Um estudo sobre Mineração de Dados focado em Análise de Sentimentos em textos de língua portuguesa nas redes sociais

Vitor Senna SILVERIO, Almir Rogerio CAMOLESI
vithorsenna07@gmail.com.br, camolesi@femanet.com.br

RESUMO: Esse artigo tem como objetivo apresentar os principais temas investigados e trabalhados durante o desenvolvimento dos estudos sobre extração de dados da Web e análise de sentimentos textuais em língua portuguesa. Foram laborados estudos sobre inteligência artificial, aprendizagem de máquina, análise de sentimentos e desenvolveu-se algumas aplicações na linguagem Python visto ser necessário pôr em prática o que estava sendo estudado. Ao final será apresentado a proposta para futuros trabalhos provenientes dessa pesquisa.

PALAVRAS-CHAVES: Análise de Sentimentos; Python; Inteligência artificial, Aprendizagem de Máquina; Mineração de Dados.

ABSTRACT: This article aims to present the main themes investigated and worked during the development of studies on web data extraction and analysis of textual feelings in Portuguese. Studies were conducted on artificial intelligence, machine learning, sentiment analysis and some applications were developed in the Python language since it was necessary to put into practice what was being studied. At the end, the proposal for future work from this research will be presented.

KEYWORDS: Sentiment Analysis, Python, Artificial Intelligence, Machine Learning, Data Mining, Web Scraping.

1.Introdução

As últimas duas décadas nos mostraram um avanço extraordinário com a tecnologia, principalmente na forma como as pessoa e empresas se conectam. Uma das mais utilizadas sem dúvidas são as redes sociais, visto que nelas se encontram dados de forma tão massiva quanto em qualquer outro lugar, sendo eles vídeos, áudios, fotos e principalmente, textos. Nos textos se encontram os mais variados tipos de informações, desde simples comentários a críticas e elogios.

Em sua pesquisa sobre redes sociais, Costa (2018) diz que apesar de terem nascido com caráter pessoal, as Redes Sociais Virtuais - RSV reúnem mais do que usuários, agrupam pessoas em grupos distintos, por afinidades que apresentam atributos sociais. Desta maneira, estudos científicos precisam contemplar esse tipo de plataforma, a qual vem se consolidando como mais um canal de interação com os usuários da Web (COSTA, 2018, p.2).

Analisado o meio mercadológico atual e as principais tendencias tecnológicas, viabiliza-se o desenvolvimento de uma ferramenta que extraia esses dados de forma automática e realize uma análise específica sobre eles.

Uma dessas análises é a "análise de sentimentos textuais", onde se torna possível saber se o conteúdo textual presente em um comentário é positivo, negativo ou neutro. As ferramentas disponíveis atualmente que realizam tal tarefa, são de uso, limitado ou especifico e algumas delas necessitam de um conhecimento aprofundado para sua utilização.

Com essa análise se torna possível, por exemplo, saber a satisfação de um determinado público sobre um produto, taxa de aprovação, reprovação, previsões de mercado, entre outros.

Costa e Picchi (2017) perceberam que a internet está provocando mudanças na forma em que as pessoas geram opiniões e na forma como as empresas lidam com os consumidores. Desse modo, as informações disponíveis na internet influenciam, por isso eles buscam na própria internet informações e consultam avaliações de consumidores e rankings de produtos/serviços antes de decidir por uma compra, embora saibam que na internet rumores, meias verdades, especulações e hipóteses podem se tornar afirmativas e causar diversos danos (COSTA, B. R. L.; PICCHI, R.; 2017, p. 16-17).

2.Análise de Sentimentos

Análise de sentimentos é um dos ramos da ciência de dados, onde é utilizado processamento de linguagem natural para extração e realizar uma ou mais análises aprofundadas. No geral, essa análise se passa por quatro estágios, que são reconhecer, extrair, dimensionar e estudar as polaridades sentimentais dentro de um conjunto de dados.

“Existem duas principais abordagens para o problema de extração de sentimentos em textos. A primeira delas é embasada nos conceitos de aprendizagem de máquina partindo da definição de características que permitam distinguir entre sentenças com diferentes sentimentos, treinamento de um modelo com sentenças previamente rotuladas e utilização do modelo de forma que ele seja capaz de identificar o sentimento em sentenças até então desconhecidas. A segunda abordagem não conta com treinamento de modelos de aprendizado de máquina e, em geral, são baseadas em tratamentos léxicos de sentimentos que envolvem o cálculo da polaridade de um texto a partir de orientação semântica das palavras contidas neste texto” (BENEVENUTO, F.; RIBEIRO, F.; ARAÚJO, M.; 2015, p.6).

2.1 Inteligência Artificial e Machine Learning

Inteligência artificial, de forma objetiva, trata-se de algoritmos que tem o objetivo de fazer uma “máquina” (software) realize determinadas ações como se fosse um humano, porém de forma muita mais precisa, eficiente e em quantidades maiores. Já Machine Learning está diretamente ligada aos conceitos de I.A, pois se trata de algoritmos que são capazes de identificar padrões e “aprender” junto a eles.

“Basicamente, nas origens dos sistemas de computação, os algoritmos precisavam ser programados para desempenharem determinada tarefa e deviam ser descritos minunciosamente. Com o Big Data, tornou-se possível que o algoritmo desenvolvesse a si mesmo, dispensando programação específica e, na verdade, o próprio Big Data ganhou valor e utilidade por conta dos algoritmos” (OLIVEIRA, B. P. G; 2020, p. 7).

Um grande exemplo de ferramenta de aprendizagem de máquina e I.A é o preenchimento de texto automático do Google Chrome. Quando digitamos algo no buscador, ele já nos oferece algumas sugestões sobre as palavras ou frases mais buscadas em relação ao que digitamos. Outro exemplo muito conhecido e que poucas pessoas tem conhecimento que

se trata de aplicações em I.A, é o buscador do Youtube. Ao digitarmos apenas a letra de uma música, mesmo que de forma incorreta, o algoritmo é capaz muitas vezes que nos retornar o que procurávamos.

2.2 Ciência de Dados

O termo “ciência de dados” (ou “data science”) já vem sendo discutido desde da década de 1960, mas sua notoriedade firmou-se a partir de 2010, quando em grandes empresas foram se formando equipes para tal estudo e desenvolvimento.

Por mais que seja um tema trabalhado a décadas, é uma área da ciência e tecnologia que ainda precisa de muito aperfeiçoamento e avanço.

Como citado anteriormente, um dos ramos da ciência de dados é a análise de sentimentos (Sentiment Analysis), a qual utiliza princípios de aprendizagem de máquina (Machine Learning), inteligência artificial (I.A) e estatística para ser abordado e desenvolvido. Há várias limitações e desafios a serem enfrentados nesse tipo de análise, principalmente quando falamos de língua portuguesa já que a linguagem utilizada nas redes sociais é majoritariamente informal. Abreviações, gírias, erros gramaticais, ironias e até mesmo variações linguísticas como as diastráticas, diatópicas, diacrônicas e diafásicas se tornam fortes barreiras durante a análise.

2.3 Mineração de Dados

Mineração de dados se trata de uma forma de extração de dados, que no caso desse trabalho, são dados provenientes da Internet. Existem várias formas de realizar essa extração, as mais utilizadas e conhecidas são através de métodos de *webscraping* ou por um acesso direto a API de alguma rede social. As duas formas possibilitam manipular dados diretamente através do código da aplicação.

2.4 Web Scraping

Como citado anteriormente, *web scraping* se trata de uma das formas de extração de dados da Internet. Essa forma de extração é muito eficiente e funciona em qualquer site, por ser uma forma mais “robusta” e “informal” de extrair os dados à “força”. Acessa diretamente o código HTML do site e assim realiza uma série de buscas através de suas TAGs para encontrar e extrair o que se deseja.
Por mais que seja funcional e eficiente, essa técnica possui desvantagens. Como ela precisa acessar diretamente uma TAG específica dentre do HTML, caso a estrutura do site mude, o código responsável pelo *web scraping* precisará passar por uma atualização para se adequar a nova estrutura.

3. Mercado

Um cientista de dados, atualmente é indispensável, principalmente para grandes empresas que necessitam de uma forma de automatizar suas análises, sejam elas quais forem.
“A quantidade de usuários ativos e o volume de dados criados diariamente nessas redes é impressionante. Uma plataforma popular nos dias de hoje é o Twitter que, sozinho, possui mais de 200 milhões de usuários, que compartilham cerca de 400 milhões de tweets por dia. Nesse contexto, pesquisadores e empresas conseguem coletar esses dados para análises de conteúdo em grande escala” (ARAÚJO, M.; GOLÇALVES, P.; BENEVENUTO, F.; 2013, p. 2).

4. Estudo de Caso

Os estudos iniciaram com a leitura de livros e artigos sobre processamento de linguagem natural, análise de sentimentos e mineração de dados. Após a melhor contextualização e entendimento sobre o assunto, iniciaram algumas pequenas aplicações dos conceitos aprendidos, afim de colocar em prática o que estava sendo estudado.
O avanço das investigações permitiu o desenvolvimento de aplicações WEB. A linguagem de programação escolhida foi Python, para a coleta de dados, e o Framework Django para a parte de desenvolvimento WEB.

4.1 Proposta

A proposta era escolher em quais sites iriam ser extraídas as informações. Foram escolhidos quatro sites: G1, R7, UOL e Twitter. A forma de extração foi igual em ambos os sites, porém a forma para ter acesso ao código HTML precisou ser diferente, devido a forma como o código de cada uma é gerado em cada página.

4.2 Desenvolvimento

No total foram três aplicações WEB realizadas. A primeira teve foco na extração de notícias no G1, UOL e R7. Para a mineração desses dados foi utilizada a biblioteca BS4 do Python, onde a mesma faz uma conexão com o navegador para acessar o código HTML. Abaixo, se encontram os códigos desenvolvidos para a extração de dados.

1. Algoritmo para requisição do site no navegador e retorno de seu conteúdo HTML

```
def get_html_content(urlsite):
    import requests
    USER_AGENT = "Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36, like Gecko)
Chrome/44.0.2403.157 Safari/537.36"
    LANGUAGE = "en-US,en; q=0.5"
    session = requests.Session()
    session.headers['User-Agent'] = USER_AGENT
    session.headers['Accept-Language'] = LANGUAGE
    session.headers['Content-Language'] = LANGUAGE

    if urlsite == 'g1':
        html_content=session.get('https://g1.globo.com/').text
    elif urlsite == 'r7':
        html_content = session.get('https://www.r7.com/').text
    elif urlsite == 'uol':
        html_content = session.get('https://www.uol.com.br/').text

    return html_content
```

Após isso, é necessário realizar a extração dos dados desejados:

1. Aplicação desenvolvida para extração de notícias

De acordo com o botão pressionado na página, será retornado as respectivas notícias de cada um.

2. Algoritmo para extração

```
from bs4 import BeautifulSoup
urlsite= ''
if 'r7' in request.GET:
urlsite = 'r7'
html_content = get_html_content(urlsite)
source = BeautifulSoup(html_content, 'html.parser')
noticiasR7_title = source.findAll('h3', attrs={'class': 'r7-flex-title-h5'})

for i in noticiasR7_title:
  noticiasR7_txttag1title.append(i.text)

  contreg=int(len(noticiasR7_txttag1title))
```

Já as segunda e terceira aplicações, a mineração foi realizada na plataforma Twitter, porém de forma um pouco divergente da primeira. O Twitter, devido ao seu código em JavaScript, só gera o código HTML se o navegador for aberto, caso contrário não é possível (levando em consideração o método Web Scraping). Devido a isso, foi necessário utilizar outra biblioteca do Python, a Selenium. Além de possuir excelentes ferramentas para mineração de dados, ela permite que o navegador seja automatizado, ou

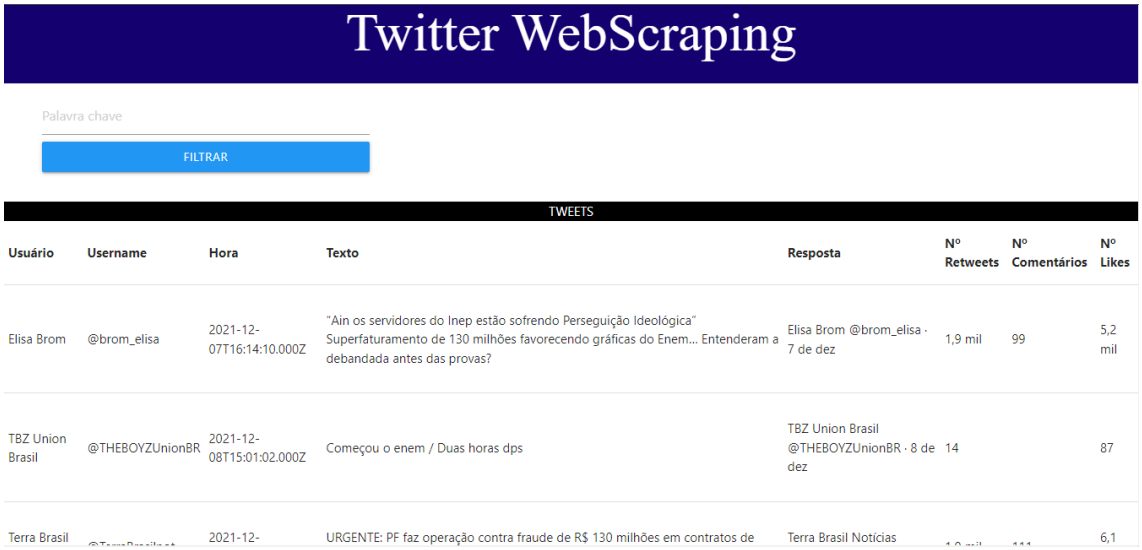
seja, ele é aberto de forma automática, acesso ao site do Twitter, e dessa forma o código HTML é gerado e armazenado no Python.

3. Algoritmo para acesso ao navegador e pesquisar a palavra desejada

```
naveg = webdriver.Chrome(options=options)
naveg.get(f'https://twitter.com/search?q={filtro}&src=typed_query')
sleep(5)
naveg.execute_script('window.scrollTo(0, document.body.scrollHeight);')
sleep(2)
```

Dessa forma, com o navegador sendo aberto na rota correta da palavra que desejamos usar como chave para a busca de tweets.

Página da aplicação onde são retornados os tweets encontrados



2. Página da aplicação onde são retornados os tweets encontrados

Podemos realizar diversas buscas, que em todas elas a página mostrará os dados correspondentes.

4. Algoritmo para identificação do tweet na página e separação de seus elementos

```
publicacao=naveg.find_elements_by_xpath('//article[@data-testid="tweet"]')
j = 0
for i in publicacao:
        tweet_usuario.append(i.find_element_by_xpath('.//span').text)
        tweet_username.append(i.find_element_by_xpath('.//span[contains(text(),
        "@")]').text)
        tweet_hora.append(i.find_element_by_xpath('.//time').get_attribute('datetime'))
        tweet_answer.append(i.find_element_by_xpath('.//div[2]/div[2]/div[1]').text)
        tweet_texto.append(i.find_element_by_xpath('.//div[2]/div[2]/div[2]/div[1]').text)
        tweet_qtdretweets.append(i.find_element_by_xpath('.//div[@data-
        testid="retweet"]').text)
        tweet_qtdcomments.append(i.find_element_by_xpath('.//div[@data-
        testid="reply"]').text)
        tweet_qtdlikes.append(i.find_element_by_xpath('.//div[@data-testid="like"]').text)
        tweet.append(TweetTouple(tweet_usuario[j], tweet_username[j],  tweet_hora[j],
        tweet_answer[j],tweet_texto[j], tweet_qtdretweets[j],  tweet_qtdcomments[j],
        tweet_qtdlikes[j]))

j = j+1
return (tweet)
```

Dessa forma, não exibimos os elementos do tweet de uma vez como um único objeto, mas sim, separando elemento por elemento e armazenamos esses dados em uma lista.

5. Conclusão

As aplicações tiveram foco na extração de dados enquanto os estudos teóricos focaram sobre inteligência artificial, processamento de dados e análise de sentimentos, tornando o andamento do projeto melhor definido. Houve um grande amadurecimento sobre aplicações WEB e com isso motiva mais os autores a continuarem o trabalho.

De várias formas as ferramentas estudadas fazem parte do nosso cotidiano e há uma grande procura de profissionais e ferramentas que realizem as mais diversas análises.

6.Trabalhos Futuros

Sobre uma futura pesquisa e/ou trabalho de conclusão de curso, o foco será laborar estudos e aplicações estatísticas sobre textos extraídos do Twitter para que, enfim, seja realizada a análise de sentimentos.

O método de extração utilizado será pelo o acesso direto a API do Twitter, possibilitando a mineração de uma maior quantidade de dados em maiores detalhes para a análise.

REFERÊNCIAS BIBLIOGRÁFICAS

BENEVENUTO, Fabricio. **Uma Abordagem Multilíngue para Análise de Sentimentos**. In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING (BRASNAM), 4., 2015, Recife. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2015, p. , ISSN 2595-6094. DOI: https://doi.org/10.5753/brasnam.2015.6767.

BENEVENUTO, F.; RIBEIRO, F.; ARAÚJO, M. **Métodos para Análise de Sentimentos em mídias sociais**. Universidade Federal de Minas Gerais. Disponível em <https://homepages.dcc.ufmg.br/~fabricio/download/webmedia-short-course.pdf>. Acesso em: 05 abr. 2021.

COSTA, Barbara Regina Lopes. **Bola de Neve Virtual: O Uso das Redes Sociais Virtuais no Processo de Coleta de Dados de uma Pesquisa Científica**. Revista Interdiciplinar de Gestão Social, Salvador, v. 7, n. 1, p. 15-37, jan./abr. 2018. Disponível em: <https://periodicos.ufba.br/index.php/rigs/article/view/24649/16131>. Acesso em: 12 nov. 2021.

FRANÇA, Thiago C. de; OLIVEIRA, Jonice. **Análise de Sentimento de Tweets Relacionados aos Protestos que ocorreram no Brasil entre Junho e Agosto de 2013**. BrasNAM – III Brazilian Workshop on Social Networks Analysis and Mining. Programa de Pós-Graduação em Informática da Universidade Federal do Rio de Janeiro (UFRJ). Disponível em: <http://www.each.usp.br/digiampietri/BraSNAM/2014/p11.pdf>. Acesso em: 12 ago. 2021.

K. H. Manguri, R. N. Ramadhan, and P. R. Mohammed Amin, **"Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks"**, Kurdistan Journal of Applied Research, vol. 5, no. 3, pp. 54-65, May 2020.

MUELLER, John Paul; MASSARON, Luca. **Inteligência Artificial para leigos**. 1ª ed. Rio de Janeiro: Editora Alta Books, 2019.

REIS, Julio; GONÇALVES, Pollyanna; ARAÚJO, Matheus; PEREIRA, Adriano César;

OLIVEIRA, Bruna Pinotti Garcia. **Inteligência Artificial e Proteção de Dados: Sobre a Autodeterminação Informativa e a Manipulação Informacional por Machine Learning**. HUMANIDADES & TECNOLOGIA (FINOM), Paracatu, v. 26, n. 1, p. 162-183, jul./set. 2020. Disponível em: <http://revistas.icesp.br/index.php/FINOM_Humanidade_Tecnologia/article/view/1356/1013>. Acesso em: 26 jun. 2021.