RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

USO DA ANÁLISE DE DADOS EM CENÁRIOS EMPRESARIAIS:
Incrementado à Estratégia de Marketing Digital

Eduardo Henrique da Silva
Graduado em Engenharia de Software – Uni-FACEF
edu.slhenrique@gmail.com.br

Prof. Me Geraldo Henrique Neto
Docente – Uni-FACEF
geraldo.henriqueneto@gmail.com

Resumo
Com a crescente utilização de mídias digitais, surge um novo perfil de consumidor, que valoriza a experiência de compra e uso do produto ou serviço além da qualidade do mesmo. Por sua vez, os fornecedores procuram meios de se aproximar deste novo perfil de cliente, criar identificação de sua marca com seu público-alvo, colaborando para que a experiência de compra seja positiva e que quando o *stakeholder* publicar algo sobre a solução em redes sociais, apresente de forma positiva a empresa. Com base nisso, identificamos a oportunidade de fornecer uma solução que, por meio de mineração de dados de redes sociais, colete publicações que citam a organização, utilize de técnicas de ciência de dados e aprendizado de máquina para normalizar e analisar o que foi coletado, fornecendo uma dashboard para que profissionais de marketing e gestão possam avaliar o posicionamento do consumidor quanto a marca bem como colaborar para a identificação de forças e fraquezas. Assim, possibilitamos a tomada de decisões baseada em dados e que a empresa tenha vantagem competitiva quanto aos demais concorrentes que utilizam o modelo tradicional.

Palavras-chave: Ciência de dados, aprendizado de máquina, mineração de texto, tomada de decisões, marketing.

Abstract
With the increasing use of digital media, a new consumer profile has emerged, one that values the buying experience and the use of the product or service, as well as its quality. Suppliers, in turn, are looking for ways to get closer to this new customer profile, to create brand identification with their target audience, helping to ensure that the buying experience is positive and that when the stakeholder posts something about the solution on social networks, it presents the company in a positive way. Based on this, we identified the opportunity to provide a solution that, by means of social network data mining, collects posts that mention the organization, uses data science and machine learning techniques to normalize and analyze what was collected, providing a dashboard so that marketing and management professionals can evaluate the consumer's positioning regarding the brand as well as collaborate to identify strengths and weaknesses. Thus, we enable data-driven decision making and give the company a competitive advantage over other competitors that use the traditional model.

Keywords: *Data science, machine learning, text mining, decision making,* marketing.

1 Introdução

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

O perfil do consumidor mudou com a disseminação da internet e redes sociais. A opinião sobre uma marca ou produto deixou de levar em consideração somente o serviço oferecido, valorizando principalmente a experiência de uso e identificação com a marca.

Conforme a Resultados Digitais (2021), o consumidor 4.0 é pouco fidelizado, busca praticidade e agilidade, trocando facilmente de fornecedor caso outro promova ações que gerem identificação e proximidade com o cliente, tenha problemas com o produto ou serviço consumido atualmente ou levando em consideração avaliações em sites e publicações de redes sociais.

Observando a mudança constante do perfil do consumidor, as empresas precisam criar metodologias de acompanhamento e estratégias comerciais para posicionamento de marca, buscando se destacar do mercado em que está inserido e aumentar a fidelização dos clientes.

Com base nesta necessidade, iremos utilizar de *data mining, data science* e *machine learning* para buscar dados online e fornecer uma solução que permita à empresa uma estratégia de tomada de decisões baseada em dados. Dessa forma, o presente trabalho irá executar a mineração de dados na base do *twitter* e gerar um relatório que forneça dados para um profissional avaliar o posicionamento da marca e as avaliações de consumidores.

A metodologia utilizada foi descritiva e quantitativa, utilizando de um estudo de caso para demonstrar o processo a ser executado para alcançar a solução proposta. A pesquisa esclarece os conceitos de mineração e análise de dados com base em pesquisa bibliográfica e os aplica em um *dataset* gerado a partir de publicações do *twitter*.

Dessa forma, o presente trabalho tem como objetivo geral fornecer um método de acompanhamento das publicações de consumidores, gerando um relatório para profissionais de marketing e gestores, com a finalidade de colaborar com o processo estratégico de tomada de decisão. E como objetivos específicos descrever e demonstrar o processo de coleta, limpeza, padronização, processamento e visualização de dados.

2 Referencial Teórico

Neste tópico serão contextualizados os principais tópicos que envolvem a solução, com o intuito de esclarecer o cenário atual do *marketing* e como o uso de dados pode ser uma vantagem competitiva para uma empresa.

2.1 Big Data

Segundo a Oracle (2021), *big data* pode ser definido com um conjunto de dados maior e complexo coletado especialmente de novas fontes. Este cenário possibilita resolver problemas de negócios que não eram possíveis antes, porém os dados são tão abundantes que os *softwares* tradicionais de processamento não conseguem gerenciá-los.

Conforme Petry (2013), essa imensidão tem uma pequena parte formada por dados estruturados — limpos, corretos. E a outra parte, essa maior

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

considerada o motivo de o *big data* ser transformador, formada pelos dados não estruturados — incompletos e caóticos.

O *big data* abrange 5 V's: variedade, velocidade, volume, valor e veracidade. Conforme explica Marr (2014):

- O volume se refere à quantidade de informações produzidas em e-mails, mensagens, comentários etc.
- A velocidade está relacionada com a rapidez que as informações são geradas e analisadas sem a necessidade de armazenamento prévio.
- A variedade diz respeito aos diversos tipos de dados que podem ser coletados simultaneamente, os estruturados e não estruturados.
- A veracidade está associada ao acerto das ideias geradas através dos dados.
- O valor, é a utilidade e os seus benefícios para as empresas.

O aumento do uso da tecnologia faz com que o número de dados se expanda exponencialmente, conforme Carter (2021) afirma-se que em 2025 serão gerados mais de 463 exabytes de dados. Nesse contexto, com essa enorme quantidade de informações em grande parte não estruturadas, a exploração da área de *big data* por profissionais de tecnologia e empresas está cada vez maior.

Mas o que possibilita com que todos esses dados sejam armazenados, processados e analisados? Conforme Petry (2013), o custo de armazenar dados caiu de forma notável, 1 gigabyte foi de 1000 dólares na década passada para 6 centavos em 2013. Paralelamente o poder de processamento aumenta próximo ao previsto na Lei de Moore — dobra a cada 18 meses.

Este baixo custo de armazenamento aliado ao alto poder de processamento e aos dados abundantes fazem com que a exploração e análise de dados ganhe espaço. Conforme Papazissis (2018), o *big data* possibilita transformar os negócios tradicionais e oferece vantagem competitiva a quem o utilize como insumo na tomada de decisões.

2.2 Text Mining

Conforme Loh(2014), estima-se que 80% das informações de uma empresa são textuais, espalhadas nas diversas fontes tais como: *emails*, *blogs*, redes sociais, arquivos eletrônicos entre outros. A grande quantidade de dados, aliada a complexidade das informações neste formato impossibilita a análise manual.

2.2.1. Definição

Text Mining ou Mineração de Texto ou *Knowledge Discovery in Text* (KDT) pode ser considerado uma extensão do *data mining* - descoberta de padrões em bases de dados. *C*onsiste na extração de informação implícita e útil por meio de aplicação de estatística em textos, frases ou até mesmo somente palavras. Sendo considerada uma evolução da área de recuperação de informações por possibilitar a obtenção de ideias implícitas que os métodos tradicionais não permitem extrair. (LOH, 2014).

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

2.2.2. Etapas

O processo de mineração de texto é resumido nas etapas representadas na Figura 1, explicados conforme (JUNIOR, 2007):

- coleta: antecede as etapas de *Text Mining* e consiste em criar um *Corpus* que sirva como insumo para a aquisição de conhecimento;
- pré-processamento: por se tratar de dados não estruturados, e por poder levar a problemas de interpretação por conta de um mesmo assunto ser relatado com diferentes palavras (sinônimos, variações linguísticas) este tipo de técnica necessita de um processamento prévio antes de se aplicar estatística. Nessa etapa os dados são trabalhados, transformados e organizados para melhorar sua qualidade e possibilitar análise;
- indexação: encontra os textos que podem conter a informação desejada, os organiza de forma a facilitar o seu acesso, similar ao índice de um livro, visa organizar os dados para mapear onde se encontra cada parte;
- mineração: são aplicadas técnicas de *Text Mining* e também técnicas de *Data Mining* tradicionais sobre os dados pré-processados, tais como classificação, *clustering*, associação entre outras.
- pós-processamento: análise dos resultados obtidos na última etapa, realizadas por um ser humano utilizando de ferramentas de análise para geração de gráficos.

2.2.3. Aplicabilidades

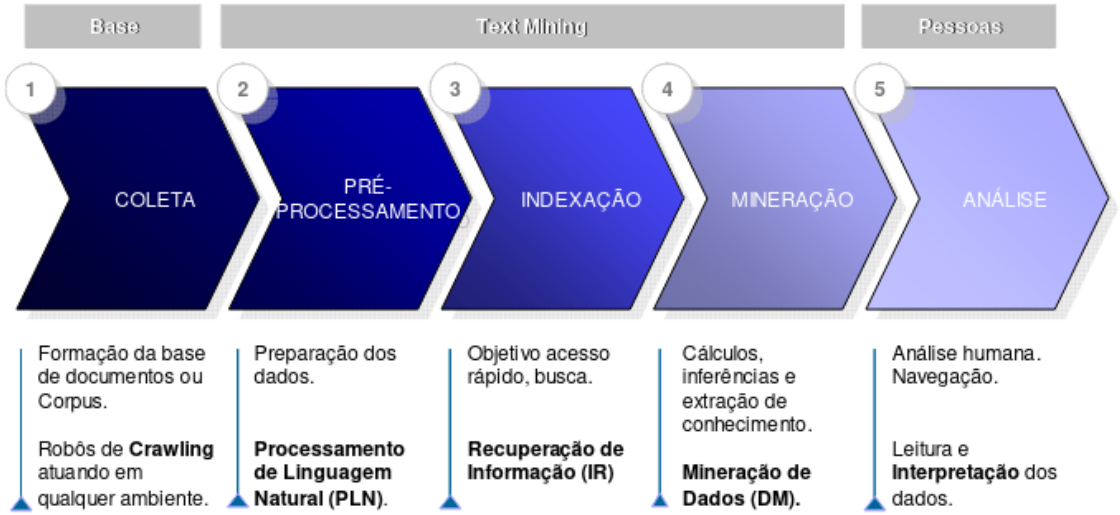
O *Text Mining* vem sendo utilizado por diversas empresas, que fornecem *software* como serviço ou do inglês *Software as a Service* (SaaS). Por exemplo:

Utilizando da técnica de *Webibliomining* - método de definição de um referencial bibliográfico inicial para pesquisas por meio de palavras chaves (COSTA, 2010). O site Medline utiliza essa técnica para fornecer dados referentes a um determinado assunto na área médica, sendo útil para contextualização e referências em novos artigos e pesquisas na área.

Utilizando de análise de sentimentos ou mineração de opinião, a empresa canadense *Keatext AI* fornece serviço de *text mining* por meio de integração por API (*Application Programming Interface*) apresentando como principais funcionalidades: análise de sentimentos em *feedbacks* e comentários de clientes, abertura de chamado para verificação em caso de críticas.

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

Figura 1 - Etapas do processo de Text Mining



Fonte: Junior (2007)

2.3 Aprendizado de Máquina

Algoritmos são uma sequência de passos finita com o objetivo de oferecer uma resposta para um determinado problema, ou seja, é constituído de uma entrada, o processamento e uma saída. Tradicionalmente, para um computador realizar uma ação, é necessário que um programador escreva um algoritmo detalhado e o implemente em uma linguagem de programação (DOMINGOS, 2017).

O processo de aprendizado de máquina ocorre quando são fornecidas somente as entradas e saídas esperadas para o computador, sendo o resultado o algoritmo que transforma um em outro. Para isso, é necessário dados de aprendizado, quanto mais dados fornecidos no treinamento maior a tendência do algoritmo gerado ter acerto. (DOMINGOS, 2017)

Segundo a IBM (2020), *machine learning* é um ramo da inteligência artificial que visa utilizar dados para simular o aprendizado humano melhorando gradualmente sua precisão. O principal benefício é a capacidade de lidar com problemas complexos que um programador teria dificuldades.

Aliar o aprendizado de máquina as técnicas de ciência de dados é essencial para empresas que procuram um diferencial competitivo.(IBM, 2020)

> Machine learning é um componente importante do crescente campo da ciência de dados. Por meio do uso de métodos estatísticos, os algoritmos são treinados para fazer classificações ou previsões, revelando os principais insights em projetos de mineração de dados. Esses insights subsequentemente conduzem a tomada de decisões em aplicativos e negócios, impactando de forma ideal as principais métricas de crescimento. Conforme o big data continua a se expandir e crescer, a demanda do mercado por cientistas de dados aumentará, exigindo que eles auxiliem na identificação das questões de negócios mais relevantes e, posteriormente, os dados para respondê-las. (IBM, 2020)

2.4 Tomada de Decisões

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

"[...] Tomar decisões é identificar e selecionar um curso de ação para lidar com um problema específico ou extrair vantagens em uma oportunidade. [...]"(SILVA, 2015).

O processo de tomada de decisão utilizado tradicionalmente se caracteriza pela intuição e experiência dos decisores, e ao longo do tempo vem demonstrando imperfeições por diversos motivos sendo os principais: informações incompletas, más ferramentas de decisão e tendências irracionais. (CAMPOS, 2018)

2.4.1. Inteligência de Negócios

O termo inteligência de negócios ou do inglês *Business Intelligence* (BI) surgiu na década de 50 sendo utilizado por Hans Peter Luhn, em um artigo intitulado "*A business intelligence system*" onde era proposto um sistema automático para indexar, processar e disseminar informações. (RIGO BOTELHO; FILHO, 2014).

"[...] pode-se dizer que Inteligência de Negócios é o nome que se atribui a iniciativas para analisar informações complexas de um contexto organizacional e de apresentar o resultado de tal análise de maneira sintetizada e simples, que facilite a interpretação pelo(s) gestor(es)." (SILVA, 2015).

Atualmente, o uso de ferramentas de BI tem como objetivo principal agilizar o processo de tomada de decisão, possibilitando a obtenção de melhores resultados, facilitando a transformação dos dados e que relatórios e análises sejam gerados com maior qualidade e rapidez, guiando o time a ser mais eficiente (MARA, 2020).

2.4.2. Data Driven Decision Making

Conforme Gutierrez (1999), o desenvolvimento tecnológico acelerado, aliado a fatores políticos e sociais, faz com que as empresas busquem um novo fluxo na tomada de decisão, onde é preciso lidar com dados que até pouco tempo atrás não eram relevantes.

A tomada de decisões com base em dados, ou do inglês *data-driven decision making* (DDDM), é o uso de métricas e dados para orientação de decisões conforme objetivos da empresa. O seu processo é definido em quatro passos por Silva (2015):

- Aquisição de dados - processo de obtenção de dados por meio de formulários *online*, consultas a base de dados, monitoramento em tempo real do contexto entre outras;
- Mineração de dados - processo de descoberta do conhecimento onde é feita a análise dos dados obtidos. Nesta fase são aplicadas técnicas como classificação, agrupamento, associação entre outras;
- Consolidação e comunicação de dados - utilização de recursos de visualização de dados como *dashboards* e *heatmaps* para resumir os dados em artefatos que possam facilitar o processo decisório;
- Tomada de decisão - processo de tomada de decisão levando em consideração os três processos anteriores: os dados brutos, o conhecimento obtido através da mineração e com os artefatos gerados.

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

Apesar das empresas reconhecerem as vantagens da tomada de decisão baseada em dados, a aplicação de DDDM não é sempre executada, seja por falta de tempo, informação ou outros recursos. Essa técnica não isenta os riscos ao tomar uma decisão, entretanto auxilia para otimizar e auxiliar a produzir maiores consequências positivas e reduzir as negativas (SILVA, 2015).

2.5 Digital Marketing Analytics

O grande volume de dados disponível aliados à evolução tecnológica em um cenário de grande competitividade torna essencial que empresas utilizem de novas técnicas para a tomada de decisão para que consigam se destacar.

"[...] em ambientes onde a competição é mais acirrada, aqueles que possuem uma grande quantidade de informação em mãos para a tomada de decisão provavelmente se sobressaíram" (POPOVIČ et al., 2012).

2.5.1. Definição

Conforme Proença (2021) o *Digital Marketing Analytics* (DMA) é a intersecção entre:

- *Digital Analytics* - Aquisição de conhecimento sobre os consumidores na internet que, diferente do *web analytics*, considera todas as atividades *online* deles (sites, mídias digitais, e-mail e etc.).
- *Marketing Analytics* - Aquisição de conhecimento sobre o comportamento dos consumidores (*online* e *offline*) para a melhoria das atividades de marketing.

Ou seja, podemos definir o DMA como a obtenção de conhecimento de consumidores de forma *online* para o aprimoramento de estratégias de *marketing*.

2.5.2. Benefícios

Segundo Proença (2021), está técnica colabora com os pilares da tomada de decisão em *marketing*:

- Conhecimento do consumidor - é importante conhecer o seu consumidor demográfica e psicologicamente. Dessa forma, as técnicas de mineração de dados tais como o *Text Mining* colaboram para análise principalmente quando se tem um alto volume de dados.
- Conhecimento específico da empresa - é o conhecimento prévio já obtido pela empresa. Nesse aspecto o DMA colabora com a análise de ações anteriores da organização, uma vez que permite a análise independente se estava sendo aplicado ou não a coleta e análise de dados no momento de implementação de uma estratégia.
- Conhecimento de mercado - é obter informação do comportamento do segmento da organização. Por exemplo, é possível utilizar a mineração e análise para identificar eventos externos que possam alterar significativamente o fluxo de consumidores ou verificar a concorrência e mapear quais estratégias de *marketing* tem funcionado ou não.

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

- Conhecimento da rede em que a organização está inserida - é o conhecimento da rede em que a organização está inserida, incluindo parceiros.

Entretanto, além da utilização da DMA para conseguir uma vantagem competitiva, as empresas dependem da capacidade absortiva da organização.

> Uma crise, por exemplo, pode significar uma oportunidade de aprendizado para uma organização que consegue ter um estímulo para inovar, ou a falência para outra, que não consegue extrair os aprendizados e se adaptar.
> (KIM,, 1998)

2.6 Python

O Python é uma linguagem de programação criada em 1991 por Guido van Rossum recebeu esse nome por conta de um grupo de comédia chamado *Monty Python* do qual o seu criador era fã. O projeto foi liderado até 2018 por Guido van Rossum e de 2019 até atualmente é mantido pelo conselho da Python *Software Foundation* (PSF) - organização sem fins lucrativos criada em 2001 pelo próprio criador do Python para mantê-lo.

O objetivo da linguagem é fornecer produtividade e legibilidade. Comparando com outras linguagens de programação a dificuldade e o tempo de aprendizado do Python é menor, permitindo que a mesma seja utilizada sem tanto conhecimento de programação (*Insight Lab,* 2019).

Conforme o *Insight Lab* (2019) é a linguagem mais utilizada para fins de análise por ser poderosa e fácil de usar, ter diversas opções de bibliotecas, ser escalável e por fornecer diversas formas de visualização de dados por meio das suas bibliotecas.

Conforme a *PyScience* Brasil algumas das características da linguagem que destacam os seus propósitos são:

- baixo uso de caracteres especiais, o que torna a linguagem muito parecida com pseudocódigo executável;
- o uso de indentação para marcar blocos;
- quase nenhum uso de palavras-chave voltadas para a compilação;
- coletor de lixo para gerenciar automaticamente o uso da memória.
- multiplataforma,
- multiparadigma, permitindo a utilização de programação funcional, procedural e orientada a objetos.

3 Coleta e Análise de Dados

Nos tópicos a seguir, serão apresentadas as técnicas utilizadas para implementação da solução proposta aplicando o *text mining* e *data mining* a partir da base de dados do *Twitter* e utilizando como filtro a empresa *IFood* - aplicativo de *delivery* de entrega sobre demanda.

3.1 Coleta de Dados

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

Para a coleta de dados foi utilizada a biblioteca *tweepy*, uma facilitadora para as chamadas da API (*Application Programming Interface*) do Twitter. Para evitar de ter que fazer múltiplas requisições ao gerar o *data frame* a partir do retorno da API, os dados foram armazenados em um arquivo CSV (*Comma Separated Values*).

O termo de pesquisa definido foi o próprio nome da empresa IFood filtrando por conteúdo em português. Foi limitado as buscas até exceder 10.000 registros e a partir da biblioteca foi possível buscar por métricas públicas de cada postagem, bem como dados do usuário.

As Figura 2 e Figura 3, demonstram o código fonte responsável por realizar as requisições até a limitação de registros, preencher o *data frame* e gerar o CSV correspondente.

Todo o código fonte estará disponível no GITHUB, referenciado no final do artigo.

Figura 2 - Trecho de geração de data frame

```
48  def generate_data_frame(query_lang: str, limit: int = 10000) -> pd.DataFrame:
49      """
50          Retorna um dataframe com base na pesquisa informada e o limite de registros,
51          para evitar que tenha que        ser executado diversas vezes é armazenado
52          um .csv do dataframe
53          :param query_lang: str
54          :param limit: int por padrão 10000
55          :return: df: pandas.DataFrame
56      """
57
58      end_time: datetime = datetime.now()
59      data: list[dict] = []
60      response: tweepy.client.Response = get_data(query_lang, end_time)
61      users: list = response.includes['users']
```

Fonte: Autores (2022)

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

Figura 3 - Continuação de trecho de geração de data frame

```
while len(data) <= limit:
    for user, tweet in zip(users, response.data):
        row: dict = {
            "id_tweet": tweet.id,
            "user": user.name,
            "text": tweet.text,
            "created_at": tweet.created_at,
            "retweet_count": tweet.public_metrics["retweet_count"],
            "reply_count": tweet.public_metrics["reply_count"],
            "like_count": tweet.public_metrics["like_count"],
            "quote_count": tweet.public_metrics["quote_count"],
        }
        data.append(row)

    try:
        response = get_data(query_lang, end_time, response.meta["next_token"])
        users = response.includes['users']
    except KeyError:
        break

df: pd.DataFrame = pd.DataFrame(data)
df.to_csv(f"csvs/{query_lang.split(' ')[0]}_{end_time}.csv", ",")
return df
```

Fonte: Autores (2022)

3.1.1. Dicionário de Dados

Com o objetivo de esclarecer e facilitar o entendimento do *data frame* gerado, logo a seguir está o dicionário de dados, representado na Tabela 1, com as colunas, tipos e descrição dos campos.

Quadro 1 - Dicionário de dados

Coluna	Tipo	Descrição
id	int64	Identificador único auto-incrementável
user	object	Nome do usuário que executou a publicação
id_tweet	int64	Identificador de publicação
text	object	Conteúdo textual da postagem
created_at	object	Data da publicação
retweet_count	int64	Quantidade de compartilhamentos, não incluídos os com novos comentários.
reply_count	int64	Quantidade de comentários

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

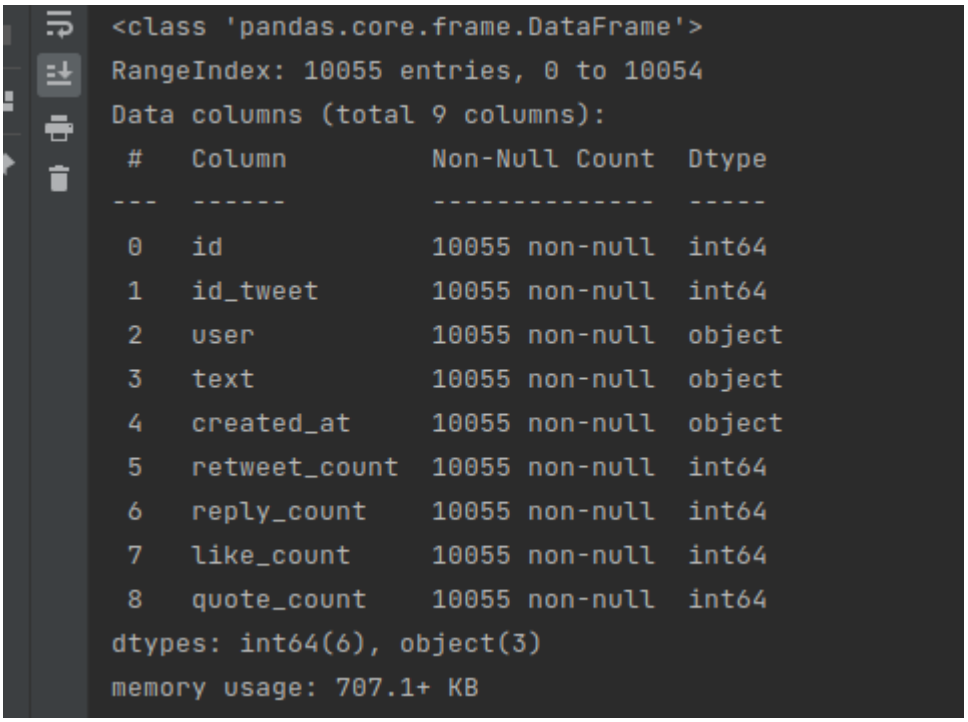
like_count	int64	Quantidades de gostei na publicação
quote_count	int64	Quantidade de compartilhamentos com novos comentários

Fonte: Autores (2022)

3.2 Pré Processamento

Utilizando da função "*.info*" da biblioteca Pandas, é possível obter dados sobre as colunas do *data frame* gerado anteriormente, como visualizado na Figura 4.

Figura 4 - Retorno da função .info



Fonte: Autores (2022)

Nota-se que foi gerado uma coluna chamada id que conforme o dicionário de dados é auto incrementável, por isso tal coluna pode ser desconsiderada do processo de análise de dados, podendo ser removida.

Além disso, foi possível perceber que o *data frame* não possui valores nulos, dessa forma dispensa esse tipo de pré-processamento.

3.2.1. Remoção de Pontuação

Por se tratar de dados não estruturados e sem padronização, os textos necessitam de ser convertidos para minúsculos e, também, da remoção das pontuações e caracteres especiais como forma de deixar a comparação entre eles mais assertiva. Para isso foi utilizada a função 'apply' passando como *callback* funções lambda, conforme linhas 115 a 124 da Figura 5.

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

3.2.2. Remoção de Dados Indesejados

Inicialmente é necessário verificar se existem valores duplicados na coluna 'text' do *data frame*. Utilizando da função '*value_counts()*' foi verificado que no *data frame* há diversas repetições, quando ocorre o compartilhamento da publicação sem novos comentários é gerado uma nova porém com o prefixo 'rt', mudando o usuário e data, preservando os demais dados.

A remoção desses dados é necessária para que não afete os resultados de análise, como por exemplo somas dos campos discretos ou então duplicações na análise de texto.

Dessa forma, foi mantida somente a publicação original e removidas as com o prefixo de compartilhamento. Para isso foi utilizada a função '*drop_duplicates()*' passando como parâmetro a coluna '*text*' do *data frame*.

Além disso, a coluna *id* também foi removida, conforme Figura 5, por não agregar no processo de análise. Com isso, ao utilizar a função '*shape()*' para visualizar as dimensões do *data frame*, passamos de 10.055 linhas e 9 colunas para 8.018 linhas e 8 colunas.

Figura 5 - Padronização e remoção de dados

```
109 def pre_processing(data_frame: pd.DataFrame) -> pd.DataFrame:
110     """
111     Recebe um dateFrame e o retorna ele mesmo pré-processado
112     :param data_frame: pandas.DataFrame
113     :return: pandas.DataFrame
114     """
115     df: pd.DataFrame = data_frame
116
117     # converte todas as letras para minúsculo
118     df["text"] = df["text"].apply(lambda x: x.lower())
119
120     # remove números e caracteres especiais
121     df["text"] = df["text"].apply(lambda x: re.sub('[0-9]|,|\.|/|\(|\)|-|\+|:|·|\$', ' ', x))
122
123     # remove acentos
124     df["text"] = df["text"].apply(
125         lambda x: ' '.join([word for word in x.split() if word not in string.punctuation]))
126
127     # remove textos duplicados
128     df = df.drop_duplicates(subset=["text"])
129
130     # remove coluna id
131     df.drop(df.columns[df.columns.str.contains('unnamed', case=False)], axis=1)
132
133     return df
```

Fonte: Autores (2022)

3.2.3. Tokenização

A tokenização é o processo de gerar uma lista de palavras onde cada uma é representada por um *token*, visando normalizar um texto para possibilitar análise morfológica.

Para esse processo foi utilizado a biblioteca NLTK que fornece a classe *RegexpTokenizer,* conforme Figura 6. Para isso, é informado a expressão regular

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

(*regex)* com os delimitadores do conteúdo dos tokens na instância da classe, conforme linha 142, foi considerado qualquer *substring* iniciada com uma letra, seguida de qualquer caractere alfanumérico incluindo o *underscore (‘_’)*, após isso o texto é passado como parâmetro para o método *‘tokenize’* responsável por gerar a lista de palavras sem repetições com base no *regex* definido.

Figura 6 -Tokenização

```
136  def tokenize(df: pd.DataFrame) -> list[str]:
137      """
138      Retorna coluna 'text' do dataframe tokenizada em uma lista de palavras
139      :param df: pd.DataFrame
140      :return: list[str]
141      """
142      tokenizer: RegexpTokenizer = RegexpTokenizer(r'[A-z]\w*')
143      tokens: list[str] = tokenizer.tokenize(df['text'].to_string())
144      return tokens
```

Fonte: Autores (2022)

3.2.4. Stopwords

As *stopwords* são palavras consideradas irrelevantes, sem valor para a análise, como por exemplos os artigos: a, o, um etc. e preposições: de, por, para entre outros.

A remoção desses dados é essencial pois são as palavras que normalmente são mais utilizadas em qualquer texto, dessa forma removê-las reduz a distorção dos dados analisados.

Foi implementado uma função que importa a lista de *stopwords* do idioma português a partir da biblioteca NLTK, adicionando a sigla ‘*rt’*, a preposição informal ‘pra’, e o termo de pesquisa utilizado inicialmente para buscar os dados, conforme Figura 7 e Figura 8, uma vez que essas palavras são constantes nas publicações e irrelevantes para análise. Em seguida essa função foi utilizada para limpar a lista de *tokens* gerada no tópico anterior.

Figura 7 - Remoção de stopwords

```
178  def remove_stopwords(list_words: list[str], query_lang: str) -> list[str]:
179      """
180      Retorna lista de palavras sem as stopwords
181      :param list_words: list[str]
182      :param query_lang: str
183      :return: list[str]
184      """
185      tokens_without_sw: list[str] = [word for word in list_words
186                                      if word not in get_stopwords(query_lang)]
187      return tokens_without_sw
```

Fonte: Autores (2022)

Figura 8 - Remoção de stopwords

```
190  def get_stopwords(query_lang: str) -> list[str]:
191      """
192      Retorna lista de stopwords incluindo o termo de pesquisa
193      :param query_lang: str
194      :return: list[str]
195      """
196      stop_words: list[str] = stopwords.words('portuguese')
197      stop_words.append('rt')
198      stop_words.append('pra')
199      stop_words.append('https')
200      stop_words.append(query_lang.split(' ')[0])
201      return stop_words
```

Fonte: Autores (2022)

3.2.5. Lematização

Visando padronizar as palavras e ter maior assertividade no processo de análise realizamos o processo de lematização, que consiste em agrupar palavras com o mesmo sentido em somente uma, identificando variações de palavras e retornando o seu lema - normalmente a palavra no infinitivo, porém não é uma regra.

Para realização desse processo foi utilizada a biblioteca *spacy* que fornece a função para as palavras em português, conforme Figura 9 dessa forma aplicamos o processo de lematização na lista de *tokens*, já trabalhada nos tópicos anteriores.

Figura 9 - Lematização

```
204  def generate_lemma(list_words: list[str]) -> list[str]:
205      """
206      Recebe uma lista de palavras e retorna nova lista, unindo as palavras
207      similares com o seu lemma
208      :param list_words:list[str]
209      :return: list[str]
210      """
211      list_lemma: list[str] = []
212      for word in nlp(' '.join(list_words)):
213          list_lemma.append(word.lemma_)
214      return list_lemma
```

Fonte: Autores (2022)

3.3 Frequência de Termos

Após as etapas de pré-processamento de dados, foi gerado uma matriz de frequência com as palavras. Utilizamos a função '*FreqDist*' do NLTK para calcular

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

a frequência dos termos a partir da lista de *tokens*, com isso geramos um *data frame* com as palavras e respectivas frequências.

Também foram calculadas as obtenções dos atributos por palavra e concatenado em um único *data frame*, gerando o exposto na Figura 10.

Figura 10 - 30 maiores frequências

	word	freq	likes	rt	reply	quotes
0	pedir	1001	6646.00000	1245.00000	1138.00000	113.00000
1	fazer	380	2359.00000	1938.00000	465.00000	33.00000
2	pagar	312	6118.00000	2561.00000	258.00000	77.00000
3	querer	283	1128.00000	292.00000	277.00000	15.00000
4	cupom	234	320.00000	46.00000	95.00000	5.00000
5	tarar	233	28.00000	0.00000	57.00000	10.00000
6	boleto	220	28.00000	0.00000	57.00000	10.00000
7	comprar	219	1162.00000	56.00000	228.00000	16.00000
8	mandar	192	850.00000	54.00000	177.00000	15.00000
9	achar	162	8954.00000	1469.00000	306.00000	78.00000
10	aqui	158	2658.00000	7497.00000	490.00000	926.00000
11	hoje	155	1149.00000	10296.000...	131.00000	7.00000
12	dar	155	1287.00000	779.00000	256.00000	34.00000
13	dia	155	1629.00000	6782.00000	163.00000	13.00000
14	https	154	8515.00000	3249.00000	1021.00000	1045.00000
15	colocar	150	819.00000	49.00000	393.00000	927.00000
16	entregar	137	3932.00000	2297.00000	283.00000	44.00000
17	entregador	135	8708.00000	6100.00000	244.00000	86.00000
18	vir	130	1705.00000	12132.000...	477.00000	925.00000
19	caro	130	5320.00000	527.00000	525.00000	949.00000
20	todo	129	717.00000	137.00000	153.00000	14.00000
21	lanchar	128	949.00000	65.00000	93.00000	12.00000
22	ver	128	810.00000	7385.00000	141.00000	10.00000
23	real	126	738.00000	202.00000	168.00000	12.00000
24	comer	122	2434.00000	12438.000...	353.00000	21.00000
25	descontar	118	213.00000	17.00000	25.00000	2.00000
26	ficar	118	1146.00000	484.00000	163.00000	16.00000
27	comido	116	392.00000	86.00000	115.00000	7.00000
28	galodeluta	116	0.00000	0.00000	0.00000	0.00000
29	bom	115	438.00000	61.00000	110.00000	5.00000
30	chegar	114	631.00000	233.00000	206.00000	12.00000

Fonte: Autores (2022)

3.4 Nuvem de Palavras

Após a geração da matriz com as frequências, geramos as respectivas nuvens de palavras para os atributos, possibilitando assim melhor visualização de quais palavras são mais curtidas, geram mais compartilhamentos, comentários e citações. As nuvens de palavras visam fornecer melhor visualização da relação de atributo/frequência, deixando as palavras com maior frequência em destaque. Na Figura 11 e Figura 12 são expostas utilizando os atributos de RT e Likes.

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

Figura 11 - Nuvem de palavras vs RT

Fonte: Autores (2022)

Figura 12 - Nuvem de palavras vs Likes

Fonte: Autores (2022)

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

3.5 Análise de Sentimentos

Para o processo de análise de sentimentos utilizamos um *dataset* de comentários da IMDB (*Internet Movie Database*), disponível no Kaggle e referenciado no fim do artigo, para treinamento e testes de um classificador. Quanto à estrutura, o *dataset* possui o texto original em inglês, o texto traduzido para português e a classe a que pertence, podendo ser 'neg'(negativo) ou 'pos' (positivo).

O classificador utiliza o algoritmo *Naive Bayes*, que se baseia no teorema de Bayes, fórmula matemática para definir a probabilidade de algo acontecer, com base em conhecimento anterior. (MANNING et al., 2009)

Para a classificação de texto, o algoritmo fornece duas formas de aprendizado, conforme MANNING et al.(2009):

- *Multinomial Naive Bayes* - trabalha com a probabilidade de uma palavra ocorrer em uma frase de determinada classe, levando em consideração a quantidade de ocorrências.
- *Multivariate Bernoulli Model* - difere do anterior pois trabalha somente com valores booleanos, não levando em consideração a quantidade de ocorrências em uma mesma frase.

A biblioteca *sklearn* fornece a implementação do algoritmo, dessa forma, a importamos para uso e testamos a precisão de ambos os algoritmos. Conforme a linha 20 da Figura 13, a base foi dividida na proporção de 90, 10 para o treinamento e testes respectivamente.

Na Figura 13, a técnica de classificação a ser utilizada pelo algoritmo é definida na linha 17 e as métricas obtidas são exibidas no print da linha 22. Ao utilizar o *MultinomialNB* foi atingido a precisão de 81% para comentários negativos e 87% para positivos, já utilizando o *BernoulliNB* a precisão subiu para 83% e 88% respectivamente. Por isso, a utilização do *BernoulliNB* foi mantida e os dados de treinamento salvos em arquivos com a biblioteca *pickle* que aplica a serialização no objeto.

Para a aplicação do classificador no dataset foi necessário a desserialização dos objetos salvos anteriormente. Dessa forma, adicionamos uma nova coluna no *dataset* com a classificação de cada tweet e obtivemos 4704 (58,67%) de comentários positivos e 2705 (41,33%) de negativos.

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

Figura 13 - Geração de classificador de sentimentos

```python
import pandas as pd
from sklearn import metrics
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import cross_val_predict
from sklearn.naive_bayes import BernoulliNB
import pickle


dataset: pd.DataFrame = pd.read_csv('csvs/treino-sentimentos.csv', encoding='utf-8')
dataset['text_pt'] = dataset['text_pt'].fillna("")
dataset['sentiment'] = dataset['sentiment'].fillna("")
text: list = dataset['text_pt'].values

labels: list = dataset['sentiment'].values
vectorizer: CountVectorizer = CountVectorizer(analyzer="word", ngram_range=(1, 2))
freq: list = vectorizer.fit_transform(text)
model: BernoulliNB = BernoulliNB()
model.fit(freq, labels)

results: list = cross_val_predict(model, freq, labels, cv=10)
metrics.accuracy_score(labels, results)
print(metrics.classification_report(labels, results))

with open('vector.pkl', 'wb') as file:
    pickle.dump(vectorizer, file)

with open('classifier.pkl', 'wb') as file:
    pickle.dump(model, file)
```

Fonte: Autores (2022)

4 Resultados da Solução Proposta

Nos tópicos a seguir, serão apresentados os resultados obtidos após a coleta e análise de dados. Visando facilitar a visualização e interpretação para os *stakeholders* bem como descrever os *insights* adquiridos no estudo de caso executado com a empresa *IFood*.

4.1 Visualização de Dados

Na etapa de visualização de dados, foi utilizada a biblioteca *PyDash,* na qual é baseada na *LoDash* do *JavaScript* e utiliza do *micro framework* do Python, *Flask* para a criação de páginas web.

A visualização de dados foi composta de duas páginas, conforme Figura 14 e Figura 15. A primeira tela (Figura 14) tem o objetivo de facilitar a interpretação do *stakeholder*, agrupando os dados mais relevantes para a tomada de decisão como os totais gerais, as *word clouds* e também fornece ao mesmo a possibilidade de inserir manualmente uma frase ou publicação para ser classificada.

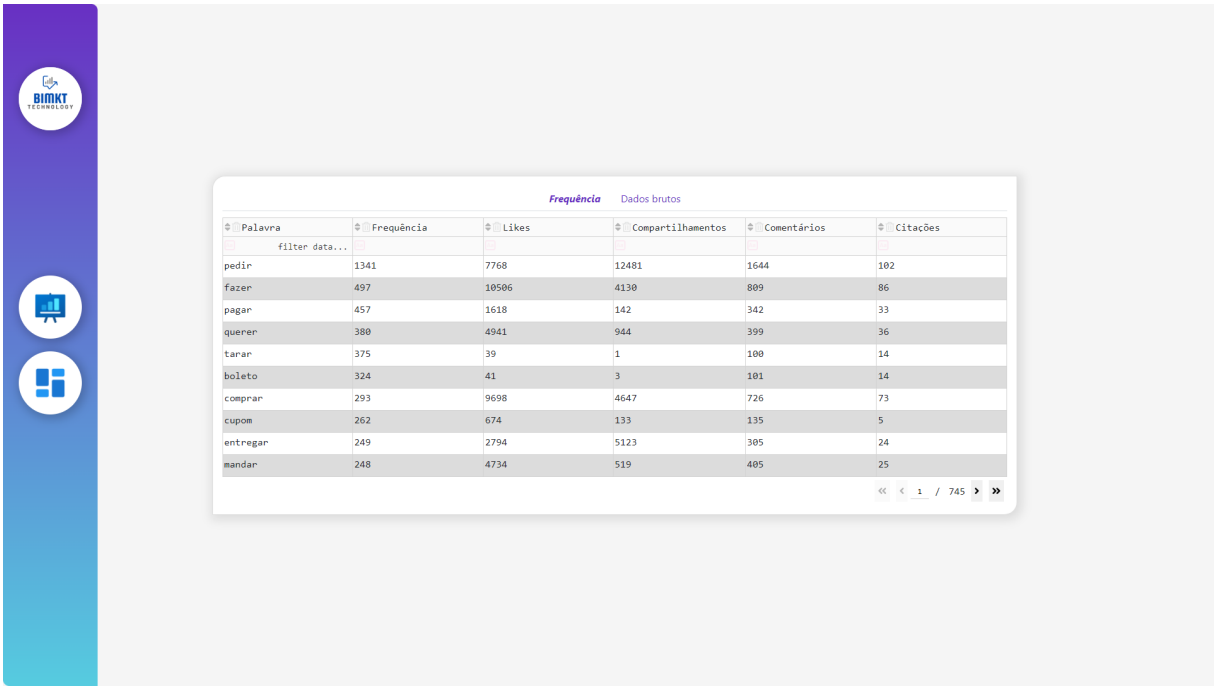
Já a segunda tela (Figura 15), permite ao usuário uma visualização mais ampla dos dados minerados e analisados, por isso agrupamos as tabelas de frequência e a de dados brutos, permitindo ao usuário filtrar como desejar.

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

Figura 14 - Home da dashboard

Fonte: Autores (2022)

Figura 15 - Tabelas na dashboard



Fonte: Autores (2022)

4.2 Insights Obtidos

Em análise dos resultados obtidos com a o processo realizado utilizando a empresa *IFood* destaca-se os seguintes pontos:

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

- A porcentagem de comentários negativos é alta, filtrando nota-se que que tal classificação está correlacionada principalmente com a falta de suporte em reembolsos e cancelamentos, bem como problemas na entrega;
- Nota-se que tem um alto número de compartilhamentos, comparando com os outros índices;
- As palavras entrega e entregador são constantes nas *word clouds,* demonstrando a importância desse processo para o sucesso do app;
- A matriz de frequências e *word clouds* demonstram que as palavras cupom, valor, boleto são frequentes. Ao buscar por elas na base, é perceptível que valor e boleto estão relacionado a muitas postagens relatando os altos gastos por comprar muito no aplicativo, enquanto há muita publicidade com a palavra cupom;
- Filtrando pelos dados positivos percebe-se que o uso do aplicativo para supermercado e farmácias vem crescendo e é destacado como algo positivo pelos usuários;
- Filtrando pelos dados negativos é perceptível a insatisfação dos usuários com a funcionalidade de pedido mínimo para cupons.

5 Conclusão

Os resultados obtidos analisando as publicações do *twitter* que citam o aplicativo de delivery *IFood*, validam como o processo de análise de dados pode contribuir para conhecer melhor o consumidor, verificar como a marca é vista, apurar repercussão de campanhas e fatores externos, bem como colabora para identificar oportunidades, forças, fraquezas e ameaças.

Por isso, conclui-se que a coleta e análise de dados provenientes de redes sociais é essencial para as empresas que buscam vantagem competitiva sobre os concorrentes, possibilitando estratégias e métricas orientadas a dados.

5.1 Implementações Futuras

Nos tópicos a seguir, será exposto melhorias da solução proposta com a finalidade de torná-la uma aplicação comercial.

5.1.1. Melhoria do Classificador de Sentimentos

Realizar melhoria no classificador, possibilitando que mais sentimentos sejam identificados, tais como raiva, surpresa, decepção etc. Para isso, seria necessário mudar a base de treinamento e verificar se o algoritmo implementado continua com uma precisão aceitável.

5.1.2. Interface Gráfica para Cadastro

Desenvolver telas de cadastro de empresa e usuários, possibilitando que cada usuário possua vínculo com uma ou mais empresas. Com isso, os termos utilizados na etapa de mineração tornam-se parametrizáveis e através do login é

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

implementado uma camada de segurança para não deixar qualquer pessoa ter acesso às informações.

5.1.3. Mineração de Dados Constante

Utilizar de um *SGDB* (Sistema de Gerenciamento de Banco de Dados) para persistir as informações, possibilitando ao *stakeholder* visualizar dados conforme período desejado, bem como criar um histórico para facilitar as métricas de progresso.

Referências

CARTER, R. *A lista definitiva de estatísticas de Big Data para 2022*. Find Stack, 5 dez. 2021. Disponível em: <https://findstack.com/pt/big-data-statistics/>. Acesso em: 20 fev. 2022.

CHOO, C. W. *A organização do conhecimento: como as organizações usam a informação para criar significado, construir conhecimento e tomar decisões*. 2 ed. São Paulo : SENAC São Paulo, 2006.

COSTA, Helder. *Modelo para webibliomining: proposta e caso de aplicação*. Rev. FAE, Curitiba, v13, n.1, p. 115-126, jan./jun.2010.

CAMPOS, Carlos Rogério de Rezende Campos. *R. MESTRADO GESTÃO E ESTRATÉGIA INDUSTRIAL TRABALHO FINAL DE MESTRADO DISSERTAÇÃO A IMPORTÂNCIA DO BIG DATA E DO CRM PARA O GESTOR DE PRODUTO*. [s.l: s.n.]. Disponível em: <https://www.repository.utl.pt/bitstream/10400.5/17658/1/DM-CRRC-2018.pdf>. Acesso em: 6 mar. 2022.

DOMINGOS, P. O Algoritmo Mestre: *Como a busca pelo algoritmo de machine learning definitivo recriará nosso mundo*. [s.l.] Novatec Editora, 2017.

GUTIERREZ, G. L. *Gestão comunicativa: maximizando criatividade e racionalidade. Rio de janeiro* : Qualitymark, 1999.

IBM. *O que é machine learning*? Disponível em: <https://www.ibm.com/br-pt/cloud/learn/machine-learning>. Acesso em: 24 ago. 2022.

IMDB PT-BR. Disponível em: <https://www.kaggle.com/datasets/luisfredgs/imdb-ptbr>. Acesso em: 7 ago. 2022.

LAB, I. *Por que o Python é a Linguagem mais adotada na área de Data Science* ? Disponível em: <https://insightlab.ufc.br/por-que-o-python-e-a-linguagem-mais-adotada-na-area-de-data-science>. Acesso em: 07 ago. 2022.

LOH, Stanley. *31 tipos de sistemas de informação - 31 maneiras de a tecnologia da*

informação ajudar as organizações. Porto Alegre, 2014.

LOH, Stanley. *BI na era do big data para cientistas de dados - indo além de cubos e*

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

dashboards na busca pelos porquês, explicações e padrões. Porto Alegre, 2014.

Mailchimp. *What Is Marketing Analytics? Definition and Examples* . Disponível em: <https://mailchimp.com/marketing-glossary/marketing-analytics/>. Acesso em: 6 mar. 2022.

MARA. *Business Intelligence: entenda o que é e como o BI agiliza sua tomada de decisão.* Disponível em: <https://resultadosdigitais.com.br/marketing/business-intelligence-bi/>. Acesso em: 26 mar. 2022.

MARR, B. Big Data: *The 5 Vs Everyone Must Know*. Linkedin, 6 mar. 2014. Disponível em: <https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know/>. Acesso em: 20 fev. 2022.

MANNING, C. D. et al. *Introduction to information retrieval.* Cambridge: Cambridge University Press, 2009.

ORACLE. *What is big data? | Oracle*. Disponível em: <https://www.oracle.com/big-data/what-is-big-data/>. Acesso em: 20 fev. 2022.

PAPAZISSIS, P. *Empresas data-driven e o uso do big data como diferencial competitivo.* Linkedin, 13 jun. 2018. Disponível em: <https://www.linkedin.com/pulse/empresas-data-driven-e-o-uso-do-big-data-como-papazissis-matuck/?originalSubdomain=pt>. Acesso em: 22 fev. 2022.

PETRI, A. *O berço do Big Data*. Veja, 15 maio 2013. Disponível em <https://www.cin.ufpe.br/~cjgf/TECNOLOGIA%20-%20material%20NAO-CLASSIFICADO/BIG%20DATA%20revista%20Veja%202013.pdf>. Acesso em: 22 fev. 2022.

Python: O que é? Por que usar? - PyScience-Brasil. Disponível em: <http://pyscience-brasil.wikidot.com/python:python-oq-e-pq>. Acesso em: 12 mar. 2022.

POPOVIČ, Aleš et al. Towards business intelligence systems success: Effects of maturity and culture on analytical decision making. Decision support systems, v. 54, n. 1, p. 729-739, 2012.

PROENÇA, Marina. A influência do uso do digital marketing analytics em decisões efetivas de marketing mediada pela capacidade absortiva.

RESULTADOS DIGITAIS. *Marketing 4.0: saiba mais sobre esse conceito e as estratégias envolvidas.* Disponível em: <https://resultadosdigitais.com.br/marketing/marketing-40/>. Acesso em: 26 mar. 2022.

RIGO BOTELHO1, F.; FILHO2, E. *CONCEITUANDO O TERMO BUSINESS INTELLIGENCE: ORIGEM E PRINCIPAIS OBJETIVOS.* [s.l: s.n.]. Disponível em: <http://www.iiisci.org/Journal/pdv/risci/pdfs/CB793JN14.pdf>. Acesso em: 26 mar. 2022.

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

SILVA, L. *Tomada de Decisão Baseada em Dados (DDDM) e Aplicações em Informática em Educação.* p. 21–46, 2015.

SILVA, Eduardo. Disponível em: <https://github.com/Eduardo681/tcc>. Acesso em: 07 ago. 2022.

TOKIO SCHOLL. *A história do Python. As versões de uma linguagem única.* Disponível em: <https://tokioschool.pt/noticias/historia-python/>. Acesso em: 24 ago. 2022.