Dados geográficos em redes sociais

Caio Vaz Machado, Carlos Augusto Lana de Mello, Denis Junio Barbosa Sales, Gabriel Lucas Silva Machado, João Paulo Castanheira

Departamento de Ciência da Computação – Universidade Federal de Minas Gerais, Brasil

caiovazmachado@gmail.com, carlosaldmello@gmail.com, denis.junio@gmail.com, gabriellucas1366@gmail.com, castanheira.jp@gmail.com

Abstract. *This document describes the work about the theme "Geografic data in social media". It is made a brief introduction about the theme, what is the motivation, what is the focus of the study, how the data was obtained, what tools were used and how they were used. In this study we also explain the analysis on the obtained data. Finally, the conclusions are presented.*

Resumo. *Este documento descreve a realização do trabalho sobre o tema Dados geográficos em redes sociais. É feita uma breve introdução à proposta do trabalho, qual a motivação, qual o foco dos estudos, como os dados foram coletados, quais ferramentas foram utilizadas e a análise realizada sobre os dados reunidos. Por fim, são apresentadas as conclusões que chegamos a partir das etapas descritas.*

1. Introdução

Nos últimos anos, o mundo se mostra vulnerável a diversos ataques de ódio que acabam causando grande impacto em vários locais. Tais ataques, como o ocorrido em Orlando no dia 12 de junho de 2016, acabam gerando grande repercussão nas mídias sociais, gerando uma quantidade maciça de dados.

O atentado mencionado foi causado por um homem de 29 anos, que entrou armado na boate *Pulse*, popular entre membros da comunidade LGBT, tirando a vida de cerca de 50 pessoas e ferindo diversas outras. Como o fluxo de informação acontece rapidamente nas redes sociais e na internet em geral, não demorou para que o assunto logo se tornasse um dos mais comentados.

Inicialmente, o trabalho seria realizado através de coleta de dados gerais em redes sociais, principalmente do Twitter, de forma à gerar uma visualização dos dados geográficos encontrados. No entanto, devido a grande repercussão do assunto após o atentado e com a possibilidade de fazer uma coleta mais específica para ser tratada, decidimos direcionar o trabalho para uma análise sobre o acontecimento.

O trabalho aqui descrito consiste, então, na coleta, visualização e análise de dados referentes a *Tweets* georreferenciados sobre Orlando, coletados no período de 12/06/2016 a 18/06/2016.

2. Decisões de projeto

2.1. Por que a escolha do Twitter

Diversas redes sociais, como Facebook, Foursquare e Twitter, disponibilizam API's que permitem que sejam coletados grandes amostras de dados para análise. Para esse trabalho, decidimos utilizar as API's do Twitter. A escolha do Twitter como fonte de dados se deu por alguns motivos, sendo alguns deles a facilidade de obtenção de dados através de sua API, a quantidade de dados disponíveis e o "leque" maior de possibilidades sobre as outras redes sociais.

Inicialmente, cogitamos realizar a coleta de dados usando, também, a API do Facebook (Facebook Graph), porém recentes mudanças na política de segurança da rede social reduziram as possibilidades de coleta de dados, disponibilizando menos informações em massa. Sendo assim, descartamos a sua utilização.

2.2. Desenvolvimento

Antes de explicar as implementações feitas para a coleta dos dados e como eles foram analisados, vamos apresentar brevemente a maneira como os dados são disponibilizados. Basicamente, o Twitter oferece aos desenvolvedores duas API's para coleta de *tweets*, a REST API e a Streaming API. Ambas são baseadas em requisições enviadas aos servidores com os filtros desejados, recebendo dos mesmos objetos JSON que possuem dados que correspondem à busca realizada.

Para este trabalho, os filtros utilizados foram a palavra-chave Orlando e uma *bounding box* que cobre toda a extensão do planeta([-180,-90],[180,90]). O uso dessa ferramenta como filtro faz com que o servidor do Twitter só retorne *tweets* georreferenciados, ou seja, *tweets* que apresentam um par de coordenadas contendo latitude e longitude em seu objeto.

Ainda sobre o georreferenciamento dos *tweets*, uma das dificuldades encontradas durante a realização do trabalho foi a quantidade de *tweets* que possuem informações de posicionamento. Como é necessário que o usuário configure manualmente a opção de permitir que a localização de onde o *tweet* foi enviado seja anexada ao seu post, principalmente por preocupações relacionadas à privacidade dos usuários, a quantidade de amostras com esses dados é relativamente muito pequena, chegando a ser cerca de 2% da massa total de *tweets*. Esse fator, somado a filtragem por palavra-chave, acabou reduzindo a quantidade de dados obtidos, restringindo as análises feitas posteriormente.

Voltando às duas API's, REST e Streaming. A REST é baseada em uma busca semelhante, mas não igual, à busca padrão disponível no próprio Twitter. Com ela, o foco é na relevância dos dados, e não na completude dos mesmos, fazendo com que alguns *tweets* talvez não sejam retornados. Além disso, só é possível consultar *tweets* publicados nos últimos sete dias. Nos programas implementados, a REST foi utilizada para buscar dados que não conseguimos coletar em tempo real.

Já a outra coleta foi feita via Streaming API. Ela permite uma conexão em tempo real com os servidores do Twitter, de forma a retornar uma quantidade maior de dados relacionados à busca desejada. Grande parte dos dados coletados vieram da utilização desse método, principalmente os *tweets* do dia após o atentado.

Dois protótipos foram implementados utilizando as API's citadas como principal fonte de funcionamento, utilizando, principalmente, Java e Python. Embora funcionem de maneiras um pouco diferentes, ambos os projetos foram utilizados para coletar os dados do servidor do Twitter e armazená-los em arquivos para análise futura.

No protótipo em Java, é possível escolher qual das API's vai ser utilizada e quais os parâmetros desejados para os filtros, além de permitir a geração de uma visualização em tempo real dos pontos extraídos de *tweets* georreferenciados em um mapa, via *Google Static Maps API*. Dos dados coletados, são retirados apenas os dados que consideramos relevantes para nossa análise, como nome do usuário, o texto contido no *tweet* e as coordenadas geográficas anexadas. Esses dados são então armazenados em arquivos **.txt*, separados por ";", de forma à facilitar a posterior leitura. Também consideramos utilizar a data e a hora em que os *tweets* foram enviados, porém quando os dados não são coletados em tempo real encontramos problemas e não foi possível confiar na sua utilização.

Já a aplicação que utiliza Python permite a mineração de dados do twitter e gera um arquivo JSON com informações em GeoJSON com a seguinte estrutura:'

```
{
    "type": "FeatureCollection",
    "features": [
        {
            "type": "Feature",
            "geometry": {
                "type": "Point",
                "coordinates": [some_latitude, some_longitude]
            },
            "properties": {
                "text": "This is sample a tweet",
                "created_at": "Sat Mar 21 12:30:00 +0000 2015"
            }
        },
    ]
}
```

A aplicação em Python mantem a conexão aberta com o servidor, coletando as informações relativas ao assunto durante a execução. No caso especifico desta aplicação, a coleta de dados ocorre com todos *tweets* relativo ao assunto requerido, utilizando a biblioteca Tweepy, que implementa a interface Cursor. Um segundo *script* lê o arquivo com todos os *tweets* e gera o arquivo com os tweets em GeoJSON, apresentado acima. Uma visualização para web foi implementada utilizando OpenStreetMap e a bliblioteca Javascript Leaflet, que usa o arquivo fornecido pelo segundo *script*. Este formato é facilmente importado para banco de dados, inclusive NoSQL.

3. Análise dos Dados

3.1. Análise Quantitativa

Durante todo o período foram coletados 4986 *tweets* georreferenciados utilizando a seguinte palavra chave para realizar a pesquisa:

- Orlando

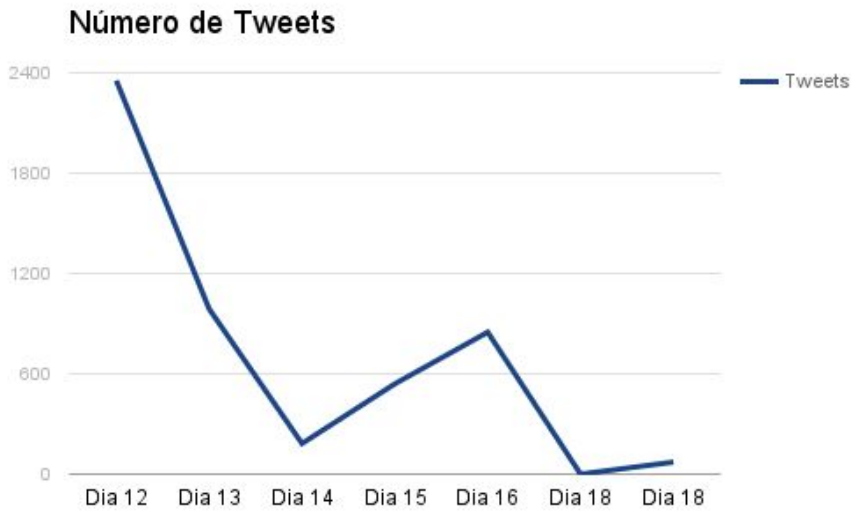

Figura 1. Gráfico referente ao número de tweets coletados.

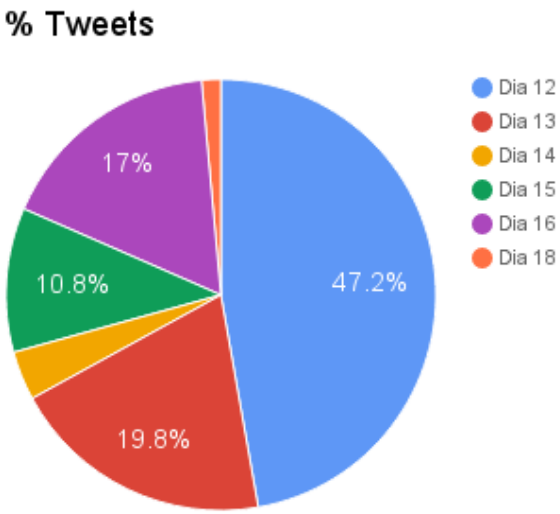

Figura 2. Distribuição dos tweets coletados para cada dia de coleta.

No próprio dia 12/06/2016 (dia do atentado terrorista) , o Estado Islâmico assumiu autoria dos ataques à boate em Orlando. Dentro dos *tweets* coletados, estes foram o usuários que mencionaram as palavras:

- ISIS
- ISLAM ou ISLAMIC

- Terrorist ou Terrorism
- Homophobes, Homophobic ou Homophobia

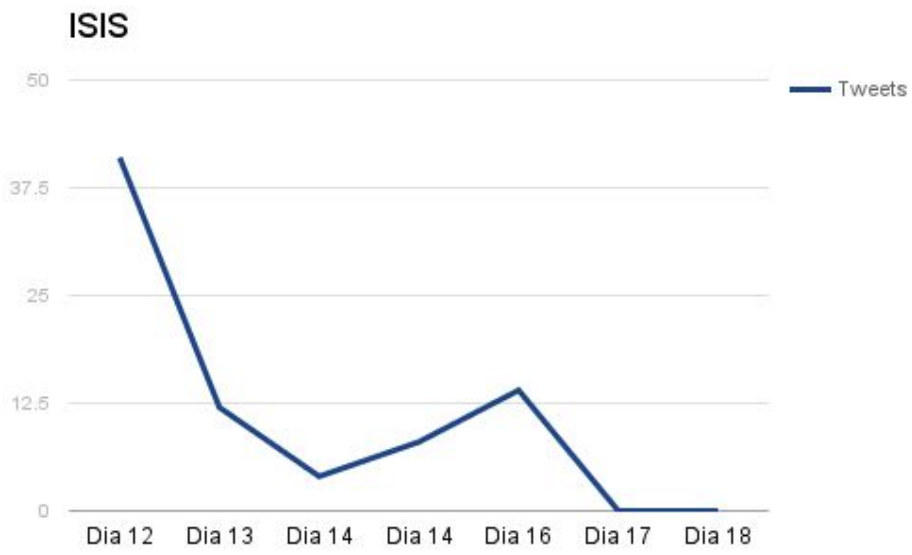

Figura 3. Tweets coletados com a tag ISIS por dia.

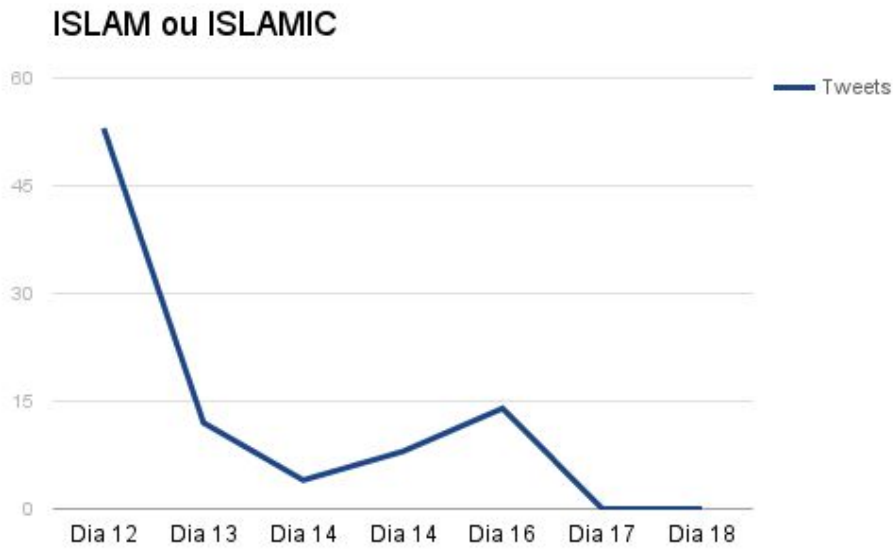

Figura 4. Tweets coletados com as tags ISLAM/ISLAMIC por dia.

Figura 5. Tweets coletados com as tags Terrorist/ Terrorism por dia.

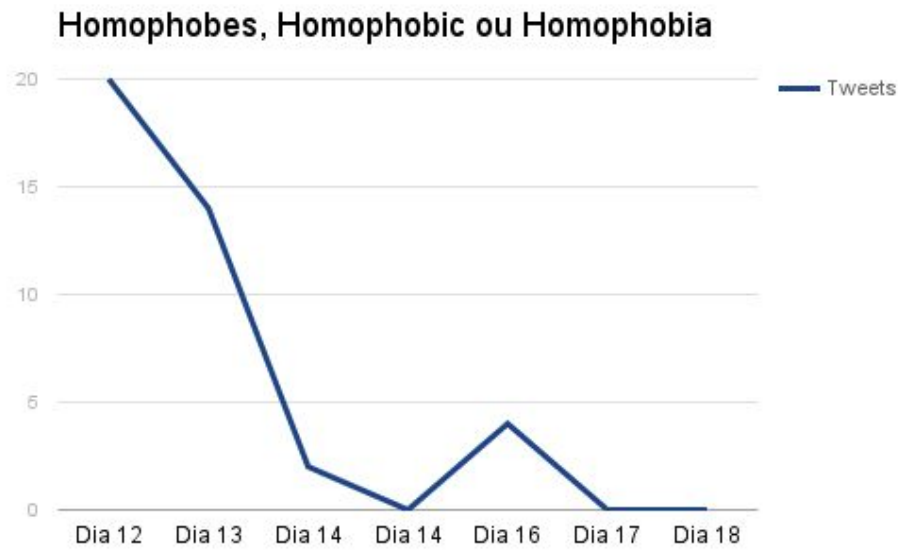

Figura 6. Tweets coletados com as tags Homophobes/Homophobic/Homophobia por dia.

Conforme figuras 7 e 8 verifica-se o padrão esperado de acordo com os dados apresentados acima. No primeiro momento o grande furor dos acontecimentos fazem com que o o número de tweets seja mais espalhado pelo globo (figura 7). Entretanto, nota-se a

tendência de que o assunto torne-se menos comentado com o tempo, dessa maneira como pode ser visto na figura 8, o número de país torna-se muito menor.

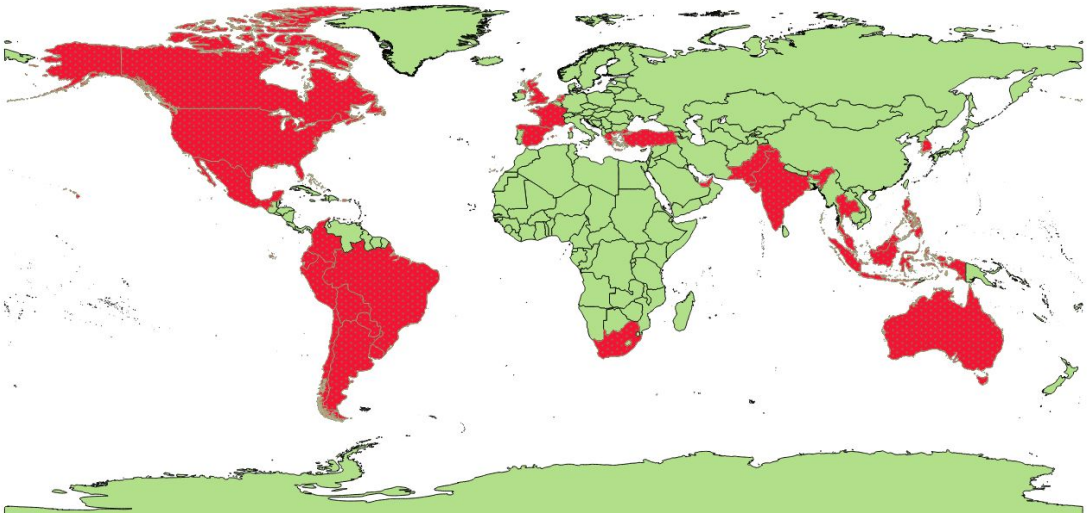

Figura 7. Países com tweets nos dias 12 e 13 de junho.

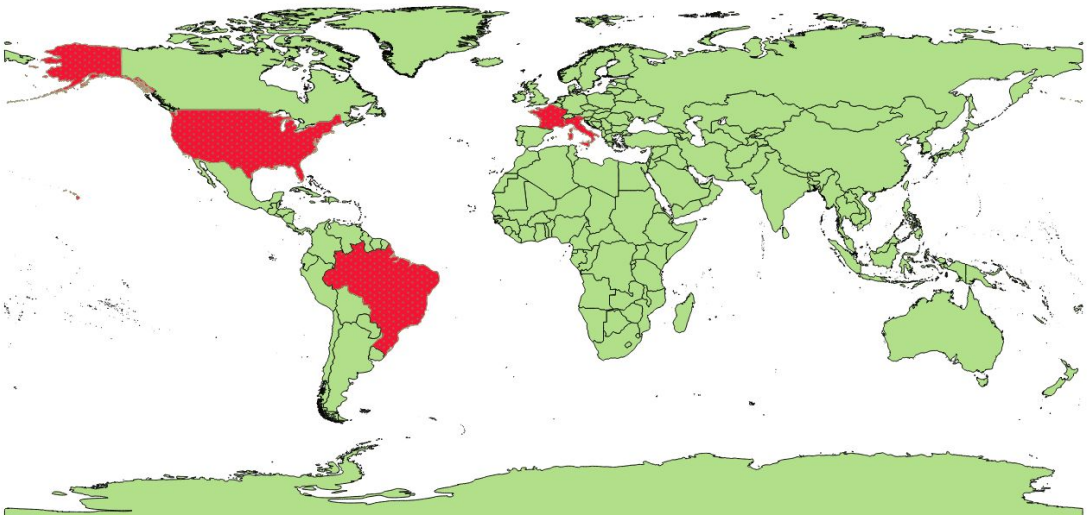

Figura 8. Países com tweets no dia 18 de junho.

País	Tweets
Brasil	3
EUA	67
França	1
Itália	1

Tabela 1. Tweets por país nos dias 18 de junho.

País	Tweets	País	Tweets
África do Sul	1	França	6
Argentina	7	Grécia	1
Austrália	6	Holanda	1
Bahamas	1	Índia	2
Bolívia	4	Indonésia	1
Brasil	6	Inglaterra	9
Canadá	11	Malásia	6
Chile	3	México	18
Colômbia	3	Paquistão	1
Coréia do Sul	7	Paraguai	1
E. Árabes Unidos	1	Peru	3
Equador	3	Porto Rico	2
Escócia	2	Singapura	1
Espanha	2	Tailândia	3
EUA	206	Turquia	2
Filipinas	3	Uruguai	1

Tabela 2. Tweets por país nos dias 12 e 13 de junho.

Naturalmente os Estados Unidos lideram o ranking de tweets relacionados ao ataque em orlando. Porém vários países diferentes puderam ser notados nas coletas de dados, principalmente nos primeiros dias. Além disto, podemos perceber uma queda rápida nas citações aos ataques, tanto em volume de tweets quanto em número de países diferentes abordando o assunto.

O assunto relacionado ao terrorismo, homofobia e estado islâmico parece ter desaparecido após o dia 17/06/2016. A palavra “Orlando”, mesmo em contextos genéricos, foi ficando menos popular. Os tweets encontrados no dia 18/06/2016 parecem não estar relacionados aos ataques, fato que iremos discutir na conclusão deste estudo. Lembrando sempre de quão filtradas são nossas buscas, pela API e pelo baixo número de tweets georreferenciados.

As conclusões que chegamos dizem respeito ao dados coletados, e não podem ser traduzidos para toda a comunidade da rede social, mas a aleatoriedade das coletas podem nos dar uma ideia em escala reduzida, de como a rede social se comporta.

4. Conclusão

Neste trabalho, abordamos o assunto Dados geográficos em redes sociais, com um foco especial no Twitter, e observamos que as redes são uma fonte maciça de dados, permitindo estudos detalhados sobre, dentre outros tópicos, o comportamento dos usuários dado um acontecimento marcante.

Apesar das limitações encontradas e descritas anteriormente, acreditamos que cumprimos boa parte dos objetivos propostos em relação a análise dos dados e a proposta do trabalho.

Este trabalho foi muito importante para o aprofundamento dos conceitos vistos em sala de aula e outros conhecimentos gerais, como acesso à uma API. Sendo assim, a realização do trabalho permitiu que trabalhássemos com diversas ferramentas visando a coleta, tratamento e análise de *tweets* georreferenciados.

Talvez pela pelas restrições impostas pela API e pelo baixo número de *tweets* georreferenciados, a pesquisa apresentou alguns dados destoantes. No dia 17/06 foram feitas diversas tentativas de se coletar os dados referentes aos atentados no Twitter. Porém não obtivemos sucesso.

Quantitativamente, podemos observar um súbito aumento no número de *tweets* no dia 16/06/2016, tanto com a palavra originalmente pesquisada "Orlando", quanto nas demais palavras que foram processadas nos textos dos *tweets*. Ao realizar algumas pesquisas, não conseguimos explicar exatamente o que ocorreu, podendo ser apenas fruto de uma aleatoriedade na aquisição dos dados. Entretanto no dia 16/06/2016 o atual presidente dos Estados Unidos, Barack Obama decretou que todas as bandeiras de estabelecimentos não privados, fossem hasteadas à meia altura em todo o país até o por do sol. Além disso o presidente também visitou a cidade de Orlando para prestar sua homenagem às vítimas e visitar as famílias das mesmas.

Finalmente, no último dia da coleta de dados 18/06/2016 percebemos uma forte queda em comparação ao período. Representando pouco mais de 1% dos *tweets*, num período de sete dias. Este dado talvez não se torne tão estranho se observarmos as incidências das palavras processadas nos textos do Twitter. No dia 18/06/2016 já não se falava mais a respeito do atentado em Orlando. Na maioria dos casos os *tweets* abordavam assuntos como viagens, faculdades e principalmente futebol.

Mesmo levando em consideração o estreitamento dos dados causados pelas restrições da API e pela baixa incidência de *tweets* georreferenciados, podemos perceber um caráter de efemeridade do engajamento dos usuários nas redes sociais. O chamado efeito manada se manifestou com muita força nos primeiros cinco dias, e rapidamente perdeu força.

Tendo dito isto, cogitamos também uma outra hipótese para a ausência de dados coletados no dia 17/06/2016. Como no dia 18/06/2016 não existiu nenhum assunto relacionado aos atentados, inferimos que o assunto deixou de existir do dia 17 em diante, o que explica a ausência de dados. O pequeno número obtido no dia 18/06/2016 com a palavra "Orlando" de deu principalmente ao fato de ter ocorrido um jogo de futebol entre os times Orlando City contra SJ Earthquakes na cidade de Orlando. Partida que terminou empatada em 2 a 2.

Referências Bibliográficas:

Bozzanini, Marco, Mining Twitter Data with Python. Disponível em <https://marcobonzanini.com/2015/03/02/>, acessado em 14 de Junho de 2016;

IETF Geographic JSON Working Group. GeoJSON. Disponível em <http://geojson.org>, acessado em 14 de Junho de 2016.

Twitter API Overview. Disponível em

<https://dev.twitter.com/overview/api>