

Projeto de Análise e Mineração de Mídias Sociais

Trabalho Prático 1: Coleta de Dados

Pedro Victor de Sousa Lima¹

¹Departamento de Ciência da Computação – Instituto de Ciências Exatas
Universidade Federal de Minas Gerais (UFMG)
Av. Antônio Carlos, 6627 – Pampulha – 31.270-901 – Belo Horizonte - MG – Brazil

Resumo. *Este projeto tem como objetivo desenvolver um coletor de dados para a plataforma GETTR, com o propósito de identificar o uso de linguagem ofensiva e discursos de ódio em várias formas de expressão. A coleta abrangeu conteúdos direcionados a questões de raça, cor, nacionalidade e orientação sexual. Para alcançar esse objetivo, foi utilizada a base de conhecimento do Hatebase.org, que é um compilado de termos (lexicons) curados em relação ao grau de ofensividade dos termos para posterior detecção de linguagem ofensiva. Durante a coleta de dados, foram obtidos 56.000 posts que possuíam termos ofensivos e identificados 3.400 usuários que potencialmente utilizam linguagem ofensiva e podem estar envolvidos em redes de discurso de ódio presentes na plataforma.*

1. Introdução

Com o advento das redes sociais e plataformas de mídia digital, a análise de conteúdo online tornou-se uma área de grande relevância para compreendermos as dinâmicas sociais e culturais que permeiam a internet. É crucial abordar essa questão de forma respeitosa e responsável, promovendo um ambiente online mais saudável e inclusivo para todos os usuários, utilizando métodos que levem em consideração a detecção da linguagem ofensiva e dos discursos de ódio, comumente propagados em diversas redes sociais.

Um dos principais desafios na detecção automática de discurso de ódio em plataformas de mídia social é a tarefa de distinguir o discurso de ódio de instâncias de linguagem ofensiva, que podem não necessariamente constituir discurso de ódio. Métodos de detecção lexical, que se baseiam apenas em termos ou palavras-chave específicas, muitas vezes apresentam baixa precisão porque categorizam todas as mensagens que contêm esses termos como discurso de ódio[Davidson et al. 2017].

Este projeto aborda a criação de um coletor de dados voltado para a plataforma GETTR, com o objetivo de investigar e identificar o uso de linguagem ofensiva e discursos de ódio orientados a questões de raça, cor, nacionalidade e orientação sexual. Em um cenário em que a disseminação de conteúdos prejudiciais pode afetar negativamente a convivência online, esta iniciativa se propõe a analisar textos e interações em busca de padrões que indiquem comportamentos ofensivos para posterior análise de redes de usuários que fazem utilização de discurso de ódio.

2. Obtenção de lexemas e termos ofensivos

Durante a execução inicial do trabalho, uma parte crucial do desenvolvimento envolveu a consulta à base de conhecimento Hatebase[Hatebase.org]. Essa base de dados é um

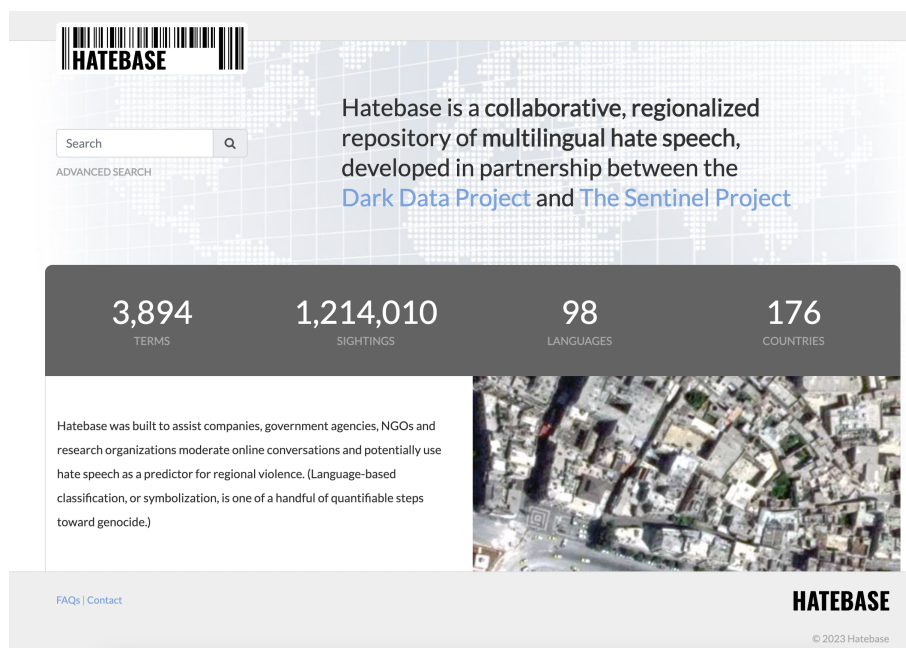


Figura 1. Hatebase.org

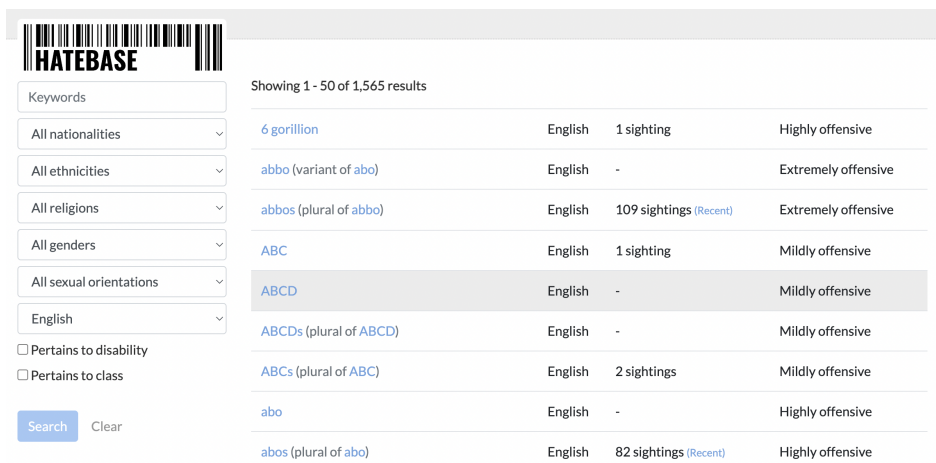
repositório abrangente de termos e lexemas linguísticos, cuidadosamente classificados de acordo com seu nível de ofensividade. Essa consulta possibilitou obter insights valiosos para identificar e categorizar, a priori, o uso de linguagem ofensiva na plataforma GETTR.

Através do Hatebase.org, foi possível extrair e coletar uma variedade de termos (*lexicons*) que representam diversas manifestações linguísticas. Esses termos são classificados por uma equipe de curadores do site com base em seu potencial ofensivo. O site determina quatro parâmetros essenciais: o próprio termo em questão, o número de vezes que ele foi visualizado, o idioma em que o termo está e, por fim, sua classificação que varia entre *suavemente ofensivo*, *moderadamente ofensivo* até *extremamente ofensivo*. Termos que não se enquadram em nenhuma das categorias de classificação são rotulados com o símbolo (-) para indicar a ausência de classificação.

Um script de web scraping foi desenvolvido com o objetivo de extrair os termos pejorativos em inglês do Hatebase.org de maneira eficiente e precisa. Além disso, o programa foi projetado para capturar não apenas os termos em si, mas também as informações essenciais relacionadas a cada termo, como as classificações de ofensividade. É importante destacar que essa implementação não deixou de fora nenhum dos 1.565 termos disponíveis no momento da extração, tendo finalmente sua representação final em um dataset no formato CSV.

3. Coletor de posts

O coletor de posts foi desenvolvido em Python, aproveitando um encapsulamento da API da plataforma GETTR, que oferece uma API pública para consulta e aquisição de dados. Este coletor apresenta duas funções essenciais: a primeira é capaz de realizar uma busca geral na API do GETTR, levando em consideração as limitações de paginação de resultados da plataforma. A segunda função se concentra na obtenção de dados de um usuário específico, incluindo suas informações, seguidores e os 200 posts mais recentes,



The screenshot shows the Hatebase.org search results page. On the left, there are filters for Keywords, Nationalities, Ethnicities, Religions, Genders, Sexual Orientations, and Language. Below these are checkboxes for 'Pertains to disability' and 'Pertains to class'. A 'Search' button and a 'Clear' link are at the bottom of the filters. The main area displays a table of results, showing terms like '6 gorillion', 'abbo', 'abbos', 'ABC', 'ABCD', 'ABCDs', 'ABCs', 'abo', and 'abos' with their respective sighting counts and offensiveness levels.

Term	Language	Sightings	Offensiveness
6 gorillion	English	1 sighting	Highly offensive
abbo (variant of abo)	English	-	Extremely offensive
abbos (plural of abbo)	English	109 sightings (Recent)	Extremely offensive
ABC	English	1 sighting	Mildly offensive
ABCD	English	-	Mildly offensive
ABCDs (plural of ABCD)	English	-	Mildly offensive
ABCs (plural of ABC)	English	2 sightings	Mildly offensive
abo	English	-	Highly offensive
abos (plural of abo)	English	82 sightings (Recent)	Highly offensive

Figura 2. Exemplo de resultado de busca no Hatebase.org

organizados em ordem cronológica reversa.

O processo de obtenção de dados foi cuidadosamente projetado para categorizar e coletar exclusivamente os posts que contenham termos pejorativos classificados como *extremamente ofensivos* ou *altamente ofensivos*. Para isso, a partir do conjunto de dados contendo os termos pejorativos obtidos na etapa anterior (a partir do Hatebase.org), foi realizado um filtro para selecionar apenas termos correspondentes, resultando em 525 termos extremamente ofensivos ou altamente ofensivos. Esses termos foram então utilizados nas chamadas à API de busca do GETTR.

4. Limitações da coleta

Assim como acontece em muitas plataformas sociais, a API pública do GETTR apresenta diversas limitações e desafios para os desenvolvedores. A falta de uma interface de aplicação bem definida dificulta o acesso aos dados da plataforma. Além disso, a plataforma possui mecanismos de segurança para detectar atividades automatizadas, como a coleta intensiva de dados por bots.

Para contornar essas limitações, o coletor foi projetado para operar com intervalos de tempo aleatórios entre as solicitações à API. Essa abordagem ajuda a evitar que a plataforma identifique o coletor como um robô, dessa forma, a paginação de resultados foi implementada para coletar posts de forma organizada e em ordem cronológica inversa, do mais recente ao mais antigo. Cada termo pejorativo é processado em iterações separadas, com um limite de 500 posts por termo, a cada consulta. Essa estratégia ajuda a controlar a quantidade de dados coletados em cada etapa e a não ser bloqueado a curto prazo.

Foram realizados testes no tocante a verificação da limitação de chamadas, porém, como não há nenhuma documentação oficial, há uma crença em comunidades on-line de desenvolvedores que o máximo de posts capazes de serem retornados por usuário é de 200. Foi possível verificar que há uma camada de segurança e interrupção de conexões caso haja um grande número de chamadas por janelas de cinco minutos. Após isto, o IP da máquina de origem é bloqueado por um firewall web da Imperva/Incapsula WAF(Web Application Firewall)[Imperva.com], deixando o coletor incapaz de retornar resultados por diversas horas e impactando a dinâmica da coleta de dados.

Existem documentações disponíveis sobre burlar o sistema[Scrapfly.io], porém, a estratégia utilizada foi implementar uma arquitetura de máquinas virtuais em nuvem, para que a cada bloqueio de origem detectado no processo de coleta, seja gerada uma nova máquina virtual com um novo IP de origem para continuar a coleta. Esta mudança de contexto entre máquinas bloqueadas e novas foi realizada manualmente e de forma supervisionada, utilizando para tal uma conta privada no Google Cloud Platform, para criar as máquinas virtuais e o banco de dados utilizado para manter os dados da coleta.

5. Modelagem e persistência de dados

A modelagem dos dados se deu de forma a manter somente um conjunto reduzido das informações retornadas pela API do GETTR. Uma vez que os dados da API consistem em vários objetos agrupados em uma única solicitação (por exemplo, informações parciais sobre o usuário, tópico e título do discurso), optou-se por selecionar apenas os seguintes atributos-chave: *txt*, *username*, *udate*, *nickname* e *lang*. Por fim, foi utilizado o banco de dados Postgres para realizar o armazenamento dos posts coletados em uma tabela, com cada atributo mapeado para um tipo de dados exclusivo. A representação desses atributos na tabela a seguir demonstra sua importância, uso e seu tipo de dados mapeado:

Tabela 1. Descrição dos Parâmetros e Tipos de Dados

Parâmetro	Descrição	Tipo de Dados
txt	Representa o conteúdo textual de uma postagem ou mensagem. Contém o texto real ou a mensagem compartilhada por um usuário.	TEXT
username	Refere-se ao identificador único ou nome de usuário associado à conta de um usuário do GETTR. Ajuda a identificar o autor de uma postagem ou mensagem.	VARCHAR(255)
udate	Significa "data de atualização" e representa o carimbo de data e hora quando uma postagem ou mensagem foi criada ou atualizada pela última vez. É retornado como um UNIX Timestamp.	BIGINT
nickname	Denota o nome de exibição ou pseudônimo escolhido por um usuário em uma plataforma. Pode ou não ser o mesmo que seu nome real e é usado para identificação e interação.	VARCHAR(255)
lang	Abreviação de "idioma", este parâmetro indica o idioma no qual a postagem ou mensagem está escrita.	VARCHAR(255)

6. Resultados da coleta

Na tabela abaixo estão contabilizadas as quantidades relativas ao número de usuários coletados e o número de posts coletados da plataforma GETTR.

Número de Usuários Coletados	Número de Posts Coletados
10365	30094

Referências

- Davidson, T., Warmley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11.
- Hatebase.org. Um repositório colaborativo e regionalizado de discurso de ódio multilíngue, desenvolvido em parceria entre o projeto dark data e o projeto sentinel. [Acesso em: 29-09-2023].
- Imperva.com. A imperva fornece segurança cibernética completa, protegendo o que realmente importa – seus dados e aplicativos – seja no local ou na nuvem. [Acesso em: 29-09-2023].
- Scrapfly.io. O scrapfly é uma api de web scraping que fornece proxies residenciais, um navegador sem interface para extrair dados e contornar sistemas de captcha e fornecedores anti-bots. [Acesso em: 29-09-2023].