B3 FINANCIAL ASSETS DATA COLLECTION ON TWITTER: A PYTHON TEXT MINING ALGORITHM

COLETA DE DADOS DE ATIVOS FINANCEIROS DA B3 NO TWITTER: UM ALGORITMO DE MINERAÇÃO DE TEXTO EM PYTHON

Fábio Renato Maturana
Mestre e Doutorando em Controladoria e Contabilidade pela FEA/USP

Luiz Paulo Fávero
Doutor em Administração pela FEA/USP
Professor Titular da FEA/USP

Fabiano Guasti Lima
Doutor em Administração pela FEA/USP
Professor Associado da FEARP/USP

Robson Braga
Doutor em Controladoria e Contabilidade pela FEA/USP
Professor da Universidade do Estado da Bahia (UNEB)

Marco Aurélio dos Santos
Doutor em Controladoria e Contabilidade pela FEA/USP
Professor da EAESP/FGV

Abstract:
This paper brings up an innovative element when seeking to develop a text mining algorithm in Python able to monitor and gather a certain set of posts on the Twitter social network made by groups of users who post, analyze, rank and discuss data and information about B3 financial assets. Considering such objective and the development process of such algorithm, the study adds a contribution to field of behavioral and corporate finance, as well as to the development of text mining algorithms, in order to make replications possible. The construct presented herein differs from the others in the literature since the scripts are demonstrated step-by-step, what serves as a guide to researchers, analysts, consultants, investors and programmers.

Keywords: text mining; algorithms; Python; B3; Twitter.

CONTABILOMETRIA

Resumo do processo editorial:
a) Submissão em: 07/06/2022.
b) Envio para avaliação em: 25/07/2022.
c) Término da avaliação em 01/08/2022.
d) Correções solicitadas em: 01/08/2022.
e) Recebimento da versão ajustada em: 10/08/2022.
f) Aprovação final em: 16/08/2022.

1 Introduction

In the middle of the last century, several models were developed aiming to understand the behavior of price variations in the financial market, such as, for instance, Fama´s (1970) Efficient Market Hypothesis (EMH), fundamental concept both for theory and the practice of modern finance, and the Capital Asset Pricing Model (CAPM), presented by Sharpe (1964) and Lintner (1965).

As highlighted by Rabelo (2004), at about the beginning of the 21st. century, due to the increase of volatility, greater market diversification, and what Lo (2011) would call the advent of a "new world order", other theories began contesting the presumptions which support such dominant economic and financial models: the Behavioral Finance. The need of exploring and investigating empirical evidence of the influence of those new information sources in the market and in the investors´ behavior, along with the emergence of new technologies in the computing science field raises new questions in this research environment.

Such investigations suggest, according to Agarwal *et al.* (2019), that the investors know how to interact among themselves, such as the information recipients or producers, through online discussions, matters, shared information, for the search of online advice, making, thus, more informed decisions, whatever its quality, and becoming an important source of information. More and more researchers try to capture the influence of the information on the internet from social media, especially Twitter, for instance.

Bollen *et al.* (2011), for example, used machine learning techniques to extract mood states from Twitter and, through the sentiment analysis techniques, seeking predictions of changes in the stock market over time. According to Agarwal *et al*. (2019), little attention was given to investigations on how the investors´ behavior changed in the light of information and its source, which maybe points to the reasons for failing to explain the casual relationship between the mood of general public and the activities of the stock market.

Zhang et al (2011) sought to identify, via a "bag of words", how emotional aspects of publications on Twitter can provide elements which work as the basis to foresee the behavior of the market in the following day. Mao *et al.* (2012), on the other hand, showed results which point to a positive correlation between the daily number of tweets and the closing prices of the S&P 500.

Regarding the Brazilian stock market, the Souza´s (2020) study, counting on *Google Cloud*, *Natural Language* and APT, associates the volume of messages and the business volume in the stock market, showing that the greater the number of messages containing negative feelings, the greater the traded volume. Carosia *et al.* (2020) present results showing the positive association between feeling and price variation of the stocks, with more consistence for long time windows, by means of machine learning techniques.

Considering the relevance of the virtual environment, especially in the social network and Twitter specifically, aiming to explore events and phenomena belonging to the Brazilian financial market, this study presents the main goal of offering the investors, students and researchers a convenient and clear construct of text mining programmed in Python (Version 3.9.7), which is able to capture publications in the Twitter social network, filtering them by a group of keywords, but which can be directed based on simple alterations and according to the desired objectives.

Thus, this paper brings up an innovative element when seeking to develop a text mining algorithm in Python able to monitor and gather a certain set of posts on the Twitter social network, analyze, and rank the groups of users who use this social network to post, discuss and inform about the B3 assets.

Considering such objective and the development process of such algorithm, the study adds a contribution to field studies related to behavioral finances, the application of text mining techniques, and the replication capacity of the elaboration of the algorithm which operates in the Twitter social network.

The construct presented herein differs from the others which exist and are available in manuals prepared in Python, or another language, such as R, since a script is demonstrated, basically step-by-step for obtaining data, besides the tests and presentations of the necessary path for data analysis. All the passages are organized in the script, including comments on the code itself which can be a guide to a user who would be starting programming being able to analyze how each step was produced. And there is also the contribution for advanced and experienced researchers: a code capable of being used in other routines for related research.

2 Theoretical Benchmark

Examining the diversity of papers within this research line, we noticed there are at least three well-set research branches studying the relationships between Twitter and the Stock Exchange. The first one aims to check how the companies make use of this social platform to

engage with their investors. The second branch seeks to demonstrate how the information on Twitter can foresee the market variation trends, which is the case of the works of Bollen *et al.* (2011) and Mittal *et al.* (2012). And the third one checks the users´ posts in relation to the expectations and news on financial returns (RUIZ *et al*., 2012; BING *et al*., 2014; BARTOV *et al.*, 2018).

Bollen *et al.* (2011) extracted almost 10 million tweets to carry out the machine learning technique, known as the sentiment analysis, and, through that, correlate the public mood expressed on Twitter with the Dow Jones Industrial Average (DJIA) index. The researchers used two ways of gathering data; one was performed via OpinionFinder, which classifies the moods found in the tweets as negative or positive, and the other one by the Google-Profile of Mood (GPOMS), analyzing six mood states: 'calm', 'alert', 'sure', 'vital', 'kind' and 'happy'.

The correlation hypothesis based on Granger causality analysis were submitted to a Self-Organizing Fuzzy Neural Network (SOFNN) test to check whether the accuracy of DJIA predictions could be improved from the inclusion of the measures of public mood obtained. The results of the SOFNN showed that the predictions based on the "calm' feeling has greater correlation, confirming the testing by Granger causality. The accuracy of the correlations was measured in Mean Absolute Percentage Error (MAPE) and by the directional, high or fall accuracy. This last one reaching 87.6% of accuracy only when the calm measure is used in the correlation. Calm also had the second lowest percentage of absolute error, 1.83%.

Mittal *et al.* (2012) have also tested the hypothesis of behavioral economics in which the financial market can be foreseen if there is previous knowledge of public mood. The data were collected from Twitter and filtered through a list of words developed based on the Profile of Mood States (POMS) questionnaire. The results of the sentiment analysis were compared to the DJIA fluctuations and correlated by Granger causality. Four types of machine learning algorithms were tried out in the research, involving linear regression, logistic regression, Support Vector Machine (SVM) and SOFNN, the last one having presented the best results. The research of Mittal *et al*. (2012) confirms the results of Bollen *et al*. (2011).

Sprenger & Welpe (2011) have also found associations between the feelings expressed in the social network and the returns of the stock market, besides correlations between the volume of tweets in a certain day and the volume of operations in the following day. In this paper, more than 250 thousand tweets were analyzed and by analyzing these data, it was possible to extract the investors´ feelings regarding all the companies of the S&P 500 index. It was also possible, through the results obtained, to notice the users who provide investment advice above the average are retweeted (that is, they have their posts replicated) more often and they have more followers, which widens their participation and influence in microblogs/forums.

Chen & Lazer (2011) tested the relationship between the Twitter sentiment analysis and the movement of the assets of the stock exchange. For this purpose, they applied the GPOMS and the SentiWordNet in a pre-defined dataset which had already been used in previous projects. The simulation showed that, even with a much simpler sentiment analysis method, a correlation between the Twitter feeling data and the movement of the stock market could be observed.

Ruiz et. al. (2012) studied the correlation between the activities of Twitter and the events of the stock market focusing on the price changes and the volumes of stocks traded. In order to do so, they gathered market data of 150 companies of the S&P 500 in the first half of 2010, using the Yahoo! Finance database, during the same time that they manually filtered about 30 relevant Twitter messages for each one of the chosen companies. Their results showed the number of components connected by interaction presents good correlation, being the negotiated volume stronger than the stock price. They also checked, via simulation, that even the smallest correlation between the platform data and the prices can be explored in establishing the investment strategies.

Bing *et al.* (2014) opted to analyze the possibility of prediction of asset prices taking into account the possibility that the asset prices of certain companies can be more predictable than those of others from the public sentiment analysis of Twitter. For this purpose, they used the SentiWordNet 3.0 data mining algorithm, analyzed 200 million tweets from October /2011 to March/2012, from a WebCrawler system applied to Twitter, storing them in the databank of the MongoDB platform, and, afterwards, correlating them to the asset prices of 30 companies listed in the NASDAQ and NYSE available at Yahoo! Finance. With the proposed algorithm, they found out that the prediction of some companies is possible with a 76.2% average accuracy.

Smailovic *et al.* (2015) opted for an approach in which they sought to analyze the best text processing to check whether the feelings expressed by Twitter messages indicated changes in the stock prices. Thus, they used, in an active learning strategy, the sentiment sorting using the SVM technique, which categorized the gathered messages by the Twitter API in three categories (positive, negative, and neutral), and then, correlated them to the stocks of the chosen NASDAQ-100 companies. By implementing the active learning of the applicative of sentiment analysis, it was shown that the initial results indicate that increasing a negotiation strategy considering the probability values of positive feelings, the returns can be improved.

On the other hand, Risius *et al.* (2015) opted to use a strategy which, based on SentiStrength2, considered the multidimensional structure of the human emotions instead of a binary valence (positive-negative). For this purpose, they analyzed about 5.5 million of messages acquired by the Twitter API about 33 S&P 100 companies, focusing on their respective stock prices informed by the Yahoo! Finance in a 3-month period. The results pointed to three main findings: differentiated emotions are more strongly associated with the changes in prices of specific company stocks than the undifferentiated average feelings; negative emotions usually have a greater explanatory power, and, finally, that, particularly, the strength of the emotions regarding specific events (depression and happiness) are responsible for a greater price variation.

Dickinson & Hu (2015) concluded that there is no uniform connection between feeling and price in all the companies. For that purpose, they used the Twitter API along with the Twitter4j Java library to filter tweets related to companies belonging to the DJIA. That, on one hand, had its data collected from the Yahoo! Finance stock API. After analyzing the tweets from the Stanford NLP Sentiment Classifier algorithm, the result was correlated to each company asset prices. The final data showed that some companies have a strong positive correlation suggesting that costumer-oriented companies are differently affected compared with other companies in general.

Nofer & Hinz (2015) replicated some feeling measuring strategies from the 100 million tweets taken from the Twitter API from January/2011 to November/2013, using, for the sentiment analysis, the German version of POMS, correlating them, afterwards, to the data of stock return of the DAX index, composed of 30 publicly held companies of best financial performance in Germany by the Frankfurt Stock Exchange. Despite the fact that a significant relationship between feeling and stock market has not been found, the results pointed the need of considering the propagation of these mood states among the Twitter users by contagious as a factor. In a later simulation, a portfolio created by the authors from the results of the investigation increased its performance in 36% after a 6-month period.

He *et al.* (2016) investigated the correlation between feelings taken from the tweets and the changes in asset prices of the 7 largest companies of finance services in the U.S. To do so, they used the Lexalytics algorithm, from Twitter API database and they concluded that the negative feeling analyzed with a one-day delay on Twitter has a significant negative effect, both statistically and economically, in the stock price.

Zhang *et al.* (2011) also chose to look into the best machine learning strategies in the search for better effectiveness in the production of sentiment data from Twitter messages for the correlation with stock prices. Thus, they processed data taken from the Twitter Search API to test the effectiveness of three Machine Learning techniques: Naive Bayes classification, Maximum Entropy Classification and SVM, comparing them, later, with data of stocks of technology companies taken from the Yahoo! Finance API. The results showed that, even intra-day, the strategy was able to extract information or clues of significant events and keywords which indicated changes in the stock prices. They concluded, from the result, that words which refer to negative feelings are the most relevant ones for the correlation with stock prices. Nevertheless, they pointed to the need of creating a most elaborated list of words correlated to the stocks from a database with greater possibility to obtain better prediction indices.

Ruan *et. al.* (2018) opted to focus on a filter through the Twitter data sentiment analysis, for 8 months and using the SentiStrength as a filter of the database of the Twitter API and the Twitter4J Library, aiming to measure the "trust between the users", that is, the strength or the reputation of the authors in their community, expecting to amplify the correlation between the Twitter sentiment valence on eight specific companies, chosen from the S&P 500 index, with abnormal return data of stock market during the prediction. The results showed that, using such measuring methods of the weight of the messages, the sentiment valences reflect the abnormal stock returns better than dealing with all the authors homogeneously or differentiating them only by numbers of followers.

Using an extensive database, with 870 thousand tweets, collected from 2009 to 2012, Bartov *et al.* (2018) found out, through the summary statement of three research questions, that the opinion of a collective of Twitter users can successfully foresee the quarterly results of the companies, such results were shown to be valid even in the cases of companies with little information availability. According to the authors, such discoveries highlight the importance of considering the aggregated opinion of the publications present in that social network, besides the traditional future perspectives of the stock exchange and valuation.

Bernardo (2014) investigated the impact of Twitter on the financial market, focusing, mainly, in the field of technology. The study concluded that the prediction capability of Twitter can be directly related both with the data time dimension and the way in which the Twitter data are grouped, just like the characteristics of the companies, which can make them more sensible to analysis or not, and the technology companies are the ones of the most interesting cases.

Santos *et al.* (2015) and Santos (2016) investigated the relationship between the posts on Twitter and the Brazilian stock exchange. The results showed patterns of users´ behavior and the frequency of posts. Besides having checked that events and news regarding the stock market can generate post peaks, the authors observed that such frequency follows the opening of the trading floor of the stock exchange and is kept for about three hours after its closing.

Alves (2015) proposes an analysis of the relationships between the Twitter messages and the data of 8 companies of the Brazilian stock market from a system of support to decision making. The results showed how much the Petrobras shares (PETR3 and PETR4) stood out as the most cited ones out of all the companies and shares investigated, the decision-making simulation work from, with the help of crowd analysis, significant profit for both the stocks from the indices obtained.

Louseiro Lima (2016) proposed a prediction model for Bovespa. From the opinion mining and machine learning techniques and using the Natural Language Processing (NLP) and SVM from the Sentiment140 and Weka, Louseiro Lima (2016) develops a model aiming to help decision-making processes of buy-sell stocks. Souza (2020) sought to check how the messages posted on Twitter are associated with the movements in the Brazilian stock market. It was confirmed that there is association between the volume of messages which are posted daily on Twitter and the volume of business in the Brazilian stock market. Moreover, it was

checked that the greater the number of messages containing a negative feeling, the greater the volume traded is as well. Carosia *et al.* (2020) carried a comparison between different machine learning techniques to perform the sentiment analysis in tweets, reinforcing the relevance of studies involving Twitter.

3 Criteria Adopted for Building up the Algorithm

This section presents the (1) procedures carried out for obtaining the Twitter developer account, (2) the data collection at API, (3) database structuring and information treatment, (4) analysis of the data obtained, (5) breakdown of the kind of machine learning technique used, (6) considerations regarding the limitations of this work.

The data collector is an algorithm developed in the Python programming language, in which the functionalities of its libraries: 'tweepy', 'json', 'datetime' and 'pymongo' were used. These are able to receive and structure the files in the *JavaScript Object Notation* (JSON) format sent by the API Twitter server. Finally, as mentioned above, to consume the streaming content of the applicative programming interface or API of Twitter, an app was created and registered in the API site of the Twitter company.

For the composition of our database, only the tweets mentioning a selection of Tickers of the stocks of companies belonging to the Brazilian stock exchange were used. For instance, tweets which contained within their text body sequences of characters, such as PETR4, ITUB4, IRBR3 or EMBR3 were collected and added to the database for later analysis.

To inform the Twitter API which tweets we wanted to receive, it was necessary to create a list of keywords, or BOW, so that a filter took place in the whole Twitter data streaming and just the tweets which corresponded to the citations of the main tickers of the B3 assets were returned. Figure 1 contains the list of the tickers we used as BOW in this selection of tweets. Such list of keywords was chosen according to the occurrence of those words within Twitter itself; analyzing the tweets received, we came to the final list with the most mentioned tickers on this social network.

Figure 1 - List of Tickers which are part of the bag-of-words of the collector algorithm

```
["IBOV", "ALSO3", "ALPA4", "ABEV3", "ASAI3", "AMER3","AZUL4", "BTOW3", "B3SA3",
 "BBSE3", "BRML3", "BBDC4", "BBDC4","BRAP4", "BBAS3", "BRKM5", "BRFS3", "BPAC11",
 "CRFB3", "CCRO3", "CMIG4", "CNTO3", "CESP6", "HGTX3", "CIEL3","CLSA3","COGN3",
 "CSMG3", "CPLE6", "CSAN3", "CPFE3", "CVCB3", "CYRE3", "DTEX3", "ECOR3", "ELET3",
 "ELET6", "EMBR3", "ENBR3", "ENGI11","ENEV3", "EGIE3", "EQTL3", "EZTC3", "FLRY3",
 "GGBR4", "GOAU4", "NTCO3", "HAPV3", "HYPE3","IGTA3", "MEAL3", "GNDI3", "IRBR3",
 "ITSA4", "ITUB4", "JBSS3", "JHSF3", "KLBN4", "LIGT3", "LINX3", "RENT3","LCAM3",
 "LWSA3", "LAME3", "LAME4", "AMAR3", "LREN3", "MDIA3", "MGLU3", "MRFG3", "BEEF3",
 "MOVI3", "MRVE3","MULT3", "NEOE3", "PCAR3", "PETR3", "PETR4", "BRDT3", "PRIO3",
 "PSSA3", "QUAL3", "RADL3", "RAPT4", "RAIL3", "SBSP3","SAPR11", "SANB3", "SANB4",
 "SANB11","CSNA3","SULA11", "SUZB3", "TAEE4","TAEE11", "VIVT3", "VIVT4","TIMS3",
 "TOTS3", "TRPL4", "UGPA3", "USIM3", "USIM5", "VALE3", "VVAR3", "WEGE3", "WEGE4",
 "YDUQ3", "PMAM3", "OGXP3", "BVMF3", "PDGR3","GFSA3", "MMXM3", "RSID3", "TIMP3",
 "GOLL4", "DTCY3", "MAPT3",  "ESTR4", "MNPR3","OSXB3", "TEKA3", "TEKA4", "GPCP3",
 "TELB4", "HAGA4", "HETA4", "RANI3", "HOOT4", "RCSL4", "TXRX4", "IGBR3","BRPR3",
 "BOBR4", "PTNT3", "PTNT4", "WIZS3", "UNIP6", "SLCE3", "AALR3", "CEAB3", "GUAR3",
 "OIBR3", "OIBR4"]
```

Source: research data. Created by the authors.

Our tweet collector algorithm consumes the data streaming provided by the Twitter API. Such posts are received in the file format known as JSON, whose feature is the dictionary mapping, famous for its "key-value", that is the feature named and its respective associated value; such values are stored in our database.

In fact, a simple tweet can have more than 150 features. For our analyses, we show, in Table 1, a description of the main keys contained in a tweet JSON file; nevertheless, to obtain more details, it is possible to check the complete mapping in the development documentation page of the Twitter.

Table 1 — Keys and types of variables collected and stored from the tweet JSON files.

Key	Types	Description
created_at	String	It denotes the moment the tweet was created at UTC (Coordinated Universal Time).
id	Inteira	Number which denotes unique record of each tweet.
text	String	Text contained in the tweet.
in_reply_to_user_id	String	There will be a record if the tweet is the answer to another user.
user	String	Dictionary containing information regarding the user who posted, as for instance: username, location, number of followers and other information.
quote_count	whole	Number of citations this tweet received.

Source: Twitter.

For the development of the algorithm, object of this study, a sequence of detailed processes was performed from the search keys and the variables collected from the JSON files. For better understanding this sequence and, also allowing that the respective algorithm to be replicated in future studies, we present the commands and codes developed in Python next.

4 Text Mining Algorithm

In this section, the blocks of codes in Python of the algorithm will allow the understanding of the process which leads to the Twitter database collection and a later procedure which leads to the most relevant findings of this study. Considering the sequence of the steps presented, we recommend the novice user to use the Jupyter Notebook for running this code. Moreover, passwords and token provided by the site of the Twitter API will be needed for the execution. To obtain them, the approval of the registration for Twitter developer is necessary.

Figure 2 – First block of code of the collector algorithm.

```
import termcolor
from asyncio.log import logger
from tweepy.streaming import StreamListener
from tweepy import OAuthHandler
from tweepy import Stream
from datetime import datetime
import json
import tweepy
import time
import asyncio
from urllib3.exceptions import ReadTimeoutError
```

Source: Created by the authors.

The collection process follows the process exposed in Figure 2, which imports data, according to the algorithm configuration, complemented by Figures 3 to 10. The processes carried out in this stage of the work, as well as the codes of the stocks of the Brazilian stock market, the time lag and the users´ profile are used as reference for the validation of the algorithm consistency, considering such choices as limitations of the present study. That is, the algorithm can be configurated, from this study, for different data and time periods.

Figure 3 – Second block of code of the collector algorithm.

```
cor = 'red'
from termcolor import colored
print(colored('hello', cor), colored('world', 'green'))
```

Source: Created by the authors.

Figure 4 – Third block of code of the collector algorithm.

```
consumer_key = "here goes your consumer_key"
consumer_secret = "here goes your consumer_secret"
access_token = "here goes your access_token"
access_token_secret = "here goes your accesss_token_secret"
auth = OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
```

Source: Created by the authors.

Figure 5 – Fourth block of code of the collector algorithm.

```
# Creating a listener class to capture the Twitter data stream and store it at MongoDB
colorcont = 1
class MyListener(StreamListener):
    def on_data(self, dados):
        tweet = json.loads(dados)
        ################################################################################

        # Here is all the information you want to collect at the time of streaming,
        # however, just have tweet id to conduct the tweet later.

        created_at = tweet["created_at"]                          # indicates the moment of creation of the tweet
        id_str = tweet["id_str"]                                  # id string is like a general record of each tweet
        screen_name = tweet["user"]["screen_name"]                # name of the user who published
        followers_count = tweet["user"]["followers_count"]        # count the number of followers of that user
        # Here we check if there is text longer than 140 characters
        try:
            text = tweet["extended_tweet"]["full_text"]
        except:
            text = tweet["text"]
        lang = tweet["lang"]
        location = tweet["user"]["location"]
        in_reply_to_screen_name = tweet["in_reply_to_screen_name"]
```

Source: Created by the authors.

Figure 6 – Continuation of the fourth block of code of the collector algorithm.

```
        # obj is the object that will be printed on the user's screen, to follow the progress of the feed

        obj = {"created_at": created_at, "id_str": id_str,
               "screen_name": screen_name, "followers_count": followers_count,
               "text": text, "lang": lang, "location": location,
               "in_reply_to_screen_name": in_reply_to_screen_name, }
        tweetind = col.insert_one(obj).inserted_id
        #print(obj)
        global colorcont
        #print(colorcont)
        if colorcont == 1:
            print(colored(obj,'red'))
        else:
            if colorcont == 2:
                print(colored(obj,'green'))
            else:
                if colorcont == 3:
                    print(colored(obj,'magenta'))
                else:
                    if colorcont == 4:
                        print(colored(obj,'cyan'))
                        colorcont = 0
        colorcont += 1
        return True
```

Source: Created by the authors.

Figure 7 – Continuation of the fourth block of code of the collector algorithm.

```
# instantiating
mylistener = MyListener()
mystream = Stream(auth, listener=mylistener, tweet_mode='extended', timeout=60)

# Bringing MONGODB (DATABASE)
from pymongo import MongoClient

cliente = MongoClient('localhost', 27017)
dbb = cliente.twitterdb
col = dbb.tweets
```

Source: Created by the authors.

Figure 8 – Fifth block of code of the collector algorithm.

```
keywords = ["IBOV",  "ALSO3", "ALPA4", "ABEV3", "ASAI3", "AMER3","AZUL4", "BTOW3", "B3SA3", "BBSE3", "BRML3", "BBDC4", "BBDC4",
            "BRAP4", "BBAS3", "BRKM5", "BRFS3", "BPAC11","CRFB3", "CCRO3", "CMIG4", "CNTO3", "CESP6", "HGTX3", "CIEL3","CLSA3",
            "COGN3", "CSMG3", "CPLE6", "CSAN3", "CPFE3", "CVCB3", "CYRE3", "DTEX3", "ECOR3", "ELET3", "ELET6", "EMBR3",
            "ENBR3", "ENGI11","ENEV3", "EGIE3", "EQTL3", "EZTC3", "FLRY3", "GGBR4", "GOAU4", "NTCO3", "HAPV3", "HYPE3",
            "IGTA3", "MEAL3", "GNDI3", "IRBR3", "ITSA4", "ITUB4", "JBSS3", "JHSF3", "KLBN4", "LIGT3", "LINX3", "RENT3",
            "LCAM3", "LWSA3", "LAME3", "LAME4", "AMAR3", "LREN3", "MDIA3", "MGLU3", "MRFG3", "BEEF3", "MOVI3", "MRVE3",
            "MULT3", "NEOE3", "PCAR3", "PETR3", "PETR4", "BRDT3", "PRIO3", "PSSA3", "QUAL3", "RADL3", "RAPT4", "RAIL3",
            "SBSP3","SAPR11", "SANB3", "SANB4", "SANB11","CSNA3","SULA11", "SUZB3", "TAEE4","TAEE11", "VIVT3", "VIVT4",
            "TIMS3", "TOTS3", "TRPL4", "UGPA3", "USIM3", "USIM5", "VALE3", "VVAR3", "WEGE3", "WEGE4", "YDUQ3", "PMAM3",
            "OGXP3", "BVMF3", "PDGR3","GFSA3", "MMXM3", "RSID3", "TIMP3", "GOLL4", "DTCY3", "MAPT3",  "ESTR4", "MNPR3",
            "OSXB3", "TEKA3", "TEKA4", "GPCP3", "TELB4", "HAGA4", "HETA4", "RANI3", "HOOT4", "RCSL4", "TXRX4", "IGBR3",
            "BRPR3", "BOBR4", "PTNT3", "PTNT4", "WIZS3", "UNIP6", "SLCE3", "AALR3", "CEAB3", "GUAR3", "OIBR3", "OIBR4"]
```

Source: Created by the authors.

Figure 9 – Sixth block of code of the collector algorithm.

```
try:
    print(datetime.now())
    print("####################\nConectando#-0-\n####################\n")
    mystream.filter(track=keywords)
finally:
    try:
        time.sleep(300)
        print(datetime.now())
        print("####################\nReconectando#-1-\n####################\n")
        mystream.filter(track=keywords)
    finally:
        try:
            time.sleep(300)
            print(datetime.now())
            print("####################\nReconectando#-2-\n####################\n")
            mystream.filter(track=keywords)
        finally:
            try:
                time.sleep(300)
                print(datetime.now())
                print("####################\nReconectando#-3-\n####################\n")
                mystream.filter(track=keywords)
            finally:
                try:
                    time.sleep(300)
                    print(datetime.now())
                    print("####################\nReconectando#-4-\n####################\n")
                    mystream.filter(track=keywords)
```

Source: Created by the authors.

According to the sequence of Figures 3 to 10, the algorithm collects the data of interest of the researcher from the Twitter database. In that case, the algorithm collected information regarding the posts on companies traded in the Brazilian Stock Market (B3) considering the ticket of trading in the market (Figure 8) and the profile of the users who perform posts regarding those stocks. It is observed that some procedures configurated in the algorithm organize the data and improve the presentation of the data in colors and graphs according to Figures 3, 6 and 7, for instance.

Figure 10 – Continuation of the sixth block of code of the collector algorithm.

```
            finally:
                try:
                    time.sleep(300)
                    print(datetime.now())
                    print("####################\nReconectando#-5-\n####################\n")
                    mystream.filter(track=keywords)
                finally:
                    try:
                        time.sleep(300)
                        print(datetime.now())
                        print("####################\nReconectando#-6-\n####################\n")
                        mystream.filter(track=keywords)
                    finally:
                        try:
                            time.sleep(300)
                            print(datetime.now())
                            print("####################\nReconectando#-7-\n####################\n")
                            mystream.filter(track=keywords)
                        finally:
                            try:
                                time.sleep(300)
                                print(datetime.now())
                                print("####################\nReconectando#-8-\n####################\n")
                                mystream.filter(track=keywords)
                            finally:
                                time.sleep(300)
                                print(datetime.now())
                                print("####################\nReconectando#-9-\n####################\n")
                                mystream.filter(track=keywords)
```

Source: Created by the authors.

Next, we present the sequence which allowed the algorithm to recover JSON files of tweets. The blocks of codes in Python for tweet recovery from their identification number ('id') are arranged next. That is, with this algorithm, it is possible, from the tweet identification number, to request the API server any tweets or tweet list which you have the numbers, and which are still available in the server. The main function of this algorithm was to generate the '.pkl' file which will be used to carry out the treatment and the data analysis afterwards.

Figure 11 – First block of code of the recovery algorithm.

```
from twython import Twython
import datetime
import time , sys
import pandas as pd
import termcolor
import pymongo
from pymongo import MongoClient
import matplotlib.pyplot as plt
import re
import xlrd # package that makes the interface with Excel
```

Source: Created by the authors.

In this first block of commands, we imported all the Python libraries which we will use during this algorithm.

Figure 12 – Second block of code of the recovery algorithm.

```
# Excel in Python list - first file
workbook = xlrd.open_workbook('relatorio_maio.xls') #use your file name
worksheet = workbook.sheet_by_name('Sheet') #use the name of your file flap
worksheet = workbook.sheet_by_index(0)

lista_excel01 = []  #  creates the empty list
for i in range(worksheet.nrows):
    valor = worksheet.cell_value(i, 3) # takes the Excel values
    lista_excel01.append(valor) # inserts the values in the list
print(len(lista_excel01), "tweet records")
```

Source: Created by the authors.

The 'maio.xls' report file contains a sheet in which each line represents a tweet picked out from March 28th, 2020 to May 25th, 2020. The whole table has a total of 27,533 tweets, and in its content, there are the columns 'created_at', 'text' and 'id', which give information about time and date in which the tweet was posted, and the identification number given to that post.

Figure 13 – Third block of code of the recovery algorithm.

```
# Excel in Python list - second file
workbook = xlrd.open_workbook('relatorio_julhoid.xls') #use your file name
worksheet = workbook.sheet_by_name('Sheet') #use the name of your file flap
worksheet = workbook.sheet_by_index(0)

lista_excel02 = []  # creates the empty list
for i in range(worksheet.nrows):
    valor = worksheet.cell_value(i, 3) # takes the Excel values
    lista_excel01.append(valor) # inserts the values in the list
    lista_excel02.append(valor) # inserts the values in the list
print(len(lista_excel01), "tweet records")
print(len(lista_excel02), "tweet records")
```

Source: Created by the authors.

The 'relatorio_julhoid.xls' file shows the identification numbers of each tweet posted from May 25th, 2020 to July 24th, 2020. In this second collection lag, we collected the 7 variants included in each tweet´s JSON files, which are 'created_at', 'text', 'screen_name' and 'followers', 'in_reply_to_screen_name', 'tweet_id' and 'location'. The two exceeding variables bring information on the number of users´ followers at the moment of the post (followers), whether the tweet is an answer to some other user and who this user would be, and the location of this user, in case such information is available on his or her profile.

Figure 14 – Fourth block of code of the recovery algorithm.

```
# Excel in Python list - second file
workbook = xlrd.open_workbook('relatorio_agostoid.xls') #use your file name
worksheet = workbook.sheet_by_name('Sheet') #use the name of your file flap
worksheet = workbook.sheet_by_index(0)

lista_excel03 = []  # creates the empty list
for i in range(worksheet.nrows):
    valor = worksheet.cell_value(i, 3) # takes the Excel values
    lista_excel01.append(valor) # inserts the values in the list
    lista_excel03.append(valor) # inserts the values in the list
print(len(lista_excel01), "tweet records")
print(len(lista_excel02), "tweet records")
print(len(lista_excel03), "tweet records")
```

Source: Created by the authors.

Finally, the 'relatorio_agostoid.xls' file brings us the remaining 30,677 entries within July 24th, 2020 to August 28th, 2021. This first stage of the data collection accumulates the total of 118,209 posts collected. After August 28th, 2020, we began using the JSON file created by the MongoDB computer program, a database with has been used in our work.

Figure 15 – Fifth block of code of the recovery algorithm.

```
lista_ids01 = lista_excel01
```

Once the list of the tweet ids saved at Excel were done, let´s proceed to the tweets saved at the MONGODB database

```
cliente = MongoClient('localhost', 27017)
dbb = cliente['twitterdb']
collection = dbb['tweets']
```

```
print(collection.estimated_document_count(), "tweets coletados de 28 de agosto até", datetime.datetime.now())
```

Source: Created by the authors.

The command lines above access the database created at mongoDB, the entry of commands 7 shows how many tweets were gathered in the whole period.

Figure 16 – Fifth block of code of the recovery algorithm.

```
dataframe = pd.DataFrame(list(collection.find()))

lista_ids02 = dataframe['id_str'].tolist()

# Gathering the list obtained from the Excel with the one obtained from MongoDB
print(len(lista_ids01))
print(len(lista_ids02))
lista_ids = lista_ids01 + lista_ids02
len(lista_ids)
```

Source: Created by the authors.

The 8, 9 and 10 command entries convert the id sequence of all the 364,178 tweets collected from August 28th, 2020 to October 31st, 2020 into a list-type data structure, which makes up a total of 482,387 tweets collected in this work.

Figure 17 – Sixth block of code of the recovery algorithm.

```
lista_ids # visualizes the list of tweet ids to be recovered
```

```
['1244018173533437952',
 '1244018913509363713',
 '1244021641883025409',
 '1244024163397963776',
 '1244025981633003522',
 '1244032097280499716',
 '1244032330110509064',
 '1244032784282324992',
 '1244035176105197572',
 '1244035182904123399',
 '1244036937696129025',
 '1244037457198415889',
 '1244039637187866631',
 '1244043617326575616',
 '1244046629528887298',
 '1244047644651028482',
 '1244047743787687942',
 '1244048908252217356',
 '1244049459396427777',
```

Source: Created by the authors.

Figure 18 – Seventh block of code of the recovery algorithm.

```
APP_KEY = "IOIBI1Agws16F58VdpN8UOJtO"
APP_SECRET = "PseLyhqU37By8vwFzPsDKmkgvMHyviVLCp0PwWAckrLLCj0t2A"
OAUTH_TOKEN = "1217353829273960448-wESswmV11Dw9XB4ndekJyakzGsKofd"
OAUTH_TOKEN_SECRET = "mZQRJUJQWR3oxiEdw5gK8vvBsMwk1og9ZJ46G8bSGUsPw"
twitter = Twython(APP_KEY, APP_SECRET,
                  OAUTH_TOKEN, OAUTH_TOKEN_SECRET)
id_of_tweet = '1273638403221880832'
tweet = twitter.show_status(id=id_of_tweet)
tweet
```

```
lista_respescagem1 = [] #resetting the lists
lista_ditados = []
ids_respescagem1 = []
```

Source: Created by the authors.

Considering all the development stages of the data collection algorithm, tests were performed to validate their consistency, having the choice of data on the stocks of the companies listed in the B3 as reference.

Figure 19 – Eighth block of code of the recovery algorithm.

```
# Tweet List fishing crawler, just inform the id list you wish to fish

cor = 'red'
j = 0
d1 = datetime.datetime.now()
dinicio = d1
print("\n","Baldejamento número =", j+1 ,"\n", d1)
while j < 1428:
    i = j*250
    while i < ((j+1)*250):
        try:
            id_of_tweet = lista_ids[i]
            tweet = twitter.show_status(id=id_of_tweet, tweet_mode='extended')
            print(i," ", end="")
            lista_ditados.append(tweet)
            i += 1
        except:
            print(termcolor.colored(i,cor))
            lista_respescagem1.append(i)
            ids_respescagem1.append(lista_ids[i])
            i += 1
    #time.sleep(300)
    print("\n","Fim do Baldejamento numero =", i)
    for k in range(0, 300):
        sys.stdout.write("\r{}".format(k))
        sys.stdout.flush()
        time.sleep(1)
    j += 1
    d1 = datetime.datetime.now()
    print("\n","Baldejamento número =", j+1 ,"\n",d1)
```

Source: Created by the authors.

Figure 20 – Ninth block of code of the recovery algorithm.

```
print(len(lista_ditados))
print(len(lista_respescagem1))

dataframe_recuperado_total = pd.DataFrame(lista_ditados)
dataframe_repescagem_total = pd.DataFrame(lista_respescagem1)

dataframe_recuperado_total.columns

dataframe_recuperado_total.to_pickle('recuperados303624.pkl')

dataframe_repescagem_total.to_pickle('respescagem01de08-07-2021')

dataframe = dataframe_recuperado_total
```

Source: Created by the authors.

In our test and validation of the algorithm, through the analysis of the data collected, we could test the research hypotheses proposed by a quite simple method of analysis. It was also possible to verify there is a collective of users engaged in discussing, informing, and posting regarding the assets of the Brazilian stock Exchange on Twitter. Moreover, it was also possible to classify groups within this collective, according to numbers of variables of the social network platform. We also determined that a great part of the tweets which directly deal with the B3 assets are posted in Portuguese.

The ostensive use of this algorithm allowed us, during the 20 months of data collection, to store almost half a million tweets, which made it possible to assess a behavior pattern similar to the weekend effect, in which the users of this social network post more when it is closer to the electronic trading floor time, and they reduce their posts drastically on weekends and holidays.

We could also ascertain, analyzing the posts stored in our database, that the most cited asset, the Petrobras preferred shares, holders of the PETR4 ticker, have, in their daily turnover, a positive correlation of 72% with the number of tweets posted the day before. It is a satisfactory and positive correlation which indicated that the use of this algorithm can be an excellent tool for the investor since it catches relevant information for the market.

It is necessary to consider the tests were carried out from a record which includes a certain group of keywords comprising the most cited tickers on Twitter posts, and its scope is not enough for a broader prediction of the market movements. In any case, the aim of this study is not to foresee the market in a totally precise way, which is impossible, but to present a technical construct, via a replicable algorithm, of how to use the Twitter public data to build a tool of analytic support to decision-making, using machine learning and text mining techniques in Python language, which can contribute to literature in the area and inspire new studies.

5 Discussion and Final Considerations

The main objective of this study was to offer the investors, students, and researchers a practical and clear text mining construct in Python, able to catch posts of the Twitter social network, filtering them through a keyword group.

Lots of paths are possible from the investigation results. There is still the need to further investigate the causal relationship between investors and market. Another important question is to determine and reconstitute the relationships between homogeneous groups and their individual characteristics so that its identification and its heuristics is more accurate. It also comes to us as a need to understand, in the future, the possible behavioral influence of the companies on the investors in the national market since the stake in the possible information feedback per Twitter user and their capability to form speculative bubbles.

Thus, we recommend the future works deal with other networks, such as Facebook, Instagram, YouTube and any other platform having user niches posting and discussing on the financial market, as it was the environment of the study underlying to the tests we carried out with the algorithm. Other future study possibilities involve the analysis, together with data from other emerging markets, such as the Argentina and Mexico stock markets in order to continue the studies on financial markets.

Besides, this text mining construct allows the implementation of the machine learning technique known as sentiment analysis, widely used in works of this research spectrum. Its use along with clustering techniques seems to be an auspicious way to be investigated in the future.

We hope that the production and publication of an open algorithm stimulate future works which can use it, whether for this same research architype, or even to test other hypotheses of the researcher´s interest, using the same data set or another set, collected in different time from that we worked with.

References

AGARWAL, S.; KUMAR, S.; GOEL, U. Stock market response to information diffusion through internet sources: A literature review, **International Journal of Information Management**, [*s. l.*], v. 45, p. 118-131, ISSN 0268-4012, https://doi.org/10.1016/j.ijinfomgt.2018.11.002., 2019.

ALVES, Deborah Silva. **Uso de técnica de computação social para tomada de decisão de compra e venda de ações no mercado brasileiro de bolsa de valores**. 2015. Tese de Doutorado Faculdade de Tecnologia – Universidade de Brasília. 2015.

ARJOON, V. Microstructures, financial reforms, and informational efficiency in an emerging market. **Research in International Business and Finance**,[*s. l.*] , v.36, p. 112-126, 2016.

ASSIS, C.; MACHADO, E. J.; PEREIRA, A.; CARRANO, E. G. Hybrid deep learning approach for financial time series classification. **Revista Brasileira De Computação Aplicada,** Passo Fundo, v.10, n.2, p.54–63, doi:10.5335/rbca.v10i2.7904, 2018.

BARTOV, E.; FAUREL, L.; MOHANRAM, P. S. Can Twitter help predict firm-level earnings and stock returns? **The Accounting Review**, [*s. l.*], v. 93, n. 3, p.25-57, 2018.

BERNARDO, I.S.P.B. **A era de um mercado social:** a relação entre o Twitter e o mercado acionista**.** 2014. Dissertação de Mestrado - Instituto Superior de Estatística e Gestão de Informação, Universidade de Lisboa, 2014.

BING, L.; CHAN, K. C. C.; OU, C. Public sentiment analysis in Twitter data for prediction of a company's stock price movements, *In:* 2014 IEEE 11TH INTERNATIONAL CONFERENCE ON E-BUSINESS ENGINEERING, [*s. l.*], **Anais** […], 2014, p. 232-239, doi: 10.1109/ICEBE.2014.47, 2014.

BOLLEN, J.; MAO, H.; ZENG, X. Twitter mood predicts the stock market. **Journal of Computational Science**, [*s. l.*], v. 2, n. 1, p.1 – 8, 2011.

BURNHAM, T. C. Toward a neo-Darwinian synthesis of neoclassical and behavioral economics. **Journal of Economic Behavior & Organization**, ,[*s. l.*], v. 90, p. S113-S127. 2013.

BUSINESSWEEK, StockTwits may change how you trade, **BusinessWeek, Online Edition** (author Max Zeledon), [*s. l.*], February 11, 2009.

CAROSIA, A.E.O.; COELHO G.P.; SILVA A.E.A. Analysing the brazilian financial market through portuguese sentiment analysis in social media. **Applied Artificial Intelligence**, [*s. l.*], v. 34, n.1, p. 1-19, DOI: 10.1080/08839514.2019.1673037, 2020.

CHEN, R; LAZER, M. Sentiment analysis of twitter feeds for the prediction of stock market movement. Available at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.374.9053&rep=rep1&type=pdf. 2011.

COFFEE & STOCKS. "Por que estou short em Via Varejo": Renoir Vieira e um debate sobre o efeito manada na bolsa. **InfoMoney. [Blog]** Recuperado de https://www.infomoney.com.br/stock-pickers/por-que-estou-short-em-via-varejorenoir-vieira-e-um-debate-sobre-o-efeito-manada-na-bolsa/, 24 Jul. 2020.

DICKINSON, B.; HU, W. Sentiment analysis of investor opinions on Twitter. **Social Networking,** [*s. l.*], v. 4, p. 62-71. doi: 10.4236/sn.2015.43008, 2015.

FAMA, E. F. Efficient capital markets: a review of theory and empirical work. **The Journal of Finance**, [*s. l.*], v. 25, n.2, p. 383-417. 1970.

FRADKOV, A. L. Early history of machine learning. **IFAC-PapersOnLine**, [*s. l.*], v. 53, n. 2, p. 1385-1390, 2020.

GIPPEL, J. K. A revolution in finance? **Australian Journal of Management**, [*s. l.*], v.38, n. 1, p. 125-146, 2013.

HE, W.; GUO, L.; SHEN, J.; AKULA, V. Social media-based forecasting: a case study of tweets and stock prices in the financial services industry. **Journal of Organizational and End User Computing (JOEUC)**, [*s. l.*], v. 28, n. 2, p. 74-91, 2016.

JAFFE, J.; WESTERFIELD, R. The week-end effect in common stock returns: The International Evidence. **Journal of Finance,** [*s. l.*], v. 40, n. 2, p. 433-454, 1985.

KAHNEMAN, D.; TVERSKY, A. Prospect theory: an analysis of decision under risk. **Econometrica**, [*s. l.*], v. 47, n. 2 (Mar., 1979), p. 263-292, 1979.

KUHN, T. The Structure of Scientific Revolutions. Chicago, IL: University of Chicago Press, 1962.

KWAK, H.; LEE, C.; PARK, H.; MOON, S. What is twitter, a social network or a news media? *In:* PROCEEDINGS OF THE 19TH INTERNATIONAL CONFERENCE, [*s. l.*], **Anais** […], p. 591–600. doi/pdf/10.1145/1772690.1772751, 2010.

LINTNER, J. The valuation of risk assets and the selection of risk investments in stock portfolios and capital budgets. **Review of Economic and Statistics**. [*s. l.*], v. 47, p. 13-37, feb. 1965.

LIU, L.; WU, J.; LI P.; LI, Q. A social-media-based approach to predicting stock comovement, **Expert Systems with Applications**, [*s. l.*], v. 42, n. 8, p. 3893-3901, 2015. ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2014.12.049 . 2015.

LO, A. W. The adaptive markets hypothesis. **The Journal of Portfolio Management**, [*s. l.*], v. 30, n. 5, p. 15-29, 2004.

LO, A. W Adaptive markets and the new world order. Available at: http://papers,ssrn.com/sol3/paperscfm?abstract_id=1977721, 2011.

LO, A. W Adaptative markets: financial evolution at the speed of thought. Princeton University Press, 2017.

LOUSEIRO LIMA, Milson. **Um Modelo para Predição de Bolsa de Valores Baseado em Mineração de opinião**. 2016. Dissertação (Mestrado) – Programa de Pós-graduação em Engenharia de eletricidade, Universidade Federal do Maranhão, São Luís, 2016.

MANAHOV, V.; HUDSON, R. A note on the relationship between market efficiency and adaptability–New evidence from artificial stock markets. **Expert Systems with Applications**, [*s. l.*], v. 41, n.16, p. 7436-7454, 2014.

MAO, Y.; WEI, W.; WANG, B.; LIU, B. Correlating S&P 500 stocks with Twitter data. *In*: Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research (HotSocial '12). **Anais** […] Association for Computing Machinery, New York, NY, USA, p. 69–72. https://doi.org/10.1145/2392622.2392634, 2012.

MITTAL, A.; GOEL, A. Stock prediction using twitter sentiment analysis. Stanford University **Working Paper**. Available at: http://cs229.stanford.edu/proj2011/GoelMittal-StockMarket PredictionUsingTwitterSentimentAnalysis.pdf, 2012.

NOFER, M; HINZ, O. Using twitter to predict the stock market: where is the mood effect? DOI: 10.1007/s12599-015-0390-4, 2015.

OLIVEIRA, N.; CORTEZ, P.; AREAL, N. The impact of microblogging data for stock market prediction: using Twiter to predict returns, volatility, trading volume and survey

sentiment indices. **Expert systems with Applications**, [*s. l.*], v. 73, p. 125-144. https://doi.org/10.1016/j.eswa.2016.12.036, 2017.

RABELO, T. S.; IKEDA, R.H. Mercados eficientes e arbitragem: um estudo sob o enfoque das finanças comportamentais. **Revista Contabilidade & Finanças USP**, São Paulo, v. 34, n.1. http://dx.doi.org/10.1590/S1519-70772004000100007, 2004.

RISIUS, M.; AKOLK, F.; BECK, R. "Differential mmotions and the stock market - the case of company-specific trading" (2015). **ECIS 2015 Completed Research Papers**. Paper 147. ISBN 978-3-00-050284-2. Available at: https://aisel.aisnet.org/ecis2015_cr/147, 2015.

RUAN, Y.; DURRESI, A.; ALFANTOUKH, L. Using Twitter trust network for stock market analysis. **Knowledge-Based Systems**, [*s. l.*], v. 145, p. 207-218, 2018.

RUBINSTEIN, M. A history of the theory of investments. Hoboken, 2006.

RUIZ E.J.; HRISTIDIS V.; CASTILLO C.; GIONIS A.; JAIMES A. Correlating financial time series with micro-blogging activity. *In:* **Proceedings of the fifth ACM international conference on Web search and data mining**, **Anais** […] ACM (2012), p. 513-522 https://doi.org/10.1145/2124295.2124358, 2012.

SANTOS, H. S.; LAENDER, A.H.F; PEREIRA, A.C.M. Uma Visão do Mercado Brasileiro de Ações a partir de Dados do Twitter. Departamento de Ciência da Computação - Universidade Federal de Minas Gerais. Belo Horizonte, 2014.

SANTOS, Hugo Silva. **Um estudo sobre o mercado brasileiro de ações a partir de dados do twitter**. 2016. Dissertação de Mestrado (Departamento de ciência da Computação) - Universidade Federal de Minas Gerais, 2016.

SANTOS, H. S. *et al*. Uma Visão do Mercado Brasileiro de Ações a partir de Dados do Twitter, *In:* 4. BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING. [*s. l.*], **Anais** […] https://doi.org/10.5753/brasnam.2015.6770, 2015.

SANTOS, Marco Aurélio dos. **Hipótese de mercados adaptativos e fatores econômico-institucionais: uma abordagem multinível**. 2018. Tese (Doutorado em Controladoria e Contabilidade) Faculdade de Administração, Economia e Contabilidade - Universidade de São Paulo, São Paulo, 2018.

SOUZA, Dyliane Mourí Silva de. **Efeito do sentimento do investidor manifesto via Twitter sobre os retornos e o volume negociado no mercado acionário brasileiro**. 2020. Dissertação – Universidade Federal da Paraíba, João Pessoa, 2020

SHARPE, W. Capital asset prices: a theory of market equilibrium under conditions of risk. **Journal of Finance**, [*s. l.*], v.19, n.3 p. 425-442, 1964.

SMAILOVIC J.; KRANJC J.; PODPEČAN V.; GRČAR M.; ŽNIDARŠIČ M.; LAVRAČ N. Active learning for sentiment analysis on data streams: methodology and workflow implementation in the ClowdFlows platform, **Information Processing & Management**, [*s. l.*], v. 51, n. 2, p. 187-203, ISSN 0306-4573, https://doi.org/10.1016/j.ipm.2014.04.001, 2015.

SPRENGER, O.; WELPE, M. Tweets and Peers: Defining Industry Groups and Strategic Peers based on Investor Perceptions of Stocks on Twitter (February 26, 2011). **Algorithmic**

Finance , [*s. l.*], v. 1, n.1, p. 57-76, Available at SSRN: https://ssrn.com/abstract=1770582, 2011.

SUROWIECKI, J. The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business. Economies, Societies and Nations. New York, NY: Anchor Books, 2004.

THALER, R. Anomalies: The Ultimatum Game. **The Journal of Economic Perspectives**, [*s. l.*], v.2, p.195-206, 1988.

URQUHART, A.; GEBKA, B.; HUDSON, R. How exactly do markets adapt? Evidence from the moving average rule in three developed markets. **Journal of International Financial Markets, Institutions and Money**, [*s. l.*], v. 38, p. 127-147, 2015.

URQUHART, A.; MCGROARTY, F. Are stock markets efficient? Evidence of Adaptive Markets Hypothesis. International. **Review of Financial Analysis**, [*s. l.*], v. 47, p. 39-49, 2016.

VALORINVESTE. As 10 ações mais recomendadas para comprar em Janeiro. [Site] Available at: https://valorinveste.globo.com/mercados/rendavariavel/noticia/2021/01/05/as-10-acoes-para-comprar-em-janeiro.ghtml, 2021.

ZHANG, X.; FUEHRES, H; GLOOR, P. Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear", **Procedia - Social and Behavioral Sciences**, [*s. l.*], v. 26, p. 55-62, ISSN 1877-0428, https://doi.org/10.1016/j.sbspro.2011.10.562, 2011.