

Faculdade

XPe



RELATÓRIO

PROJETO
APLICADO

PÓS-GRADUAÇÃO

XP Educação
Relatório do Projeto Aplicado

Análise de Sentimento para o Mercado de Ações: Uma Visão sob a Ótica da Rede Social Twitter

Fernando Guedes de Campos Júnior

Orientador(a): Daniel Viana

25 de setembro de 2023



FERNANDO GUEDES DE CAMPOS JÚNIOR

XP EDUCAÇÃO

RELATÓRIO DO PROJETO APLICADO

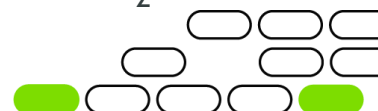
Análise de Sentimento para o Mercado de Ações: Uma Visão sob a Ótica da Rede Social Twitter

Relatório de Projeto Aplicado
desenvolvido para fins de conclusão do
curso de pós-graduação lato sensu MBA
em Ciência de Dados

Orientador (a): Daniel Viana

Brasília

25 de setembro de 2023

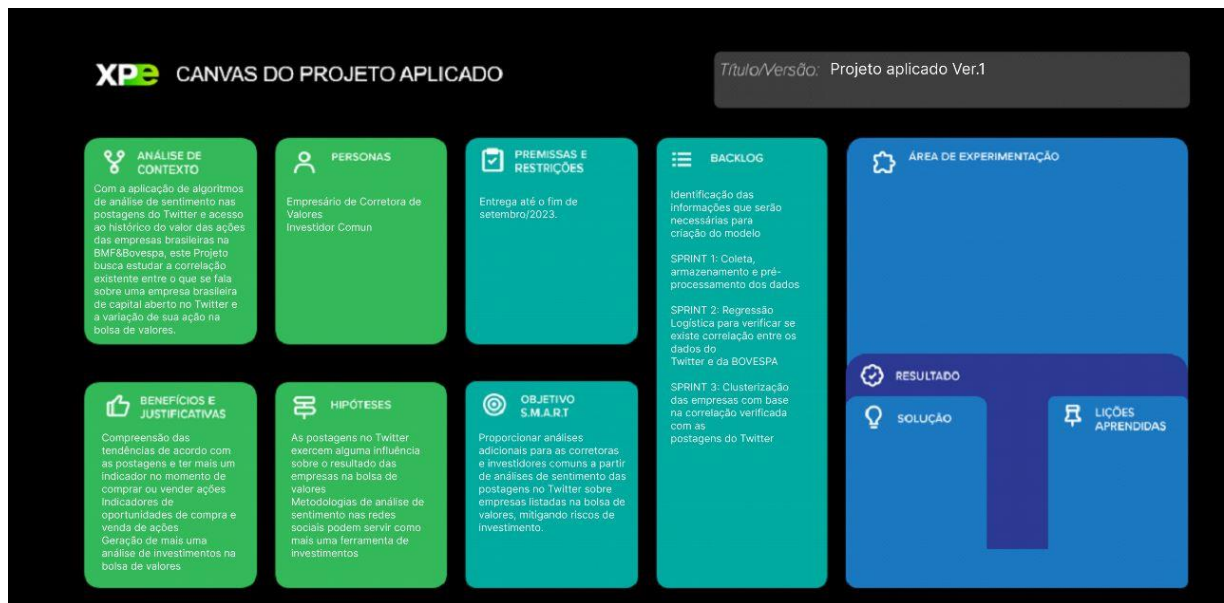


Sumário

1. CANVAS do Projeto Aplicado	4
1.1 Desafio	5
1.1.1 Análise de Contexto	5
1.1.2 Personas	7
1.1.3 Benefícios e Justificativas	9
1.1.4 Hipóteses	10
1.2 Solução	12
1.2.1 Objetivo SMART	12
1.2.2 Premissas e Restrições	12
1.2.3 Backlog de Produto	13
2. Área de Experimentação	14
2.1 Sprint 1	14
2.1.1 Solução	14
2.1.2 Lições Aprendidas	18
2.2 Sprint 2	18
2.2.1 Solução	18
2.2.2 Lições Aprendidas	23
2.3 Sprint 3	24
2.3.1 Solução	24
2.3.2 Lições Aprendidas	31
3. Considerações Finais	31
3.1 Resultados	31
3.2 Contribuições	32
3.3 Próximos passos	32



1. CANVAS do Projeto Aplicado



1.1 Desafio

1.1.1 Análise de Contexto

Com o massivo uso das redes sociais, a sociedade passou a remodelar várias características importantes das relações humanas. O acesso à informação das pessoas e das instituições, trouxeram novas formas de relacionamento, modificando assim, vários aspectos da vida em sociedade. As influências dessas mudanças são mais perceptíveis em algumas áreas, como: marketing, comunicação, jornalismo, publicidade, propaganda, política, comércio, prestação de serviço, pesquisa etc., mas, quando nos referimos ao mercado de capitais, ainda não é possível dimensionar os impactos deste uso no seu funcionamento.

Tendo em vista a quantidade de informações geradas diariamente pelos usuários das redes sociais, já existem softwares que monitoram essas informações em tempo real para as mais diversas finalidades. Hoje em dia, várias empresas têm setores exclusivamente dedicados para realizar este tipo de monitoramento, visando, de forma geral, saber o que os usuários da rede estão comentando sobre a instituição nas redes sociais.

Diante da imensa quantidade de informações processadas diariamente, torna-se imprescindível para as empresas tentar estruturar e analisar esses dados, visando formar alicerces e diretrizes para tomada de decisões. Mas os usuários comuns também podem aproveitar essa quantidade de dados disponíveis, para, por exemplo, utilizá-los no momento de investir no mercado de capitais. Pois além das análises, fundamentalista e gráfica, um mapeamento do que vem sendo processado nas redes sociais, pode também ser de grande importância para a elaboração de uma estratégia de investimentos.

Tendo em vista tantas variáveis e contextos, a principal questão que o estudo aborda é: As postagens dos usuários no Twitter afetam, de alguma forma, no valor das ações de uma empresa de capital aberto na bolsa de valores brasileira?

Com a aplicação de algoritmos de análise de sentimento nas postagens do Twitter e acesso ao histórico do valor das ações das empresas brasileiras na BMF&Bovespa, este Projeto busca estudar a correlação existente entre o que se fala sobre uma empresa brasileira de capital aberto no Twitter e a variação de sua ação na bolsa de valores.

Para entender melhor sobre a contextualização do problema, será utilizado a matriz CSD (certezas, suposições e dúvidas) no intuito de buscar uma solução mais assertiva.



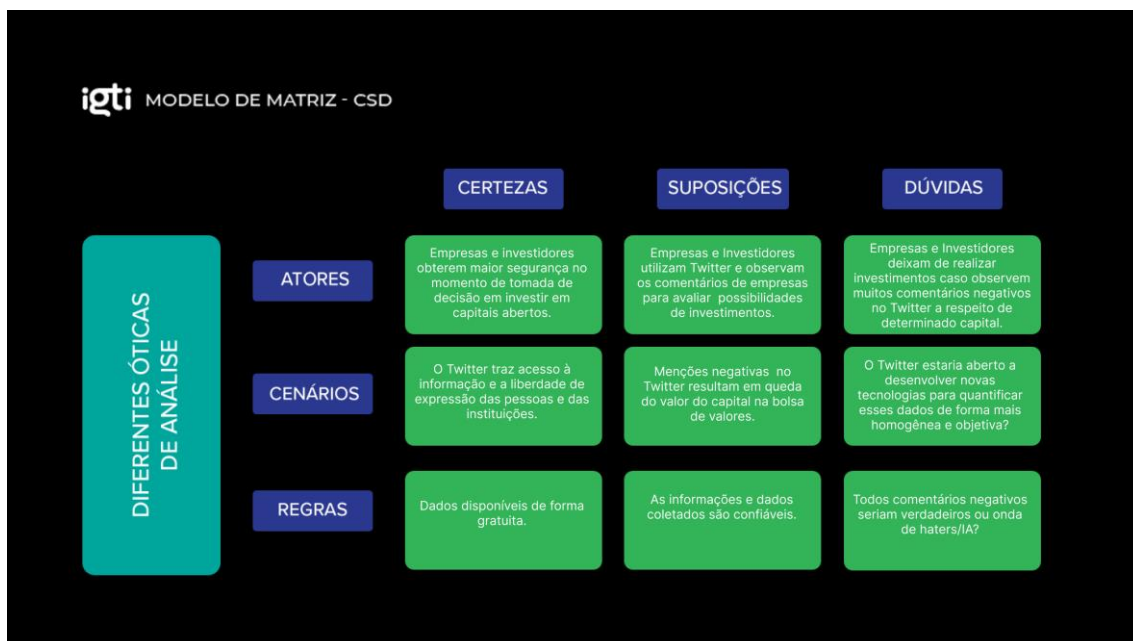


Figura 1. Matriz CSD

Para complementar a análise, utilizaremos a ferramenta de análise do contexto do problema POEMS.



Figura 2. POEMS

1.1.2 Personas

Para que possamos analisar mais a fundo o problema, precisamos entender as necessidades do cliente e, para isso, foram definidos as Personas que representam o perfil do usuário ideal do produto.

PERSONA 1:

Anderson é empresário, tem 34 anos, casado, e trabalha em uma Corretora de valores. Sua paixão pela economia o fez se formar nesta área e realizar um MBA em Investimentos e Finanças.

Para superar a concorrência no mercado, Anderson busca criar um diferencial competitivo para sua empresa, obtendo um mais uma ferramenta de análise de investimentos para promover maior segurança para seus investidores. Ele está sempre aberto ao diálogo e à colaboração, buscando envolver todos os colaboradores em suas estratégias de negócio.

Anderson sabe que o mercado da Bolsa de Valores é altamente competitivo, dinâmico e rotativo, portanto, busca constantemente novas estratégias para se destacar da concorrência. Ele acredita que a tecnologia pode ser um grande diferencial para sua empresa, mas enfrenta desafios para implementação devido ao alto custo e mão de obra específica.

Ele está em busca de soluções que possam melhorar a gestão de sua carteira de renda variável, para assim dar maior segurança aos seus clientes e investidores.

PERSONA 2:

Matheus tem 28 anos, solteiro e trabalha em um escritório de contabilidade em São Paulo. Sempre teve muito interesse em investimentos e bolsa de valores pensando em construir sua própria aposentadoria e fazer com que seu dinheiro tenha um rendimento mais eficaz.

Matheus trabalha de segunda a sexta 9h por dia, e ao chegar em casa se sente muito cansado para estudar sobre investimentos e entender melhor como e onde aplicar. Infelizmente ele não possui condição financeira para pagar um Consultor, e assim, acaba por muitas vezes deixar passar boas oportunidades ou acaba investindo em ações que não geram bons rendimentos.

Ele busca uma solução prática e intuitiva para ajudá-lo a investir com maior assertividade, que possa lhe indicar possíveis oportunidades de compra de ações na bolsa de valores ou alertar acerca de possível necessidade de venda.



Para entender melhor as necessidades e expectativas dos clientes, foram utilizados mapas de empatia para considerar os pensamentos e sentimentos do empresário e do investidor comum. O objetivo era identificar aspectos mais específicos desse processo, apontar problemas e, por fim, identificar dores como perdas financeiras, baixos rendimentos, falta de investimentos e tensão no momento de investir.

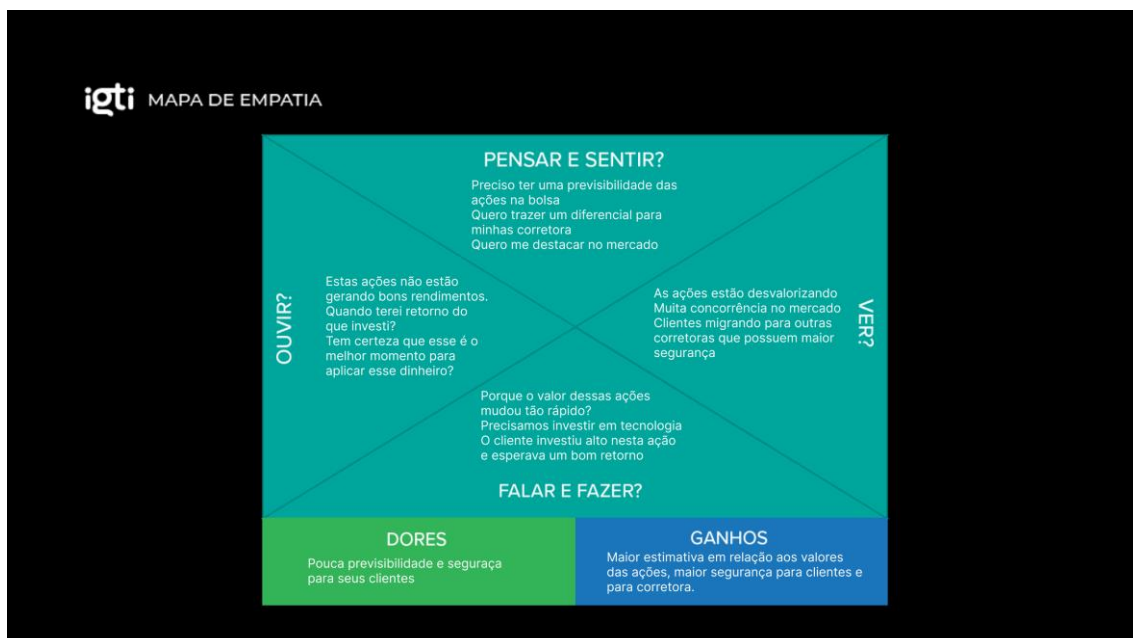


Figura 3. Mapa de empatia Persona 1

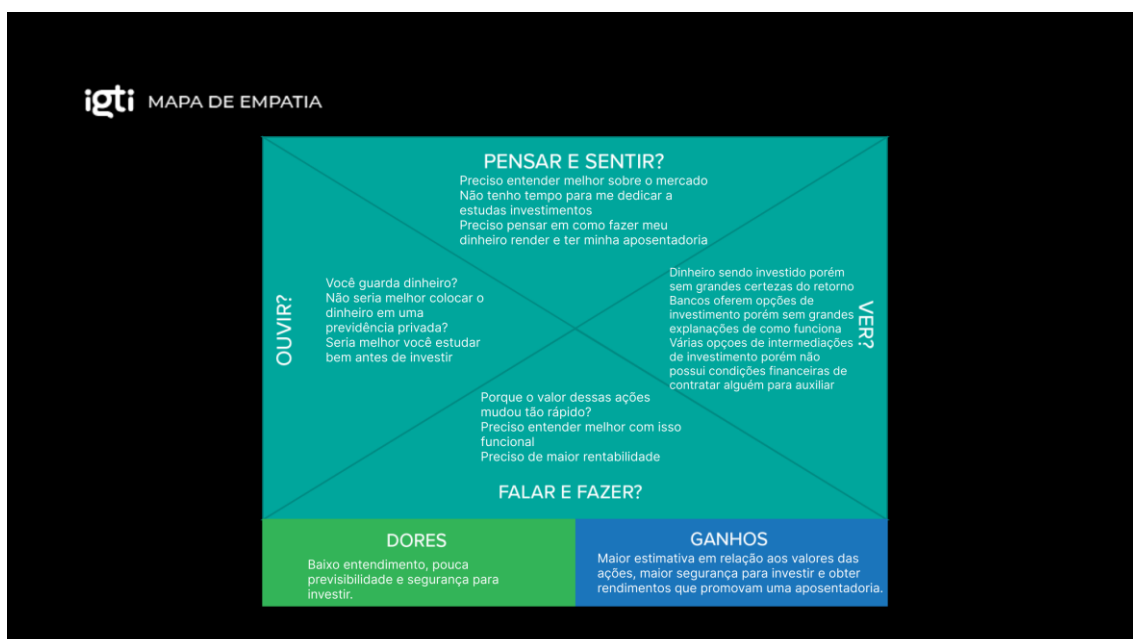


Figura 4. Mapa de empatia Persona 2

1.1.3 Benefícios e Justificativas

Pensando em destacar quais são os benefícios e as justificativas para propor uma solução que sane ou reduza as dores do cliente, ajudando as empresas e investidores no enfrentamento de seu problema, será utilizado a metodologia Business Design Blueprint, que é uma ferramenta que analisa a rotina do cliente, identificando os elementos-chave da estratégia para solucionar os desafios propostos. Utilizando dessa metodologia, entender e desenvolver propostas que solucionem o problema e que realmente atendam às necessidades urgentes do cliente se torna mais simples e prático.

A tabela a seguir foi preenchida utilizando esta abordagem, onde pode ser observado uma breve descrição das ações empreendidas para conduzir uma análise mais detalhada. Foi realizado o preenchimento de uma tabela utilizando a metodologia e na Tabela 1 consta um pedaço do que foi feito para realizar essa análise mais aprofundada.

Ações do Cliente	Estimar custos
Objetivos	Calcular a correlação existente entre o sentimento de postagens sobre uma determinada empresa no Twitter com o valor da sua ação na bolsa de valores
Atividades	Coleta de dados do Twitter e da BMF&Bovespa
Questões	Existe correlação entre o que se posta sobre uma empresa no Twitter e o valor da sua ação na bolsa de valores?
Barreiras	Incertezas na veracidade das postagens do Twitter
Funcionalidades	Mais um indicador para análise de compra e venda de ações na bolsa de valores
Interação	Geração de mais uma análise de investimentos na bolsa de valores
Mensagem	Compreensão das tendências de acordo com as postagens e ter mais um indicador no momento de comprar ou vender ações
Onde ocorre	Aplicativo de desktop ou móvel, plataforma baseada na web
Tarefas aparentes	Gerar indicadores
Tarefas escondidas	Análise de sentimento, Clusterizações e Regressão Logística
Processos de Suporte	Melhorias do modelo
Saída desejada	Indicadores de oportunidades de compra e venda de ações

Tabela 1. Business Design Blueprint

As análises realizadas para o entendimento das tarefas necessárias para se realizar a construção do modelo apontam para um grau de complexidade relevante, tendo em vista que envolve diversas fases.



Pensando em como explicar essa proposta, preenchemos uma ferramenta muito útil chamada “Explicação de Proposta de Valor”, contendo assim as dores, remédios, o criador de ganho dessa proposta e qual será o produto e as tarefas que ele realizará.

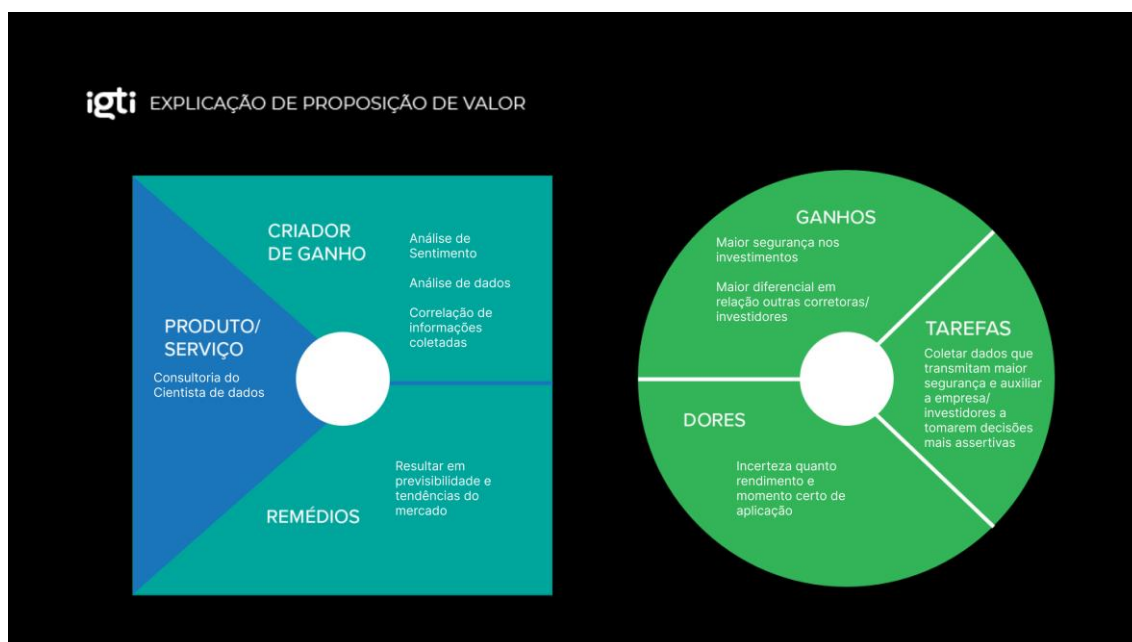


Figura 5. Explicação de Proposição de valor

1.1.4 Hipóteses

Aplicações em ações que desvalorizam rapidamente geram enormes implicações econômicas e financeiras de acordo com montante aplicado. Portanto, é um aspecto ao qual as empresas e investidores dão muita atenção. O retorno dos investimentos, sendo a médio ou a longo prazo é de extrema importância.

O fornecimento de ferramentas que ajudam a identificar possíveis oportunidades de compra ou venda de ações mitiga o risco na tomada de decisão em operações na bolsa de valores. Para expor melhor as hipóteses, a tabela 2 apresenta observações e hipóteses a serem testadas.

Observação	Hipótese
Aplicações em ações sem a correta avaliação levam a prejuízos financeiros	As postagens no Twitter exercem alguma influência sobre o resultado das empresas na bolsa de valores
Corretoras e investidores possuem resistência a mudanças de metodologia	Metodologias de análise de sentimento nas redes sociais podem servir como mais uma ferramenta de investimentos
A maioria dos investidores comuns não entendem a complexidade do mercado financeiro e de investimentos	Estudar economia e investimentos não é algo rápido e intuitivo, é necessário dedicação e tempo
Indicadores de sentimento nas redes sociais são pouco usados para entender o mercado	O conteúdo das redes sociais não é confiável o suficiente para subsidiar as decisões
Dados importantes são disponibilizados gratuitamente e são subutilizados	Falta de mão de obra/tecnologia específica que entenda da análise de dados e sua complexidade

Tabela 2. Tabela de Observações e Hipóteses

Com as hipóteses expostas e todo o contexto do problema analisado, vamos listar abaixo as ideias para a sua resolução e priorizar aquelas mais possíveis. Com o método de Matriz de Priorização BASICO (Benefícios, Abrangência, Satisfação, Investimento, Cliente, Operações) podemos priorizar quais soluções devem ser adotadas e, assim, criar um ranking.

Soluções	B	A	S	I	C	O	Total	Priorização
Calcular a correlação existente entre o sentimento de postagens sobre uma determinada empresa no Twitter com o valor da sua ação na bolsa de valores	5	5	4	3	3	3	23	1
Criação de indicador de compra e venda de ações com base nas postagens do Twitter	4	4	4	4	3	2	21	2
Modelo de predição do preço das ações de acordo com as postagens no Twitter	5	5	5	1	1	1	18	3

Tabela 3. Matriz BASICO - priorização de ideias.

A partir da análise das ideias, priorizamos as aquelas com as maiores pontuações, e com isso “Calcular a correlação existente entre o sentimento de postagens sobre uma determinada empresa no Twitter com o valor da sua ação na bolsa de valores” e “Criação de indicador de compra e venda de ações com base nas postagens do Twitter”.



1.2 Solução

1.2.1 Objetivo SMART

O objetivo SMART (específico, mensurável, atingível, relevante e temporal) do projeto é criar mais uma ferramenta de análise de investimento na bolsa de valores utilizando dados de postagens no Twitter.

Este objetivo é específico, pois aborda o desafio direto de proporcionar análises adicionais para as corretoras e investidores comuns a partir de análises de sentimento das postagens no Twitter sobre empresas listadas na bolsa de valores.

É mensurável, pois o sucesso do projeto pode ser quantificado pela precisão e confiabilidade das previsões de tendências de valores produzidas pela correlação.

É atingível, dado que as técnicas e tecnologias necessárias para realizar previsões baseadas em dados disponíveis são acessíveis e utilizáveis. A relevância do objetivo reside na solução que oferece a um problema prático e significativo enfrentado por grandes corretoras e investidores comuns.

Por fim, é temporal, pois o projeto tem um prazo claramente definido para a conclusão e as previsões produzidas pelo modelo devem ser atualizadas regularmente para manter sua eficácia e relevância.

1.2.2 Premissas e Restrições

Premissas do projeto:

- Máximo de duas horas de disponibilidade por dia;
- Utilização da linguagem Python;
- Dados utilizados serão públicos e suas fontes são Twitter e Bolsa de Valores - BMF&Bovespa.

Restrições do projeto:

- Prazo até o fim de setembro/2023.



1.2.3 Backlog de Produto

As etapas necessárias para a realização da solução proposta neste trabalho, de análise de sentimento de postagens sobre empresas listadas na bolsa de valores na rede social Twitter estão representadas na Figura 6, que apresenta o backlog do produto utilizando o aplicativo Trello.

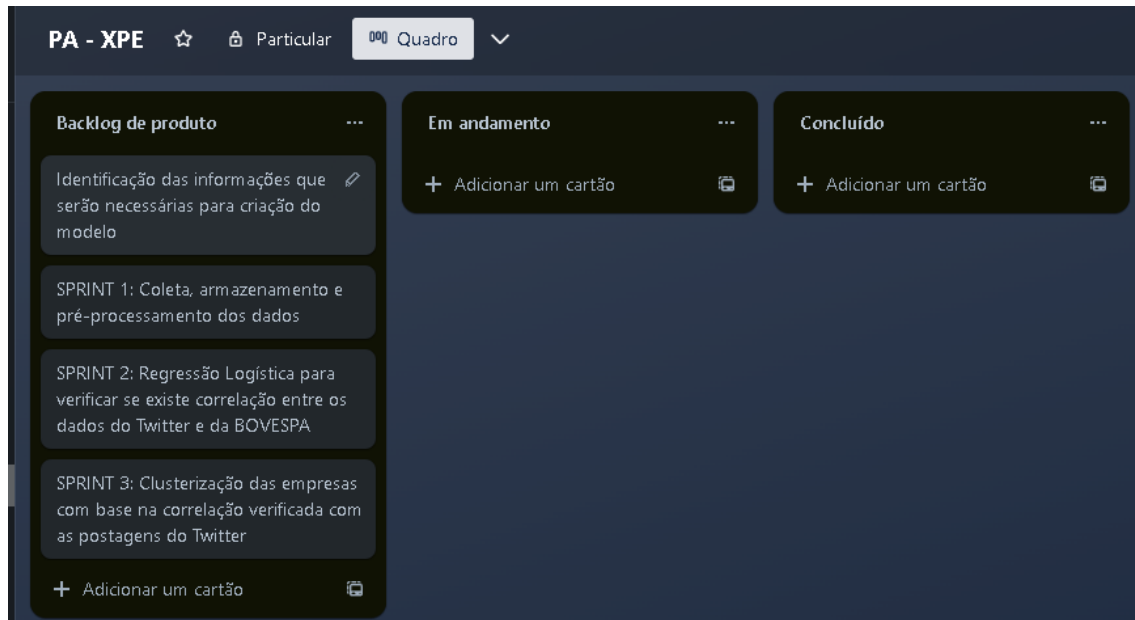


Figura 6. Backlog de Produto

2. Área de Experimentação

Esta seção apresenta a descrição da execução deste projeto, detalhando todas as etapas realizadas para atingir o objetivo SMART, conforme Figura 6, apresentada na seção anterior.

2.1 Sprint 1

2.1.1 Solução

▪ Evidência do planejamento:

Planejamento e execução da Sprint 1 apresentados na Figura 7, que possui como foco principal a coleta e o armazenamento dos dados, conforme descrito no BackLog do produto (Figura 6).

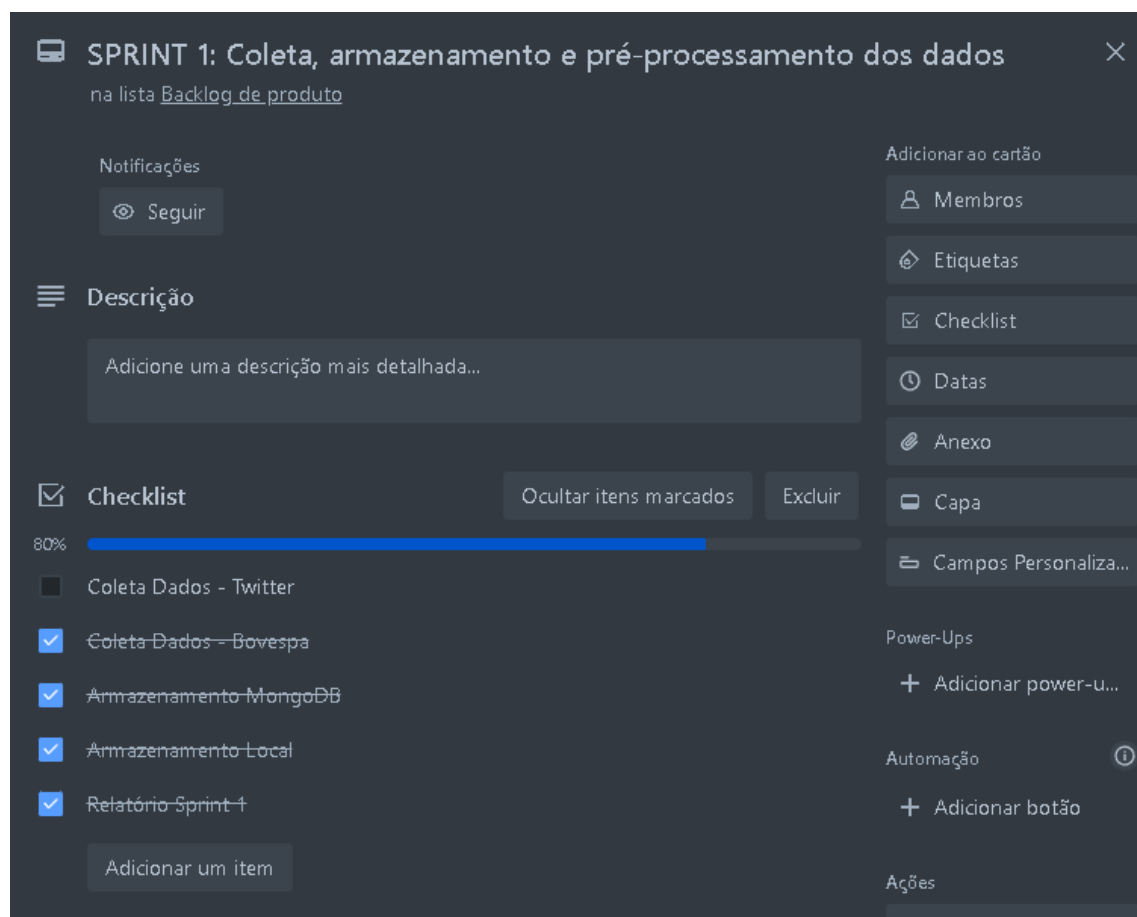
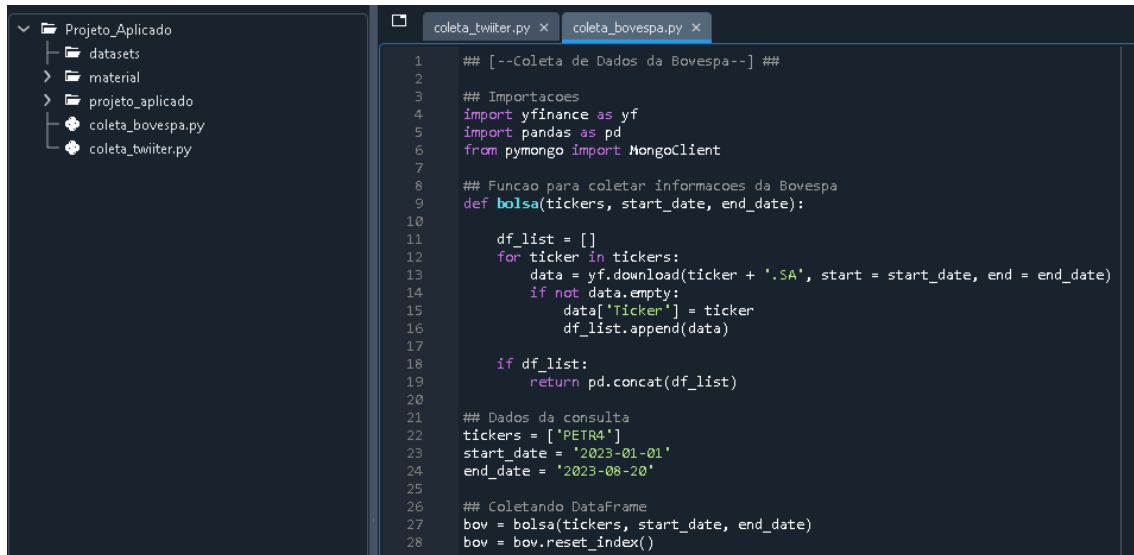


Figura 7. Planejamento e execução da Sprint 1

- Evidência de execução de cada requisito:

A Figura 8 apresenta a evidência de coleta dos dados da Bovespa relativos a empresa Petrobrás, que será utilizada como referência para o projeto por conta das limitações de acesso aos dados do Twitter.

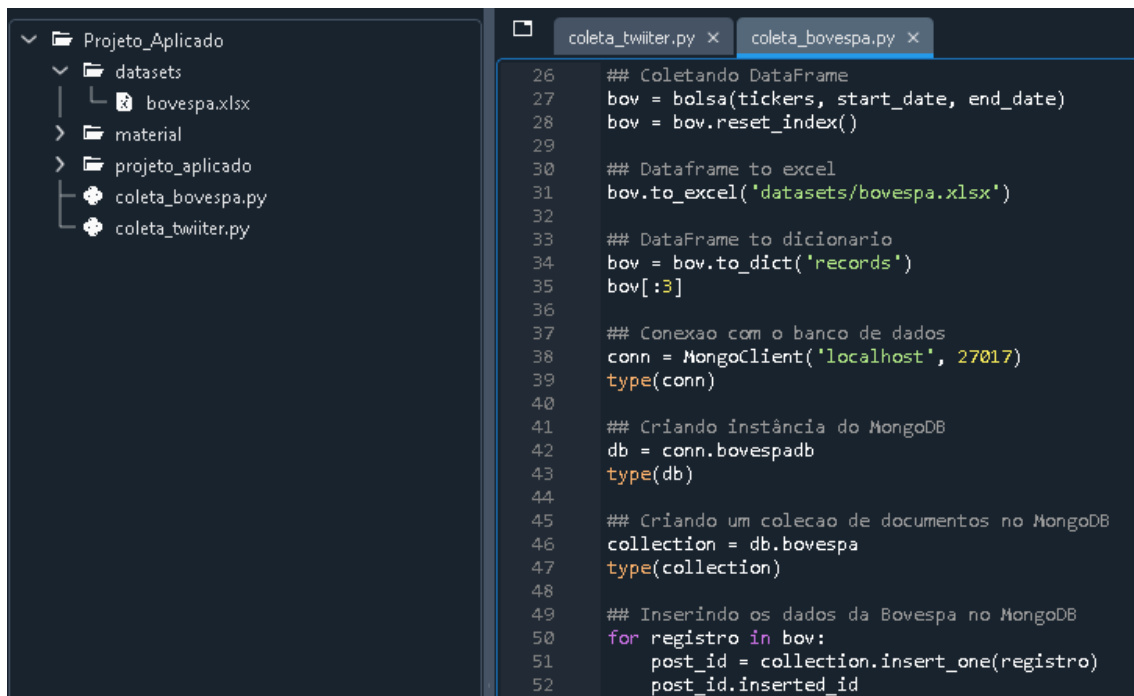


```

1  ## [--Coleta de Dados da Bovespa--] ##
2
3  ## Importacoes
4  import yfinance as yf
5  import pandas as pd
6  from pymongo import MongoClient
7
8  ## Funcao para coletar informacoes da Bovespa
9  def bolsa(tickers, start_date, end_date):
10
11      df_list = []
12      for ticker in tickers:
13          data = yf.download(ticker + '.SA', start = start_date, end = end_date)
14          if not data.empty:
15              data['Ticker'] = ticker
16              df_list.append(data)
17
18      if df_list:
19          return pd.concat(df_list)
20
21  ## Dados da consulta
22  tickers = ['PETR4']
23  start_date = '2023-01-01'
24  end_date = '2023-08-20'
25
26  ## Coletando DataFrame
27  bov = bolsa(tickers, start_date, end_date)
28  bov = bov.reset_index()
  
```

Figura 8. Coleta de Dados da Bovespa - Petrobrás

A Figura 9 apresenta a evidência do armazenamento dos dados da Bovespa no disco local e no Mongo DB.



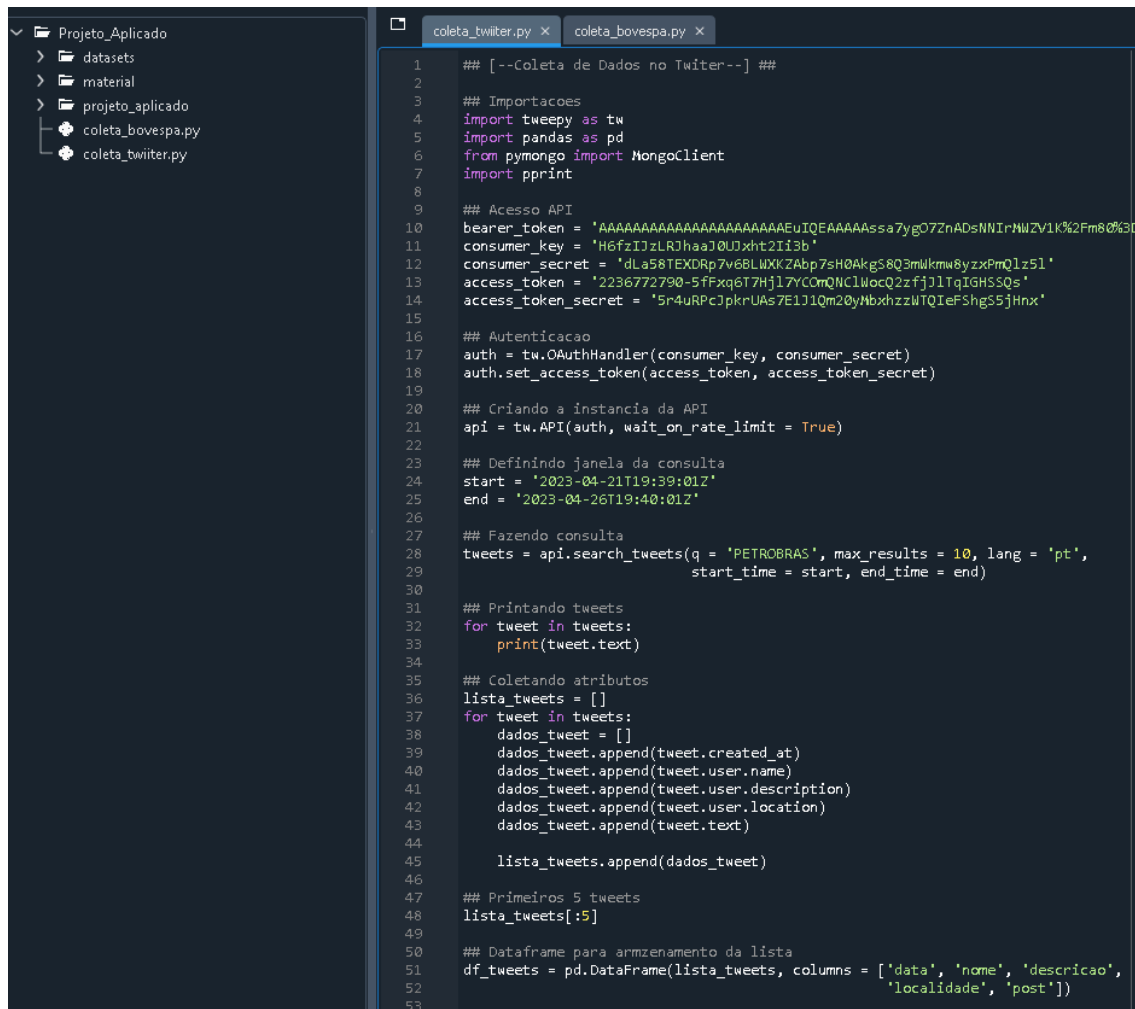
```

26  ## Coletando DataFrame
27  bov = bolsa(tickers, start_date, end_date)
28  bov = bov.reset_index()
29
30  ## Dataframe to excel
31  bov.to_excel('datasets/bovespa.xlsx')
32
33  ## Dataframe to dicionario
34  bov = bov.to_dict('records')
35  bov[:3]
36
37  ## Conexao com o banco de dados
38  conn = MongoClient('localhost', 27017)
39  type(conn)
40
41  ## Criando instancia do MongoDB
42  db = conn.bovespadb
43  type(db)
44
45  ## Criando um colecao de documentos no MongoDB
46  collection = db.bovespa
47  type(collection)
48
49  ## Inserindo os dados da Bovespa no MongoDB
50  for registro in bov:
51      post_id = collection.insert_one(registro)
52      post_id.inserted_id
  
```

Figura 9. Armazenamento dos Dados da Bovespa



As Figuras 10 e 11 apresentam a evidência do script construído para coletar as informações do Twitter. No entanto, as credenciais utilizadas para acesso à API não possuíam mais cargas de consultas no momento da execução.

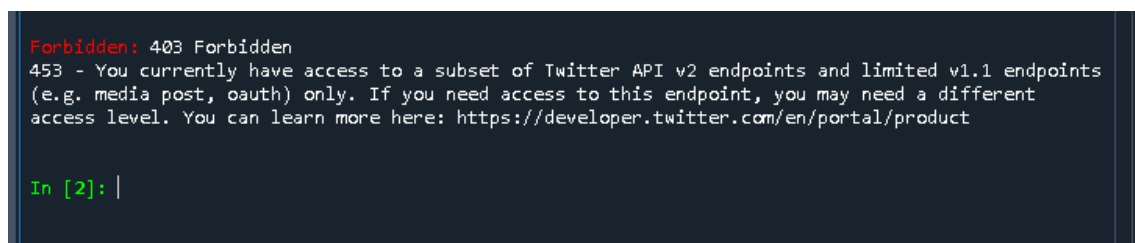


```

1  ## [--Coleta de Dados no Twitter--] ##
2
3  ## Importacoes
4  import tweepy as tw
5  import pandas as pd
6  from pymongo import MongoClient
7  import pprint
8
9  ## Acesso API
10 bearer_token = 'AAAAAAAAAAAAAAAAAAAAEuIQEAAAAAssa7ygO7ZnADsNNIrMWZV1K%2Fm80%3DA'
11 consumer_key = 'H6fzIJzLRJhaa10UJxht2Ii3b'
12 consumer_secret = 'dLa58TEXDRp7v68LWXKZAbp7sH0AkG58Q3mWkmw8yzxPmQ1z51'
13 access_token = '2236772790-5fFxa6T7Hj17YCOmQNC1WocQ2zfj1LTqIGHSSQs'
14 access_token_secret = '5n4uRPeJpkrUAs7E1J1Qm20yNbxhzzWTQIEFShgS5jHnx'
15
16 ## Autenticacao
17 auth = tw.OAuthHandler(consumer_key, consumer_secret)
18 auth.set_access_token(access_token, access_token_secret)
19
20 ## Criando a instancia da API
21 api = tw.API(auth, wait_on_rate_limit = True)
22
23 ## Definindo janela da consulta
24 start = '2023-04-21T19:39:01Z'
25 end = '2023-04-26T19:40:01Z'
26
27 ## Fazendo consulta
28 tweets = api.search_tweets(q = 'PETROBRAS', max_results = 10, lang = 'pt',
29                             start_time = start, end_time = end)
30
31 ## Printando tweets
32 for tweet in tweets:
33     print(tweet.text)
34
35 ## Coletando atributos
36 lista_tweets = []
37 for tweet in tweets:
38     dados_tweet = []
39     dados_tweet.append(tweet.created_at)
40     dados_tweet.append(tweet.user.name)
41     dados_tweet.append(tweet.user.description)
42     dados_tweet.append(tweet.user.location)
43     dados_tweet.append(tweet.text)
44
45     lista_tweets.append(dados_tweet)
46
47 ## Primeiros 5 tweets
48 lista_tweets[:5]
49
50 ## Dataframe para armazenamento da lista
51 df_tweets = pd.DataFrame(lista_tweets, columns = ['data', 'nome', 'descricao',
52                                                  'localidade', 'post'])
53

```

Figura 10. Tentativa de Coleta de Dados do Twitter



```

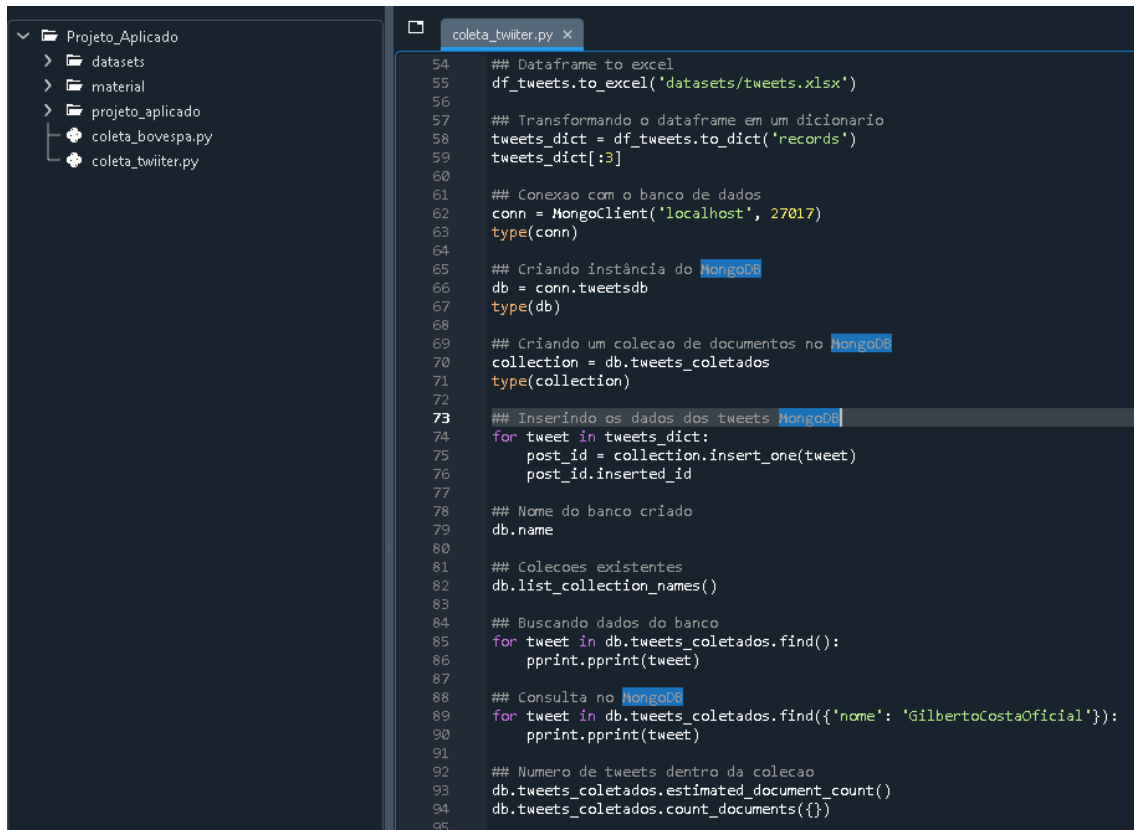
Forbidden: 403 Forbidden
453 - You currently have access to a subset of Twitter API v2 endpoints and limited v1.1 endpoints
(e.g. media post, oauth) only. If you need access to this endpoint, you may need a different
access level. You can learn more here: https://developer.twitter.com/en/portal/product

In [2]: |

```

Figura 11. Erro API

Ainda que o acesso aos dados do Twitter não esteja funcionando, as linhas de código para o armazenamento local e no Mongo DB já estão prontas, conforme evidenciado na Figura 12.



```

54  ## Dataframe to excel
55  df_tweets.to_excel('datasets/tweets.xlsx')
56
57  ## Transformando o dataframe em um dicionario
58  tweets_dict = df_tweets.to_dict('records')
59  tweets_dict[:3]
60
61  ## Conexao com o banco de dados
62  conn = MongoClient('localhost', 27017)
63  type(conn)
64
65  ## Criando instância do MongoDB
66  db = conn.tweetsdb
67  type(db)
68
69  ## Criando um coleção de documentos no MongoDB
70  collection = db.tweets_coletados
71  type(collection)
72
73  ## Inserindo os dados dos tweets MongoDB
74  for tweet in tweets_dict:
75      post_id = collection.insert_one(tweet)
76      post_id.inserted_id
77
78  ## Nome do banco criado
79  db.name
80
81  ## Coleções existentes
82  db.list_collection_names()
83
84  ## Buscando dados do banco
85  for tweet in db.tweets_coletados.find():
86      pprint.pprint(tweet)
87
88  ## Consulta no MongoDB
89  for tweet in db.tweets_coletados.find({'nome': 'GilbertoCostaOficial'}):
90      pprint.pprint(tweet)
91
92  ## Numero de tweets dentro da coleção
93  db.tweets_coletados.estimated_document_count()
94  db.tweets_coletados.count_documents({})
95

```

Figura 12. Evidência do Script de Armazenamento Local e no Mongo DB dos dados do Twitter

▪ Evidência dos resultados

A Figura 13 apresenta a evidência do resultado obtido com o armazenamento dos dados da Bovespa no Mongo DB, restando apenas o acesso aos dados do Twitter para a realização da coleta e respectivo armazenamento também no Mondo DB. A chave de ligação entre os dois bancos será a data.

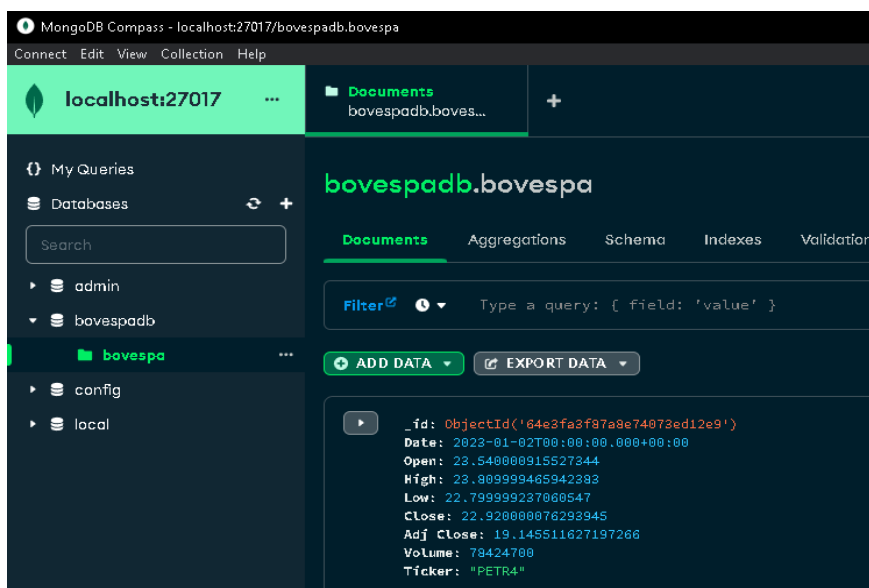


Figura 13. Armazenamento dos Dados da Bovespa no Mongo DB



2.1.2 Lições Aprendidas

Tendo em vista a dificuldade encontrada para acessar os dados do Twitter, foi necessário mudar a estratégia inicial de utilizar dados de mais uma empresa para o estudo. Devido à restrição de acesso aos dados, o estudo irá se concentrar apenas nas menções realizadas para a Petrobrás, empresa estatal e de capital aberto.

No entanto, ainda que o acesso aos dados permaneça restrito à uma única empresa, ainda persiste o problema de acesso a qualquer informação do Twitter por enquanto. Continuarei na busca por solução no decorrer do Sprint 2.

2.2 Sprint 2

2.2.1 Solução

▪ Evidência do planejamento:

Planejamento e execução da Sprint 2 apresentados na Figura 14, que possui como foco principal a organização dos scripts do projeto, a análise de sentimento dos tweets coletados, a exportação dos datasets e as análises iniciais.

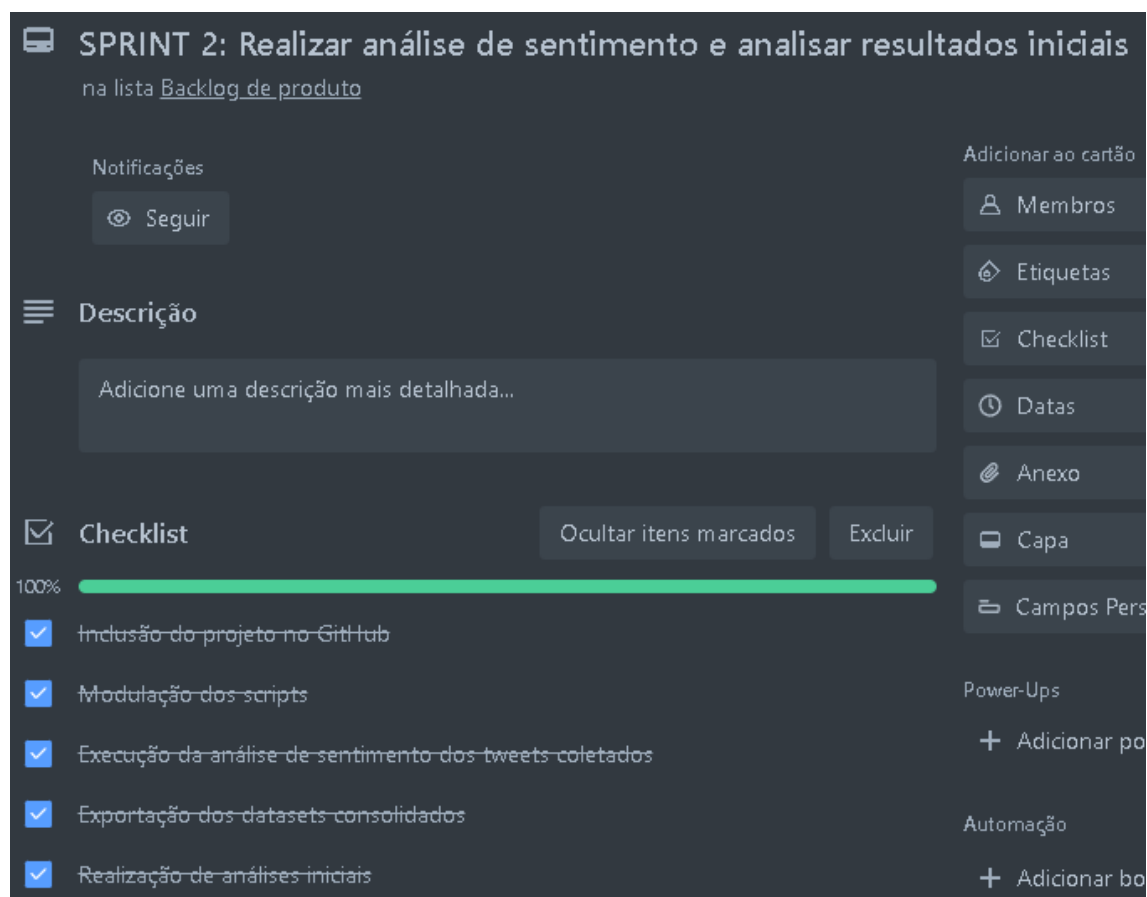


Figura 14. Planejamento e execução da Sprint 2

▪ **Evidência de execução de cada requisito:**

Antes de iniciar a execução da Sprint 2 foi necessário realizar a atividade remanescente da Sprint 1, de coletar um dataset de tweets, tendo em vista que a API utilizada não tinha mais disponibilidade.

Portanto, para fins exclusivos de modelagem do projeto, o estudo utiliza um dataset (.jsonl) disponível no zenodo, que contém mais de 5 milhões de tweets coletados entre 6 de dezembro de 2013 e 30 de junho de 2019.

DataSet: <https://zenodo.org/record/5068253>

A Figura 15 apresenta a atualização do checklist de atividades da Sprint 1, evidenciando a conclusão da respectiva etapa:

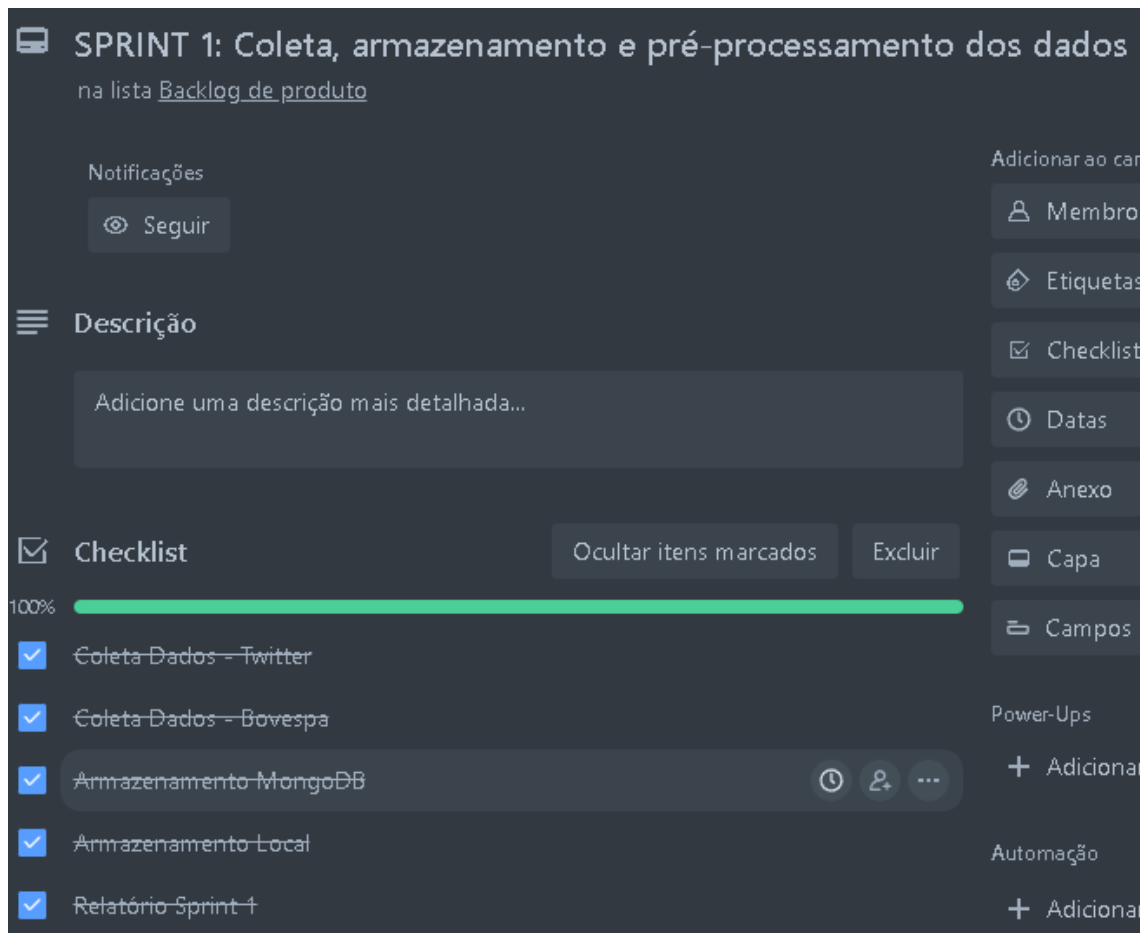


Figura 15. Planejamento e execução da Sprint 1

No entanto, por se tratar de um dataset de tweets norte-americanos, foi necessário ajustar a estratégia para a utilização de papéis negociados na Nasdaq ao invés da BOVESPA.

Partindo para as evidências da Sptint 2, a Figura 16 apresenta a inclusão do projeto no GitHub <https://github.com/guedesf/tweets_market>

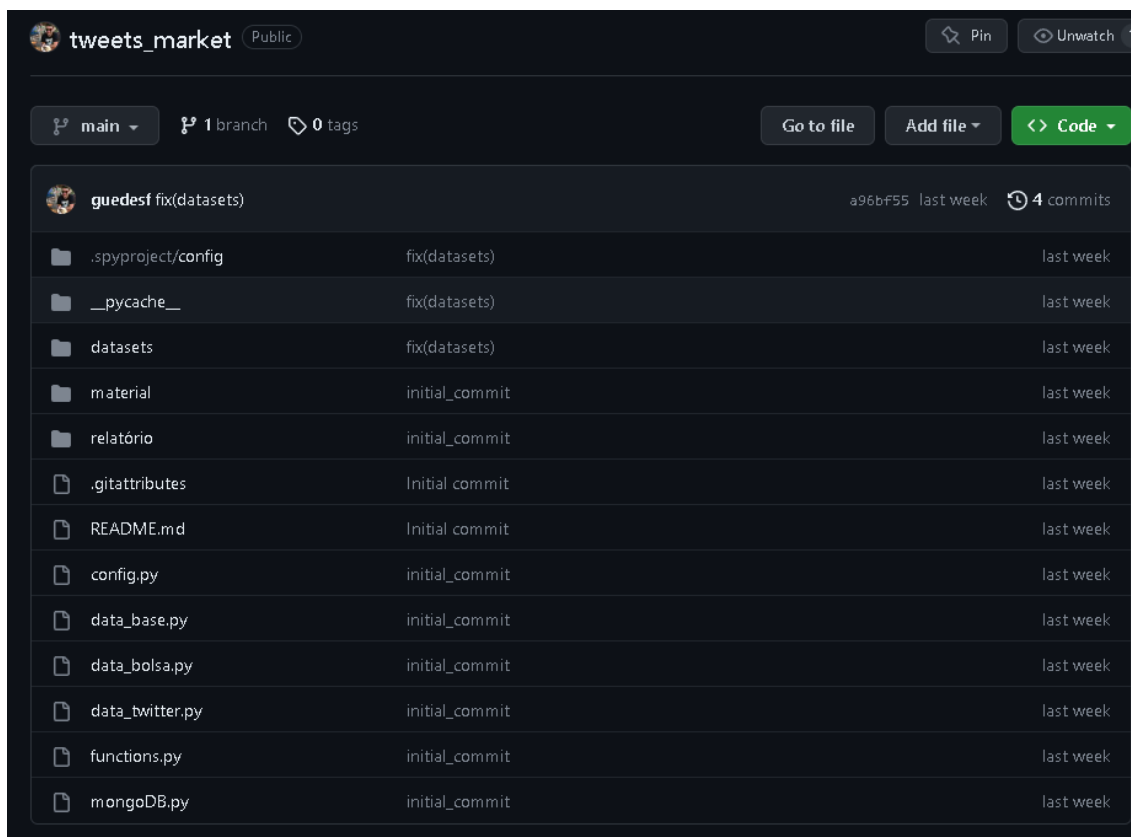


Figura 16. Inclusão do projeto no GitHub

A Figura 17 apresenta a modulação dos scripts do projeto na IDE Spyder, com a criação de módulo específico para as funções que são utilizadas no projeto, separação dos scripts que coletam dados da Nasdaq e do arquivo .jsonl com os tweets, script específico para inserir dados no MongoDB e script de inputs e parâmetros.

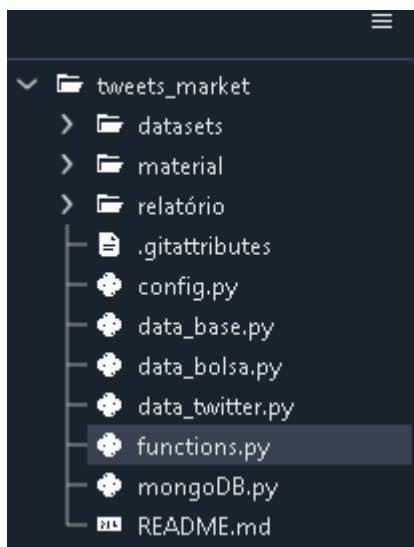
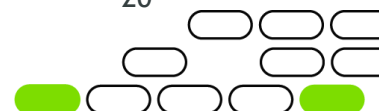


Figura 17. Modulação do projeto



A Figura 18 apresenta a função que realiza a análise de sentimento dos tweets contidos no dataset utilizando o pacote TextBlob.

```
1  ## [--DataBase--] ##
2
3  ## Importacoes
4  import pandas as pd
5  import yfinance as yf
6  from textblob import TextBlob
7
8  ## Coleta informacoes NYSE e NASDAQ
9  def bolsa(tickers, start_date, end_date):
10
11      df_list = []
12      for ticker in tickers:
13          data = yf.download(ticker, start = start_date, end = end_date)
14
15          if not data.empty:
16              data['Ticker'] = ticker
17              df_list.append(data)
18
19      if df_list:
20          return pd.concat(df_list)
21
22  ## Analise de sentimento
23  def sentiment(tweet_text):
24      analysis = TextBlob(tweet_text)
25      sentiment_score = analysis.sentiment.polarity
26
27      if sentiment_score > 0:
28          return 'positivo'
29
30      elif sentiment_score < 0:
31          return 'negativo'
32
33      else:
34          return 'neutro'
```

Figura 18. Execução da análise de sentimento

Após a modulação do projeto, foram realizadas alterações para o cruzamento entre os dados de uma empresa específica negociada na Nasdaq e os tweets utilizados. Para essa etapa foram utilizadas as seguintes empresas:

- Apple;
- Google;
- Meta/Facebook;
- Microsoft;
- Tesla.

Os dados foram consolidados apenas para os dias em que houve pelo menos um tweet publicado com menção à empresa analisada. Após o cruzamento dos dados, verificou-se os seguintes quantitativos de amostra de dados por empresa:



Empresa	Amostra
Apple	818
Google	1.197
Meta	1.257
Microsoft	317
Tesla	160

Tabela 4. Amostras do DataSet

A Figura 19 evidencia a exportação dos datasets individuais (tweets e Nasdaq) e a consolidação e exportação em .xlsx do dataset para análise para cada empresa.

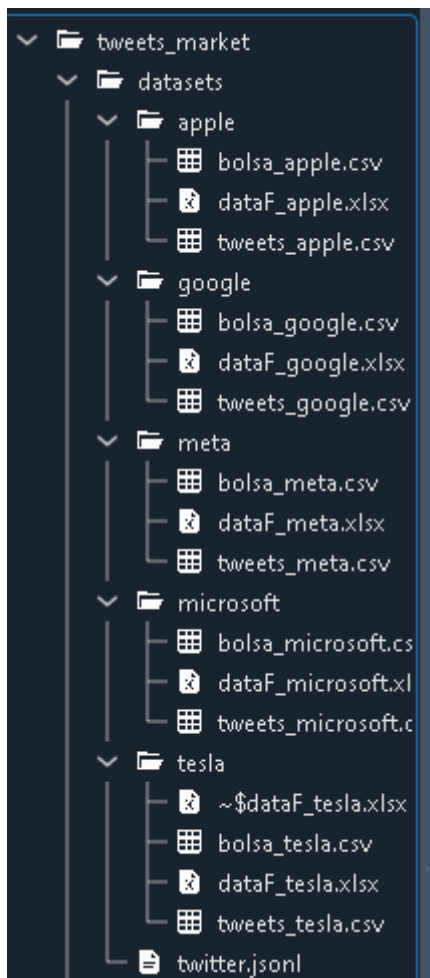


Figura 19. Exportação dos datasets consolidados

As análises iniciais apontam para uma correlação linear de Pearson inexistente entre o resultado das empresas na Nasdaq e a quantidade de tweets (positivos e negativos), bem como com o volume diário de negociações.

A Figura 20 apresenta evidência da análise inicial de calor dos datasets consolidados para cada empresa:

	A	B	C	D	E	F	G
1	Id_Date	Volume	Retorno	Quantidade	Positivo	Negativo	Neutro
233	20150105	41.182.000	-1,91%	495	159	47	289
234	20150106	54.456.000	-2,47%	566	196	77	293
235	20150107	46.918.000	-0,29%	587	157	88	342
236	20150108	73.054.000	0,35%	482	149	93	240
237	20150109	42.000.000	-1,22%	450	143	43	264
238	20150112	57.138.000	-0,73%	630	183	79	368
239	20150113	60.958.000	0,95%	584	149	46	389
240	20150114	52.800.000	0,82%	617	193	72	352
241	20150115	51.068.000	-0,38%	545	126	69	350
242	20150116	49.658.000	1,28%	474	123	42	309
243	20150120	46.796.000	-0,10%	340	112	35	193
244	20150121	46.356.000	2,05%	348	127	27	194
245	20150122	56.068.000	3,25%	388	107	30	251
246	20150123	45.966.000	0,87%	317	88	31	198
247	20150126	30.932.000	-0,97%	351	106	39	206
248	20150127	39.148.000	-2,89%	386	189	34	163
249	20150128	35.822.000	-1,68%	306	97	47	162
250	20150129	79.018.000	0,16%	547	274	57	216
251	20150130	121.108.000	4,74%	408	155	43	210
252	20150202	75.378.000	-1,00%	2	1	0	1

Figura 20. Evidência de análise inicial

▪ Evidência dos resultados

O projeto do GitHub <https://github.com/guedesf/tweets_market> contém todo o conteúdo do projeto até a finalização da Sprint 2, comprovando a execução de todas as etapas propostas e apresentadas.

Antes da postagem da Sprint 3, será realizado novo commit no projeto para que seja possível acompanhar as alterações de código e os resultados obtidos.

2.2.2 Lições Aprendidas

Considerando a dificuldade em se obter a coleta de dados de tweets, foi necessário readequar o projeto por diversas vezes como forma de ajustamento aos desafios reais, mostrando que por muitas vezes o inicialmente projetado pode não ser possível naquele momento.

A obtenção de um dataset insuficiente para a confiabilidade dos resultados, ainda que não seja o melhor cenário, destravou o projeto para a continuidade da modelagem da análise, sendo necessário, quando possível, a substituição do dataset utilizado no momento.



2.3 Sprint 3

2.3.1 Solução

- **Evidência do planejamento:**

Planejamento e execução da Sprint 3 apresentados na Figura 21, que possui como foco as análises iniciais dos resultados.

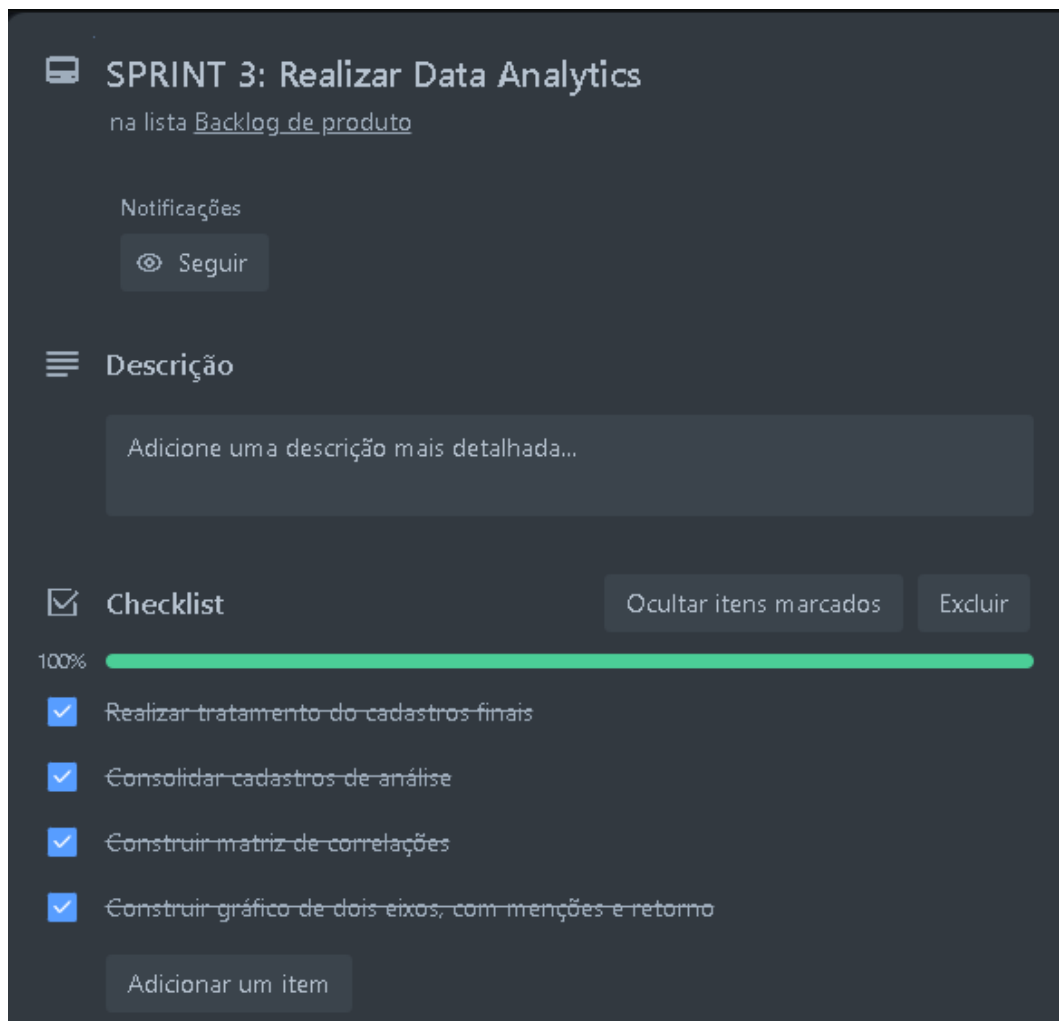


Figura 21. Planejamento e execução da Sprint 3

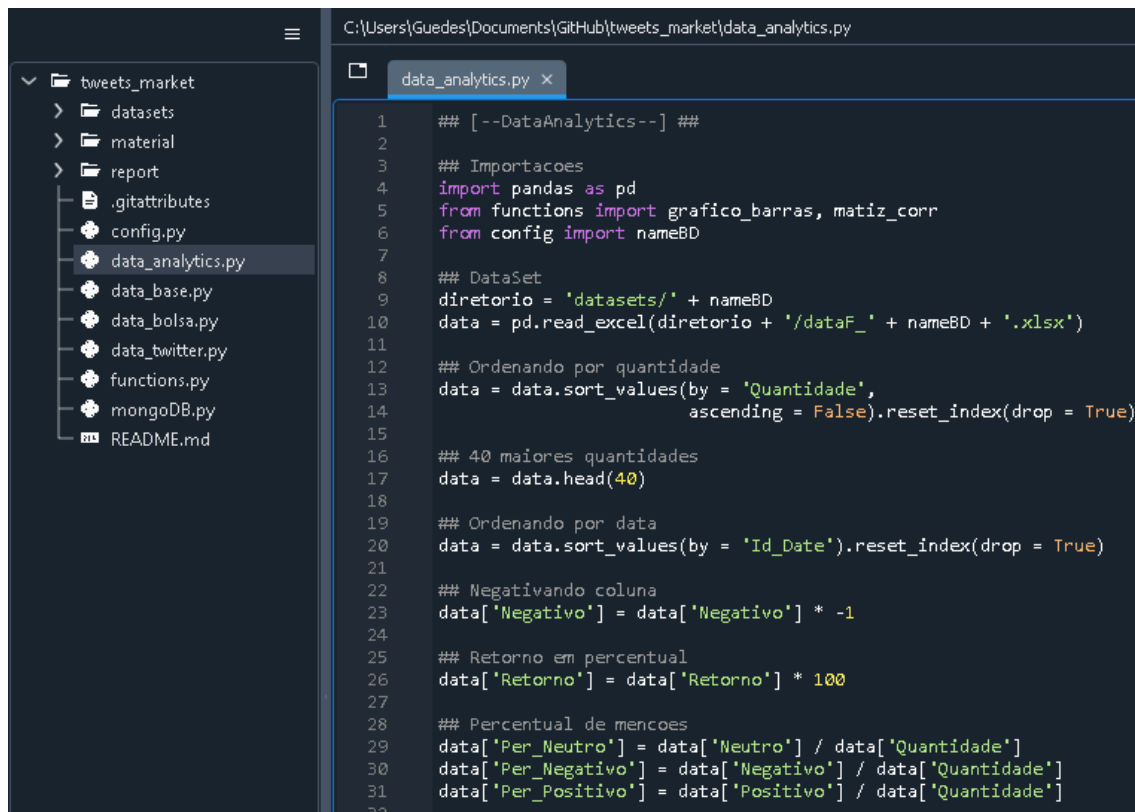
Em decorrência das alterações ocorridas no decorrer do projeto, frente às dificuldades detalhadas nas seções anteriores, as atividades da Sprint 3 precisaram ser ajustadas, de forma a compreender as modificações efetuadas.

- **Evidência de execução de cada requisito:**

A Figura 22 apresenta pequenos ajustes realizados nos cadastros obtidos na Sprint2 para melhoria da visualização dos resultados. Também apresenta a criação de novas variáveis, ordenamento de acordo com as datas que possuem a maior quantidade de



menções e consolidação do cadastro para análise com os 40 dias com as maiores quantidades observadas.



```

1  ## [--DataAnalytics--] ##
2
3  ## Importacoes
4  import pandas as pd
5  from functions import grafico_barras, matiz_corr
6  from config import nameBD
7
8  ## DataSet
9  diretorio = 'datasets/' + nameBD
10 data = pd.read_excel(diretorio + '/dataF_' + nameBD + '.xlsx')
11
12 ## Ordenando por quantidade
13 data = data.sort_values(by = 'Quantidade',
14                         ascending = False).reset_index(drop = True)
15
16 ## 40 maiores quantidades
17 data = data.head(40)
18
19 ## Ordenando por data
20 data = data.sort_values(by = 'Id_Date').reset_index(drop = True)
21
22 ## Negativando coluna
23 data['Negativo'] = data['Negativo'] * -1
24
25 ## Retorno em percentual
26 data['Retorno'] = data['Retorno'] * 100
27
28 ## Percentual de mencoes
29 data['Per_Neutro'] = data['Neutro'] / data['Quantidade']
30 data['Per_Negativo'] = data['Negativo'] / data['Quantidade']
31 data['Per_Positivo'] = data['Positivo'] / data['Quantidade']
32
  
```

Figura 22. Tratamento dos cadastros

A Figura 23 apresenta as funções criadas para a geração das matrizes de correlação e do gráfico de barras com dois eixos para visualização das quantidades de menções positivas, negativas e o retorno da empresa na Nasdaq.



```

## Grafico de barras
def grafico_barras(data):
    # Espacamento e index
    bar_width = 0.2
    index = np.arange(len(data['Id_Date']))

    # Tamanho do grafico
    plt.figure(figsize=(12, 6))

    plt.bar(index - bar_width, data['Positivo'], bar_width,
            label = 'Positivo', color = 'b', alpha = 0.7)
    plt.bar(index + bar_width, data['Negativo'], bar_width,
            label = 'Negativo', color = 'r', alpha = 0.7)

    # Orientacao vertical
    plt.xticks(index, data['Id_Date'], rotation = 'vertical')

    # Posicao legenda y1
    plt.legend(loc = 'upper left')

    # Legenda y1
    plt.ylabel('Quantidade de Mencoes')

    # Criando eixo y2
    ax2 = plt.twinx()

    # Retorno no y2
    ax2.plot(index + bar_width, data['Retorno'],
            marker = 'o', color = '.50', label = 'Retorno')

    # Posicao legenda y2
    ax2.legend(loc = 'upper right')

    # Legenda y2
    ax2.set_ylabel('Retorno (%)')

    # Título
    plt.title('Analise ' + nameBD)
    plt.grid(True)

    # Salvando
    plt.savefig(diretorio + '/analise_' + nameBD + '.png',
            dpi = 300, bbox_inches = 'tight')

## Matriz de correlacao
def matiz_corr(data):
    # Matriz
    matriz_corr = data.corr()
    plt.figure(figsize = (8, 6))
    sns.heatmap(matriz_corr, annot = True,
            cmap = 'coolwarm', vmin = -1, vmax = 1)

    # Salvando
    plt.title('Matriz de Correlacao')
    plt.savefig(diretorio + '/correlacao_' + nameBD + '.png',
            dpi = 300, bbox_inches = 'tight')
  
```

Figura 23. Funções de criação dos gráficos de análise



As alterações efetuadas no projeto podem ser verificadas com maiores detalhes no commit 'sprint3' realizado no GitHub <https://github.com/guedesf/tweets_market>

Evidência dos resultados

Os resultados obtidos para cada empresa analisada encontram-se apresentados nos tópicos a seguir.

Apple

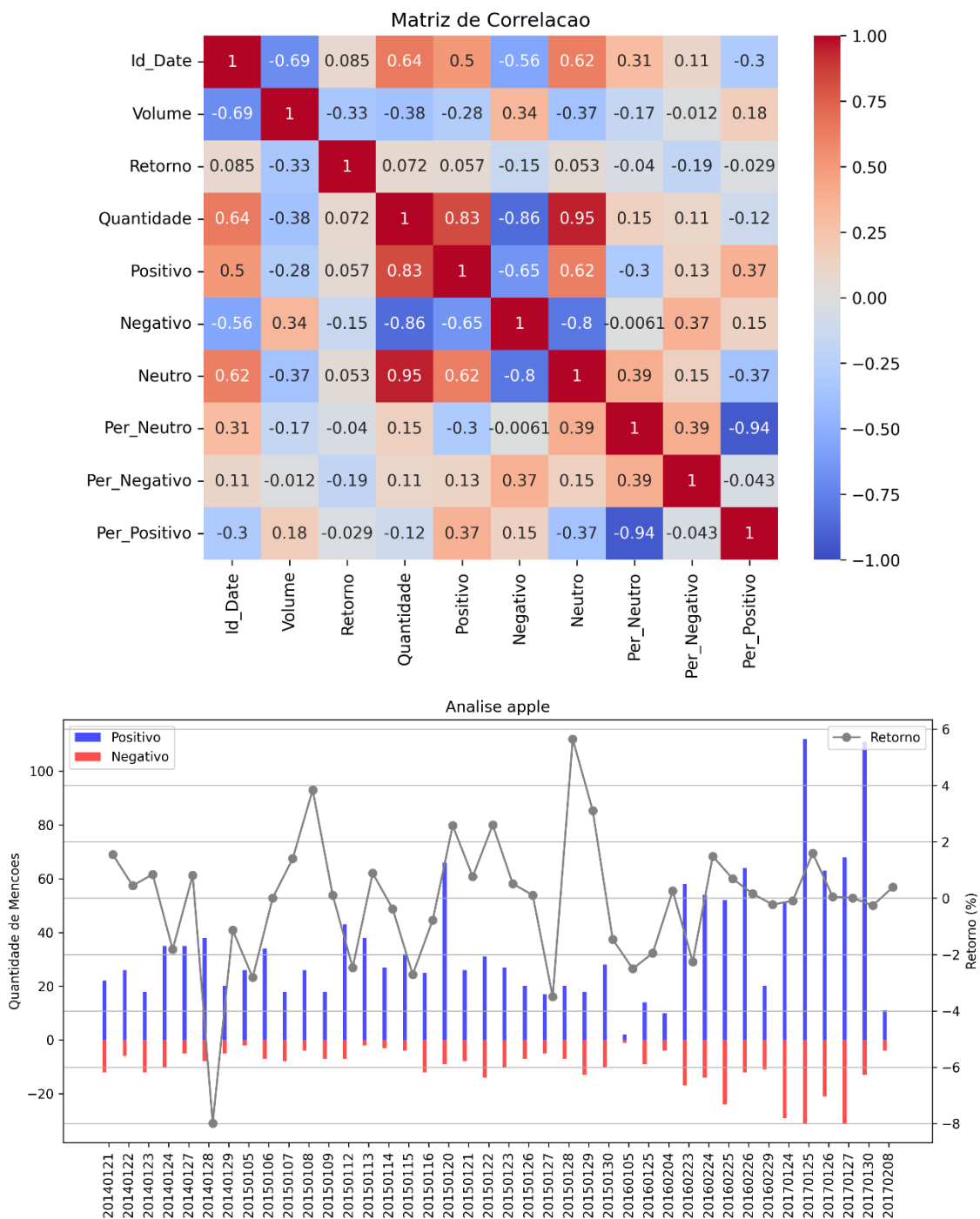


Figura 24. Resultados Apple

Google

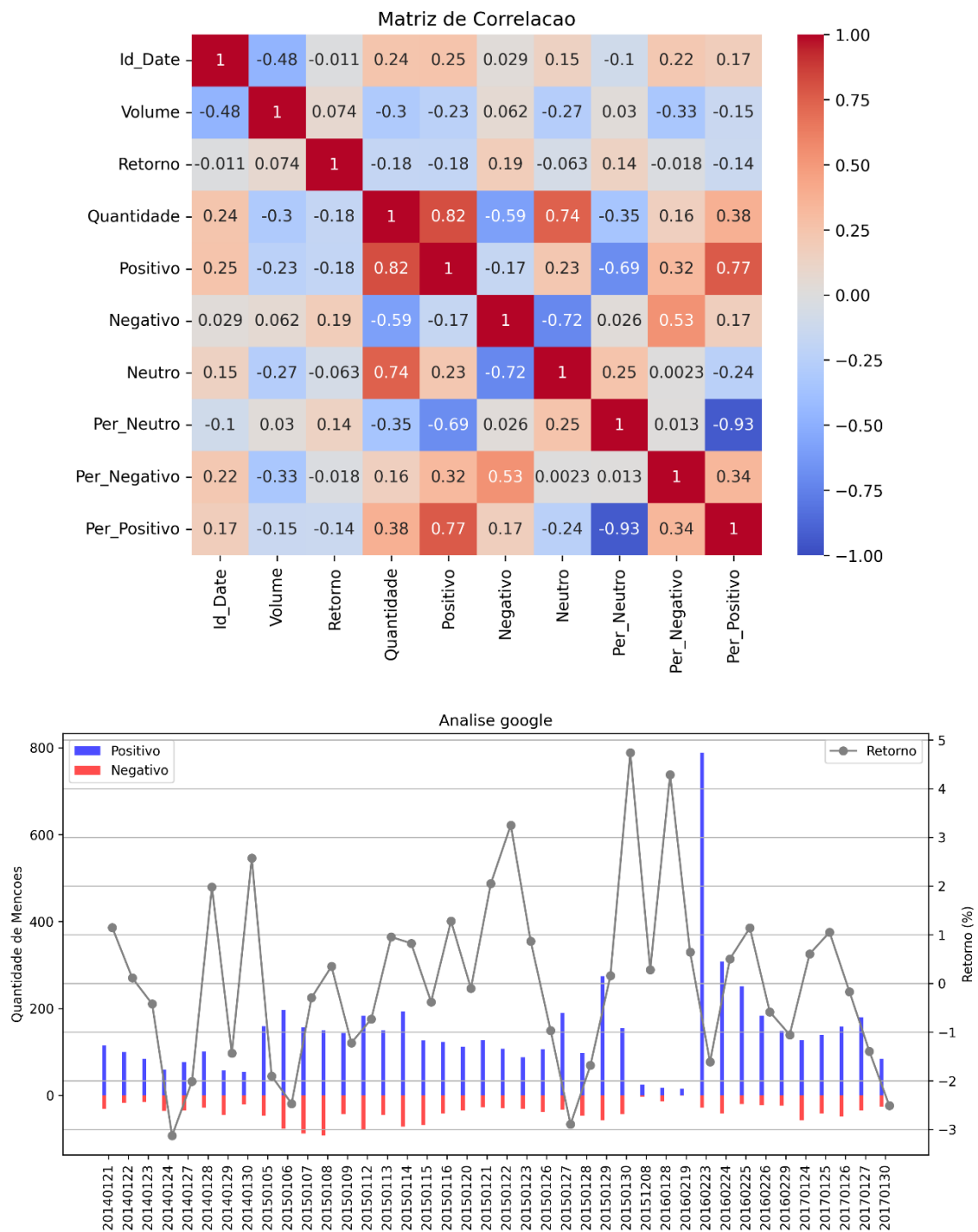


Figura 25. Resultados Google

Meta/Facebook

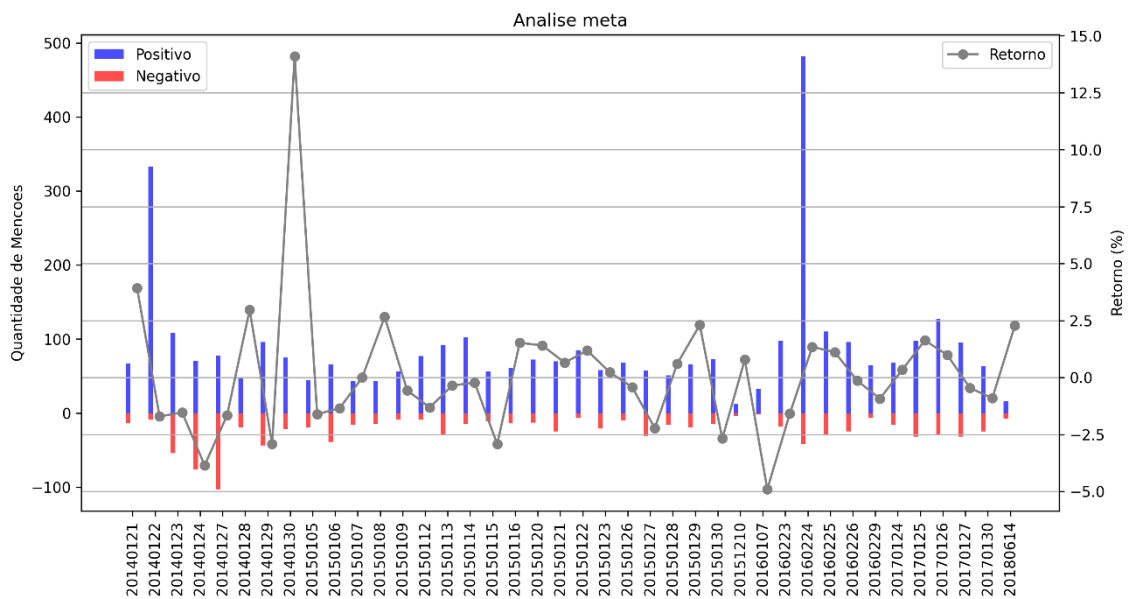
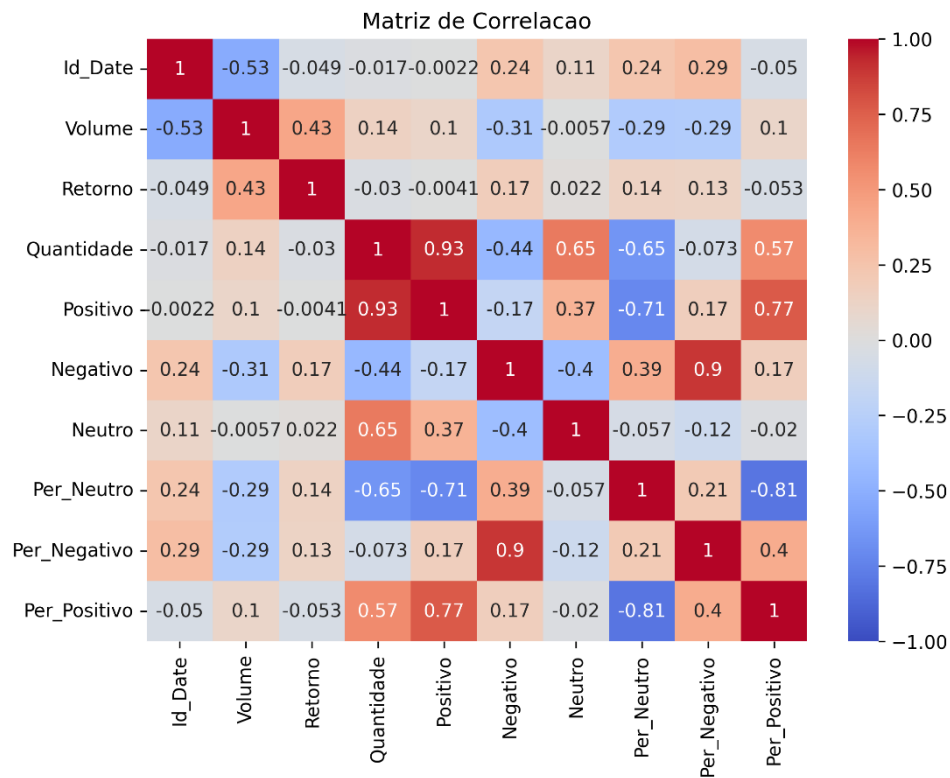


Figura 26. Resultados Meta/Facebook

Microsoft

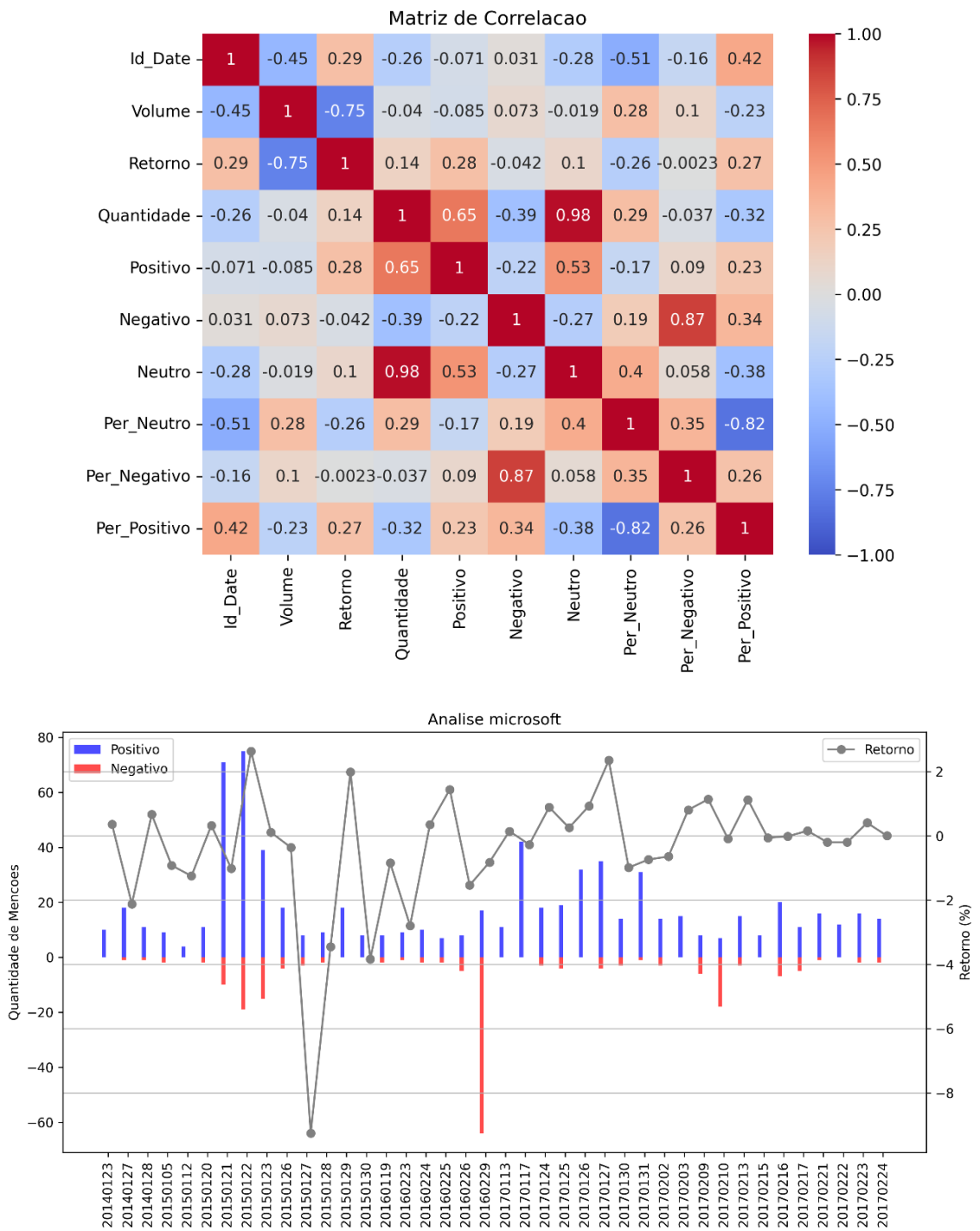


Figura 27. Resultados Microsoft

▪ Tesla

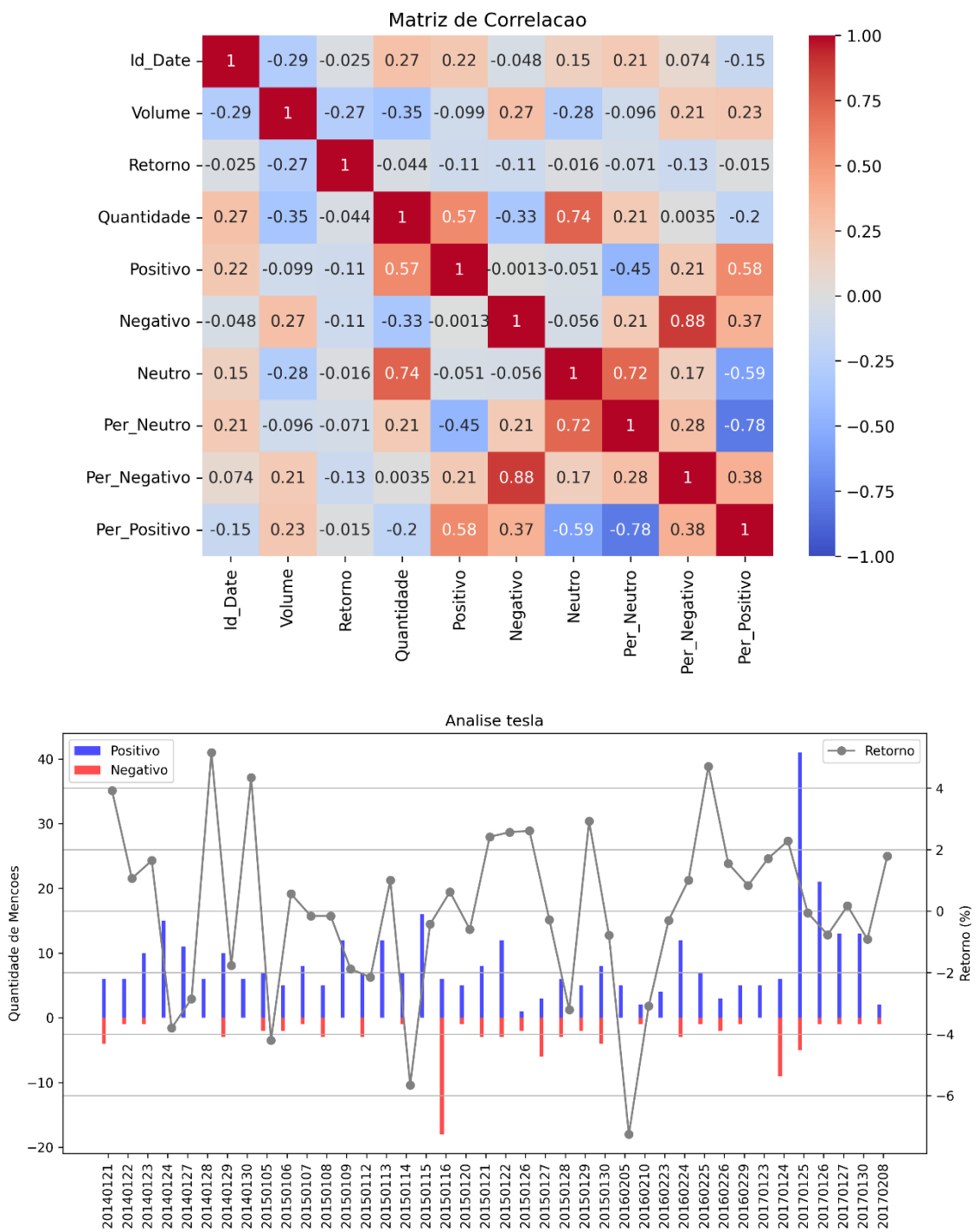


Figura 28. Resultados Tesla

2.3.2 Lições Aprendidas

Frente a todas as dificuldades enfrentadas na obtenção dos dados, sendo necessária a utilização de um dataset aleatório para fins didáticos, é possível concluir que não existe nenhum indicativo de correlação entre os tweets que mencionam as empresas listadas na Nasdaq, disponíveis nos dados considerados, e o seu respectivo retorno.

Permanece a necessidade de obtenção de um dataset mais confiável ou mais correlacionado com o estudo, de forma a verificar eventual utilidade do modelo.

3. Considerações Finais

3.1 Resultados

O objetivo deste trabalho foi o de analisar a existência de eventual correlação entre o conteúdo de tweets que mencionam empresas de capital aberto, postados na rede social Twitter, e o desempenho dessas empresas na bolsa de valores.

Inicialmente o estudo buscou analisar empresas negociadas na Bovespa, mas essa premissa precisou ser ajustada em decorrência das dificuldades enfrentadas na coleta de dados via API do Twitter.

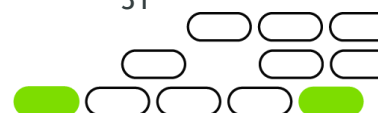
Utilizando um dataset de tweets, disponível no site Zenodo¹, que contém 5.024.756 tweets postados no Twitter entre 6 de dezembro de 2013 e 30 de junho de 2019, foi possível modelar a versão inicial do estudo, sendo necessária a posterior substituição para um dataset mais confiável e relacionado com a matéria.

Portanto, considerando que o dataset utilizado é composto por tweets postados nos EUA, tornou-se necessária a coleta de dados de empresas negociadas na Nasdaq, realizada por meio do pacote yfinance.

Na Sprint 1 foram consolidadas a análise de contexto, as personas, os benefícios, justificativas e as hipóteses do projeto, bem como a definição do objetivo SMART, backlog do produto e o planejamento da execução. Adicionalmente, foram coletados e armazenados os dados da bolsa de valores no Mongo BD.

As atividades desenvolvidas na Sprint 2 consistiram na modulação do projeto em Python, publicação no GitHub, criação das funções de análise de sentimento dos tweets

¹ <<https://zenodo.org/record/5068253>>



e de coleta de dados da Nasdaq, automatização do armazenamento no Mongo DB, seleção das empresas, consolidação do dataset final e as análises preliminares.

Por fim, na Sprint 3, foram analisadas as correlações existentes entre os tweets e o retorno das empresas na Nasdaq e a consolidação dos resultados obtidos.

3.2 Contribuições

Em decorrência das dificuldades encontradas no desenvolvimento do projeto, principalmente no que se refere à coleta de dados via API do Twitter, entende-se que os resultados obtidos neste relatório são inconclusivos.

No entanto, cabe ressaltar que a utilização de um dataset de tweets mais consistente e convergente com a matéria, pode apontar para resultados mais conclusivos, ainda que seja para demonstrar uma inexistência total de correlação entre as variáveis.

3.3 Próximos passos

Permanece a necessidade de obtenção de um dataset mais confiável ou mais correlacionado com o estudo, de forma a verificar eventual utilidade do modelo, sendo o ponto primordial para qualquer continuidade de desenvolvimento do projeto.

Superada a barreira do dataset de tweets, frente aos novos resultados obtidos, podem ser aprofundados os estudos sobre a metodologia para a análise de sentimentos dos tweets e de métricas para apuração de eventual acurácia do modelo.

