Introdução

Boa tarde, me chamo Arthur e vou apresentá-los o plano de trabalho que realizei na Iniciação Científica:

ElasticSearch e linguagem R para análise de violência política na plataforma Telegram

Objetivos da Apresentação

Meus objetivos nessa apresentação são três:

- Breve contextualização da pesquisa, para que seja possível tratar do meu plano de trabalho em si.
- Expôr rapidamente a “evolução” dos métodos utilizados na pesquisa e, por fim,
- Apresentar as potencialidades do R e do ElasticSearch numa pesquisa em ciências sociais computacionais, visto a ênfase do plano de trabalho na questão metodológica.

Resumo da Pesquisa

A pesquisa surge da observação de determinados fenômenos ocorridos pela ou na plataforma Telegram, que já tem o histórico de ser uma plataforma que propicia processos de radicalização e diversas atividades ilegais visto seus mecanismos de automação e privacidade, mas o interesse principal da pesquisa surge pelo impacto do Telegram nos anos de 2020 e 2021 no Brasil, como a presença do aplicativo em mais da metade dos celulares no país [^1] unida à movimentação ou retorno ao app por figuras da direita [^2], a suspenção das contas de Donald Trump no Instagram e no Facebook [^3] que acabou minando a imagem do WhatsApp, as atualizações de privacidade e compartilhamento realizadas na época… enfim, uma série de fatores que influenciaram na movimentação da direita e da extrema-direita no Telegram. E a pesquisa começa oficialmente em junho de 2021!

A ideia central da pesquisa é compreender o papel do Telegram na produção e compartilhamento de desinformação em chats brasileiros de extrema-direita.

Tive que arrancar uma parte da apresentação por conta do tempo, mas é importante destacar que estamos tratando de dados não estruturados, que englobam textos, vídeos, áudios e imagens. Então, é interessante que façamos a pesquisa numa abordagem multimetodológica, quanti e quali, e, na análise quantitativa, temos o `R`.

Análise Quantitativa com R

R é um sistema gráfico e estatístico composto de duas partes: a própria linguagem R e um ambiente de software

Mas não interessa tanto discutir a definição do `R`, mas entender que é ele quem permite, por exemplo, a

Mineração de Dados

Que é o estudo da coleta, limpeza, processamento e análise dos dados, que culmina no processo de

ETL

- Extração: Seja por raspagem da web ou requisições via API
- Transformação ou manipulação dos dados: que diz respeito ao tratamento dos dados ou construção de variáveis que contribuam à análise do objeto, ou seja, a construção de categorias analíticas, sendo uma decisão metodológica e não meramente computacional;
- Armazenamento: que também leva em conta uma série de questões computacionais, como o tamanho dos dados, e, de novo, metodológicas, pois o formato dos dados refletem neles mesmos e, a depender, o formato implica algumas dificuldades numa próxima análise.

Visualização com R

O R comporta o `ggplot2`, um sistema de criação declarada de gráficos, e o `Shiny`, uma pacote voltada a construção de dashboards interativos ou webApps.

Resumo do R

Então, o R entra precisamente nessa parte da análise exploratória, visualizar medidas de posição (“são utilizadas para resumir, em um único número, o conjunto de dados observados da variável em estudo”) e dispersão (“variabilidade” dos dados), na mineração de dados e na visualização gráfica.

O R também permite análise qualitativa mas não é a melhor ferramenta pra isso.

Análise Qualitativa com Atlas.ti

Na análise qualitativa temos o `Atlas.ti`, que é um

- Software de análise qualitativa de dados com o auxílio do computador, que lhe permite analisar grandes volumes de dados
- Seus principais recursos, como as codificações, as famílias e as “redes de conexões” dialogam diretamente com a abordagem da teoria fundamentada.

Porém, dada as condições da pesquisa, foram necessárias algumas melhorias nos métodos, tanto no ETL quanto na análise qualitativa.

Melhorias no método: ElasticSearch

E para entendermos isso vou destacar duas

Particularidades do objeto

Em primeiro temos

- “Apagões” frequentes, isto é, a partir do momento em que o discurso de ódio ou o conteúdo desinformativo cumpriu seu “papel social”, produziu o efeito que se pretendia, estas mensagens são apagadas, e isso é uma prática que ocorre constantemente, algo próprio da dinâmica desses públicos

E também o

- Nascimento e morte de grupos e canais que, quando não são deletados mesmos, frequentemente assumem uma postura de “zumbi”, como chamamos na pesquisa. Ou seja, determinado chat para de enviar mensagens, cessa o funcionamento, e “nasce”, um novo que contêm nome ou conteúdo semelhante, sendo isso uma forma de driblar a vigilância do conteúdo compartilhado e o rastreamento dessas pessoas

Então, como se não bastasse lidarmos com grandes volumes de dados, produzidos diariamente, são dados efêmeros, voláteis a ponto de coletarmos um chat hoje e amanhã nada daquilo existir.

Disso temos as

- Limitações do `R` e do `Atlas.ti`

No `R` resumidamente, para tanto o ETL quanto qualquer análise exploratória , é necessário um mínimo de conhecimento que, para quem nunca programou, é um grande obstáculo; No sentido do `Atlas.ti` , por mais que ele seja voltado à análise de grandes bases de dados, ele não te oferece uma visualização tão boa desses dados de forma geral, e ele náo é a ferramenta ideal para lidarmos com dados tão efêmeros e que exigem não só a coleta e o armazenamento, mas a análise em “tempo real”.

Para demonstrar rapidamente essas limitações, eu separei dois exemplos:

- Esse é o código necessário pra um gráfico de barras minimamente bonito no R
- E uma imagem genérica da interface do Atlas, que tem ao lado os documentos primários selecionados e à direita os trechos codificados, com essa nuvem de palavras

Retomando a questão, é justamente na visualização dos dados, na forma na qual os acessamos que o Atlas encontra limitações

Assim chegamos no uso do

ElasticSearch & Kibana

`ElasticSearch` é um bancos de dados não estruturados, que funciona conjuntamente ao `Kibana` , responsável pela construção de gráficos e dashboards.

E o que eu quero mostrar aqui é como o ElasticSearch (e o Kibana) dão conta, em boa medida, dessas limitações que falei, e mostrar a potencialidade do ElasticSearch numa pesquisa em Ciências Sociais Computacionais, expondo os impactos dessas ferramentas, o que elas possibilitam para análise de violência política no Telegram, por exemplo.

Bem, a estrutura do Elastic permite uma

- Melhoria considerável no processo de ETL, especialmente no armazenamento em tempo real, que supera ao problema da inserção dos dados no Atlas.ti para análise qualitativa

E isso facilita “mapear”, com precisão, o nascimento e morte de grupos, canais e de determinados tópicos que estejam sendo discutidos na Internet; e uma

- visualização e anaálise qualitativa otimizada

- Visualização otimizada

Essa é uma imagem que compila os recursos ofertados pelo ElasticSearch e o Kibana, mostrando que você pode fazer aquele mesmo gráfico de barras como aquele com alguns cliques; Além disso, a interface da plataforma lhe permite modificar o tipo de gráfico, os parâmetros, legendas, cores, tudo isso, na mesma tela sem grandes dificuldades, qualquer pessoa consegue utilizar o ElasticSearch e o Kibana sem qualquer conhecimento de **programação**.

- Análise qualitativa otimizada

E, no âmbito da análise qualitativa, o ElasticSearch facilita a visualização dos dados, como é possível perceber nessa seção (*data explore*), onde mostra os dados e já lhe oferece gráficos de barras simples; Na barra de busca, você pode pesquisar por expressões específicas em toda a base de dados e, a depender das variáveis que foram construídas no processo dito anteriormente, você pode buscar por links, imagens, nome de grupos e canais… Temos também a ferramenta de Períodos de tempo, algo essencial para a análise e que, pelo `R` exige conhecimentos computacionais e pelo `Atlas.ti` uma nova seleção dos documentos primários.

Guilherme Boulos

E, finalmente, pra destacar as potencialidades do ElasticSearch em análise de violência política no Telegram, eu separei um exemplo usando os dados da pesquisa da seguinte forma:

- Busquei pela *query* `all_text: "guilherme boulos" or boulos`

all_text é uma variável que comporta todos os tipos de textos numa mensagem, então tanto textos isolados quanto legendas e links são considerados, e o nome divido em duas expressões. Um detalhe sobre isso é que o ElasticSearch não é “caso sensitivo”, ele não diferencia maiúsculas e minúsculas.

- No período de tempo, achei interessante buscar por essa expressão uma semana antes e depois das eleições municipais

A busca, nessas condições, me retornou 7.121 mensagens.

Visualização básica no ElasticSearch

De forma geral, uma visualização básica pelo elasticsearch envolve três etapas:

- Seleção das variáveis de interesse
- Quando clicamos na variável aparece essa janela que, por padrão, mostra os top cinco valores daquela variável, então são as cinco mensagens mais compartilhadas
- E quando clicamos em `Visualize` somos direcionados ao `Lens` um editor de arrastar e soltar que faz uma mediação do Elastic e o Kibana

No canto superior esquerdo temos o tipo de visualização, aqui a nossa tabela com a variável `all_text` e a quantidade de compartilhamentos de cada mensagem, e na direita temos as configurações da tabela.

Mas isso é muito básico, é bem o começo da análise, o que mais interessa aqui é a construção de um dashboard, isso que nos direciona bem na análise, então eu construí um dashboard do Boulos nesses `chats`:

Dashboard!

A combinação das ferramentas

E, só pro R não parecer todo xoxo, o elastic é uma boa ferramenta para construir a base da análise, tanto quanti quanto quali, mas operações mais complexas como análise automatizada de texto, análise de sentimento, codificação (organizada) das mensagens, a visualização de medidas de posição e dispersão, um dashboard que não dependa de internet, esse tipo de coisa só podem ser feitas com o uso do r, python, atlas.ti, fora que pra chegarmos nesse nível de “facilidade” de manipulação dos dados por meio do elasticsearch, há todo um processo computacional por trás que sustenta isso, a pesquisa, por exemplo, usa de um framework em python pra interagir com api do telegram, extrai os dados, transforma, armazena no postgresql… ou seja, o elasticsearch só é benéfico dessa forma quando usado em conjunto, de forma isolada ele não é tão “uau” assim.

Contribuições da Pesquisa

Além de expôr o uso dessas ferramentas numa pesquisa, que é o que fiz no caso do Boulos, eu separei alguns trabalhos do laboratório para que todos tivessem acesso à análises mais desenvolvidas com o uso dessas ferramentas.

Conclusão

Deixarei as referências à consulta numa página separada, e me despeço e agradeço pela atenção!