Um Processo para Extração de Dados em Mídias Sociais para Detecção de Reações Adversas a Medicamentos

Luan Nascimento, Lucas Carvalho, Matheus Souza, Renato Mauro, Kele Belloze[1]

[1]Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ)
Rio de Janeiro, Brasil

{luan.nascimento,lucas.carvalho.2,matheus.felipe}@aluno.cefet-rj.br

{renato.mauro, kele.belloze}@cefet-rj.br

Abstract. *Pharmacovigilance is an area responsible for, among many topics, detecting adverse drug reactions (ADRs). Such reactions can be reported in systems designed for this purpose. However, social media such as Twitter can reveal various notifications. This work aims to present a process for extracting data from social media to support the detection of ADRs. This work could help future research related to analyzing textual data extracted from social media to detect ADRs for the benefit of Brazilian pharmacovigilance.*

Resumo. *Em farmacovigilância, estuda-se, dentre muitos tópicos, reações adversas à medicamentos (RAMs). Tais reações podem ser notificadas em sistemas próprios para esta finalidade. Contudo, as notificações podem ser visualizadas sendo feitas em mídias sociais como o Twitter. Este trabalho possui como objetivo apresentar um processo para extração de dados em mídias sociais para apoiar a detecção de RAMs. Almeja-se que esse trabalho auxilie pesquisas futuras relacionadas à analise dos dados textuais extraídos de mídias sociais para detecção de RAMs em benefício da farmacovigilância brasileira.*

1. Introdução

O desenvolvimento de novos medicamentos evolui a todo instante. É um processo rigoroso que envolve diversas etapas até assegurar que um medicamento é seguro para utilização pela população. Este processo é divido em duas grandes fases, a pré-clínica e a clínica [Lombardino and Lowe 2004]. A fase pré-clínica se inicia desde a concepção do projeto do medicamento em questão e se estende passando por testes *in vitro* (ensaios laboratoriais sem o uso de seres vivos) até os testes em animais.

A fase clínica é divida em quatro fases. Na primeira fase, o medicamento é testado pela primeira vez em humanos em um grupo pequeno de voluntários, geralmente sadios, para testar a segurança do medicamento em questão. Na segunda fase é utilizado um número maior de pessoas, mas uma amostra ainda considerada pequena. Nessa fase são utilizadas pessoas com alguma enfermidade ou condição patológica para verificar a segurança a curto-prazo e a eficácia do medicamento. Utilizando grupos de pacientes grandes e variados, a fase três visa determinar o risco/benefício a curto e longo prazo do medicamento assim como estabelecer um comparativo com os padrões já existentes. Mesmo após a comercialização, um medicamento deve continuar sendo pesquisado e estudado, possibilitando a observação de seus efeitos a longo prazo. A fase quatro, consiste em observar o surgimento de sintomas provenientes da utilização do medicamento

em questão [CNS 1997]. Estes sintomas não previstos nos testes clínicos, são chamados reações adversas a medicamentos (RAM) [Pirmohamed et al. 1998].

Atualmente, a Anvisa (Agência Nacional de Vigilância Sanitária) no Brasil, assim como diversos países membros do Programa de Monitoramento Internacional de Medicamentos pelo *Uppsala Monitoring Center (UMC)*, centro colaborador da OMS [Anvisa 2019], utiliza o VigiFlow, chamado aqui no Brasil de VigiMed, sistema que recolhe notificações de possíveis RAMs. O sistema é mais utilizado por profissionais da área da saúde, contudo, ele é pouco conhecido pela população em geral. Entre dezembro de 2018 a dezembro de 2019 o maior índice de utilização do VigiMed foi por farmacêuticos; cerca de 78% dos relatos foram feitos por eles [Vogler et al. 2020], mostrando assim uma baixa utilização por parte dos cidadãos comuns.

Embora sistemas de notificação sejam o principal meio de monitoração de eventos adversos e sejam praticamente indispensáveis na detecção de RAMs, ainda há clara necessidade de um modo de vigilância que se mostre mais ativo [Organization et al. 2002]. Em hospitais e similares, onde a observação de sintomas pode ser mais fácil, pode ocorrer como parte de um processo interno, porém, ao usuário comum que desconhece a possibilidade e importância deste relato, a informação observada pode acabar se perdendo. Tendo isso em vista, para aumentar a captura de informações relevantes sobre RAM, as mídias sociais podem ser uma grande fonte de informação útil, com milhões de usuários espalhados pelo mundo inteiro compartilhando informações pessoais a todo momento, abrindo assim a possibilidade de utilizar este grande alcance em favor da farmacovigilância.

Dado este cenário, este projeto apresenta um processo para extração de dados em mídias sociais para apoiar a detecção de RAMs. É construído um processo de extração, transformação e carga (ETL - *Extract, Transform and Load*) aplicado na rede social Twitter com foco na língua portuguesa do Brasil. Além do processo de ETL, este trabalho apresenta como contribuição a organização do *dataset* de medicamentos disponibilizado pela Anvisa. Espera-se que este trabalho apoie pesquisas futuras na área de modo a encontrar informação relevante nas mídias sociais no contexto da população brasileira.

O restante deste artigo está organizado como segue. As seções 2 e 3 apresentam de maneira resumida os principais conceitos utilizados no trabalho e os trabalhos relacionados, respectivamente. A seção 4 descreve o processo proposto. A seção 5 expõe a avaliação experimental realizada. Por fim, a seção 6 apresenta a conclusão e trabalhos futuros.

2. Fundamentação Teórica

2.1. Reações Adversas a Medicamentos

As reações adversas a medicamentos, são definidas pela Organização Mundial da Saúde (*World Health Organization*) como qualquer efeito nocivo, não intencional, resultante da utilização de medicamentos em doses permitidas com objetivo de prevenir, diagnosticar e tratar alguma condição do usuário [Saff 2018] [Zhang et al. 2020].

2.2. Medicamentos Registrados no Brasil

A ANVISA disponibiliza, por meio do portal de dados abertos do Governo Federal um *dataset* composto de todos os medicamentos registrados no Brasil [Anvisa 2020]. Esse

dataset, utilizado neste trabalho, contém o registro histórico de medicamentos cadastrados junto a entidade reguladora, com dados como data de cadastro e validade do registro, categoria e empresa detentora do registro do fármaco.

2.3. Twitter API

Twitter API[2] é a API do twitter que permite a extração de dados da plataforma de diversas formas como em tempo real ou busca histórica. Esta API permite a obtenção de *tweets* públicos de maneira oficial. A extração desses dados deve ser filtrada e com regras estabelecidas durante as chamadas para atender a demanda do sistema que está sendo desenvolvido. Para acessar esta API é necessária uma conta de desenvolvedor e autorização concedida pela equipe do Twitter.

2.4. Extração, Transformação e Carga

Extração, Transformação e Carga (ETL, do inglês *Extraction, Transform and Load*) é um processo cujo objetivo é extrair dados de uma ou múltiplas fontes externas à sua aplicação, realizar transformação ou limpeza dos dados aplicando regras de negócio ou remapeando informações e finalmente a carga dos dados tratados em *Data Lakes*, *Data Warehouses* ou outras formas de armazenamento desse resultado [Ferreira et al. 2010].

3. Trabalhos relacionados

Esta seção apresenta a busca por trabalhos relacionados, a qual permitiu conhecer as metodologias para lidar com diferentes *datasets*, além da extração e análise de dados. É dado o enfoque para a extração dos dados. Grande parte dos trabalhos analisados apresentam justificativas de suas escolhas tanto de ferramentas como de abordagens e métodos utilizados durante seu desenvolvimento. Desta forma recorremos à experiência destes trabalhos para levantar o estado da arte e os pontos positivos e negativos de cada método por eles utilizados, e assim traçar o caminho desta pesquisa baseado nestes fatores.

Como base para essa pesquisa foram utilizados 44 artigos relacionados, pesquisados na base Scopus em março de 2021 a partir da seguinte string de busca: *pharmacovigilance AND ("adverse drug reaction"OR "adverse drug event") AND (twitter OR tweet)*.

Sobre o processo de extração de dados, algumas divergências foram encontradas principalmente em relação aos trabalhos mais recentes. A grande maioria dos casos cita a API do Twitter [Koutkias et al. 2017], [Duval and da Silva 2019], [Plachouras et al. 2016].

Existem algumas estratégias abordadas no processo de extração do Twitter que são fornecidas pela própria API como Regras. São semelhantes a um filtro de busca avançada na própria página do twitter, porém com uma gama maior de possibilidades[TwitterDev]; Há uma lógica para a manipulação dessas regras utilizadas em buscas dentro de uma conta de desenvolvedor. Elas visam atuar como um filtro para as informações coletadas, de modo a minimizar os dados indesejados e diminuir o trabalho de processamento destes dados. Caso não seja realizado este processo, *tweets* com propagandas, replicações de *tweets* por outras pessoas, e *tweets* com *links* de propagandas, podem ser recuperados no momento da extração.

[2]https://developer.twitter.com/en/docs

4. Metodologia

A metodologia tem seu início na obtenção e limpeza do *dataset* de medicamentos registrados no Brasil, disponibilizado pela Anvisa. A limpeza teve como objetivo remover os medicamentos redundantes, medicamentos com outras semânticas além de medicamentos e colunas sem utilidade para o trabalho. O processo de implementação do projeto foi definido como pode ser observado na Figura 1. O projeto foi desenvolvido em Java, com o *framework* Spring e utiliza Kafka[1]. O Kafka foi escolhido vislumbrando futuros usos, para uma eventual busca e monitoramento no twitter em tempo real, com performance e escalabilidade. Na sequência, é detalhada como foi implementada cada etapa do processo; O código fonte deste trabalho pode ser encontrado em repositório do GitHub[2].

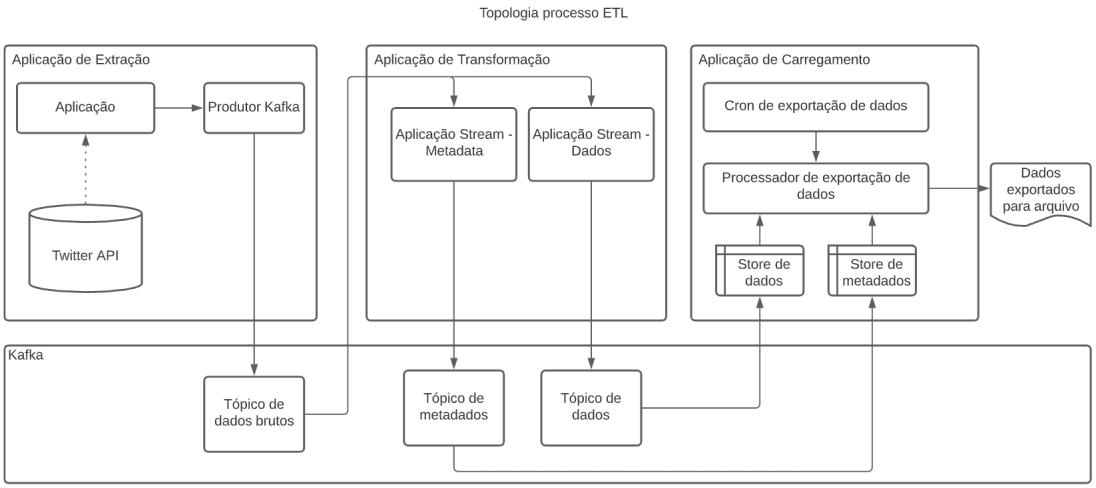

Figura 1. Processo do projeto referente a cada etapa do processo de ETL.

4.1. Camada de Extração (*Extraction*)

A camada de extração do projeto tem como principal objetivo a obtenção de dados. Muitas mídias sociais disponibilizam ferramentas e APIs para a extração de dados. No caso dessa aplicação, foi utilizada inicialmente a API do Twitter para busca recente, que permite consulta nos *tweets* em até uma semana anterior. A API do Twitter tem diversas limitações que encorajam o uso mais assertivo da API, como limitações de extrações mensais e de requisições por intervalos. Essas limitações são devido ao possível alto volume de dados e poder de processamento necessário para disponibilizar os *tweets* a todos os clientes da API. Neste trabalho foi definida para a consulta um filtro básico baseado na presença de palavras chaves no corpo do *tweet*. Essa filtragem inicial permite dados mais relevantes para as aplicações que consomem os dados, considerando o alto volume de *tweets*. Otimizar a consulta para não consumir a franquia de *tweets* extraídos com aqueles que potencialmente não serão relevantes é essencial. O diagrama de sequência exibido na Figura 4.1 sumariza a camada de extração.

[1]https://kafka.apache.org/

[2]https://github.com/LucasPLTC/TCC_CEFET_RJ_ADR_EXTRACTION_PROCESS

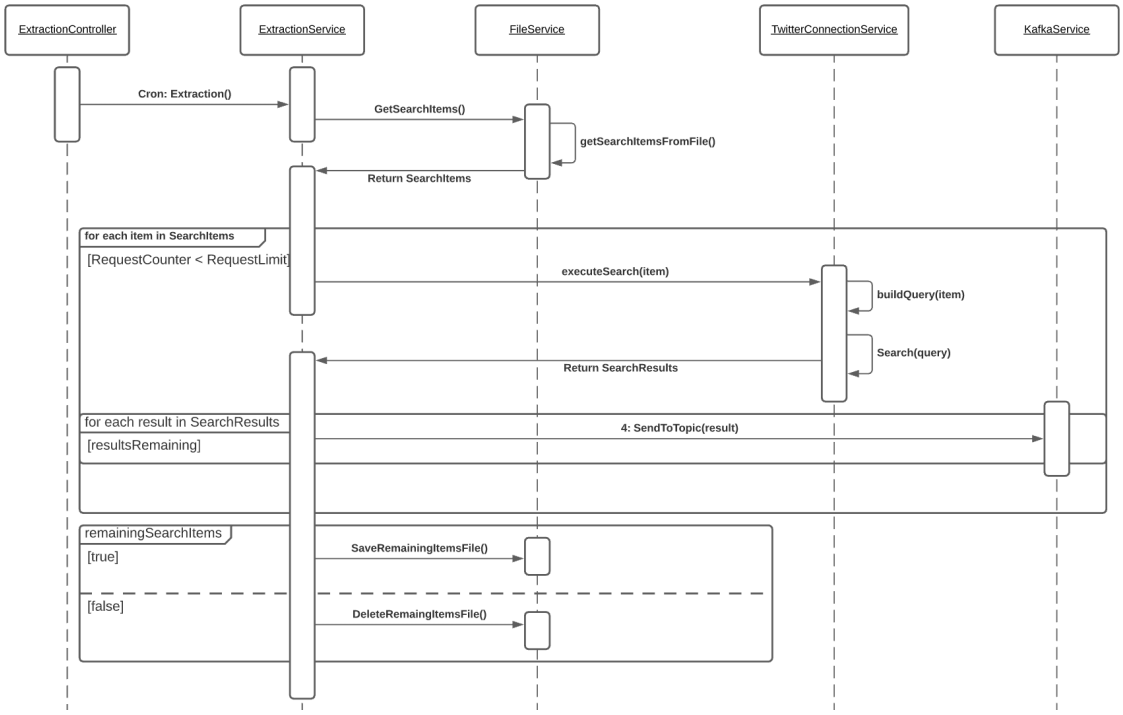

Figura 2. Diagrama de sequência do processo de extração.

1. Início do processo
   O processo inicia com uma tarefa agendada que ocorre a cada 20 minutos a partir do ExtractionController. O motivo para esse tempo é devido a API do Twitter possuir uma taxa máxima de 450 chamadas a cada 15 minutos, e é necessário aguardar o *reset* desta taxa antes de iniciar a extração novamente. Em alguns casos, esse *reset* pode levar de 1 a 2 minutos para ser propagado.
2. Medicamentos para pesquisa
   Os medicamentos são importados após a limpeza descrita previamente. Como são muitos medicamentos, é impossível realizar a consulta com todos os medicamentos de uma única vez. O arquivo de *checkpoint* contém os medicamentos ainda não pesquisados. Ao decidir de onde lerá os dados, o FileService retorna ao ExtractionService uma lista com os medicamentos para consulta.
3. Consulta a API do twitter
   A API do twitter tem diversas restrições, como quantidade de *tweets* extraídos mensalmente e intervalo entre as chamadas, o que torna necessário acompanhar a quantidade de chamadas que serão realizadas conforme o limite dependendo do tipo de conta de desenvolvedor associada ao projeto .Além da restrição de chamadas a API, também é importante observar a restrição de tamanho da resposta, de 10 a 100 *tweets* por página da resposta. Essas restrições impõem a condição da pesquisa resultar entre 200 a 300 medicamentos pesquisados na API a cada 20 minutos. Por isso o FileService precisa verificar a existência de um *checkpoint* antes de retornar ao ExtractionService. A consulta ocorre pelo nome do medicamento acompanhado da palavra “tomei”, um de cada vez, com as restrições de serem *tweets* em português e que não sejam *retweets*. Novamente, o motivo dessas restrições é a quantidade de dados que podem ser importados via API. Os dados

obtidos para os medicamentos presentes na consulta são então enviados ao tópico de dados extraídos do Kafka pelo KafkaService. Uma vez alcançado o limite de chamadas ou não restando mais medicamentos a serem pesquisados, o processo de consulta é terminado.

4. *Checkpoint* de medicamentos
   Conforme a execução das consultas é terminada, é importante verificar a existência de medicamentos que não foram utilizados para consulta. Como é impossível fazer mais de 450 chamadas dentro do limite de tempo da API, foi necessário criar um arquivo de *checkpoint* para que a próxima execução iniciasse a partir destes medicamentos que ainda não foram pesquisados. Para tal, é criado um arquivo de medicamentos temporários que tem removidos os itens que já foram consumidos para consulta. Também é importante remover esse arquivo após todos os medicamentos terem sido corretamente pesquisados.

4.2. Camada de Transformação (*Transform*)

Esta etapa do ETL tem como objetivo realizar a transformação dos dados vindos do tópico de dados extraídos. Ela consiste de duas aplicações *stream* que reagem a entradas no tópico, gerando metadados e textos para exportação na etapa de carga. Todo esse processo é feito reativamente pelo Kafka, não sendo necessário nenhum comando para que seja realizada a transformação dos dados. Também é importante destacar que como esta aplicação não depende diretamente da aplicação de extração, é possível realizar a extração em um momento e a transformação em outro. Isso é importante para que dados não sejam perdidos por indisponibilidade da aplicação.

1. Tópico de dados extraídos
   Esse tópico kafka é utilizado como entrada para ambos os processadores de dados e metadados. Este tópico tem como chave o ID dos *tweets*, e como valor as informações com texto dos *tweets*, o medicamento que levou aquele *tweet* a ser encontrado e a data de extração, em formato JSON.
2. Processador de dados
   O processador de dados tem como principal função extrair os textos dos *tweets* para facilitar a exportação. Para tal ele obtém os dados advindos do tópico de dados extraídos e os transforma em um novo par chave-valor que tem como chave o ID do *tweet* e como valor apenas o texto do *tweet*, em formato *string*. A distinção entre este tópico e o tópico de dados extraídos, é que um possui apenas o texto do *tweet* e outro possui dados como a data da extração e os medicamentos encontrados no *tweet* extraído; O que facilita na exportação, pois limita os dados da exportação sem precisar de lógica relacionada a leitura na etapa de carga.
3. Processador de metadados
   Este processador foi adicionado posteriormente e utilizado principalmente para métricas relacionadas com a quantidade de medicamentos exportados. Ele utiliza como entrada o tópico de dados extraídos, porém esse tópico salva valores do tipo *long* e usa os nomes dos medicamentos como chave. O motivo para isso é facilitar a contagem de *tweets* obtidos por medicamento.

4.3. Camada de Carga (*Load*)

A aplicação de carga consiste na exportação dos dados para CSV. Nesta etapa, os tópicos de dados e de metadados são consolidados de modo a gerar dois arquivos, um de da-

dos, que contém os textos dos *tweets* obtidos, e o de metadados que é uma agregação e contagem da quantidade de *tweets* obtidos por medicamento.

Esta aplicação tem também os controladores para exportação dos dados e metadados. Os metadados foram usados apenas para controle e avaliações na busca dos dados, e portanto não são exportados automaticamente. Os dados no entanto, são exportados semanalmente. Com todas as aplicações rodando, o processo de obtenção de dados e a geração dos *datasets* é realizado automaticamente, desde a extração de dados advindos da API do Twitter até a exportação dos dados que também pode acontecer de maneira manual. Com a estrutura baseada em Kafka, é possível realizar cada etapa em um momento diferente ou simultaneamente.

5. Avaliação Experimental

A limpeza realizada no *dataset* de medicamentos removeu colunas não relevantes para esse trabalho permanecendo apenas as colunas ”Nome”e ”Data”. Ainda, a limpeza removeu medicamentos redundantes e medicamentos cujos nomes apresentavam outras semânticas, como nomes de pessoas, lugares ou nomes curtos equivalentes a siglas. Essa última limpeza foi realizada após a primeira execução do processo, que apresentou inúmeros ruídos por esta questão. Ao final, a lista de medicamentos foi reduzida de aproximadamente 30.000 para 5.783 registros.

Ademais, houve a construção de um *dataset* de *tweets* contendo menções a medicamentos. A partir de uma busca realizada no dia 12/10/2021 e considerando resultados de até uma semana antes, uma análise quantitativa expôs um *dataset* com 1.414 tweets, totalizando 29.008 palavras. Houve um total de 186 medicamentos encontrados. Destes, os mais populares foram: Dorflex®, Rivotril® e Vitamina com 202, 201 e 153 ocorrências, respectivamente.Nos resultados, ainda observamos a maioria relatos da ingestão, abrindo escopo para mais aplicações no futuro que possam utilizar esta pesquisa como base. Com uma breve análise dos resultados à luz dos dados utilizados para obtenção deste *dataset*, também é possível observar uma diferença clara na busca por todos os medicamentos, com poucos concentrando grande número de resultados.

A principal contribuição deste trabalho, no entanto, é o processo desenvolvido para a extração dos *tweets*. O ETL relatado nesta pesquisa, comprovou-se bastante modular e escalável, tanto em desempenho como em fontes de dados. Graças a este processo, o Twitter não é um requisito, e qualquer rede social pode ser utilizada neste processo, com a devida implementação da obtenção dos dados conforme a necessidade da mesma.

6. Conclusão e Trabalhos Futuros

Apresentamos um processo escalável para extração de RAM's de redes sociais e um dataset resultante desse processo. As técnicas aqui apresentadas servem como ponto de partida para diversas aplicações, como por exemplo monitoramento de surtos de doenças ou de lotes de medicamentos com falhas. A execução do processo é em fila. Com essa técnica, nenhum dado é perdido, limitado apenas pelo método de extração utilizado. Trabalhos futuros envolvendo a identificação de RAM's por meio de processamento de linguagem natural, aprendizado de máquina ou inteligência artificial (e combinações entre elas), levariam uma vantagem enorme por poupar esforços utilizando o processo apresentado nesta pesquisa como ponto de partida.

Referências

[Anvisa 2019] Anvisa (2019). Vigimed: Sistema de notificação de eventos adversos no uso de medicamentos. http://antigo.anvisa.gov.br/documents/33868/399600/VigiMed+-+Perguntas+e+respostas/04c8d69a-0650-4edd-8a20-83bbe5dbba05. Acessado em: abril/2021.

[Anvisa 2020] Anvisa (2020). Portal brasileiro de dados abertos: Medicamentos registrados no brasil. https://dados.gov.br/dataset/medicamentos-registrados-no-brasil. Acessado em: Setembro/2021.

[CNS 1997] CNS (1997). Regulamento para a realização de ensaios clínicos com medicamentos no brasil. https://bvsms.saude.gov.br/bvs/saudelegis/cns/1997/res0251_07_08_1997.html. Acessado em: abril/2021.

[Duval and da Silva 2019] Duval, F. V. and da Silva, F. A. B. (2019). Mining in twitter for adverse events from malaria drugs: The case of doxycycline. *Cadernos de Saude Publica*, 35(5).

[Ferreira et al. 2010] Ferreira, J., Miranda, M., Abelha, A., and Machado, J. (2010). O Processo ETL em Sistemas Data Warehouse. *INForum 2010 - II Simpósio de Informática*, pages 757–765.

[Koutkias et al. 2017] Koutkias, V., Lillo-Le Louët, A., and Jaulent, M.-C. (2017). Exploiting heterogeneous publicly available data sources for drug safety surveillance: computational framework and case studies. *Expert Opinion on Drug Safety*, 16(2):113–124.

[Lombardino and Lowe 2004] Lombardino, J. G. and Lowe, J. A. (2004). The role of the medicinal chemist in drug discovery—then and now. *Nature Reviews Drug Discovery*, 3(10):853–862.

[Organization et al. 2002] Organization, W. H. et al. (2002). The importance of pharmacovigilance.

[Pirmohamed et al. 1998] Pirmohamed, M., Breckenridge, A. M., Kitteringham, N. R., and Park, B. K. (1998). Adverse drug reactions. *Bmj*, 316(7140):1295–1298.

[Plachouras et al. 2016] Plachouras, V., Leidner, J. L., and Garrow, A. G. (2016). Quantifying self-reported adverse drug events on Twitter: Signal and topic analysis. In *ACM International Conference Proceeding Series*.

[Saff 2018] Saff, R. (2018). *Epidemiology of Drug Allergy*. cited By 1.

[TwitterDev] TwitterDev (?). Filtered stream:how to build a rule. https://developer.twitter.com/en/docs/twitter-api/tweets/filtered-stream/integrate/build-a-rule. Acessado em: setembro/2021.

[Vogler et al. 2020] Vogler, M., Conesa, H. R., de Araújo Ferreira, K., Cruz, F. M., Gasparotto, F. S., Fleck, K., Rebelo, F. M., Kollross, B., and Gonçalves, Y. S. (2020). Electronic reporting systems in pharmacovigilance: The implementation of vigiflow in brazil. *Pharmaceutical medicine*, 34(5):327–334.

[Zhang et al. 2020] Zhang, Y., Cui, S., and Gao, H. (2020). Adverse drug reaction detection on social media with deep linguistic features. *Journal of Biomedical Informatics*, 106. cited By 4.