

Carrefour Data Challenge



DIGITAL
INNOVATION
ONE

Autor: Rodrigo S. Marinho

GitHub: <https://github.com/allons-y-rod>

LinkedIn: [linkedin.com/in/rodrigo-marinho-55a64514a/](https://www.linkedin.com/in/rodrigo-marinho-55a64514a/)

Extraindo os Trends do Twitter e armazenando-os para futura exploração de dados

1. Introdução

Trends, traduzindo livremente para o português seriam as tendências. No mundo digital são uma importante ferramenta pois mostram para o usuário os termos mais relevantes dentro da plataforma em um determinado período.

Mas qual a vantagem de explorar os trends? Eles podem gerar informações vitais para uma empresa nos dias atuais. Uma vez identificado os assuntos que passaram a ter destaque na internet, campanhas e materiais específicos podem ser construídos de forma a aumentar o engajamento da marca. Permitem refinar constantemente o seu modelo de negócio a partir da exploração de tendências dentro do seu segmento de mercado.

2. Metodologia e Resultados

Para extrair os dados do Twitter, inicialmente é necessário criar uma conta na plataforma e aplicar para obter as chaves de acesso geradas pela Plataforma de Desenvolvimento do Twitter. Após análise das intenções por trás desse pedido, o acesso é então liberado. As chaves obtidas foram utilizadas dentro do programa criado para a obtenção dos dados.

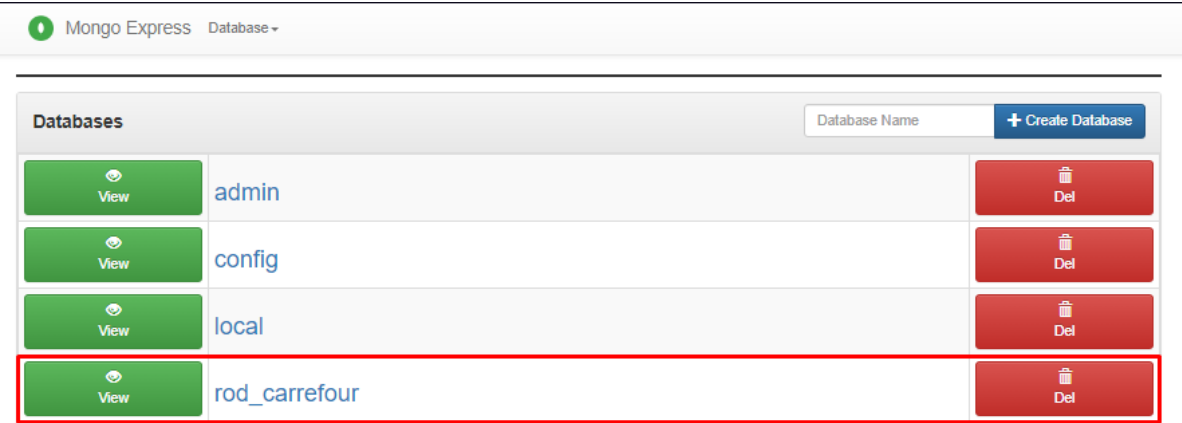
Imagem: Developer Portal

The image shows a screenshot of the Twitter Developer Portal. On the left is a dark sidebar with the Twitter logo and 'Developer Portal' text. Below it is a section titled '#UseCases' with instructions: 'First things first, let's get you the right application. Pick the use case that most closely relates to the type of work you intend to do while using the Twitter developer platform. There's some overlap between these so focus on the specifics of your main objective.' The main content area has a header 'Which best describes you?' and a sub-header 'This is how you intend to use the Twitter developer platform'. Below this are three light blue boxes with icons and labels: 'Professional' (laptop icon), 'Hobbyist' (paper plane icon), and 'Academic' (stack of books icon). At the top right of the main area are links for 'Docs', 'Community', 'Updates', and 'Support'. At the bottom is a footer with links for 'Privacy', 'Cookies', 'Twitter Terms & Conditions', 'Developer Policy & Terms', '© 2021 Twitter Inc.', 'Follow @TWITTERDEV', and 'Subscribe to Developer News'.

Fonte: <https://developer.twitter.com/en/portal/petition/use-case> (2021)

Os dados extraídos foram armazenados em um banco de dados NoSQL (Not Only SQL) para que posteriormente etapas de ETL (extração, transformação, carregamento) fossem aplicadas. O banco de dados escolhido foi o MongoDB, onde a aplicação extrai os dados e armazena-os em um database que pode ser visualizado no Mongo Express.

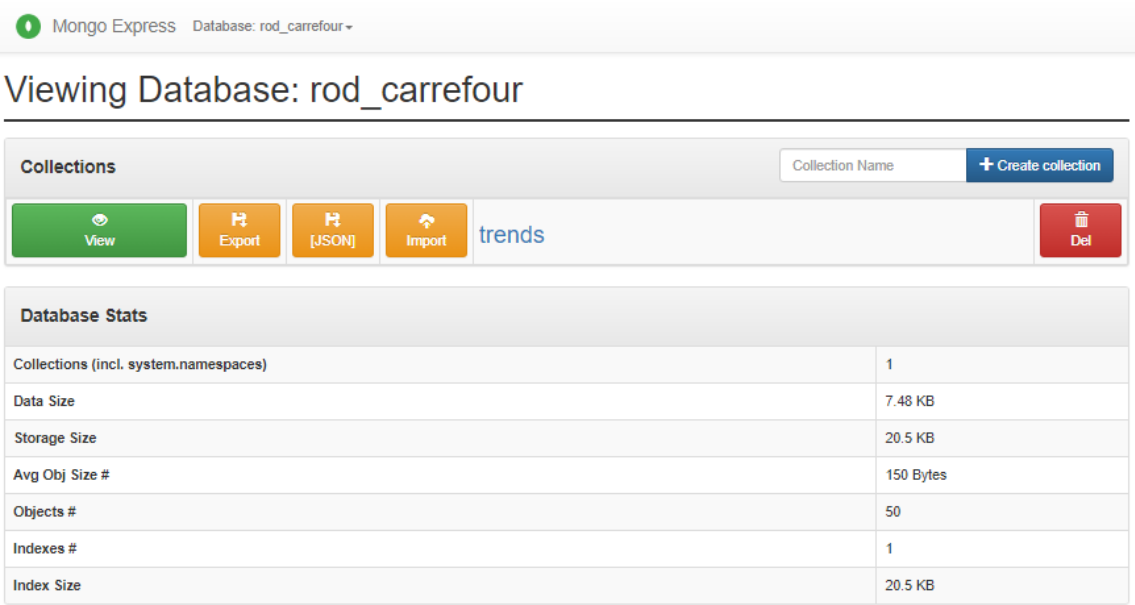
Imagem: Mongo Express



Fonte: Usuário (2021)

Ao acessar o database do desafio (rod_carrefour) podemos obter na janela as informações de número de objetos armazenados, tamanho dos arquivos armazenados, etc.

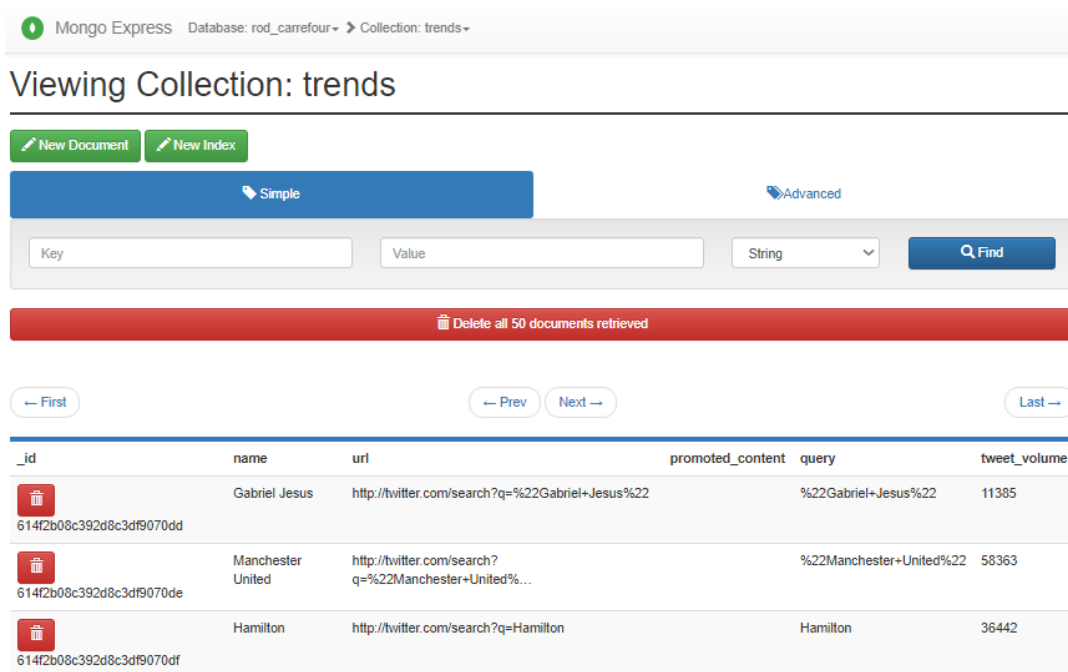
Imagem: Database: rod_carrefour



Fonte: Usuário (2021)

No nosso caso de estudo estamos armazenando os Trends Topics do Brasil, onde os 50 assuntos mais relevantes no momento da busca foram armazenados. Ao acessar os dados temos as seguintes informações:

Imagem: Database: rod_carrefour



Mongo Express Database: rod_carrefour Collection: trends

Viewing Collection: trends

New Document New Index

Simple Advanced

Key Value String Find

Delete all 50 documents retrieved

First Prev Next Last

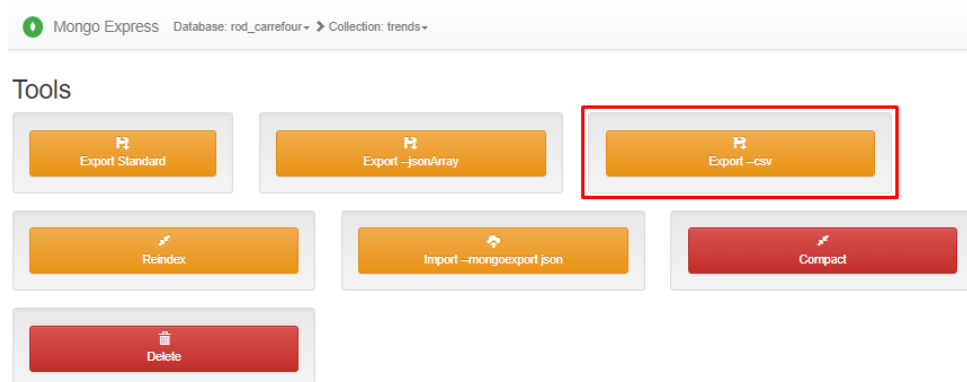
_id	name	url	promoted_content	query	tweet_volume
614f2b08c392d8c3df9070dd	Gabriel Jesus	http://twitter.com/search?q=%22Gabriel+Jesus%22		%22Gabriel+Jesus%22	11385
614f2b08c392d8c3df9070de	Manchester United	http://twitter.com/search?q=%22Manchester+United%22		%22Manchester+United%22	58363
614f2b08c392d8c3df9070df	Hamilton	http://twitter.com/search?q=Hamilton		Hamilton	36442

Fonte: Usuário (2021)

Na coluna “name” temos o assunto que se tornou Trend, e ao acessar o link da coluna “url” somos redirecionados para a plataforma do Twitter, onde vemos a interação dos usuários. Na coluna “tweet_volume” temos a quantidade de tweets sobre o assunto no momento da pesquisa.

Os dados armazenados foram exportados no formato CSV (Comma-separated values) através da ferramenta do Mongo Express e posteriormente foram inseridos no MicroStrategy para visualização dos mesmos.

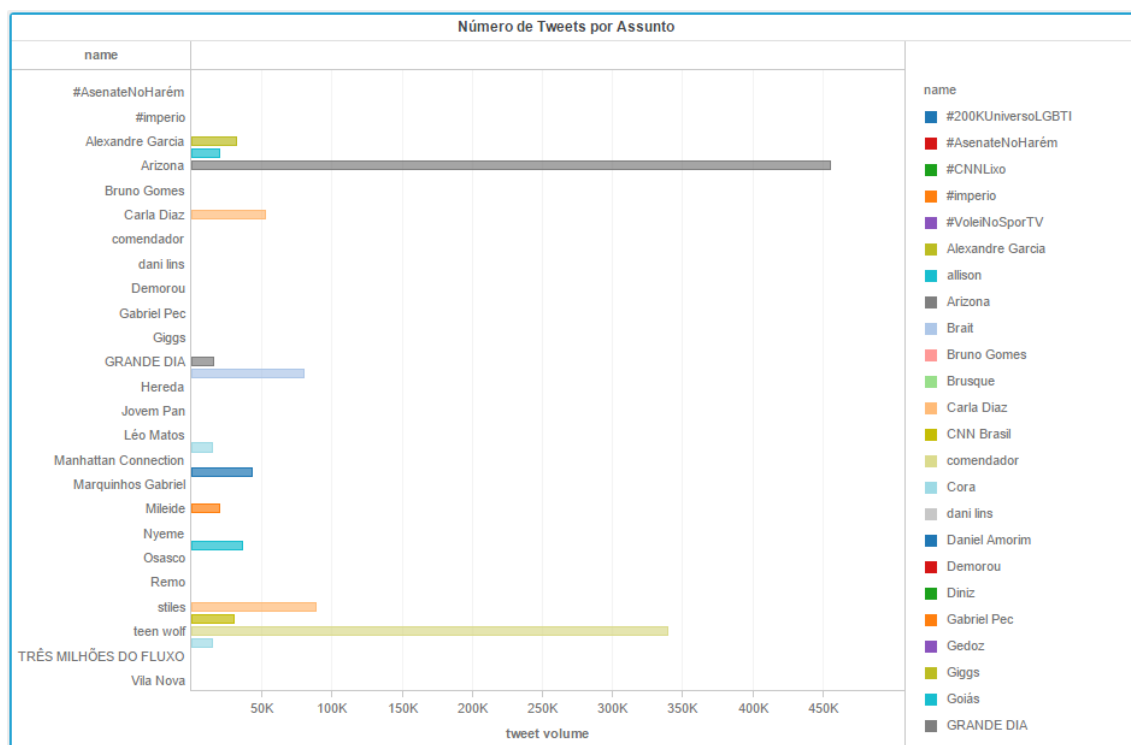
Imagem: Database: rod_carrefour



Fonte: Usuário (2021)

Utilizando o MicroStrategy para ler o arquivo CSV e plotar um gráfico de barras temos uma melhor visualização dos dados obtidos:

Imagem: MicroStrategy



Fonte: Usuário (2021)

Podemos notar que para alguns nomes nos trends o “tweet_volume” se encontra como zero. Isso se dá por conta de a aplicação não conseguir recuperar essa informação em determinadas tags no momento da busca. Poderíamos eliminar os mesmos por não termos informações sobre o volume, mas optou-se por mantê-los, uma vez que apenas estar nos trends nos traz a informação de terem engajamento na plataforma no momento da pesquisa.

Esta análise inicial nos deu uma visão geral sobre os assuntos em alta no momento e seu volume na rede social. Mas como podemos obter mais informações com isso? Pensando em responder esta pergunta foi criado um segundo programa destinado à análise exploratória dos dados. Tal programa foi criado em um Jupyter Notebook (Data Challenge - Dio - Banco Carrefour.ipynb) de forma que as etapas do processo fiquem mais claras e de fácil entendimento.

Primeiro foi escolhida apenas uma trend das extraídas anteriormente, contudo poderíamos utilizar todas ao mesmo tempo, mas o volume de dados seria muito grande para esta análise de estudo. A trend escolhida foi “teen wolf” (série que contou com 6 temporadas, exibidas entre 2011 e 2017. Onde recentemente foi anunciado um novo filme).

Utilizando o programa de análise exploratória sob a keyword “teen wolf”, extraímos os tweets para um arquivo JSON (“tweet_teenwolf.json”) que então foi importado como um dataframe utilizando a biblioteca Pandas.

Imagem: Importação do arquivo json e criação do dataframe.

```
df = pd.read_json('tweet_teenwolf.json', lines = True)
df.head(10)
```

	created_at	User_id	tweet
0	2021-09-25 15:51:00+00:00	1441792358325051392	RT @stilinskygf: eu: quero morrer \teen wolf ...
1	2021-09-25 15:50:57+00:00	1441792346815934484	RT @srxvision: ⚠️ QUERO SEGUIR MAIS FÃS DE TEE...
2	2021-09-25 15:50:55+00:00	1441792340308403328	RT @SeriesTWBZ: Um filme de Teen Wolf é confir...
3	2021-09-25 15:50:55+00:00	1441792338083434496	RT @hbitualz: época boa era quando pretty litt...
4	2021-09-25 15:50:54+00:00	1441792335507955712	RT @goticslahey: cenas de teen wolf que dá par...
5	2021-09-25 15:50:54+00:00	1441792333171892224	RT @bissexualuke: vendo a série vendo...
6	2021-09-25 15:50:53+00:00	1441792331292880192	RT @ilyella4: STANS DE TEEN WOLF ME SIGAM EU J...
7	2021-09-25 15:50:52+00:00	1441792325336780800	RT @fallingbstars: os fãs de teen wolf agr htt...
8	2021-09-25 15:50:52+00:00	1441792324778921984	RT @fallingbstars: os fãs de teen wolf agr htt...
9	2021-09-25 15:50:51+00:00	1441792321892088272	O filme de Teen Wolf é um surto mto grande

Fonte: Usuário (2021)

Um conjunto de bibliotecas para processamento simbólico e estatístico da linguagem natural foi utilizado. Realizou-se um processo de dividir uma string ou textos em uma lista de tokens, tal processo é conhecido como Tokenize. Para o twitter, o padrão utilizado foi diferente, pois a análise dos emojis também será efetuada. Visando uma busca mais eficiente, a eliminação de Stopwords, que são termos frequentes em um idioma, porém que não possuem relevância nas pesquisas, foi realizada. Tais termos como preposições, artigos, conjunções e outros.

Ao término do processo uma coluna contendo os tweets “limpos” foi criada no dataframe para armazená-los (“preprocessed”).

Imagem: tweets “limpos”.

	created_at	User_id	tweet	preprocessed
0	2021-09-25 15:51:00+00:00	1441792358325051392	RT @stilinskygf: eu: quero morrer \teen wolf ...	@stilinskygf quero morrer teen wolf voltando q...
1	2021-09-25 15:50:57+00:00	1441792346815934484	RT @srxvision: ⚠️ QUERO SEGUIR MAIS FÃS DE TEE...	@srxvision ⚠️ quero seguir fãs teen wolf ⚠️ de...
2	2021-09-25 15:50:55+00:00	1441792340308403328	RT @SeriesTWBZ: Um filme de Teen Wolf é confir...	@seriestwbz filme teen wolf confirmado 2022 pa...
3	2021-09-25 15:50:55+00:00	1441792338083434496	RT @hbitualz: época boa era quando pretty litt...	@hbitualz época boa pretty little liars teen w...
4	2021-09-25 15:50:54+00:00	1441792335507955712	RT @goticslahey: cenas de teen wolf que dá par...	@goticslahey cenas teen wolf dá ouvir

Fonte: Usuário (2021)

Por último utilizou-se uma biblioteca para realizar a análise de sentimentos nos tweets após o tratamento dos dados. Podemos ver na imagem o processo de análise em cada tweet:

Imagem: Análise de Sentimento

```
@stilinskygf quero morrer teen wolf voltando quer nada quero nada, hein Sentiment(classification='neg', p_pos=0.2991412213740452, p_neg=0.7008587786259555)
@srxvision 🐺 quero seguir fãs teen wolf 🐺 deem aqui favor Sentiment(classification='pos', p_pos=0.7249911902035067, p_neg=0.27500880979649434)
@seriestwbz filme teen wolf confirmado 2022 paramount+ Sentiment(classification='pos', p_pos=0.6340201653086353, p_neg=0.36597983469136575)
```

Fonte: Usuário (2021)

Para uma melhor visualização os dados foram organizados em novas colunas no dataframe.

Imagem: Dataframe com análise de sentimentos

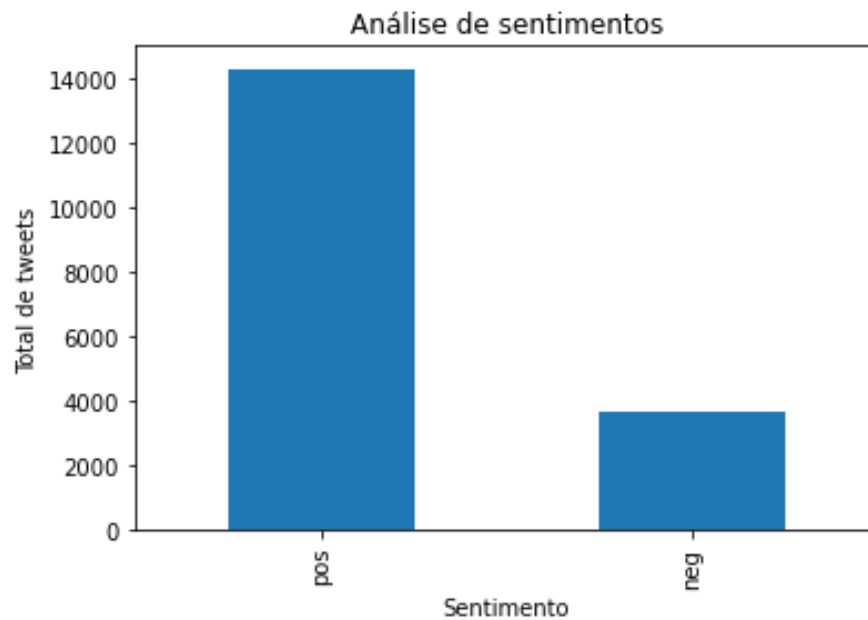
	created_at	User_id	tweet	preprocessed	classification	p_pos	p_neg
0	2021-09-25 15:51:00+00:00	1441792358325051392	RT @stilinskygf: eu: quero morrer lnteen wolf ...	@stilinskygf quero morrer teen wolf voltando q...	neg	0.299141	0.700859
1	2021-09-25 15:50:57+00:00	1441792346815934464	RT @srxvision: 🐺 QUERO SEGUIR MAIS FÂS DE TEE...	@srxvision 🐺 quero seguir fãs teen wolf 🐺 de...	pos	0.724991	0.275009
2	2021-09-25 15:50:55+00:00	1441792340308403328	RT @SeriesTWBZ: Um filme de Teen Wolf é confir...	@seriestwbz filme teen wolf confirmado 2022 pa...	pos	0.634020	0.365980
3	2021-09-25 15:50:55+00:00	1441792338083434496	RT @hbitualz: época boa era quando pretty litt...	@hbitualz época boa pretty little liars teen w...	neg	0.082032	0.917968
4	2021-09-25 15:50:54+00:00	1441792335507955712	RT @goticslahey: cenas de teen wolf que dá par...	@goticslahey cenas teen wolf dá ouvir	pos	0.561493	0.438507
...
17995	2021-09-25 11:27:21+00:00	1441726012249739264	RT @explicitmxndes: o tyler posey quando soube ...	@explicitmxndes tyler posey soube finalmente at...	pos	0.536718	0.463282
17996	2021-09-25 11:27:21+00:00	1441726009259147264	RT @serpensbarnes: exijo o meu lobão derek hal...	@serpensbarnes exijo lobão derek hale nesse fi...	pos	0.744255	0.255745
17997	2021-09-25 11:27:20+00:00	1441726007782879232	RT @sadbpineapplezz: e se o vilão do filme de t...	@sadbpineapplezz vilão filme teen wolf verdade ...	pos	0.649068	0.350932
17998	2021-09-25 11:27:19+00:00	1441726002862772224	RT @Twolfmaniac: ansioso pra ficar teorizando...	@twolfmaniac ansioso pra ficar teorizando tra...	neg	0.423051	0.576949
17999	2021-09-25 11:27:19+00:00	1441726000268629120	RT @fallingbstars: os fãs de teen wolf agr htt...	@fallingbstars fãs teen wolf agr	pos	0.561493	0.438507

18000 rows x 7 columns

Fonte: Usuário (2021)

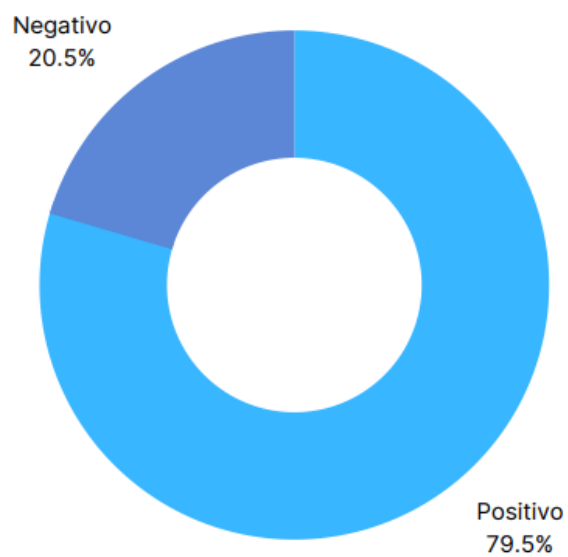
Assim podemos plotar um gráfico de barras para uma melhor visualização da classificação de sentimento.

Imagem: Análise de sentimentos



Fonte: Usuário (2021)

Análise de sentimentos



3. Conclusões

Na era digital se tornou imperativo conseguir identificar as principais tendências no mundo. Essa estratégia de marketing é vital tanto economicamente quanto operacional. A partir da análise inicial realizada, podemos ver o quanto a notícia do lançamento de um filme da franquia “Teen Wolf” movimentou a rede social com pico de mais de 300 mil tweets. Uma análise de sentimento em 18000 mil tweets mostrou que quase 80% da análise foi positiva, o que demonstra a aceitação do público brasileiro na plataforma.

Um programa desse tipo pode ser útil para uma empresa de varejo que deseja se preparar para a Black Friday por exemplo, ou uma empresa que deseja informações instantâneas sobre determinado produto/serviço que deseja lançar ou recolher feedbacks sobre os já lançados. Com isso concluímos a importância dos Trends como uma ferramenta “termômetro” fundamental na atualidade.

4. Referências

- [1] <https://developer.twitter.com/en/docs>
- [2] <https://pandas.pydata.org/docs/>
- [3] <https://matplotlib.org/stable/contents.html>
- [4] <https://textblob.readthedocs.io/en/dev/>
- [5] <https://www.nltk.org/#>
- [6] <https://youtu.be/H8mHWbzTSol>