

\_\_\_\_\_

PÓS-GRADUAÇÃO

**XP Educação**

**Relatório do Projeto Aplicado**

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

Fernando Guedes de Campos Júnior

Orientador(a): Daniel Viana

31 de julho de 2023

****

**FERNANDO GUEDES DE CAMPOS JÚNIOR**

**XP EDUCAÇÃO**RELATÓRIO DO PROJETO APLICADO

Análise de Sentimento para o Mercado de Ações Brasileiro: 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  
31 de julho de 2023**

**Sumário**

[1. CANVAS do Projeto Aplicado 4](#_Toc144754545)

[1.1 Desafio 5](#_Toc144754546)

[1.1.1 Análise de Contexto 5](#_Toc144754547)

[1.1.2 Personas 7](#_Toc144754548)

[1.1.3 Benefícios e Justificativas 9](#_Toc144754549)

[1.1.4 Hipóteses 10](#_Toc144754550)

[1.2 Solução 12](#_Toc144754551)

[1.2.1 Objetivo SMART 12](#_Toc144754552)

[1.2.2 Premissas e Restrições 12](#_Toc144754553)

[1.2.3 Backlog de Produto 13](#_Toc144754554)

[2. Área de Experimentação 14](#_Toc144754555)

[2.1 Sprint 1 14](#_Toc144754556)

[2.1.1 Solução 14](#_Toc144754557)

[2.1.2 Lições Aprendidas 18](#_Toc144754558)

[2.2 Sprint 2 18](#_Toc144754559)

[2.2.1 Solução 18](#_Toc144754560)

[2.2.2 Lições Aprendidas 23](#_Toc144754561)

## 

## 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.

Uma imagem contendo Interface gráfica do usuário

Descrição gerada automaticamente

Figura 1. Matriz CSD

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

Interface gráfica do usuário

Descrição gerada automaticamente

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.

Site

Descrição gerada automaticamente com confiança média

Figura 3. Mapa de empatia Persona 1

Uma imagem contendo Interface gráfica do usuário

Descrição gerada automaticamente

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 Twiiter |
| **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 Desing 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.

Cd com mensagem de texto

Descrição gerada automaticamente com confiança baixa

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.

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

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).

**Interface gráfica do usuário, Texto

Descrição gerada automaticamente**

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.

Texto

Descrição gerada automaticamente

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.

**Texto

Descrição gerada automaticamente**

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.

**Texto

Descrição gerada automaticamente**

Figura 10. Tentativa de Coleta de Dados do Twitter

**Texto

Descrição gerada automaticamente**

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.

Texto

Descrição gerada automaticamente

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.

**Tela de celular com aplicativo aberto

Descrição gerada automaticamente**

Figura 12. 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 13, que possui como foco principal a organização dos scritps do projeto, a análise de sentimento dos tweets coletados, a exportação dos datasets e as análises iniciais.

Interface gráfica do usuário, Texto

Descrição gerada automaticamente

Figura 13. 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 14 apresenta a atualização do checklist de atividades da Sprint 1, evidenciando a conclusão da respectiva etapa:

Tela de celular com aplicativo aberto

Descrição gerada automaticamente

Figura 14. 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 15 apresenta a inclusão do projeto no GitHub <<https://github.com/guedesf/tweets_market>>

Tela de celular com fundo preto

Descrição gerada automaticamente

Figura 15. Inclusão do projeto no GitHub

A Figura 16 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.

Texto

Descrição gerada automaticamente

Figura 16. Modulação do projeto

A Figura 17 apresenta a função que realiza a análise de sentimento dos tweets contidos no dataset utilizando o pacote TextBlob.

Texto

Descrição gerada automaticamente

Figura 17. 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 |

A Figura 18 evidencia a exportação dos datasets individuais (tweets e Nasdaq) e a consolidação e exportação em .xlsx do dataset para análise para cada empresa.

Texto

Descrição gerada automaticamente com confiança média

Figura 18. 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 19 apresenta evidência da análise inicial de calor dos datasets consolidados para cada empresa:

Tabela

Descrição gerada automaticamente

Figura 19. 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 relizado 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.