated than volume and positive feeling. This may suggest a deeper negativity feeling among investors inside the online community. Moreover, negative messages tend to be retweeted and replied more

often than the positive ones increasing again their spread. In contrast, if the percentage of positive tweets increases, the number of retweets is expected to decrease. It indicates that people are more likely to retweet negative information than the positive one, giving them more importance. Also, retweets and replies are correlated to the tweet volume. In fact, their relations are strong and significant. This seems to assert our hypothesis that they are good proxies for the weight given to messages because they cause a greater exchange of messages and potentially debate among the community. They are also very correlated to each other.

Then, the trading and tweet volumes show a significant but low link which goes in the direction of our first hypothesis. Besides, positive tweets tend to cause a decrease in the trading volume while negative messages are more likely to lead to an increase of the latter but the correlations determined are also low. Moreover, retweets and replies have a positive correlation close to ten percent to the stock trading volume while their relations to returns are small or non-significant. So, the first links on the contrary of the second ones lead to our fourth hypothesis. On the top of the trading volume aspect, the stock return and percentage of negative tweets are negatively correlated which is coherent because an increase in negative emotion should be linked to a decrease of return.

As a consequence, pairwise correlations indicate interesting relationships between our variables and even go in the direction of some of our hypotheses. However, they do not address the parameter interdependence and so the global link between tweet and market features. In order to determine whether Twitter can help explaining market movements, we conducted contemporaneous regression. The following interpretations of these regressions are made “all else being equal”.

As our data is composed of a cross sectional index (i.e. the ticker i) and a time index (i.e. the date t) which varies within the group index i, fixed-effects panel regressions were necessary. In order to account for cross-sectional differences, we used firm fixed-effects. A standard fixed-effect model applied to our model can be written as followed:

$$y_{i,t} = \alpha + u_i + \beta_1 POS_{i,t} + \beta_2 NEG_{i,t} + \beta_3 MSG_{i,t} + \beta_4 RTW_{i,t} + \beta_5 RPL_{i,t} + \beta_6 MRET_{z,t} + \beta_7 MVOL_{z,t} + \beta_8 X_{i,t} + \varepsilon_{i,t} \quad [3]$$

$Y_{i,t}$ represents the dependent variable varying in each of our regression. $X_{i,t}$ is a control variable for another stock market feature which relies on the dependent variable.

Namely in the return regressions, it consists in the stock trading volume while in the trading volume regressions it is the stock return. $MRET_{z,t}$ and $MVOL_{z,t}$ also consist in control variables but for the index z corresponding to the stock i. The other parameters were already defined in the table 1.

A global table comprising all the regressions made between tweet features as independent variables and market features as our outcome variables is exhibited in table 3 below.

**Regression results**

| | *Dependent variable:* | |
|---|---|---|
| | RETi,t | VOLi,t |
| | (1) | (2) |
| POSi,t | 0.041 | -0.019 |
| | (0.042) | (0.013) |
| NEGi,t | -0.117** | 0.008 |
| | (0.054) | (0.017) |
| MSGi,t | 0.126* | 0.037 |
| | (0.075) | (0.023) |
| RTWi,t | 0.0001 | -0.0004 |
| | (0.001) | (0.0003) |
| RPLi,t | 0.0003 | -0.001 |
| | (0.003) | (0.001) |
| MRETz,t | 1.070*** | -0.027*** |
| | (0.015) | (0.006) |
| MVOLz,t | -0.004 | 0.022*** |
| | (0.009) | (0.003) |
| VOLi,t | 0.104*** | |
| | (0.035) | |
| RETi,t | | 0.010*** |
| | | (0.003) |
| Observations | 8,546 | 8,546 |
| $R^2$ | 0.362 | 0.011 |
| Adjusted $R^2$ | 0.356 | 0.002 |
| F Statistic (df = 8; 8472) | 600.000*** | 11.300*** |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | |

*Table 3: Results of the regression of tweet parameters on market features*

In the table 3, only a few results have revealed themselves statistically significant but they will be disclosed and explained hereafter. First it seems that the negative feeling embedded in messages have a fairly deep impact on stock returns. Since the first parameter reflects a percentage value and the second one is a logreturn, we can interpret the result as the following: “If the percentage of negative messages increases by one percentage point, the logreturn will tend to decrease by 0,12 percentage point. So, we do

not reject our third null-hypothesis that negative emotions help explaining market features (in the return point of view at least). This tends to confirm the behavioural finance assumptions which assert that stock prices are influenced by emotions. Moreover, the tweet volume is also linked to the market return. Their relationship was already established in the correlation matrix but for a lower figure. Now that the link is amplified, it can no longer be ignored. So, an increase of the message volume by one percent is likely to cause a 0,13 percentage points increase of stock market returns. It suggests that if a stock is more discussed on Twitter, its return would increase. Dewally (2003) showed that the majority of posts on internet message boards are buy signals. In our sample, the proportion of positive tweets against the negative ones is higher providing a good feeling to the readers and more largely to the market. Plus, it supports Dewally (2003)'s assumption of a momentum in microblogs: people post about stocks having already known an increasing performance. Others reading these messages buy these stocks and by doing so increase even more the stock return. Yet, we found no support for our second hypothesis stating that a link should exist between positive opinion and return. This is connected with Tirunillai & Tellis (2012)'s result demonstrating that only negative UGC have an impact on returns. Nevertheless, according Tumarkin & Whitelaw (2001), it is only the positive feelings that have an effect. So results diverge on this. Furthermore, our fourth hypothesis (i.e. the number of replies and retweets impacting the stock market movements) is also rejected due to the non-significance of their corresponding coefficients.

Unfortunately, for the trading volume, we do not find evidence asserting our hypotheses. On the contrary of Sprenger et al. (2014)'s study for Twitter and Antweiler and Frank (2004) for microblogs, we reject the idea of links between message and trading volumes.

Regressions have a F statistic with a significant *p*-value which means that the models are significantly explicative. The adjusted $R^2$ of the first regression is 36 percent that is to say that it explains 36 percent of the stock return movements.

Now that the relationships between tweet and market features have been established, we can jump to the addition of corporate parameters in order to get their impact on the previous regressions realized.

## Relationships when including company features

Company features are mainly composed of the two following variables: the market capitalization and the percentage of retail investors that were already described either in the methodology or in the data description sections. Following the same structure as the "relationships between tweet and market features" point, we will first give the correlations and then exposed the regressions.

Spearman pairwise correlations between the two new variables and the previous ones are exhibited in the following table:

| | POS | NEG | VOL | MSG | RET | RTW | RPL | CAP |
|---|---|---|---|---|---|---|---|---|
| **CAP** | 0.0021 | 0.0448*** | 0.4239*** | 0.0638*** | 0.0010 | 0.1316*** | 0.0981*** | |
| **RETAIL** | -0.0282*** | 0.0043 | -0.0406*** | 0.0475*** | -0.0005 | 0.0908*** | 0.0254** | 0.2038*** |

****: p < .01, **: p < .05, *: p < .1*

*Table 4: Correlation matrix between corporate and tweet/market features*

They consolidate our ideas that market capitalization and retail percentage are linked to tweet and market features. First, as the market cap goes up, the trading volume increases sharply meaning that big firm shares tend to be traded in higher volume than smaller firm participations. Conversely, higher is the percentage of retail investors in a firm, lower is the trading volume of that stock. Through that relation, we guess that institutional investors actually lead the market but we cannot be totally sure about this as we were not able to separate individual from institutional investors in the factor "public investors" in our initial dataset from Orbis. Then, both corporate variables have positive relations to the message volume and number of replies and retweets. Therefore, a firm with a high number of retail investors is likely to demonstrate a bigger volume of messages as well as feedbacks to them and the same stands for the market capitalization. It goes in the direction of our fifth and sixth hypotheses concerning the impact of corporate features on Twitter activity. However, bigger firms tend to generate more negative message. Indeed, in social media, people can be harsh towards companies and especially towards the big ones often in the hearth of scandals. Similarly, the percentage of retail investors is negatively correlated to the percentage of positive tweets. It means that if a firm has a high number of retail investors, the proportion of positive messages on Twitter about its stock performance will decrease. This asserts the link found in the previous correlation matrix (*table 2*) and suggesting a negativity feeling inside the investor community on

Twitter. To conclude with the correlation matrix, the retailers' ratio and market cap are positively correlated with each other with a figure close to twenty percent.

Our next step consists in an analysis of the regressions including the corporate variables $RETAIL_i$ and $CAP_{i,t}$. $RETAIL_i$ being time-invariant we faced an issue when implementing the same type of regression done between tweet and market features. This is the reason why, we needed to jump from a firm-fixed regression panel to a random effects model. It relies on the same equation as fixed effects model (i.e. equation [3]) but is more restrictive by imposing that the individual specific effect (i.e. $u_i$) must be uncorrelated to the explanatory variables. Thanks to this change of formulation we obtained the new regressions resumed table 5.

**Regression results**

| | *Dependent variable:* | |
|---|---|---|
| | RETi,t | VOLi,t |
| | (1) | (2) |
| POSi,t | 0.032 | -0.024* |
| | (0.041) | (0.013) |
| NEGi,t | -0.121** | -0.003 |
| | (0.054) | (0.017) |
| MSGi,t | 0.141* | 0.045** |
| | (0.073) | (0.023) |
| RTWi,t | -0.00003 | -0.0003 |
| | (0.001) | (0.0003) |
| RPLi,t | 0.0002 | -0.001 |
| | (0.003) | (0.001) |
| CAPi,t | -0.001 | -0.007*** |
| | (0.001) | (0.001) |
| RETAILi | 0.286 | -2.060 |
| | (0.259) | (1.265) |
| MRETz,t | 1.061*** | -0.028*** |
| | (0.016) | (0.006) |
| MVOLz,t | -0.004 | 0.018*** |
| | (0.008) | (0.003) |
| VOLi,t | 0.014 | |
| | (0.022) | |
| RETi,t | | 0.010*** |
| | | (0.003) |
| Constant | -0.228 | 14.360*** |
| | (0.322) | (0.173) |
| Observations | 8,163 | 8,163 |
| $R^2$ | 0.353 | 0.023 |
| Adjusted $R^2$ | 0.352 | 0.022 |
| F Statistic | 4,445.000*** | 195.900*** |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | |

*Table 5: Results of the regression of tweet and corporate parameters on market features*

In table 5, the figures relative to tweet parameters are similar to the ones of the previous regressions. However, the significance level has changed for the link between the tweet level and the trading volume since it becomes now significant at a *p*-value comprises between 5 and 10 percent. This new link does not reject our first hypothesis that the tweet volume helps explaining the trading volume. Yet, as this result varies between the two types of regression *(table 3 and table 5)*, it cannot be considered. Unfortunately since the significance level also changes for the link between the percentage of positive tweets and the trading volume, we decide to not consider it even if it would go in the direction of our second hypothesis and confirm the result of Sprenger et al. (2004).

Then, for the added corporate variables, a statistically significant relationship exists between the market capitalization and the trading volume but it is very low. If the market cap increases by one billion euro, the trading volume is likely to decrease by 0,7 percent. This is in opposition with our previous finding in the correlation matrix. The other results for the market capitalization and retailer percentage are non-significant.

In order to test our results, multicollinearity analyses were conducted on the regressions of the table 5 and shown in table 6. Multicollinearity happened when an approximative linear relation exists between a parameter and other independent variable(s). The variance inflation factor (i.e. VIF) is a common test used to determine from how much the variance of a coefficient is increased due to its relationship with other variables. By performing a VIF on our two regressions from the table 5, we reported no significant multicollinearity between our variables since the results of the tests are all comprised between 1 and 1,6. Nevertheless, we decided to compute the regressions between corporate and tweet parameters in order to determine the impact of the former on the latter.

| | RETi,t | VOLi,t |
|---|---|---|
| **POSi,t** | 1,22 | 1,22 |
| **NEGi,t** | 1,22 | 1,22 |
| **MSGi,t** | 1,09 | 1,09 |
| **RTWi,t** | 1,30 | 1,29 |
| **RPLi,t** | 1,21 | 1,21 |
| **CAPi,t** | 1,04 | 1,01 |
| **RETAILi** | 1,05 | 1,00 |
| **MRETz,t** | 1,00 | 1,54 |
| **MVOLz,t** | 1,04 | 1,01 |
| **VOLi,t** | 1,09 | |
| **RETi,t** | | 1,55 |

*Table 6: Results of the VIF tests on the regressions from table 5*

As stated in our fifth and sixth hypotheses, a link may exist between corporate parameters and tweet volume. In order to test it more precisely than with the VIX, we conducted additional regressions on Twitter variables and exhibited in table 7.

**Regression results**

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | POSi,t | NEGi,t | MSGi,t | RTWi,t | RPLi,t |
| | (1) | (2) | (3) | (4) | (5) |
| MSGi,t | 0.007 | 0.033** | | 18.800*** | 0.251 |
| | (0.015) | (0.015) | | (0.833) | (0.311) |
| NEGi,t | -0.168*** | | 0.018** | 0.411 | -0.042 |
| | (0.010) | | (0.008) | (0.613) | (0.229) |
| POSi,t | | -0.181*** | 0.004 | 0.573 | 0.688*** |
| | | (0.011) | (0.008) | (0.637) | (0.237) |
| RTWi,t | 0.0002 | 0.0001 | 0.003*** | | 0.150*** |
| | (0.0002) | (0.0002) | (0.0001) | | (0.004) |
| RPLi,t | 0.002*** | -0.0002 | 0.0002 | 1.066*** | |
| | (0.001) | (0.001) | (0.0004) | (0.027) | |
| CAPi,t | 0.0004* | -0.00004 | 0.001*** | 0.032** | 0.001 |
| | (0.0003) | (0.0003) | (0.0002) | (0.016) | (0.003) |
| RETAILi | -0.077 | -0.039 | -0.010 | 0.250 | 0.542 |
| | (0.055) | (0.053) | (0.079) | (3.656) | (0.607) |
| RETi,t | -0.002 | -0.005** | 0.004*** | -0.040 | -0.024 |
| | (0.002) | (0.002) | (0.001) | (0.105) | (0.039) |
| VOLi,t | -0.006 | 0.002 | 0.005 | -0.030 | 0.030 |
| | (0.005) | (0.005) | (0.005) | (0.326) | (0.077) |
| AEX 25 | 0.014 | 0.008 | -0.043 | -1.180 | 0.175 |
| | (0.029) | (0.028) | (0.040) | (1.884) | (0.326) |
| CAC 40 | 0.010 | 0.047* | -0.075* | -0.489 | 0.154 |
| | (0.030) | (0.029) | (0.041) | (1.954) | (0.333) |
| Automobiles & Parts | 0.014 | -0.022 | -0.027 | -0.714 | -0.052 |
| | (0.026) | (0.025) | (0.036) | (1.683) | (0.280) |
| Banks | -0.015 | 0.046 | -0.082** | -1.565 | -0.160 |
| | (0.030) | (0.029) | (0.041) | (1.966) | (0.340) |
| Basic Resources | -0.011 | 0.007 | -0.036 | 1.049 | -0.394 |
| | (0.034) | (0.033) | (0.048) | (2.241) | (0.377) |
| Chemicals | 0.019 | 0.008 | -0.068 | -1.194 | -0.105 |
| | (0.030) | (0.029) | (0.042) | (1.967) | (0.329) |
| Construction & Materials | -0.040 | -0.014 | -0.007 | 1.392 | -0.09 |
| | (0.027) | (0.026) | (0.038) | (1.771) | (0.295) |
| Financial Services | -0.147 | -0.125 | 0.151 | -6.493 | 0.236 |
| | (0.142) | (0.137) | (0.203) | (9.387) | (1.557) |
| Food & Beverage | -0.032 | 0.088*** | -0.106** | 7.760*** | -0.192 |
| | (0.030) | (0.029) | (0.042) | (1.952) | (0.330) |
| Health Care | -0.018 | 0.010 | -0.038 | -0.442 | 0.013 |

| | | | | | |
|---|---|---|---|---|---|
| | (0.028) | (0.027) | (0.039) | (1.855) | (0.320) |
| Industrial Goods & Services | -0.023 | 0.024 | -0.021 | 0.890 | -0.376 |
| | (0.026) | (0.025) | (0.037) | (1.707) | (0.285) |
| Insurance | -0.025 | 0.039 | 0.027 | 9.422*** | 0.411 |
| | (0.029) | (0.028) | (0.040) | (1.901) | (0.334) |
| Media | -0.017 | 0.006 | -0.002 | -1.726 | 0.419 |
| | (0.027) | (0.026) | (0.038) | (1.761) | (0.294) |
| Oil & Gas | -0.077** | 0.025 | -0.124*** | 0.430 | -0.461 |
| | (0.036) | (0.035) | (0.048) | (2.349) | (0.412) |
| Personal & Household Goods | -0.025 | 0.048 | -0.067 | -2.979 | 0.05 |
| | (0.038) | (0.037) | (0.052) | (2.476) | (0.437) |
| Real Estate | -0.017 | 0.234** | -0.033 | -2.069 | 0.012 |
| | (0.123) | (0.118) | (0.175) | (8.091) | (1.340) |
| Retail | -0.067** | -0.033 | -0.036 | 1.485 | -0.269 |
| | (0.027) | (0.026) | (0.039) | (1.793) | (0.299) |
| Telecommunications | -0.017 | 0.023 | -0.078* | -1.194 | -0.146 |
| | (0.034) | (0.033) | (0.047) | (2.216) | (0.381) |
| Travel & Leisure | -0.051 | 0.027 | -0.003 | -1.790 | 2.190*** |
| | (0.033) | (0.032) | (0.047) | (2.190) | (0.363) |
| Utilities | 0.008 | -0.012 | -0.047 | -1.113 | -0.197 |
| | (0.037) | (0.036) | (0.053) | (2.456) | (0.412) |
| Constant | 0.257*** | 0.086 | 0.754*** | -11.780** | -0.836 |
| | (0.074) | (0.074) | (0.075) | (4.573) | (1.046) |
| Observations | 8,182 | 8,182 | 8,182 | 8,182 | 8,182 |
| $R^2$ | 0.036 | 0.039 | 0.082 | 0.237 | 0.200 |
| Adjusted $R^2$ | 0.032 | 0.035 | 0.079 | 0.234 | 0.197 |
| F Statistic | 301.600*** | 326.900*** | 728.200*** | 2,531.000*** | 2,035.000*** |
| *Note:* | | | | | *p<0.1; **p<0.05; ***p<0.01 |

*Table 7: Regression results of tweet, corporate and market features on tweet parameters*

First, through the regression on message volume in table 7, we do not reject our sixth hypothesis stating that companies' market capitalization influences Twitter activity. However, the coefficient is pretty low: an increase of the market cap of one billion euro will tend to increase the tweet volume by 0,1 percent. But it asserts that bigger firms have more tweets about their stock performance than the others, stressing the importance of companies size in their popularity on Twitter. This is coherent with the previous Tumarkin and Whitelaw (2001)'s study on microblogs. By the way, the market cap also exhibits a positive link with the number of retweets. These results about market

capitalization are connected with the investor recognition hypothesis. It states that a more famous company sees the public's attention going up and by doing so the information volume about its activity (i.e. tweet and retweet volume here). Therefore, some investors may decide to invest in its stocks. This also explains the previous link between market cap and trading volume.

Unfortunately, we find no support for the fifth hypothesis concerning the link between the percentage of retailers and Twitter activity.

Then, the table 7 also displays the relationships between tweet features. Among them, several were statistically significant. For instance, the link between $RTW_{i,t}$ and $MSG_{i,t}$. In fact, an increase of one percent of the tweet volume is likely to cause 19 additional retweets among them. Moreover, the pairwise correlation between retweets and replies is confirmed through the regression table. In fact, if the number of retweets increase by one unit, the number of replies will tend to go up by 0,15 unit. Similarly, the fairly high negative relation between the percentages of positive and negative tweets is verified. Plus, if the percentage of positive tweets increase by 1 percentage point, the number of replies tend to increase by 0,38 unit. It shows that users reply to positive messages but we cannot make a comparison with the negative ones as results for them are non-significant. However, according to Zhang (2011), users with messages on average more bullish have a greater reputation. As reputation and audience are linked we can deduct that positive messages are likely to spread more.

Moreover, in the regressions displayed in table 7, besides Twitter, market and corporate features, we also added binary variables. The purpose is to report differences between indexes and industries. As reference variables, we chose the BEL 20 for the stock index, since it is the smallest one, and the technology sector for the industry. According to Tumarkin and Whitelaw (2001), the Internet area should be the most reflected in microblogs. So we relied on this assumption and took the technology industry as our reference variable. Each coefficient associated to a stock index or industry in the table 7 reflects from how much this stock index/industry varies compared to its reference in respect of the dependent variable.

The results obtained from the index and industry comparisons are summarized as followed: the sectors of banks, basic resources, chemicals, insurance, media,

telecommunications and utilities have on average a higher percentage of positive messages than the technology area while the travel & leisure and financial services have less. The latter also receives a lower percentage of negative tweets compared to the technology industry whereas banks, food & beverage, oil & gas and real estate tend to get a higher percentage of negative tweets than the technology sector. On the one hand, it may reflect which industries tend to be preferred or not by consumers. On the other hand, financial services and banks show the same effect for both positive and negative tweet percentages: respectively a decrease or an increase of both $POS_{i,t}$ and $NEG_{i,t}$. So, it may indicate that the global volume of tweets increase.

However, banks, food & beverage, oil & gas and telecommunications are less hot topics compared to our reference because their coefficients in the $MSG_{i,t}$ regression are negative. So, the sign of the coefficient for banks contradicts our idea of a global tweet volume increase for this sector in comparison to the technology area. Plus, since no sector has a positive and statistically significant sign in respect of the tweet volume, it suggest than the technology industry is a well-discussed topic on Twitter. This is in line with Tumarkin and Whitelaw (2001)'s idea but is not true for the feelings embedded in tweets since some sectors have a higher negative/positive tweet percentage.

Similarly, the number of retweets is likely to increase for the food & beverage and insurance sectors in comparison to their reference. Plus, travel & leisure would show a higher number of replies than the technology area.

In the index point of view, only the CAC40 gives statistically significant results. It is likely to show higher percentages of positive and negative tweets as well as a lower tweet volume compared to the BEL20. This suggests that the CAC40 is less discussed proportionally to the BEL20 but that the emotions in tweets are more contrasted.

To conclude, we do not reject our hypothesis reporting that the market capitalization should influence the tweet volume but we reject our other assumption (i.e. $H_6$) that the percentage of retail investors in companies impacts Twitter activity. Now, as it has been already explained in the data processing part, we will do the same regressions but with $POS_{i,t}$ and $NEG_{i,t}$ reporting for the other sentiment analysis method using the Loughran-McDonald dictionary.

## Analyses based on sentiments from the Loughran-McDonald dictionary

In the following conformation, sentiments were computed using the Loughran-McDonald dictionary to account for the differences between a financial purpose and a more generic lexicon. What is more, since the Loughran-McDonald dictionary was originally created for financial statement analyses, it is interesting to reflect the change from financial documents to tweets analyses.

As a consequence, all variables stayed identical resulting in the same descriptive statistics and correlations except for $POS_{i,t}$ and $NEG_{i,t}$. For them, the descriptive statistics under the Loughran-McDonald dictionary are exhibited in appendix 14. The numbers of NA and values are the same between both methods but the number of null elements is different (i.e. value at zero either in the percent of positive or negative tweets). It is explained by the divergence in the number of neutral tweets between both models as shown in the appendix 13. Then, their means also differ: slightly for the negative emotion (i.e. 0,16 in LM against 0,18 in GI) whereas the difference is bigger for the positive feeling (i.e. 0,14 in LM against 0,46 in GI). This divergence is explained by the difference in the number of positive and negative tweets between both methods (*appendix 13*) since they do not compute emotions the same way.

In the table 8, the new Spearman correlations are displayed for $POS_{i,t}$ and $NEG_{i,t}$. They reveal changes in numbers and *p*-values of the relations between the percentage of positive and negative tweets and the other parameters. In fact, the modification is greater for the $POS_{i,t}$ variable whereas it only slightly changes for the $NEG_{i,t}$. The explanation also comes from the divergence in the number of positive tweets between the two sentiment analysis methods while the negative tweet volume remains close (*appendix 13*). So, we guess that the NLP classified the same tweets as negative in both dictionaries.

| | POS | NEG |
|---|---|---|
| **POS** | | |
| **NEG** | -0.1338*** | |
| **VOL** | -0.0050 | 0.0301*** |
| **MSG** | 0.1463*** | 0.1584*** |
| **RET** | -0.0051 | -0.0229** |
| **RTW** | 0.0236** | 0.0782*** |
| **RPL** | 0.0518*** | 0.0867*** |
| **CAP** | 0.0227*** | 0.0547*** |
| **RETAIL** | -0.0286** | -0.0131 |

*Table 8: Spearman correlations between the percentage of positive and negative tweets and the other parameters following the sentiment analysis based on the Loughran-McDonald dictionary*

First, the correlation between $POS_{i,t}$ and $NEG_{i,t}$ drops in absolute terms maybe because the global percentage of positive tweets is less important. Then the link between the tweet volume and the percent of positive messages increases from 2 to 15 percent. This strengthening of their bond modifies our previous idea of a negativity feeling inside the investing community since now the values for both $POS_{i,t}$ - and $NEG_{i,t}$ – $MSG_{i,t}$ are close to each other. An increase is also noticed for its relation to the number of replies but in a lower scale. On the contrary, the correlation between the number of retweets and $POS_{i,t}$ stays at the same absolute value but changes of sign: from negative to positive. This is not astonishing as the correlation was already low before. Then, for their links with others than Twitter features, the correlation between the market capitalization and the percentage of positive tweets increases and becomes statistically significant. Yet, the opposite effect happens for relation between $POS_{i,t}$ and return which may benefit our second hypothesis.

The regression tables are displayed on appendixes 15 and 16 as well as in table 9. We run the regressions in order to see if the sentiment dictionary change would modify our outcomes. Consequently, we can draw approximately the same conclusions from all our regressions using the Loughran-McDonald lexicon as what was previously done with the Harvard-IV dictionary expect for the last regression table hereafter in table 9.

**Regression results**

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | POSi,t | NEGi,t | MSGi,t | RTWi,t | RPLi,t |
| | (1) | (2) | (3) | (4) | (5) |
| MSGi,t | 0.007 | 0.033** | | 18.800*** | 0.251 |
| | (0.015) | (0.015) | | (0.833) | (0.311) |
| NEGi,t | -0.168*** | | 0.018** | 0.411 | -0.042 |
| | (0.010) | | (0.008) | (0.613) | (0.229) |
| POSi,t | | -0.181*** | 0.004 | 0.573 | 0.688*** |
| | | (0.011) | (0.008) | (0.637) | (0.237) |
| RTWi,t | 0.0002 | 0.0001 | 0.003*** | | 0.150*** |
| | (0.0002) | (0.0002) | (0.0001) | | (0.004) |
| RPLi,t | 0.002*** | -0.0002 | 0.0002 | 1.066*** | |
| | (0.001) | (0.001) | (0.0004) | (0.027) | |
| CAPi,t | 0.0004* | -0.00004 | 0.001*** | 0.032** | 0.001 |
| | (0.0003) | (0.0003) | (0.0002) | (0.016) | (0.003) |
| RETAILi | -0.077 | -0.039 | -0.010 | 0.250 | 0.542 |
| | (0.055) | (0.053) | (0.079) | (3.656) | (0.607) |
| RETi,t | -0.002 | -0.005** | 0.004*** | -0.040 | -0.024 |
| | (0.002) | (0.002) | (0.001) | (0.105) | (0.039) |
| VOLi,t | -0.006 | 0.002 | 0.005 | -0.030 | 0.030 |
| | (0.005) | (0.005) | (0.005) | (0.326) | (0.077) |
| AEX 25 | 0.014 | 0.008 | -0.043 | -1.180 | 0.175 |
| | (0.029) | (0.028) | (0.040) | (1.884) | (0.326) |
| CAC 40 | 0.010 | 0.047* | -0.075* | -0.489 | 0.154 |
| | (0.030) | (0.029) | (0.041) | (1.954) | (0.333) |
| Automobiles & Parts | 0.014 | -0.022 | -0.027 | -0.714 | -0.052 |
| | (0.026) | (0.025) | (0.036) | (1.683) | (0.280) |
| Banks | -0.015 | 0.046 | -0.082** | -1.565 | -0.160 |
| | (0.030) | (0.029) | (0.041) | (1.966) | (0.340) |
| Basic Resources | -0.011 | 0.007 | -0.036 | 1.049 | -0.394 |
| | (0.034) | (0.033) | (0.048) | (2.241) | (0.377) |
| Chemicals | 0.019 | 0.008 | -0.068 | -1.194 | -0.105 |
| | (0.030) | (0.029) | (0.042) | (1.967) | (0.329) |
| Construction & Materials | -0.040 | -0.014 | -0.007 | 1.392 | -0.09 |
| | (0.027) | (0.026) | (0.038) | (1.771) | (0.295) |
| Financial Services | -0.147 | -0.125 | 0.151 | -6.493 | 0.236 |
| | (0.142) | (0.137) | (0.203) | (9.387) | (1.557) |
| Food & Beverage | -0.032 | 0.088*** | -0.106** | 7.760*** | -0.192 |
| | (0.030) | (0.029) | (0.042) | (1.952) | (0.330) |
| Health Care | -0.018 | 0.010 | -0.038 | -0.442 | 0.013 |

| | | | | | |
|---|---|---|---|---|---|
| | (0.028) | (0.027) | (0.039) | (1.855) | (0.320) |
| Industrial Goods & Services | -0.023 | 0.024 | -0.021 | 0.890 | -0.376 |
| | (0.026) | (0.025) | (0.037) | (1.707) | (0.285) |
| Insurance | -0.025 | 0.039 | 0.027 | 9.422*** | 0.411 |
| | (0.029) | (0.028) | (0.040) | (1.901) | (0.334) |
| Media | -0.017 | 0.006 | -0.002 | -1.726 | 0.419 |
| | (0.027) | (0.026) | (0.038) | (1.761) | (0.294) |
| Oil & Gas | -0.077** | 0.025 | -0.124*** | 0.430 | -0.461 |
| | (0.036) | (0.035) | (0.048) | (2.349) | (0.412) |
| Personal & Household Goods | -0.025 | 0.048 | -0.067 | -2.979 | 0.05 |
| | (0.038) | (0.037) | (0.052) | (2.476) | (0.437) |
| Real Estate | -0.017 | 0.234** | -0.033 | -2.069 | 0.012 |
| | (0.123) | (0.118) | (0.175) | (8.091) | (1.340) |
| Retail | -0.067** | -0.033 | -0.036 | 1.485 | -0.269 |
| | (0.027) | (0.026) | (0.039) | (1.793) | (0.299) |
| Telecommunications | -0.017 | 0.023 | -0.078* | -1.194 | -0.146 |
| | (0.034) | (0.033) | (0.047) | (2.216) | (0.381) |
| Travel & Leisure | -0.051 | 0.027 | -0.003 | -1.790 | 2.190*** |
| | (0.033) | (0.032) | (0.047) | (2.190) | (0.363) |
| Utilities | 0.008 | -0.012 | -0.047 | -1.113 | -0.197 |
| | (0.037) | (0.036) | (0.053) | (2.456) | (0.412) |
| Constant | 0.257*** | 0.086 | 0.754*** | -11.780** | -0.836 |
| | (0.074) | (0.074) | (0.075) | (4.573) | (1.046) |
| Observations | 8,182 | 8,182 | 8,182 | 8,182 | 8,182 |
| $R^2$ | 0.036 | 0.039 | 0.082 | 0.237 | 0.200 |
| Adjusted $R^2$ | 0.032 | 0.035 | 0.079 | 0.234 | 0.197 |
| F Statistic | 301.600*** | 326.900*** | 728.200*** | 2,531.000*** | 2,035.000*** |
| *Note:* | | | | | *p<0.1; **p<0.05; ***p<0.01 |

*Table 9: Regression results of tweet, corporate and market features on Twitter parameters following the sentiment analysis based on the Loughran-McDonald dictionary*

In table 9, we see that the link between the percentages of positive and negative tweets is weaken by the use of the new lexicon : it goes from -0,6 to -0,17. The same change was revealed in their pairwise correlation. Moreover, the relationship between $POS_{i,t}$ and $VOL_{i,t}$ becomes non-significant while another one sees its *p*-value decrease until less than 10 percent to become significant: $MSG_{i,t}$ and $NEG_{i,t}$. Plus, as revealed in the correlation matrix, the link between the percentage of positive tweets and the number of replies is emphasised.

Then, almost all the previously significant links between industries and $POS_{i,t}$ are no longer statistically significant whereas two new sectors (oil & gas and retail) now tend to be lower than the technology industry in terms of percentage of positive tweets. The same phenomenon is observed for the percentage of negative tweets which sees three fifth of its previous significant relations disappears. By the way, the adjusted $R^2$ of the $POS_{i,t}$ and $NEG_{i,t}$ regressions decrease to 3 percent with the new sentiment indicators.

To conclude, variations are reported due to the sentiment dictionary change but they remain low except for some coefficients of the last regression. However, the changes do not affect the initial hypotheses we were tested. Therefore, we cannot determine the superiority of one method against the other. Since we have now achieved all our desired analyses, we can end this report by its conclusion and limits.

## Section VIII: Conclusion

In this master's thesis, our main objectives were to determine, on the one hand, whether the tweet volume and sentiment are good explanatory variables for stock returns and trading volume and, on the other hand, whether corporate variables help explaining Twitter activity. We based our analyses on a dataset comprising tweets about stocks related to the CAC 40, AEX 25 and BEL 20 from 2008 to 2018. This offered us an insight into the European market and by doing so differentiated us from previous studies.

Through this data, we generated sentiment variables reporting the percentages of positive and negative tweets about each stock at each date. We did this computation using two different dictionaries: first the Harvard-IV and then the Loughran-McDonald[50]. Moreover, we computed a measure for the stock daily volume of tweets but also the sum of daily retweets and replies for each stock in our dataset in order to reflect the tweet popularity. Then, we estimated the contemporaneous relationships between these variables and the stock return as well as trading volume.

From the regressions, we do not reject our third hypothesis according to which negative emotions help explaining stock market movements. In fact, the relationship between the former and stock returns was proven whereas the link with trading volume was not. The negative link between the percentage of negative tweets and stock returns suggests that the market reacts negatively to negative sentiments embedded in tweets. It is coherent with the behavioural finance assumptions which state that due to the arbitrageurs' limited power, noise traders impact more the market and so that stock prices can be influenced by emotions. Moreover, the impact of tweet volume on stock returns was shown. However, the link we were trying to demonstrate in our first hypothesis concerned the tweet and trading volumes but we find no evidence supporting this one. According to Dewally (2003)'s idea, people post about stocks having already known an increasing performance. If people read positive opinion about these stocks and decide to buy them, it creates a momentum effect by rising again their returns. Next, we also conclude that there is no evidence supporting our second and fourth hypothesis (i.e. positive sentiment and number of replies and retweets help explaining stock market

[50] The main results deriving from our regressions were sensibly the same using both dictionaries.

movements). This is coherent with Tirunillai and Tellis (2012)'s result that only negative UGC have an impact on returns.

Then, we extended the existing literature by computing the relationships between corporate features, represented by the market capitalization and percentage of retail investors, and Twitter parameters. Through the regressions done, we deduced that the companies' market capitalization does help explaining Twitter activity. Indeed, bigger firms get more tweets and retweets about their stocks than others. Therefore, it stresses the importance of the size and popularity of companies on Twitter. Plus, it confirms the theory of investor recognition but also extends the Tumarkin and Whitelaw (2001)'s result for microblogs to Twitter. By contrast, we do not find evidence of a link between the percentage of retail investors and Twitter activity (i.e. $H_6$). But in these regressions, we also confirm Zhang (2011)'s study showing that positive messages have a great reputation because the link between positive emotion and number of replies is positive and statistically significant. It suggests that this type of messages is appreciated or spark debate.

Furthermore, we added binary variables to reflect differences between industries and indexes. The results about tweet volume are in line with Tumarkin and Whitelaw (2001)'s theory as the technology sector is the most discussed area in our sample. Nevertheless, this is not true for emotions since some sectors show higher negative or positive tweet percentages. Then, the CAC 40 is proportionally less discussed than the BEL 20 but the emotions embedded in tweets are more contrasted.

Consequently, this master's thesis assert the explanatory power of Twitter in the European market but is under our expectations for some hypotheses. Nevertheless, the results can be useful for managers and investors of European companies as it allows them to understand that the volume as well as the opinions inside tweets are linked to stock returns.

Now that the conclusions of our study have been summarized, we still need to discuss about their limitations. This will be done in the next section.

## Section IX: Limitations & further works

Limitations to our study concern diverse topics. The first one is related to the efficiency of the market and the use of Twitter. Supposing that the number of informed investors goes up thanks to Twitter, their decisions would become more enlightened and the stock market more efficient. As a consequence, investors would get lower incentives to search for more information. However, data is never fully shared and available and so the market is not informationally perfectly efficient and prices reflect only a portion of the available information to informed investors. Therefore, the power of Twitter is not expected to decrease except if investors no longer use it. Nevertheless, as the technology world is in constant evolution, we may imagine a new communication channel replacing Twitter in the next years. Yet, as the number of Twitter users seems to stagnate for some years, its impact on the stock market is not likely to increase in the future.

The second limit refers to our dataset. As explained in the sixth section, we were not able to get one hundred percent coverage for tweets concerning our sample due to the restrictions imposed by Twitter in the data retrieval. In fact, we cannot even know the total amount of tweets for the time period and companies we investigated. The only way to be sure to recover all the tweets is to use an independent tweet provider. As a reminder, it is a company that get tweets thanks to an agreement with Twitter but their access is expensive. So, our results are likely to be modified with another sample. Nevertheless, for the outcomes confirming previous studies, we believe that they would not change. For further work, increasing the size of our dataset is a necessity.

The third limit concerns the users' advices and opinions. Despite the Wisdom of Crowd effect, consumers can be wrong in their posts. Through the "information cascades", people ignore their private information and comply with the crowd. Thus, if the crowd is wrong, this compliance will lead to bad investment decisions. The well-known fast diffusion of fake news is an example of bad quality tweets[51]. The same effect is possible with experts that people may follow blindly without having verified their sources (if there are). Therefore, the advice quality of Twitter can be questioned. Without forgetting that the website is free which makes it gather everything and anything. However as explained

[51] For instance, in 2013, hackers put a message on the Associated Press' account informing about explosions in the White House. This led to a decline in S&P500 and a change in VIX and was a proof of default of trading robots reacting to fake news (Foxman & Phillips, 2013).

in the literature review, Twitter users have an incentive to share valuable information as they want to be retweeted and followed. So we relied on this assumption for the quality of tweets.

Then, another limit is linked to the sentiment analysis. NLP when it gives tweets a rating based its choice on a dictionary. First, analysing tweets is not an easy task because they are composed of abbreviations, spelling and typing errors, slang language, etc. So, mistakes can be done if one word is taken for another (but the probability of that happening is pretty low) whereas the possibility to not consider an important word is bigger. The number of NA values in the sentiment analysis shows the limit of the NLP. Plus, in the dictionary point of view, tweets are different than financial statements (for the Loughran-McDonald dictionary) or psychological language (for the Harvard-IV lexicon). Therefore, their analysis using such dictionaries has its own limit. For the time being, no dictionary dedicated to tweet study has been created and this may be an opportunity for further researches. Another solution against these computations problems would be to use a supervised classification method but it is highly time consuming as words and tweets have to be manually assigned a rating. Besides, according to Bollen et al. (2011), binary sentiment indicators are less better predictors than more complex sentiment measures including other feelings. The calmness was, according to their study, the strongest indicator. An idea for further work would be to estimate also the score for other feelings such as fear or hope in order to reflect differences in both computation methods.

Finally, the study of Twitter impact on the stock market is not easy since stocks are influenced by a wide range of economic variables on one day. This let the possibility to several further researches. For instance, one may include the information demand, in addition to the supply, as it also influences the market. In this master's thesis, we decided not to take it into account to restrict our research. Then, for the non-rejected hypotheses and relationships found, a further work may consist in deepening these links. Their directions were not investigated here as they were already done in previous studies but a new section may be added to this report including a causality test (for instance the Granger one).

With all the elements explained in this master's thesis, we can go back to the sentence that launches our analyses and understand it completely now:

*"A slip of the foot you may soon recover, but a slip of the tongue you may never get over" Benjamin Franklin*

## Section X: Bibliography

Antweiler, W.& Frank, M.Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, 59, 1259-1294. doi: 10.1111/j.1540-6261.2004.00662.x.

Asur, S. & Huberman, B.A. (2010). Predicting the future with social media. *Proceedings International Conference on Web Intelligence and Intelligent Agent Technology*, 1, 492-499. doi: 10.1109/WI-IAT.2010.63.

Barber, B.M. & Odean, T. (2008). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *Review of Financial Studies*, 21, 785-818. doi: 10.1093/rfs/hhm079.

Behrendt & Schmidt. (2018). The Twitter myth revisited: Intraday investor sentiment, Twitter activity and individual-level stock return volatility. *Journal of Banking & Finance*. doi: 96. 10.1016/j.jbankfin.2018.09.016.

Bing, L., Keith C. C., & Carol O. (2014, november*). Public sentiment analysis in Twitter data for prediction of a company's stock price movements*. Paper presented at the 11th International Conference on eBusiness Engineering (ICEBE). doi: 10.1109/ICEBE.2014.47.

Birz, G & Lott Jr, J.R.(2011). The effect of macroeconomic news on stock returns: New evidence from newspaper coverage. *Journal of Banking & Finance*, 35, 2791-2800. doi: 10.1016/j.jbankfin.2011.03.006.

Bloomberg. (2013). *Bloomberg launches a Twitter feed optimized for trading*. Retrieved from https://www.bloomberg.com/company/announcements/bloomberg-launches-twitter-feed-optimized-trading/.

Bodie, Z., Kane, A. & Marcus, A.J. (2014). *Investments*. New York: McGraw-Hill Education.

Boursorama. (2019). *Cours des actions par marché. Accès direct aux différents marché des actions. Indice CAC 40.* Retrieved June 12, 2019 from

https://www.boursorama.com/bourse/actions/cotations/?quotation_az_filter%5Bmarket%5D=1rPCAC.

Breed, M. (n.d.). *11 Best Stock Market Investment News, Analysis & Research Sites*. Retrieved from https://www.moneycrashers.com/best-stock-market-investment-news-analysis-research-sites/

Brightplanet. (2013). *Twitter firehose vs. Twitter API: What's the difference and why should you care?.* Retrieved from https://brightplanet.com/2013/06/25/twitter-firehose-vs-twitter-api-whats-the-difference-and-why-should-you-care.

Cha, M., Haddadi, H., Benevenuto, F. & Gummadi, K.P. (2010, May). *Measuring user influence in Twitter: the million follower fallacy*. Paper presented at the AAAI 14th Conference on Weblogs and Social Media, Washington. Retrieved from https://www.researchgate.net/publication/221298004_Measuring_User_Influence_in_Twitter_The_Million_Follower_Fallacy.

Chaffey, D. (2019). *Global social media research summary 2019*. Retrieved from Smartinsights website: https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/.

Chambers, A.E. & Penman, S.H. (1984). Timeliness of reporting and the stock price reaction to earnings announcements. *Journal of Accounting Research*, 22, 21-47. doi: 10.2307/2490700.

Chevalier, J.A. & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43, 345-354. doi: 10.1509/jmkr.43.3.345.

Choffat, A. (2019). *Changement d'heure: nous sommes à l'heure d'été, et après?* Retrieved from https://www.linternaute.com/actualite/societe/1038385-changement-d-heure-l-heure-d-ete-est-arrivee-pour-la-derniere-fois.

Ciftci, K. & Ozturk, S. (2015). A sentiment analysis of twitter content as a predictor of exchange rate movements. *Rimini Centre for Economic Analysis*, 6, 132-140. doi: 10.13140/RG.2.1.1022.9201.

Clément, G. (2018). *Bourse : Tous ces tweets qui ont affolé les marchés financiers*. Retrieved from https://www.lerevenu.com/bourse/bourse-tous-ces-tweets-qui-ont-affole-les-marches-financiers.

CNN. (n.d.). *What is the Fear & Greed Index?.* Retrieved from https://money.cnn.com/investing/about-fear-greed-tool/index.html.

Coëffé, T. (2018). *Chiffres Twitter – 2018.* Retrieved from https://www.blogdumoderateur.com/chiffres-twitter.

Coeurderoy, R., Neysen, N. & Paque, B. (2018). *Open innovation*. Syllabus, Université catholique de Louvain.

Da, Z., Engelberg, J. & Gao, P. (2011). In search of attention. *The Journal of Finance*, 66, 1461-1499. doi: 10.1111/j.1540-6261.2011.01679.x.

Daniel, K., Hirshleifer, D. & Subrahmanyam, A. (2002). Investor Psychology and Security Market Under- and Overreactions. *The Journal of Finance*, 53, 1839-1885. doi: 10.1111/0022-1082.00077.

Das, S. & Chen, M.Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*, 53, 1375-1388. doi: 10.1287/mnsc.1070.0704.

De Long, J.B., Shleifer, A., Summers, L.H. & Waldmann, R.J. (1990). Noise trader risk in financial markets. *The Journal of Political Economy*, 98, 703-738. doi: 10.1086/261703.

Dewally, M. (2003). Internet Investment Advice: Investing with a Rock of Salt. Financial Analysts Journal 59, 65–77. doi: 10.2139/ssrn.206089

Dhar III, V. & Chang, E.A. (2007). Does Chatter Matter? The Impact of User-Generated Content on Music Sales. *Journal of Interactive Marketing*, 23, 300-307. doi: 10.2139/ssrn.1113536.

Edmans, A., Garcia, D. & Norli, O. (2007). Sports sentiment and stock returns. *The Journal of Finance*, 62, 1967-1998. doi: 10.1111/j.1540-6261.2007.01262.x.

Euronext. (June 2019). *CAC 40*. Unpublished document.

Euronext. (June 2019). *AEX 25*. Unpublished document.

Euronext. (June 2019). *BEL 20*. Unpublished document.

Fama, E. (1965). The behaviour of stock market prices. *Journal of Business*, 64, 34-105. Retrieved from https://www.jstor.org/stable/2350752.

Fama, E. (1970). Efficient capital markets: a review of theory and empirical work. *The Journal of Finance*, 25, 383-417. doi: 10.2307/2325486

Farnsworth, G.V. (2008). *Econometrics in R.* Unpublished document.

Farrington, R. (2018). *Best social networks for investors*. Retrieved from https://thecollegeinvestor.com/21341/best-social-networks-investors.

Fehle, F. R., Tsyplakov, S. & Zdorovtsov, V. (2005). Can companies influence investor behaviour through advertising? Super Bowl commercials and stock returns. *European Financial Management*, 11, 625-647. doi: 10.2139/ssrn.477301.

Fogel, J. & Nehmad, E. (2009). Internet social network communities: Risk taking, trust, and privacy concerns. *Computers in Human Behavior*, 25, 153-160. doi: 153-160. 10.1016/j.chb.2008.08.006.

Forbergskog, J.-O. & Blom, C.R. (2013). *Twitter and stock returns* (Master's thesis). BI Norwegian Business School, Oslo.

Foxman, S. & Phillips, M. (2013). What Happened to Stock Markets When the AP's Twitter Account Was Hacked. Retrieved from https://www.theatlantic.com/technology/archive/2013/04/what-happened-to-stock-markets-when-the-aps-twitter-account-was-hacked/275230/

Gilles, J. (2006). Internet encyclopaedias go head to head. *Nature*, 438, 900-901. doi: 10.1038/438900a.

Greenfield, R. (2011). *This is how a Twitter-based hedge fund beat the stock market*. Retrieved from https://www.theatlantic.com/business/archive/2011/08/how-twitter-based-hedge-fund-beat-stock-market/354245/.

Groß-Klußmann, A. & König, S. & Ebner, M. (2019). Buzzwords build Momentum: Global financial Twitter sentiment and the aggregate stock market. *Expert Systems with Applications*. doi: 10.1016/j.eswa.2019.06.027.

Hackernoon. (n.d.) *Text processing and sentiment analysis of Twitter data*. Retrieved from https://hackernoon.com/text-processing-and-sentiment-analysis-of-twitter-data-22ff5e51e14c.

Heather, K. (2013). *Twitter hacked; 250,000 accounts affected*. Retrieved from https://edition.cnn.com/2013/02/01/tech/social-media/twitter-hacked/index.html.

Hill, S. & Ready-Campbell, N. (2011). Expert stock picker: The wisdom of (experts in) crowds. *International Journal of Electronic Commerce*, 15 (3), 73-102 . doi: 10.2753/JEC1086-4415150304.

Hirshleifer, D. & Shumway, T. (2003). Good day sunshine: Stock returns and the weather. *The Journal of Finance*, 58, 1009-1032. doi: 10.1111/1540-6261.00556.

Hootsuite. (2019). *The Global State of Digital in 2019 Report*. Retrieved from https://hootsuite.com/pages/digital-in-2019.

Iania, L. & Nguyen, A. (2017-2018). *Foundations of investment: Course 1*. Syllabus, Université catholique de Louvain.

Jockers, M. (n.d.) *Syuzhet v1.0.4*. Retrieved from https://www.rdocumentation.org/packages/syuzhet/versions/1.0.4.

Johnson, E.J & Tversky, A. (1983). Affect, generalization, and the perception of risk. *Journal of Personality and Social Psychology*, 45, 20-31. doi: 10.1037/0022-3514.45.1.20.

Kahneman, D. and Tversky, A. (1972). Subjective Probability: A Judgment of Representativeness. *Cognitive Psychology*, 3, 430–454. doi: 10.1016/0010-0285(72)90016-3.

Kahneman, D. and Tversky, A. (1973). On the Psychology of Prediction. *Psychology Review*, 80, pp. 237–51. doi: 10.1037/h0034747.

Kahneman, D. and Tversky, A. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, 106, 1039-1061. doi: 10.2307/2937956.

Kamtra, M.J., Kramer, L.A. & Levi, M.D. (2003). Winter blues: a SAD stock market cycle. *American Economic Review*, 93, 324-343. doi: 10.2139/ssrn.208622.

Kelly, K. (2011). *Hedge fund to predict markets using Twitter*. Retrieved from https://www.cnbc.com/id/41948279

Kim, A. & Meschke, F. (2011). *CEO interviews on CNBC*. Unpublished document.

Kim, H. (2018). *Limits of the Bing, AFINN, and NRC Lexicons with the Tidytext Package in R.* Retrieved from https://hoyeolkim.wordpress.com/2018/02/25/the-limits-of-the-bing-afinn-and-nrc-lexicons-with-the-tidytext-package-in-r.

La Libre (2019). *Les débordements d'Elon Musk sur Twitter risquent de coûter cher à Tesla*. Retrieved from https://www.lalibre.be/economie/libre-entreprise/les-debordements-d-elon-musk-sur-twitter-risquent-de-couter-cher-a-tesla-5c74d5b2d8ad5878f0f2a843.

Levy, T. & Yagil, J. (2011). Air pollution and stock returns in the US. *Journal of Economic Psychology,* 32, 374-383. doi: 10.1016/j.joep.2011.01.004.

Manson, M. (2019). *C-Suite execs need to up their game on social if they want to keep investors close.* Retrieved from Brunswick website: https://www.brunswickgroup.com/digital-investor-survey-c-suite-use-social-media-to-keep-investors-close-i9475/.

Mao, H., Counts, S. & Bollen, J. (2011). *Predicting financial markets: Comparing survey, news, twitter and search engine data.* Retrieved from https://arxiv.org/abs/1112.1051.

Merton, R.C. (1987). A simple model of capital market equilibrium with incomplete information. *Journal of Finance*, 42, 483-510. doi: 10.1111/j.1540-6261.1987.tb04565.x

Miller, B. & Steyvers, M. (2011). The wisdom of crowds with communication. *Annual Meeting of the Cognitive Science Society*, 33rd, 1292-1297. Retrieved from https://cloudfront.escholarship.org/dist/prd/content/qt4jt6q62c/qt4jt6q62c.pdf.

Naldi, M. (2019). *A review of sentiment computation methods with R packages*. Unpublished document

Nofer, M.. (2014). *The value of social media for predicting stock returns: Preconditions, instruments and performance analysis.* Darmstad: Springer Vieweg. doi: 10.1007/978-3-658-09508-6.

Oh, C. & Sheng, O. (2011*). Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement*. Paper presented at the 4th International Conference on Information Systems, Shangai. Retrieved from http://www.misrc.umn.edu/workshops/2011/fall/OliviaSheng_Paper.pdf.

Oliveira & Cortez & Areal. (2016). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications*. doi: 73. 10.1016/j.eswa.2016.12.036.

Pagolu, S., Nayan Reddy Challa, K., Panda, G. & Majhi, B. (2016, october). *Sentiment analysis of Twitter data for predicting stock market movements*. Paper presented at International conference on Signal Processing, Communication, Power and Embedded System (SCOPES), Paralakhemundi. Retrieved from https://www.researchgate.net/publication/309551728_Sentiment_Analysis_of_Twitter_Data_for_Predicting_Stock_Market_Movements.

Parker, T. (2012). *10 Twitter feeds investors should follow*. Retrieved from https://www.investopedia.com/financial-edge/0712/10-twitter-feeds-investors-should-follow.aspx.

Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M. & Mozetic, I. (2015). The effects of Twitter sentiment on stock price returns. *PLoS one*, 10, e0138441. doi: 10.1371/journal.pone.0138441.

Sagar, C. (2018). *Twitter sentiment analysis using R*. Retrieved from https://dataaspirant.com/2018/03/22/twitter-sentiment-analysis-using-r.

Scott, G. (2019). *Euronext*. Retrieved from https://www.investopedia.com/terms/e/euronext.asp.

SEC. (2013). *SEC says social media OK for company announcements if investors are alerted.* Retrieved from https://www.sec.gov/news/press-release/2013-2013-51htm.

Skiera, B. & Spann, M. (2003). Internet-based virtual stock markets for business forecasting. *Management Science*, 49, 1310-1326. doi: 10.1287/mnsc.49.10.1310.17314.

Smailovic, J., Grcar, M., & Znidarsic, M. (2012). *Sentiment analysis on tweets in a financial domain.* Unpublished document.

*Social media*. (2019). Merriam-Webster (11th ed.). Retrieved from https://www.merriam-webster.com/dictionary/social%20media.

Sprenger, T.O., Tumasjan, A., Sandner, P.G. & Welpe, I.M. (2014). Tweets and trades: The information content of stock microblogs. *European Financial Management*, 20, 926-957. doi: 10.1111/j.1468-036X.2013.12007.x.

Statista. (2019). *Distribution of Twitter users worldwide as of April 2019, by age group*. Retrieved from https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users.

Statista. (2019). *Twitter: Number of monthly active users 2010-2019*. Retrieved from https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users.

Symanovich, S. (n.d.). *Privacy vs. security: what's the difference?*. Retrieved from https://us.norton.com/internetsecurity-privacy-privacy-vs-security-whats-the-difference.html.

Takahashi, S., Takahashi, M., Takahashi, H. & Tsuda, K. (2007). Analysis of the relation between stock price returns and headline news using text categorization. *Lecture Notes in Artificial Intelligence*, 4693, 1339-1345. doi: 10.1007/978-3-540-74827-4_167.

Takeda, F. & Wakao, T. (2014). Google search intensity and its relationship with returns and trading volume of Japanese stocks. *Pacific-Basin Finance Journal*, 27, 1–18. doi: 10.1016/j.pacfin.2014.01.003.

Tellis, G.J. & Johnson, J. (2007). The value of quality. *Marketing Science*, 26, 758-773. doi: 10.1287/mksc.1070.0286.

Tetlock, P.C. (2007). Giving content to investor Sentiment: The role of media in the stock market. *The Journal of Finance*, 62, 1139-1168. doi: 10.1111/j.1540-6261.2007.01232.x.

Tirunillai, S. & Tellis, G.J.; (2012); Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Science*, 31, 198-215. doi: 10.1287/mksc.1110.0682.

Trusov, M., Bucklin, R.E. & Pauwels, K. (2009). Effects of word-of-mouth versus traditional marketing: Findings from an Internet social networking site. *Journal of Marketing*, 73, 90-102. doi: 10.2139/ssrn.1129351.

Tumarkin, R. & Whitelaw, R. (2001). News or noise? Internet postings and stock prices. *Financial Analysts Journal*, 57(3), 41-51. doi: 10.2469/faj.v57.n3.2449.

Twitter. (n.d.). *Getting started*. Retrieved from https://help.twitter.com/en/twitter-guide.

Twitter. (n.d.). *Faites l'actualité. Diffusez votre message, élargissez votre audience et générez plus de trafic avec les publicités Twitter.* Retrieved from https://business.twitter.com/fr.html.

*User-generated*. (2019). Oxford Dictionaries. Retrieved from https://en.oxforddictionaries.com/definition/user-generated.

Vlastakis, N. & Markellos, R.N. (2012). Information demand and stock market volatility. *Journal of Banking & Finance*, 36, 1808-1821. doi: 10.1016/j.jbankfin.2012.02.007.

Watson W.E., Kuman, K. & Michaelsen, L.K. (1993). Cultural diversity's impact on group process and performance: Comparing culturally homogeneous and culturally diverse task groups. *The Academy of Management Journal*, 36, 590-602. doi: 10.2307/256593.

Wikipedia. (2010). *Calendar anomalies*. Retrieved Mei 22, 2019 from https://en.wikipedia.org/wiki/Market_anomaly#Calendar_anomalies.

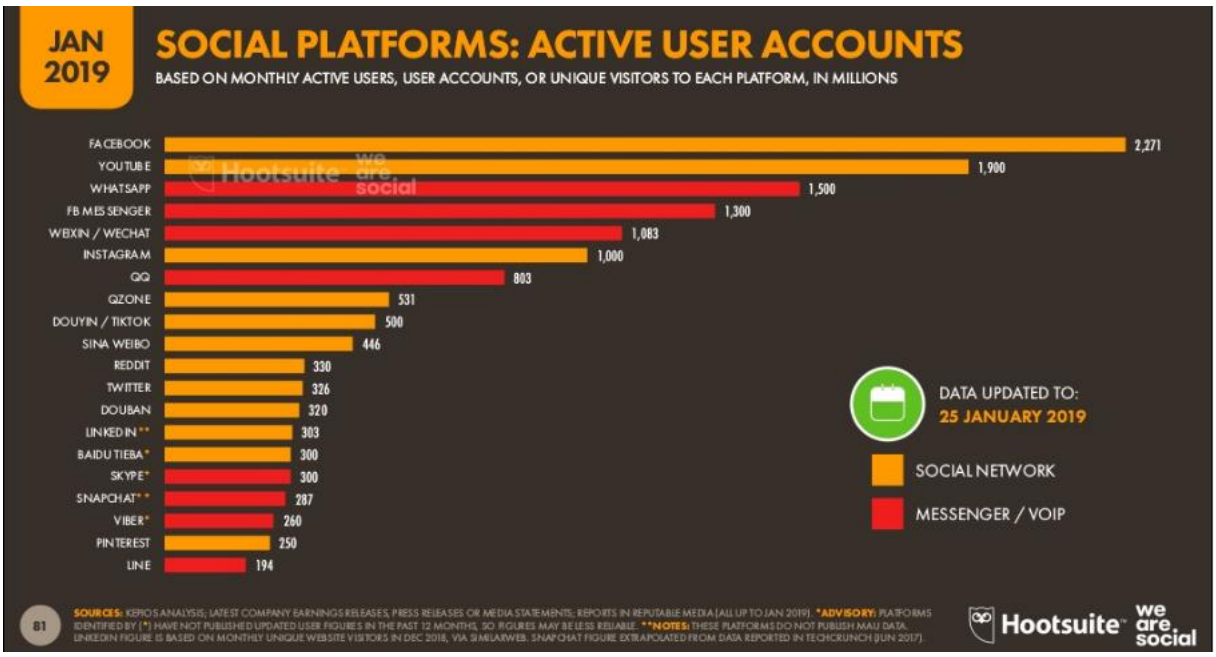
Wikipedia. (2019). *Twitter*. Retrieved Mei 18, 2019 from https://en.wikipedia.org/wiki/Twitter.

Wysocki, P. (1999). *Cheap talk on the web: The determinants of postings on stock message boards*. University of Michigan Business School Working Paper, N°: 98025SSRN. doi: 10.2139/ssrn.160170.

Zhang, X., Fuehres, H. & Gloor. P.A. (2011). Predicting stock market indicators through Twitter "I hope it is not as bad as I fear*". Procedia – social and behavioural sciences*, 26, 55-62. doi: 10.1016/j.sbspro.2011.10.562.

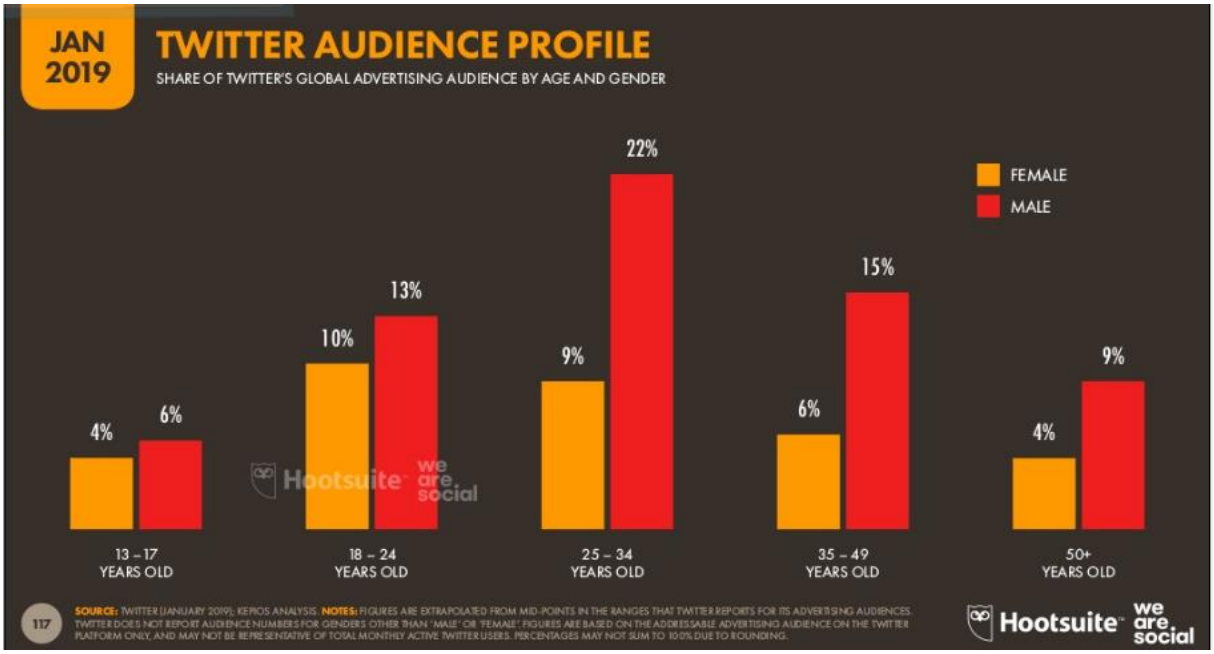
## Section XI: Appendixes

*Appendix 1: Distribution of active users through social media*



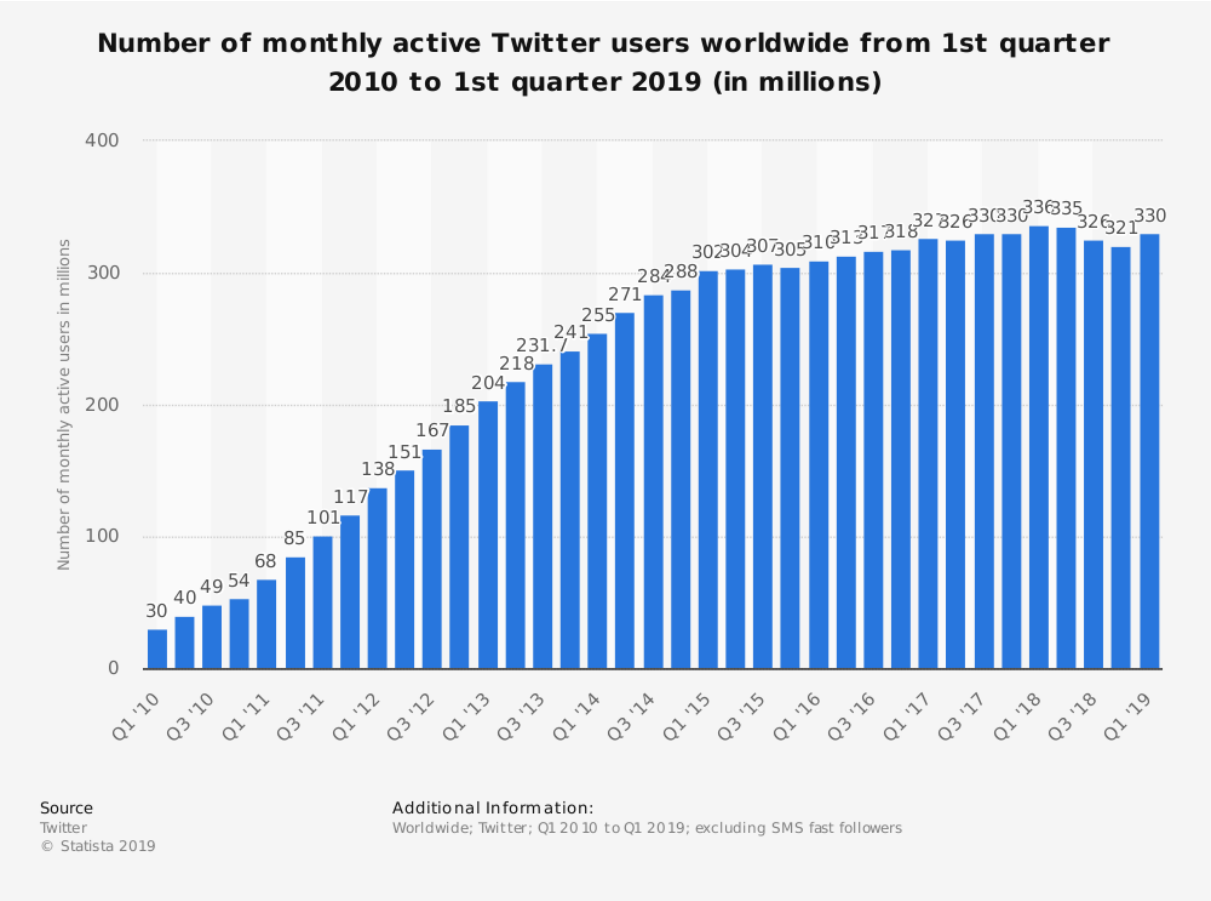

Source: Hootsuite and We are social retrieved from https://hootsuite.com/pages/digital-in-2019

*Appendix 2:* Distribution of Twitter audience




Source: Hootsuite and We are social retrieved from https://hootsuite.com/pages/digital-in-2019

*Appendix 3: Evolution of the number of Twitter users*




Source: Statista retrieved from

https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users

*Appendix 4: Kylie Jenner's tweet causing the drop in the Snapchat stock price*


*Source: Twitter retrieved from*

*https://twitter.com/larrykim/status/966865430533910528*

*Appendix 5: CAC 40 composition*

## Components

| Company | MNEMO | Cnty | Sector (ICB) | Weight (%) | Index Cap | Float |
|---|---|---|---|---|---|---|
| TOTAL | FP | FR | Oil & Gas | 10,06 | 124,72 | 0,95 |
| LVMH | MC | FR | Personal & Household Goods | 8,39 | 103,97 | 0,55 |
| SANOFI | SAN | FR | Health Care | 6,90 | 85,50 | 0,9 |
| AIRBUS | AIR | FR | Industrial Goods & Services | 5,86 | 72,59 | 0,75 |
| L'OREAL | OR | FR | Personal & Household Goods | 5,09 | 63,14 | 0,45 |
| AIR LIQUIDE | AI | FR | Chemicals | 4,25 | 52,70 | 1 |
| DANONE | BN | FR | Food & Beverage | 3,91 | 48,48 | 0,95 |
| BNP PARIBAS ACT.A | BNP | FR | Banks | 3,79 | 46,98 | 0,9 |
| SCHNEIDER ELECTRIC | SU | FR | Industrial Goods & Services | 3,73 | 46,18 | 1 |
| AXA | CS | FR | Insurance | 3,61 | 44,78 | 0,8 |
| VINCI | DG | FR | Construction & Materials | 3,47 | 43,00 | 0,8 |
| SAFRAN | SAF | FR | Industrial Goods & Services | 3,47 | 42,99 | 0,8 |
| KERING | KER | FR | Retail | 3,18 | 39,41 | 0,6 |
| ESSILORLUXOTTICA | EL | FR | Health Care | 2,81 | 34,81 | 0,7 |
| PERNOD RICARD | RI | FR | Food & Beverage | 2,60 | 32,26 | 0,75 |
| ORANGE | ORA | FR | Telecommunications | 2,23 | 27,66 | 0,75 |
| ENGIE | ENGI | FR | Utilities | 1,97 | 24,37 | 0,75 |
| VIVENDI | VIV | FR | Media | 1,78 | 22,12 | 0,7 |
| MICHELIN | ML | FR | Automobiles & Parts | 1,53 | 18,97 | 0,95 |
| DASSAULT SYSTEMES | DSY | FR | Technology | 1,48 | 18,35 | 0,5 |
| UNIBAIL-RODAMCO-WE | URW | NL | Real Estate | 1,47 | 18,22 | 1 |
| SAINT GOBAIN | SGO | FR | Construction & Materials | 1,45 | 17,99 | 0,95 |
| CAPGEMINI | CAP | FR | Technology | 1,41 | 17,54 | 0,95 |
| LEGRAND | LR | FR | Industrial Goods & Services | 1,38 | 17,16 | 1 |
| SOCIETE GENERALE | GLE | FR | Banks | 1,38 | 17,05 | 0,95 |
| HERMES INTL | RMS | FR | Personal & Household Goods | 1,35 | 16,74 | 0,25 |
| PEUGEOT | UG | FR | Automobiles & Parts | 1,03 | 12,74 | 0,65 |
| VEOLIA ENVIRON. | VIE | FR | Utilities | 0,98 | 12,12 | 1 |
| CREDIT AGRICOLE | ACA | FR | Banks | 0,98 | 12,09 | 0,4 |
| THALES | HO | FR | Industrial Goods & Services | 0,93 | 11,55 | 0,5 |
| STMICROELECTRONICS | STM | FR | Technology | 0,86 | 10,65 | 0,75 |
| RENAULT | RNO | FR | Automobiles & Parts | 0,86 | 10,63 | 0,65 |
| ARCELORMITTAL SA | MT | NL | Basic Resources | 0,84 | 10,45 | 0,65 |
| CARREFOUR | CA | FR | Retail | 0,81 | 10,05 | 0,75 |
| TECHNIPFMC | FTI | FR | Oil & Gas | 0,81 | 10,03 | 0,95 |
| PUBLICIS GROUPE SA | PUB | FR | Media | 0,79 | 9,83 | 0,9 |
| SODEXO | SW | FR | Travel & Leisure | 0,75 | 9,30 | 0,6 |
| BOUYGUES | EN | FR | Construction & Materials | 0,63 | 7,75 | 0,65 |
| ACCOR | AC | FR | Travel & Leisure | 0,62 | 7,67 | 0,7 |
| ATOS | ATO | FR | Technology | 0,57 | 7,06 | 0,9 |

*Source: Euronext retrieved from*

*https://live.euronext.com/en/product/indices/FR0003500008-XPAR*

*Appendix 6: AEX 25 composition*

## Components

| Company | MNEMO | Cnty | Sector (ICB) | Weight (%) | Index Cap | Float |
|---|---|---|---|---|---|---|
| ROYAL DUTCH SHELLA | RDSA | NL | Oil & Gas | 14,91 | 83,41 | 100% |
| ASML HOLDING | ASML | NL | Technology | 13,98 | 78,21 | 100% |
| UNILEVER DR | UNAT | NL | Personal & Household Goods | 13,96 | 78,06 | 85% |
| RELX | REN | NL | Media | 7,82 | 43,74 | 100% |
| ING GROEP N.V. | INGA | NL | Banks | 7,10 | 39,68 | 100% |
| PHILIPS KON | PHIA | NL | Health Care | 6,32 | 35,37 | 100% |
| AHOLD DEL | AD | NL | Retail | 4,19 | 23,41 | 100% |
| HEINEKEN | HEIA | NL | Food & Beverage | 4,04 | 22,61 | 40% |
| DSM KON | DSM | NL | Chemicals | 3,35 | 18,73 | 95% |
| UNIBAIL-RODAMCO-WE | URW | NL | Real Estate | 3,26 | 18,22 | 100% |
| WOLTERS KLUWER | WKL | NL | Media | 3,20 | 17,91 | 100% |
| AKZO NOBEL | AKZA | NL | Chemicals | 3,03 | 16,94 | 90% |
| NN GROUP | NN | NL | Insurance | 1,94 | 10,87 | 90% |
| ARCELORMITTAL SA | MT | NL | Basic Resources | 1,87 | 10,45 | 65% |
| KPN KON | KPN | NL | Telecommunications | 1,72 | 9,65 | 85% |
| ADYEN | ADYEN | NL | Industrial Goods & Services | 1,44 | 8,03 | 40% |
| ABN AMRO Group | ABN | NL | Banks | 1,35 | 7,53 | 85% |
| AEGON | AGN | NL | Insurance | 1,31 | 7,34 | 80% |
| RANDSTAD NV | RAND | NL | Industrial Goods & Services | 1,03 | 5,75 | 65% |
| ASR NEDERLAND | ASRNL | NL | Insurance | 0,90 | 5,04 | 100% |
| GALAPAGOS | GLPG | NL | Health Care | 0,88 | 4,94 | 80% |
| IMCD | IMCD | NL | Chemicals | 0,76 | 4,24 | 100% |
| TAKEAWAY | TKWY | NL | Retail | 0,59 | 3,28 | 65% |
| AALBERTS NV | AALB | NL | Industrial Goods & Services | 0,58 | 3,25 | 85% |
| VOPAK | VPK | NL | Industrial Goods & Services | 0,46 | 2,59 | 50% |

*Source: Euronext retrieved from*

*https://live.euronext.com/en/product/indices/NL0000000107-XAMS/market-information*

*Appendix 7: BEL 20 composition*

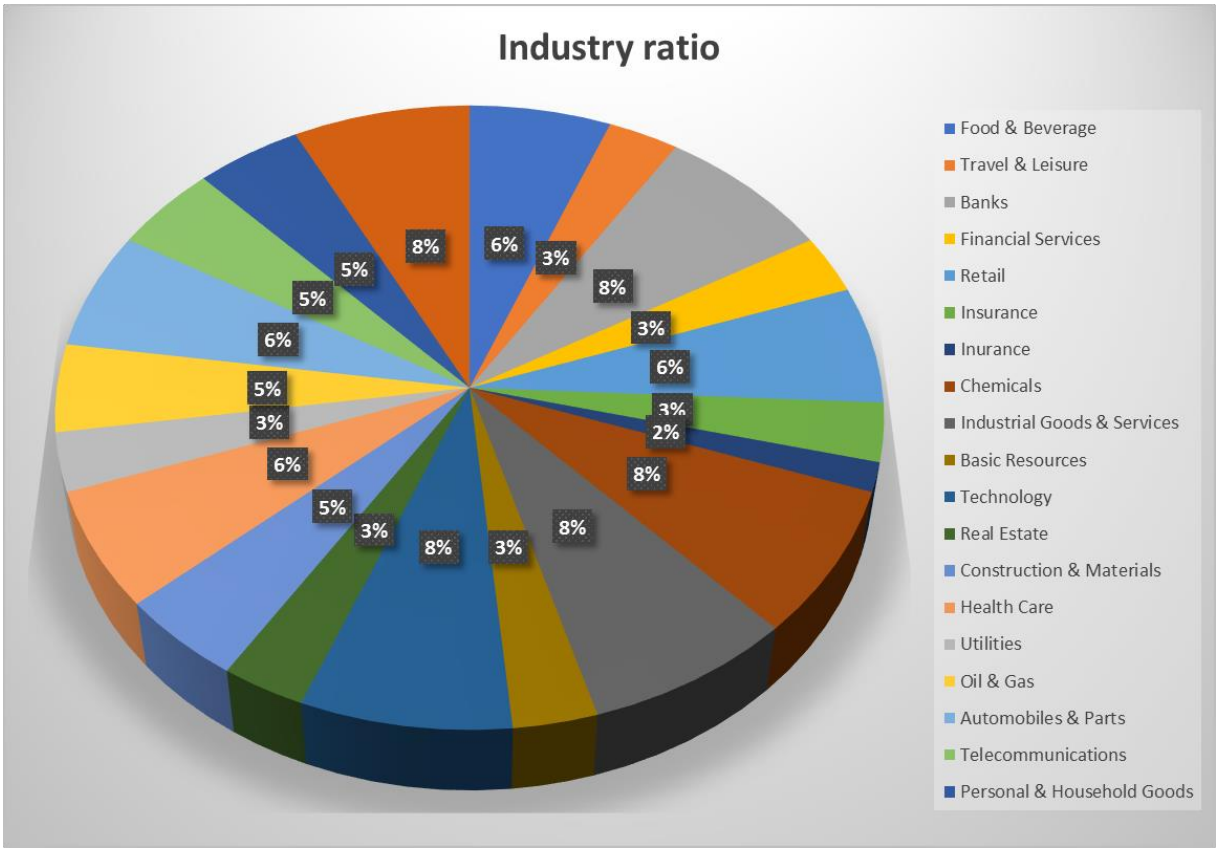
## Components

| Company | MNEMO | Cnty | Sector (ICB) | Weight (%) | Index Cap | Float |
|---|---|---|---|---|---|---|
| AB INBEV | ABI | BE | Food & Beverage | 13,32 | 14,69 | 0,45 |
| KBC | KBC | BE | Banks | 11,49 | 12,68 | 0,6 |
| ING GROEP N.V. | INGA | NL | Banks | 11,47 | 12,65 | 1 |
| UCB | UCB | BE | Health Care | 8,36 | 9,22 | 0,65 |
| AGEAS | AGS | BE | Insurance | 7,99 | 8,82 | 0,95 |
| GBL | GBLB | BE | Financial Services | 6,31 | 6,96 | 0,5 |
| SOLVAY | SOLB | BE | Chemicals | 5,68 | 6,27 | 0,65 |
| UMICORE | UMI | BE | Chemicals | 5,35 | 5,91 | 0,85 |
| GALAPAGOS | GLPG | NL | Health Care | 4,48 | 4,94 | 0,8 |
| ARGENX SE | ARGX | BE | Health Care | 4,21 | 4,64 | 1 |
| PROXIMUS | PROX | BE | Telecommunications | 3,57 | 3,94 | 0,45 |
| ACKERMANS V.HAAREN | ACKB | BE | Financial Services | 2,60 | 2,87 | 0,65 |
| SOFINA | SOF | BE | Financial Services | 2,34 | 2,58 | 0,45 |
| COLRUYT | COLR | BE | Retail | 2,32 | 2,56 | 0,35 |
| WDP | WDP | BE | Real Estate | 2,32 | 2,56 | 0,75 |
| COFINIMMO | COFB | BE | Real Estate | 2,31 | 2,55 | 1 |
| TELENET GROUP | TNET | BE | Media | 2,09 | 2,31 | 0,4 |
| BARCO | BAR | BE | Industrial Goods & Services | 1,76 | 1,94 | 0,8 |
| APERAM | APAM | NL | Basic Resources | 1,15 | 1,27 | 0,6 |
| ONTEX GROUP | ONTEX | BE | Personal & Household Goods | 0,85 | 0,93 | 0,8 |

*Source: Euronext retrieved from*

*https://live.euronext.com/en/product/indices/BE0389555039-XBRU/market-information*

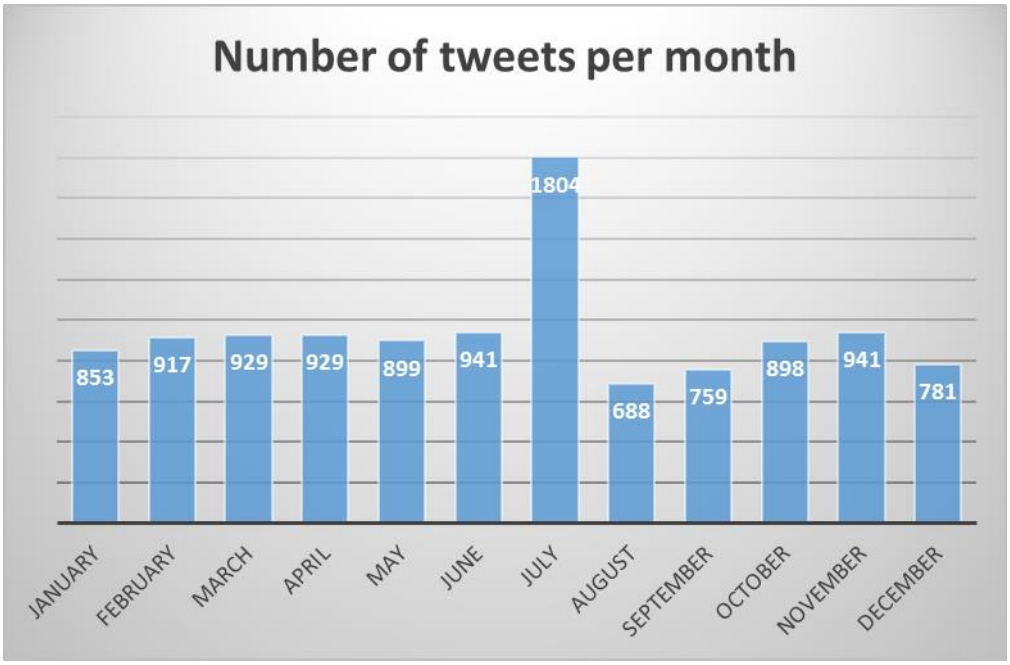
*Appendix 8: Industry composition of our tweet sample*



*Appendix 9: Tweet number per ticker in our sample*

| Ticker | Tweet number | Ticker | Tweet number |
|---|---|---|---|
| ABI | 1 | KBC | 26 |
| AC | 193 | KER | 123 |
| ACA | 188 | KPN | 79 |
| ACKB | 13 | LR | 120 |
| AD | 359 | MC | 239 |
| AGN | 718 | ML | 152 |
| AGS | 6 | MT | 342 |
| AI | 263 | ORA | 219 |
| AIR | 360 | PHIA | 79 |
| AKZA | 95 | PROX | 80 |
| APAM | 189 | PUB | 164 |
| ASML | 32 | RAND | 144 |
| ATO | 157 | RDSA | 233 |
| BN | 234 | REN | 333 |
| BNP | 189 | RI | 120 |
| CA | 277 | RMS | 111 |
| CAP | 195 | RNO | 203 |
| COFB | 3 | SAF | 134 |
| COLR | 3 | SAN | 436 |
| CS | 453 | SGO | 133 |
| DG | 394 | SOLB | 118 |
| DSM | 105 | STM | 276 |
| DSY | 92 | SU | 323 |
| EL | 315 | SW | 124 |
| EN | 324 | TNET | 24 |
| ENGI | 74 | UCB | 18 |
| FP | 124 | UG | 167 |
| FR | 213 | UMI | 80 |
| FTI | 218 | UNA | 125 |
| GBLB | 5 | URW | 15 |
| GLE | 135 | VIE | 142 |
| HEIA | 98 | VIV | 212 |
| INGA | 106 | WKL | 114 |

*Appendix 10: Monthly tweet distribution of our sample*




*Appendix 11: Sentiment value descriptive statistics*

| Min | 1st Qu. | Median | Mean | 3rd Qu. | Max | NA's |
|---|---|---|---|---|---|---|
| -1,00000 | 0,00000 | 0,00000 | 0,06751 | 0,16667 | 1,00000 | 106,00000 |

*Appendix 12: Descriptive statistics of tweet and market variables*

| | POSi,t | NEGi,t | Tweet_volumei,t | RTWi,t | RPLi,t | RETi,t |
|---|---|---|---|---|---|---|
| **nbr.val** | 8.703,0000 | 8.703,0000 | 181.830,0000 | 8.705,0000 | 8.705,0000 | 181.523,0000 |
| **nbr.null** | 4.374,0000 | 6.903,0000 | 173.125,0000 | 3.724,0000 | 6.925,0000 | 1.490,0000 |
| **nbr.na** | 173.127,0000 | 173.127,0000 | 0,0000 | 173.125,0000 | 173.125,0000 | 307,0000 |
| **min** | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | -228,9694 |
| **max** | 1,0000 | 1,0000 | 35,0000 | 693,0000 | 623,0000 | 229,6103 |
| **range** | 1,0000 | 1,0000 | 35,0000 | 693,0000 | 623,0000 | 458,5797 |
| **sum** | 3.970,6147 | 1.538,3043 | 11.339,0000 | 40.114,0000 | 6.831,0000 | 1.808,8862 |
| **median** | 0,0000 | 0,0000 | 0,0000 | 1,0000 | 0,0000 | 0,0231 |
| **mean** | 0,4562 | 0,1768 | 0,0624 | 4,6080 | 0,7847 | 0,0100 |
| **SE.mean** | 0,0051 | 0,0039 | 0,0009 | 0,2280 | 0,0813 | 0,0072 |
| **CI.mean.0,95** | 0,0101 | 0,0077 | 0,0017 | 0,4480 | 0,1594 | 0,0140 |
| **var** | 0,2288 | 0,1338 | 0,1428 | 453,7580 | 57,5672 | 9,3101 |
| **std.dev** | 0,4783 | 0,3658 | 0,3779 | 21,3020 | 7,5873 | 3,0513 |
| **coef.var** | 1,0484 | 2,0696 | 6,0594 | 4,6230 | 9,6688 | 306,1953 |

52

[52] The table was divided between this page and the next one.

| | Trading_volumei,t | MRETz,t | Trading_volumez,t | CAPi,t | RETAILi |
|---|---|---|---|---|---|
| **nbr.val** | 181.698,0000 | 181.147,0000 | 181.516,0000 | 156.940,0000 | 181.830,0000 |
| **nbr.null** | 20,0000 | 45,0000 | 9.254,0000 | 0,0000 | 65.246,0000 |
| **nbr.na** | 132,0000 | 683,0000 | 314,0000 | 24.890,0000 | 0,0000 |
| **min** | 0,0000 | -29,8471 | 0,0000 | 0,4777 | 0,0000 |
| **max** | 217.503.585,0000 | 10,5946 | 531.247.600,0000 | 183,9829 | 0,6500 |
| **range** | 217.503.585,0000 | 40,4417 | 531.247.600,0000 | 183,5052 | 0,6500 |
| **sum** | 539.956.441.633,0000 | -88,8934 | 18.477.498.338.900,0000 | 3.867.703,9166 | 9.457,9819 |
| **median** | 1.274.063,0000 | 0,0299 | 100.058.200,0000 | 14,8103 | 0,0120 |
| **mean** | 2.971.724,7400 | -0,0005 | 101.795.424,8600 | 24,6445 | 0,0520 |
| **SE.mean** | 11.317,3100 | 0,0032 | 142.380,1220 | 0,0663 | 0,0003 |
| **CI.mean.0,95** | 22.181,6700 | 0,0064 | 279.061,7720 | 0,1300 | 0,0005 |
| **var** | 23.272.158.032.041,1000 | 1,9079 | 3.679.710.358.538.570,0000 | 690,8702 | 0,0120 |
| **std.dev** | 4.824.122,5100 | 1,3813 | 60.660.616,2060 | 26,2844 | 0,1094 |
| **coef.var** | 1,6200 | -2.814,7718 | 0,5960 | 1,0665 | 2,1026 |

*Appendix 13: Table summarizing the results for the different dictionaries*

| | Syuzhet | GI | HE | LM | QDAP |
|---|---|---|---|---|---|
| **Min** | -5,40 | -1,00 | -0,50 | -1,00 | -1,00 |
| **Max** | 6,10 | 1,00 | 1,00 | 1,00 | 1,00 |
| **Mean** | 0,30 | 0,07 | 0,02 | -0,01 | 0,06 |
| **NA's** | 0,00 | 106,00 | 106,00 | 106,00 | 106,00 |
| **Number of negative words** | 2223,00 | 2014,00 | 576,00 | 2091,00 | 1789,00 |
| **Number of neutral words** | 3407,00 | 4000,00 | 8887,00 | 7522,00 | 4798,00 |
| **Number of positive words** | 5709,00 | 5219,00 | 1770,00 | 1620,00 | 4646,00 |

*Appendix 14: Descriptive statistics of the percentage of positive and negative tweets following the sentiment analysis based on the Loughran-McDonald dictionary*

| | POSi,t | NEGi,t |
|---|---|---|
| **nbr.val** | 8.703,0000 | 8.703,0000 |
| **nbr.null** | 7.205,0000 | 7.054,0000 |
| **nbr.na** | 173.127,0000 | 173.127,0000 |
| **min** | 0,0000 | 0,0000 |
| **max** | 1,0000 | 1,0000 |
| **range** | 1,0000 | 1,0000 |
| **sum** | 1.245,3165 | 1.383,7221 |
| **median** | 0,0000 | 0,0000 |
| **mean** | 0,1431 | 0,1590 |
| **SE.mean** | 0,0036 | 0,0037 |
| **CI.mean.0,95** | 0,0070 | 0,0073 |
| **var** | 0,1117 | 0,1220 |
| **std.dev** | 0,3343 | 0,3492 |
| **coef.var** | 2,3359 | 2,1965 |

*Appendix 15: Regression results of tweet parameters on market features following the sentiment analysis based on the Loughran-McDonald dictionary*

**Regression results**

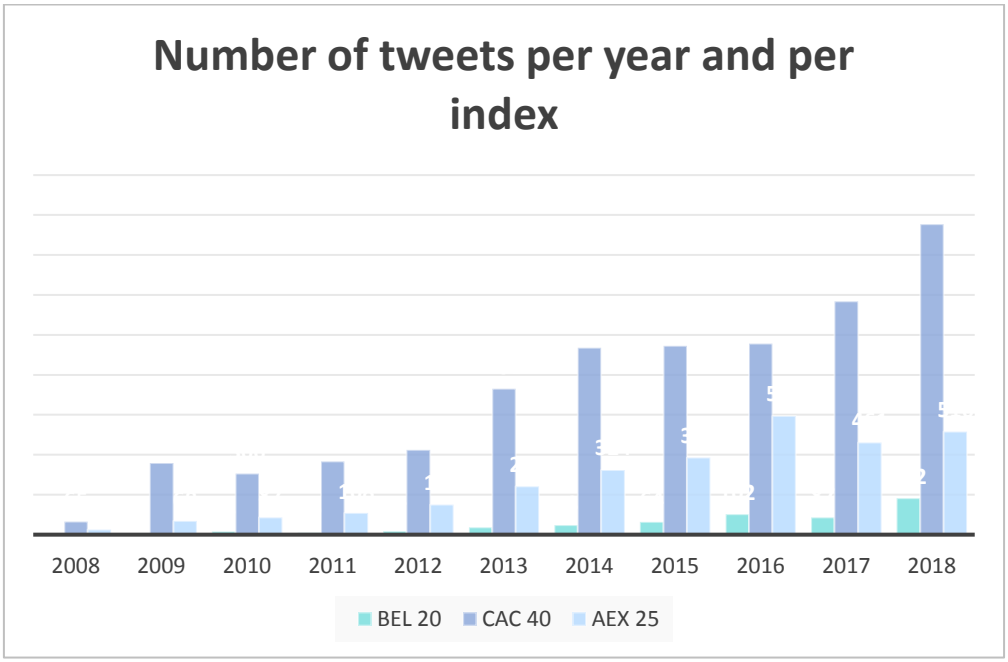
| | *Dependent variable:* | |
|---|---|---|
| | RETi,t | VOLi,t |
| | (1) | (2) |
| POSi,t | 0.010 | 0.006 |
| | (0.055) | (0.017) |
| NEGi,t | -0.131** | 0.018 |
| | (0.052) | (0.016) |
| MSGi,t | 0.130* | 0.036 |
| | (0.075) | (0.023) |
| RTWi,t | 0.00003 | -0.0004 |
| | (0.001) | (0.0003) |
| RPLi,t | 0.0003 | -0.001 |
| | (0.003) | (0.001) |
| MRETz,t | 1.069*** | -0.027*** |
| | (0.015) | (0.006) |
| MVOLz,t | -0.005 | 0.022*** |
| | (0.009) | (0.003) |
| VOLi,t | 0.103*** | |
| | (0.035) | |
| RETi,t | | 0.010*** |
| | | (0.003) |
| Observations | 8,546 | 8,546 |
| $R^2$ | 0.361 | 0.010 |
| Adjusted $R^2$ | 0.356 | 0.002 |
| F Statistic (df = 8; 8472) | 599.200*** | 10.980*** |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | |

*Appendix 16: Regression results of tweet and corporate parameters on market features following the sentiment analysis based on the Loughran-McDonald dictionary*

**Regression results**

| | *Dependent variable:* | |
|---|---|---|
| | RETi,t | VOLi,t |
| | (1) | (2) |
| POSi,t | 0.004 | 0.002 |
| | (0.054) | (0.017) |
| NEGi,t | -0.104** | 0.010 |
| | (0.052) | (0.016) |
| MSGi,t | 0.143** | 0.044* |
| | (0.073) | (0.023) |
| RTWi,t | -0.0001 | -0.0003 |
| | (0.001) | (0.0003) |
| RPLi,t | 0.0002 | -0.001 |
| | (0.003) | (0.001) |
| CAPi,t | -0.0005 | -0.007*** |
| | (0.001) | (0.001) |
| RETAILi | 0.286 | -2.059 |
| | (0.260) | (1.265) |
| MRETz,t | 1.061*** | -0.028*** |
| | (0.016) | (0.006) |
| MVOLz,t | -0.004 | 0.018*** |
| | (0.008) | (0.003) |
| VOLi,t | 0.013 | |
| | (0.022) | |
| RETi,t | | 0.010*** |
| | | (0.003) |
| Constant | -0.204 | 14.350*** |
| | (0.322) | (0.173) |
| Observations | 8,163 | 8,163 |
| $R^2$ | 0.353 | 0.023 |
| Adjusted $R^2$ | 0.352 | 0.022 |
| F Statistic | 4,438.000*** | 192.400*** |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | |

*Appendix 17: Distribution of the yearly tweet volume per index in our sample*




*Appendix 18: List of packages used and their utilities*

| Packages | Utilities |
|---|---|
| NLP | sentiment analysis |
| syuzhet | sentiment dictionary |
| tm | text mining |
| stringi | text processing |
| xlsx | excel download/upload |
| SentimentAnalysis | sentiment analysis |
| dplyr | data frame tool |
| tidyverse | facilitating the installation of other packages |
| lubridate | date processing |
| Hmisc | data analysis |
| plm | regression function |
| lm | regression function |
| pastecs | data analysis |
| stargazer | table visualization |
| car | functions in link with regressions |

**UNIVERSITÉ CATHOLIQUE DE LOUVAIN**
**Louvain School of Management**
Chaussée de Binche 151, 7000 Mons, Belgique | **www.uclouvain.be/lsm**