Joanny Eva Pacheco Monteiro
Orientadora: Deborah Maria Vieira Magalhães

Análise de palavras-chave associadas à localização de pessoas nos desastres naturais

Picos - PI
14 de Agosto de 2023

Joanny Eva Pacheco Monteiro
Orientadora: Deborah Maria Vieira Magalhães

Análise de palavras-chave associadas à localização de pessoas nos desastres naturais

Monografia submetida ao Curso de Bacharelado em Sistemas de Informação como requisito parcial para obtenção de grau de Bacharel em Sistemas de Informação.

Universidade Federal do Piauí
Campus Senador Helvídio Nunes de Barros
Bacharelado em Sistemas de Informação

Picos - PI
14 de Agosto de 2023

FICHA CATALOGRÁFICA
Serviço de Processamento Técnico da Universidade Federal do Piauí
Biblioteca José Albano de Macêdo

M775a Monteiro, Joanny Eva Pacheco
Análise de palavras-chave associadas à localização de pessoas nos desastres naturais [recurso eletrônico] / Joanny Eva Pacheco Monteiro - 2023.
38 f.

1 Arquivo em PDF
Indexado no catálogo *online* da biblioteca José Albano de Macêdo-CSHNB
Aberto a pesquisadores, com restrições da Biblioteca

Trabalho de Conclusão de Curso (Graduação) – Universidade Federal do Piauí, Bacharelado em Sistemas de Informação, Picos, 2023.
"Orientadora : Profa. Dra. Débora Maria Vieira Magalhães "

1. Base de dados. 2. Dados espaço - temporais. 3. Mineração de dados. 4. Palavras - chave. 5. Desastres naturais. I. Magalhães, Débora Maria Vieira. II. Título.

CDD 005.74

Emanuele Alves Araújo CRB 3/1290

ANÁLISE DE PALAVRAS-CHAVE ASSOCIADAS À LOCALIZAÇÃO DE PESSOAS NOS DESASTRES NATURAIS

JOANNY EVA PACHECO MONTEIRO

Monografia APROVADA como exigência parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Data de Aprovação

Picos – PI, 14 de agosto de 2023

Profa. Deborah Maria Vieira Magalhães

Prof. Leonardo Pereira de Sousa

Profa. Francisca Pâmela Carvalho Nunes

Agradecimentos

Gostaria de aproveitar este momento para expressar minha profunda gratidão a todos que contribuíram para o meu crescimento e aprendizado durante os anos de curso. Foi uma trajetória difícil, mas a perseverança e a determinação me mantiveram focada em alcançar meus objetivos. Minha orientadora, Deborah, merece um agradecimento especial. Sua dedicação e paciência foram fundamentais para guiar-me ao longo deste processo, sou grata por suas horas dedicadas, suas palavras encorajadoras e por acreditar em meu potencial. Por fim, gostaria de agradecer à minha família e amigos que conheci na faculdade, sei que as lições aprendidas e as amizades construídas durante esse período ficarão comigo por toda a vida.

Resumo

As sociedades são confrontadas com desastres naturais de modo crescente, entre eles, terremotos, furacões, tornados, e outros. Tais desastres geram prejuízos ambientais e trazem um impacto negativo e em larga escala para a economia e cultura. As redes sociais estão desempenhando um papel cada vez mais importante nos sistemas de alerta precoce e sistemas de localização, auxiliando na avaliação rápida de desastres e na recuperação pós-desastre. A identificação da localização geográfica é uma das tarefas desafiadoras, pois os campos de informações da localização, como a localização do usuário e o nome do local dos *tweets*, não são confiáveis. A extração de informações de localização do texto do *tweet* é difícil, pois contém muitos erros gramaticais, abreviações não-padrão e assim por diante. Desta forma, este trabalho visa avaliar palavras-chave na descrição de lugares para definir a localização de um indivíduo. Para esse fim, será construída uma base de dados através da Interface de Programação de Aplicações do Twitter, com o intuito de coletar dados e produzir uma análise das palavras-chave dos *tweets* coletados, auxiliando no processo de identificação da localização de pessoas nos desastres naturais.

Palavras-chaves: desastre natural, mineração de dados, palavras-chave, dados espaço-temporais.

Abstract

Societies are increasingly confronted with natural disasters, including earthquakes, hurricanes, and tornadoes. Such disasters cause environmental damage and have a large-scale, negative impact on the economy and culture. Social networks are increasingly important in early warning and location-based systems, assisting in rapid disaster assessment and post-disaster recovery. Identification of Geographic location is challenging because location information is unreliable, such as user location and place name of tweets. Extracting location information from tweet text is difficult, as it contains many grammatical errors and non-standard abbreviations. Thus, this work aims to evaluate keywords in place descriptions to define the location of an individual. To this end, a database will be built through Twitter's Application Programming Interface to collect data and produce an analysis of the keywords in the collected tweets, assisting in identifying the location of people in natural disasters.

Keywords: natural disaster, data mining, keywords, spatio-temporal data.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

API	Application Programming Interface
EUA	Estados Unidos da América
GPS	Global Positioning System
JSON	JavaScript Object Notation
LDA	Latitude Dirichlet Allocation
NLTK	Natural Language Tool Kit
PLN	Processamento de Linguagem Natural
QR	Quick Response
SIG	Sistema de Informação Geográfico
USGS	United States Geological Survey
VGI	Voluntary Geographic Informations

Sumário

1 Introdução

Em situações de desastres naturais, como grandes furacões, milhares de pessoas podem encontrar-se precisando de ajuda urgente para salvar vidas, como assistência para evacuação e/ou atendimento médico (POWERS DEVARAJ, 2023).

As plataformas de mídia social estão desempenhando papéis cada vez mais críticos na resposta a desastres e operações de resgate. Durante emergências, os usuários podem postar solicitações de resgate junto com seus endereços nas mídias sociais, enquanto os voluntários podem procurar essas mensagens e enviar ajuda. No entanto, alavancar com eficiência as mídias sociais em operações de resgate continua sendo um desafio, devido à falta de ferramentas para identificar mensagens de solicitação de resgate nas mídias sociais de forma automática e rápida. (ZHOU, 2022).

Nessas circunstâncias, os países enfrentam desafios em termos de comunicação, coordenação de operações de resgate e socorro, bem como a necessidade de acesso a informações atualizadas sobre o evento em tempo real. O Twitter é usado durante crises para comunicar-se com autoridades, fornecendo operações de resgate e socorro em tempo real. As informações de localização geográfica do evento, como de pessoas, são de vital importância em tais cenários. Diferentemente de outras formas tradicionais, as plataformas de mídia social reúnem uma escala sem precedentes de dados e quantidade de informações, registrando reações do público. De acordo com (AHN SON, 2021), durante desastres naturais, os sistemas de comunicação de emergência ficam sobrecarregados e as pessoas são forçadas a se virar nas redes sociais para fazer pedidos de ajuda.

A fim de endereçar tal desafio, neste trabalho, propomos a construção de uma base de dados espaço-temporais. Nessa base, são utilizados dados provenientes de *tweets* durante a ocorrência de um terremoto no Havaí em maio de 2018. Tal terremoto ocorreu próximo à uma região vulcânica, o que desencadeou a erupção do vulcão Kilauea. Após a aquisição dos dados, estes foram agrupados e passaram por um processo de filtragem. Onde, em seguida, foram analisadas as palavras-chave mais utilizadas pelos usuários, a fim de ajudar no processo de identificação da localização de pessoas em situação de desastre.

1.1 Objetivos

1.1.1 Objetivo Geral

O objetivo geral deste trabalho reside em auxiliar na identificação da localização de pessoas desaparecidas em desastres naturais através do uso de palavras-chave coletados dos *tweets* minerados.

1.1.2 Objetivos Específicos

1. Implementar a etapa de aquisição e mineração de dados do Twitter.
2. Investigar as palavras-chave mais utilizadas no Twitter com o intuito de ajudar na localização de pessoas de forma mais rápida e eficiente.
3. Implementar técnicas para encontrar a localização do indivíduo através do *tweet* coletado.
4. Analisar as palavras-chave coletadas e mostrar quais são as mais utilizadas na procura da localização de pessoas desaparecidas.

1.2 Contribuições

Para alcançar os objetivos citados acima, as seguintes contribuições foram realizadas:

1. Implementação da etapa de Extração de Dados do Twitter:

   Uma das contribuições fundamentais deste trabalho é a implementação bem-sucedida da etapa de extração de dados do Twitter. Foi desenvolvida uma solução utilizando a API do Twitter para coletar *tweets* relevantes relacionados a pessoas desaparecidas. Essa etapa envolveu a obtenção dos tokens de acesso necessários (API Key, API Secret Key, Access Token e Access Token Secret), a definição dos parâmetros de busca apropriados e a implementação de um processo automatizado de coleta de dados.

2. Investigação das Palavras-Chave mais utilizadas no Twitter:

   Um passo importante para a localização de pessoas desaparecidas é a identificação das palavras-chave mais utilizadas nas postagens do Twitter relacionadas a esse tema. Para isso, foi realizado um estudo de análise de frequência das palavras encontradas nos *tweets* coletados. Essa investigação permitiu identificar as palavras-chave mais relevantes, proporcionando uma base sólida para a busca e localização mais rápida e eficiente de pessoas desaparecidas.

3. Implementação de técnicas para encontrar a Localização do Indivíduo:

   Com o intuito de encontrar a localização do indivíduo através dos *tweets* coletados, foram desenvolvidas técnicas específicas. Essas técnicas envolveram a análise dos metadados dos *tweets*, como informações de localização fornecidas pelo usuário, dados de geolocalização e outros atributos relacionados. Foi implementado um processo de processamento de linguagem natural (PLN) para identificar informações relevantes sobre a localização em cada *tweet* e, assim, auxiliar na determinação da localização atual do indivíduo.

4. Análise das Palavras-Chave mais utilizadas na procura de Pessoas Desaparecidas:

   Por fim, as palavras-chave coletadas durante o processo de extração de dados foram analisadas em relação à sua frequência de uso na procura por pessoas desaparecidas. Essa análise permitiu identificar quais palavras-chave são mais comumente utilizadas pelos usuários do Twitter nesse contexto, fornecendo informações valiosas para organizações e autoridades envolvidas na busca por pessoas desaparecidas.

   Em resumo, este trabalho contribui para a área de localização de pessoas desaparecidas ao desenvolver e implementar uma abordagem que combina a coleta de dados do Twitter, a identificação de palavras-chave relevantes, a análise de metadados e o processamento de linguagem natural. Essas contribuições visam melhorar a eficiência e a rapidez na localização de pessoas desaparecidas, oferecendo uma abordagem mais eficaz para lidar com esse desafio socialmente importante.

1.3 Estrutura do Trabalho

Nesta subseção é possível encontrar a organização deste trabalho. No Capítulo 2, é apresentado o Referencial Teórico contendo conceitos para ajudar no embasamento do leitor no que se refere ao entendimento do trabalho. No Capítulo 3, apresenta-se os Trabalhos Relacionados com temáticas semelhantes a este projeto. A Proposta do trabalho, o Cenário de Desastre e a etapa de Mineração dos Dados estão presentes no Capítulo 4. O Capítulo 5 exibe os Resultados e Discussões. Por fim, o Capítulo 6 contém a conclusão e as considerações finais do trabalho.

2 Referencial Teórico

Esta seção aborda tópicos relevantes ao contexto deste trabalho: desastres naturais, processamento de linguagem natural, dados espaço-temporais e, por fim, filtro de Kalman.

2.1 Desastres Naturais

Os desastres naturais representam um conjunto de fenômenos que fazem parte da geodinâmica terrestre, portanto, da natureza do planeta. Muitos desastres têm ocorrido porque o planeta Terra está sofrendo cada vez mais com o aquecimento global e o efeito estufa, o que leva ao aumento dos desastres naturais, ocasionados pelo desiquilíbrio da natureza. O momento e a magnitude dos desastres naturais são imprevisíveis e, portanto, estocásticos.

O número de mortes e pessoas desaparecidas causadas por desastres naturais é frequentemente usado para medir a magnitude dos desastres. Milhares de pessoas desaparecem todos os dias. O sumiço se dá por diversos motivos, entre eles conflitos armados, desastres naturais, migrações e atentados. (MARTÍN; LI; CUTTER, 2017) afirmam que catástrofes naturais são custosas em termos de propriedade, estabilidade política e vidas perdidas.

Segundo (KRYVASHEYEU et al., 2016), as pessoas se voltam para plataformas de redes sociais durante eventos catastróficos com o intuito de se comunicarem. A análise do comportamento público, tal qual a maneira pela qual as pessoas se preparam e respondem às catástrofes, desempenha um papel importante no gerenciamento de crises, na resposta a desastres e no planejamento de evacuação (CHAE et al., 2014).

2.2 Processamento de Linguagem Natural

O Processamento da Linguagem Natural (PLN) trata computacionalmente os diversos aspectos da comunicação humana, como sons, palavras, sentenças e discursos, considerando formatos e referências, estruturas e significados, contextos e usos (ZVAREVASHE; OLUGBARA, 2018). Em sentido bem amplo, podemos dizer que o PLN visa fazer o computador se comunicar em linguagem humana, nem sempre necessariamente em todos os níveis de entendimento.

De acordo com (KANAKARAJ; GUDDETI, 2015), o PLN começou nos anos 50 como a subárea da inteligência artificial e da linguística, além disso, a PLN era distinta da recuperação de informação de texto que consiste em empregar técnicas altamente escalonáveis baseadas em estatísticas para indexar e pesquisar grandes volumes de texto.

A PLN analisa a estrutura da linguagem de acordo com cinco níveis: pragmático, morfológico, sintático, fonológico e semântico (MALDONADO et al., 2016). Assim, segundo (ZVAREVASHE; OLUGBARA, 2018), o resultado da análise de opinião e da mineração de opinião depende em grande parte, do uso de ferramentas para executar diferentes tarefas do PLN.

Segundo (KANAKARAJ; GUDDETI, 2015), minerar opiniões e analisar sentimentos a partir de dados de redes sociais auxiliam em vários campos, como a previsão de eventos, a análise do clima geral do público em uma questão social específica, dentre outros aspectos. Além disso, afirmam que a análise do sentimento de texto envolve a aplicação de aptidão computacional na compreensão do sentimento implícito no texto, assim, o PLN envolve a tarefa de encontrar o significado semântico do conteúdo do texto e assim, analisar a informação do texto.

2.3 Dados espaço-temporais

Atualmente as mídias sociais estão disponíveis em tecnologias móveis, geralmente com recepções integradas do Sistema de Posicionamento Global (GPS). As atuais tecnologias, tais como, GPS e sensores proveem recursos que facilitam no acesso à determinados dados, como por exemplo, o *geotagged* (MARTÍN; LI; CUTTER, 2017). De acordo com (CHAE et al., 2014), os *geotagged* oferecem uma precisão confiável em relação aos dados. A nível de software, há dois objetos JSON usados para descrever o local associado a um *tweet*: (i) coordenadas geográficas, latitude e longitude; e (ii) nome do local.

Para (AVVENUTI et al., 2014), as mídias sociais representam uma maneira poderosa de investigar preferências, gostos e atividades de grupos de usuários. Atualmente, as pessoas compartilham continuamente comentários e conteúdo multimídia sobre suas vidas, interesses, sentimentos e opiniões. Portanto, os usuários de redes sociais podem ser considerados como sensores capazes de transmitir informações valiosas sobre situações e fatos, como afirmado pelo paradigma de detecção social.

Um número crescente de pessoas estão usando serviços de redes sociais baseados em localização, como microblogs, onde são criados dados geo-localizados com marca de tempo. Os dados espaço-temporais possuem um grande potencial para melhorar a consciência situacional durante a situação de crise, fornecendo informações sobre o evento em evolução, a resposta pública e os possíveis cursos de ação (CHAE et al., 2014).

Segundo (MARTÍN; LI; CUTTER, 2017), a disponibilidade de redes sociais baseadas em localização com dados de redes sociais *geotagged* melhorou significativamente a pesquisa e a prática de gestão de emergências, através do monitoramento de desastres e da reação das pessoas. Além disso, pesquisas usando dados de redes sociais *geotagged* e a identificação de padrões espaciais de usuários de redes sociais é generalizada, ou seja, podem ser encontrados conteúdos de mídia social acerca de terremotos, incêndios florestais,

ciclones tropicais, eventos de inundação, dentre outros.

A análise desses dados, como a maneira pela qual as pessoas se preparam e respondem às catástrofes, desempenha um papel importante no gerenciamento de crises, na resposta a desastres e no planejamento de evacuação. As mídias sociais têm sido reconhecidas por muitos na pesquisa de gestão de crises e nas comunidades de prática como uma ferramenta para comunicação e coleta de informações (MACEACHREN et al., 2011).

2.4 Filtro de Kalman

O filtro de Kalman é essencialmente um conjunto de equações matemáticas que implementa um estimador de estados conhecidos como "preditor-corretor". Quando algumas condições são satisfeitas, o filtro de Kalman é considerado um estimador ótimo e minimiza a covariância do erro estimado (SANTANA, 2011).

Neste projeto, o filtro de Kalman foi utilizado para remover o que consideramos ruídos, tais como: números, caracteres especiais e stopwords, que consistem em palavras muito comuns que não são significativas para o dataset. Posteriormente foi executado um filtro para remoção de conteúdo duplicado e possíveis spams.

Sakaki, Okazaki e Matsuo (2010) propõe que o filtro de Kalman é aplicado a sistemas de controle sujeitos a ruídos cujos parâmetros não podem ser devidamente medidos. Um filtro Kalman funciona melhor em um ambiente Gaussiano linear, porque é projetado para lidar com incertezas e informações imprecisas, ajustando continuamente as estimativas do estado do sistema com base nas medições e nas previsões do próprio filtro. Ao utilizar filtros Kalman, é importante construir um bom modelo e parâmetros. As aplicações comuns dos filtros Kalman incluem orientação, navegação, sistemas de controle, sistemas de visão computacional e processamento de sinal.

De acordo com (BISHOP; WELCH et al., 2001), o filtro de Kalman é utilizado quando: (i)É possível obter medições sobre um determinado evento a uma taxa constante; (ii)As medidas têm um erro que segue uma distribuição normal; (iii)A matemática que regula a situação é conhecida; (iv)O processo que será medido pode ser descrito como um sistema linear; (v)Busca-se uma estimativa do que está realmente acontecendo.

3 Trabalhos Relacionados

Nesta seção abaixo é apresentado o protocolo de mapeamento sistemático. Na Subseção 3.2 são abordados os trabalhos correlatados, onde são comparadas as métricas: objetivos, metodologia e base de dados utilizada.

3.1 Mapeamento Sistemático

Em alguns casos, a busca de material científico é realizada sem critérios de busca explícitos, tornando o referencial teórico enviesado. Para evitar esse problema, foi realizado um mapeamento sistemático, afim de investigar materiais científicos para o embasamento teórico através do protocolo de busca e etapas de planejamento.

Na etapa de formação do protocolo foram definidas questões de pesquisa, tais como, base de dados utilizada, *string* de busca, critérios de inclusão/exclusão, aceitação/rejeição e palavras-chave. Além disso, foram formuladas as questões de pesquisa, sendo uma primária e a outra secundária.

Questão Principal(QP): Como os dados das mídias sociais podem contribuir para encontrar a localização de pessoas desaparecidas nos desastres naturais?
Questão Secundária(QS): Quais técnicas, métodos e ferramentas estão sendo utilizadas para localização baseadas em dados do Twitter?

Outra etapa significante para a especificação do protocolo é a definição das palavras-chave que irão compor a *string* de busca, sendo esta etapa muito importante, pois, tem grande impacto na seleção dos trabalhos relacionados. As palavras-chave usadas foram definidas de acordo com as questões de pesquisa, conforme a tabela abaixo:

Tabela 1 – Palavras-chave

Keywords	Palavras-chave
Missing people	Pessoas desaparecidas
Natural disasters	Desastres naturais
Spatio temporal data	Dados espaço-temporais

Dando continuidade ao desenvolvimento do protocolo, foram definidos os critérios de inclusão/exclusão que foram utilizados como fundamento para o filtro de busca da base de dados. Além disso, também foram definidos os critérios para aceitação/rejeição. Nesta fase, os trabalhos são classificados como aceitos ou rejeitados de acordo com as regras definidas do protocolo.

Tabela 2 – Critérios para filtragem

Critérios		
Inclusão	Publicações dos últimos 10 anos (2013 a 2023)	Deve estar no contexto do assunto pesquisado
Exclusão	Não serão aceitos idiomas diferentes de português ou inglês	
Aceitação	Análise de palavras-chave no cenário de desastres naturais	As palavras-chave devem aparecer no título e resumo
Rejeição	Devem estar em português ou inglês	

Concluída a etapa de formação dos critérios para filtragem, foi elaborada a string de busca baseada nas palavras-chave, sendo este um processo crucial da etapa de conclusão da revisão sistemática. A string de busca foi inserida manualmente na base de dados. Conforme a tabela 3, é notável que a sintaxe da string varia de acordo com a base de dados utilizada.

Tabela 3 – Inserção da string de busca

String	Base de dados	Resultados
(("Keywords") AND ("missing people") AND ("natural disasters"))	ACM	3
(("Keywords") AND ("natural disasters"))	IEEE	19
(("Keywords") AND ("natural disasters"))	PubMed	31
(("Keywords") AND ("missing people") AND ("natural disasters") AND ("spatiotemporal data analysis"))	Science Direct	41
(("Keywords") AND ("missing people") AND ("natural disasters") AND ("spatiotemporal data analysis"))	WorldWideScience	22

Em alguns casos foi necessário diminuir a string de busca, pois, a base de dados não retornava os resultados através da string aplicada nos demais. Após a filtragem, os critérios de aceitação/rejeição são aplicados com base no título, resumo e palavras-chave do artigo. Os artigos aceitos são lidos por completo e novamente são aplicados os critérios de aceitação/rejeição, sendo este, o último filtro. Neste trabalho foram considerados trabalhos que abordaram a análise e tratamento de dados, e técnicas para encontrar a localização através de um *tweet* no cenário de desastres naturais. O mapeamento sistemático auxiliou bastante na busca de trabalhos correlatos.

3.2 Trabalhos Selecionados

Há vários métodos para a extração de dados no Twitter. Assim, serão citados trabalhos referentes à extração de dados com foco em classificação e análise de dados relacionados ao cenário de desastres naturais.

Avvenuti et al. (2014) propõem um sistema que auxilia a definir onde concentrar as equipes de resgate e organizar uma pronta resposta de emergência. Além disso, extrai o conteúdo das mensagens associadas a um evento para descobrir o conhecimento sobre suas consequências. Os eventos detectados são transmitidos automaticamente por um determinado sistema através de uma conta dedicada do Twitter e por notificações de e-mail. O EARS (*Earthquake Alert and Report System*) utiliza técnicas de processamento de linguagem natural para a detecção do idioma italiano e descarta mensagens que não parecem estar em italiano, já que estava sendo realizada uma avaliação de danos do

terremoto na Itália. O sistema de propostas pode fornecer claramente informações úteis sobre as consequências dos eventos sísmicos. Tais informações são deduzidas de sensores sociais no solo segundos após a detecção de um terremoto. O EARS é um ambiente de suporte à decisão para gestão de crise sísmica implantado para pesquisadores e analistas do Instituto Nacional de Geofísica e Vulcanologia da Itália (INGV), ele pode ser integrado com outros sistemas já estabelecidos para ajudar a identificar as áreas em que o dano é provável. As informações descobertas por este sistema ajudam a definir onde concentrar as equipes de resgate e ou organizar uma resposta de emergência imediata.

Kryvasheyeu et al. (2016) usou o limite geográfico da região afetada pelo furacão Sandy e agrupou todas as mensagens que continham um conjunto pré-selecionado de palavras-chave, explorando postagens por meio do uso de palavras-chave relacionadas ao furacão Sandy e demonstrando a trajetória dos evacuandos. Os resultados sugerem que, durante um desastre, os funcionários devem prestar atenção aos níveis normalizados de atividade, taxas de criação de conteúdo original e taxas da retransmissão de conteúdo para identificar as áreas mais afetadas em tempo real. Imediatamente após um desastre, eles devem se concentrar na persistência nos níveis de atividade para avaliar quais áreas são susceptíveis de precisar de mais assistência. No projeto foram utilizados vários algoritmos para análise de sentimento, incluindo o algoritmo proprietário da Topsy, o Linguistic Inquiry and Word Count (LIWC) e o SentiStrength. Os autores avaliaram o algoritmo léxico de análise de sentimento da Topsy com os pesos das palavras do dicionário variando de -5 a +5. Os resultados do estudo mostraram que a atividade nas redes sociais relacionada ao furacão Sandy estava fortemente relacionada à proximidade do furacão, e que a intensidade da atividade era moderada pela intensidade da cobertura da mídia. Eles analisaram a correlação entre a atividade nas redes sociais e o número de pedidos de assistência individual da FEMA, as reivindicações de seguro relacionadas ao furacão Sandy, o número de dias de fechamento de escolas como proxy para a perda de energia, o número de chamadas para a linha de emergência estadual de postos de gasolina como proxy para a escassez de gasolina e o número de tweets relacionados ao furacão Sandy. Além disso, os autores descobriram que, no nível de mensagens individuais, as métricas de análise de sentimento estavam fortemente correlacionadas entre si, e que as tendências temporais na média de sentimentos das mensagens agregadas por hora para todas as três métricas eram semelhantes, sugerindo que todas as técnicas de classificação eram comparáveis e robustas, especialmente na análise agregada.

Martín, Li e Cutter (2017) analisam a variabilidade espaço-temporal na resposta das mídias sociais e desenvolve uma nova abordagem para alavancar *tweets* geotagged no intuito de avaliar as respostas de evacuação dos residentes. Este trabalho se mostra bastante útil pois sua abordagem envolve a recuperação de *tweets* do Twitter Stream, filtragem de diferentes conjuntos de dados e tratamento estatístico e espacial para traçar e mapear os resultados a partir da coleta de *tweets* para os locais afetados pelo desastre. A

palavra-chave foi deixada em branco para coletar a maior quantidade de *tweets* possível. A abordagem oferece uma solução para abordar várias das desvantagens das avaliações tradicionais das taxas de evacuação através de pesquisas de questionários. Estes são frequentemente demorados e caros, e as taxas de resposta são muitas vezes longe do ideal. Esta alternativa é econômica e oportuna, pois os dados podem ser coletados em tempo real, proporcionando um tamanho de amostra notável com resultados bem-sucedidos.

Becken et al. (2017) monitoraram o ambiente e sentimento humano em relação à grande barreira de corais na Austrália, eles utilizaram a API do Twitter para recuperar dados. Durante a coleta, havia restrições para capturar *tweets* com marcação geográfica postados da região de interesse. As palavras-chave foram extraídas usando uma técnica de busca insensível a maiúsculas e minúsculas, e variações da mesma palavra (por exemplo, "mergulho", "mergulhando") foram compiladas como a mesma palavra-chave. Os números de ocorrências para cada palavra-chave foram contados. Além das frequências de palavras-chave, os *tweets* foram analisados com relação à polaridade positiva ou negativa. Também foi mostrado um mapa de calor de *tweets* georreferenciados. A análise dos *tweets* mostrou que o volume de dados é muito melhorado pelos turistas que visitam a cidade/região, destacando a importância dos sensores não residentes no desenvolvimento dessas abordagens. As análises de palavras-chave e sentimentos destacaram que o número real de *tweets* relacionados ao meio marinho são pequenos e precisariam ser impulsionados por vários mecanismos.

Wu e Cui (2018) realizam uma análise hierárquica combinando dados de mídia social, perdas econômicas e geo-informação. Verificaram o papel desempenhado pelas mídias sociais antes, durante e após um desastre natural. Investigam se a combinação de mídia social e informações de localização geográfica podem contribuir para um sistema de alerta precoce mais eficiente e ajudar na avaliação de desastres. Demonstram que a gravidade do dano em uma área está positivamente correlacionada com a intensidade da atividade relacionada a desastres. Enquanto isso, as áreas costeiras e áreas próximas ao centro do furacão tendem a sofrer maiores perdas durante um desastre. As descobertas exploram o papel desempenhado pelas mídias sociais de indivíduos nas populações afetadas e como elas respondem aos desastres naturais. Os resultados mostram que as pessoas estão usando as mídias sociais para três propósitos principais durante um desastre: expressão emocional, atualizações situacionais e trocas de informações relacionadas a desastres, cujas tendências sugerem um crescimento simultâneo considerável, juntamente com a rápida expansão. Chegaram a conclusão de que a intensidade da atividade do Twitter relacionada a desastres tem um significativo coeficiente de correlação positiva com as perdas de danos.

Zhou (2022) desenvolveu e comparou modelos para identificar *tweets* de solicitação de resgate. Um total de 3.191 *tweets* relacionados a desastres rotulados manualmente, publicados durante o furacão Harvey em 2017, foram usados como conjuntos de dados de treinamento e teste. Eles avaliaram o desempenho de cada modelo pela precisão da

classificação, custo de computação e estabilidade do modelo. O resultado do trabalho foi a comparação de dez modelos de processamento de linguagem natural para classificação de tweets de resgate em três tarefas independentes. Os modelos comparados incluíram GloVe, ELMo, BERT, RoBERTa, DistilBERT, ALBERT e XLNet. Os resultados mostraram que os modelos baseados em BERT geralmente superaram o modelo de referência em cerca de 10%, com o modelo BERT com classificador CNN apresentando o melhor desempenho geral com um F1-score de 0,919 na tarefa de identificação de tweets de solicitação de resgate. Além disso, o estudo introduziu uma nova métrica de estabilidade de modelo normalizada, o Índice de Estabilidade de Modelo Normalizado (NMSI), para avaliar a estabilidade dos modelos.

Sufi (2022) projetaram e desenvolveram um modelo de inteligência artificial (IA) totalmente automatizada baseada em Sistema de Suporte à Decisão (DSS) disponível em várias plataformas, como iOS, Android e Windows. O DSS proposto usa um feed do Twitter ao vivo para obter *tweets* relacionados à desastres naturais em 110 idiomas. O sistema executa automaticamente a tradução baseada em IA, análise de sentimento e algoritmo K-Means automatizado para gerar insights direcionados para estrategistas de desastres. Durante a experimentação, a metodologia descobriu automaticamente o número correto de aglomerados em tipos de desastres, analisando vários parâmetros do Twitter. A análise detalhada do desempenho através de múltiplas métricas de avaliação, como precisão, recall e F1-Score revela que o sistema apresentado é capaz de fornecer soluções altamente precisas através da extração de palavras-chave do Twitter. Os valores alcançados foram obtidos por meio da análise de desempenho e avaliação de diferentes categorias de entidades. Os resultados mostraram uma precisão, recall, F1-Score e acurácia de 0,957, 0,969, 0,963 e 0,997, respectivamente, para o processo de extração de entidades. Além disso, a solução proposta foi implantada em várias plataformas, como dispositivos móveis e tablets, e recebeu avaliações positivas de estrategistas de desastres, com 83,33% deles considerando o aplicativo fácil de usar, eficaz e autoexplicativo.

Powers Devaraj (2023) utilizaram aprendizado de máquina e inteligência artificial para detectar, identificar e categorizar automaticamente os *tweets* relevantes para os socorristas durante o furacão Harvey. Eles organizaram um conjunto de dados de *tweets* e apresentaram um esquema de rotulagem com base na relevância e urgência, e desenvolveram modelos neurais e modelos de aprendizado de máquina não neurais para categorizar *tweets* automaticamente. Este estudo também apresenta modelos de aprendizado de máquina que classificam *tweets* relacionados ao desastre como relevantes ou irrelevantes e urgentes ou não urgentes. Outra contribuição importante é a demonstração das incorporações médias de palavras, que são uma maneira viável de produzir características de comprimento fixo sobre as quais modelos não-neurais (como SVMs e modelos de regressão logística) podem ser treinados. Além disso, o trabalho deles é o primeiro a apresentar o uso de modelos de linguagem pré-treinados como BERT e XLNet para construir classificadores

de *tweets* de desastres, e seu desempenho superior em comparação com os modelos não-neurais. Para a tarefa de classificação de urgência, os modelos neurais BERT e XLNet alcançaram F1 Score de 0.67 e 0.68, respectivamente. Para a tarefa de classificação de relevância, os modelos neurais BERT e XLNet alcançaram F1 Score significativamente mais altos do que os modelos não neurais, com valores de 0.78 e 0.77, respectivamente.

Karimiziarani (2023) extraíram e processaram mais de vinte milhões de *tweets* para descobrir os principais tópicos de discussão e relacionamento entre eles e classificaram os *tweets* em tópicos e categorias humanitárias para ajudar no gerenciamento de desastres com análise de sentimento. Eles também empregaram uma variedade de algoritmos em Inteligência Artificial para Processamento de Linguagem Natural (NLP), incluindo análise de sentimento, modelagem de tópicos e classificação de texto para assimilar o conteúdo da informação em dados maciços do Twitter. As descobertas mostraram que o estado que sofreu a maior perda econômica como resultado do furacão Ian foi onde mais *tweets* foram produzidos. Os estados com uma chance de entrar no caminho do furacão também tiveram uma frequência mais alta de tweets em comparação com outros estados na área de estudo. Eles mostram como os tweets classificados sobre o furacão Ian foram distribuídos diariamente em nível estadual nos EUA antes, durante e depois da chegada do furacão na área de estudo. Os tweets extraídos foram divididos em seis categorias, incluindo "Cuidado", "Danos", "Evacuação", "Ferimentos", "Ajuda"e "Simpatia", que representam os tópicos humanitários com os quais os usuários do Twitter na área de pesquisa estão mais preocupados. As pessoas na Flórida foram as que mais tuitaram sobre o furacão Ian, pois ele atingiu a Flórida como um furacão de categoria 4 que causou bilhões de dólares em prejuízos. Depois da Flórida, a Carolina do Sul teve o maior número de *tweets* relacionados ao furacão, mas com menos de 40% de *tweets* em comparação com a Flórida. De modo geral, os estados do sudeste foram os que mais participaram dos *tweets* sobre o furacão. "Danos"é o tópico mais frequentemente tuitado por todos os estados. Enquanto "Cuidado"é o segundo tópico mais discutido nos estados localizados no sudeste, do Alabama à Virgínia, "Ajuda"é o segundo assunto mais discutido. Esses estados do sudeste também corriam o risco de serem afetados pelo furacão Ian. O resultado também mostra a flutuação temporal de cada classe na área de estudo durante as três fases do furacão (pré-furacão, furacão e pós-furacão).

A Tabela 4 apresenta um resumo dos trabalhos relacionados que abordam a mesma temática desta proposta.

Tabela 4 – Tabela Comparativa de Trabalhos Relacionados

Trabalho	Objetivo	Metodologia
Avvenuti et al. (2014)	Utilizam os dados do Twitter para pronta resposta de emergência.	Mineração de dados, processamento de linguagem natural, aprendizado de máquina.
Kryvasheyeu et al. (2016)	Utilizam os dados do Twitter para análise de comportamento de evacuação.	Processamento de dados, análise de dados e análise visual.
Martín, Li e Cutter (2017)	Utilizam os dados do Twitter para análise de comportamento de evacuação.	Mineração de dados, processamento de dados.
Becken et al. (2017)	Utilizam os dados do Twitter para monitorar o ambiente e sentimento humano em relação à grande barreira de corais na Austrália.	Coleta de dados, análise de dados.
Wu e Cui (2018)	Utilizam os dados do Twitter para análise hierárquica de dados.	Processamento de dados, análise de dados.
Zhou (2022)	Utilizam os dados do Twitter para desenvolver e comparar modelos para identificar tweets de solicitação de resgate.	Processamento de linguagem natural, mineração de dados, análise de dados.
Sufi (2022)	Utilizam os dados do Twitter para projetar e desenvolver uma inteligencia artificial totalmente automatizada baseada em Sistema de Suporte à Decisão (DSS).	Processamento de linguagem natural, inteligência artificial.
Powers Devaraj (2023)	Utilizam os dados do Twitter para detectar, identificar e categorizar automaticamente os tweets relevantes para os socorristas durante o furacão Harvey.	Aprendizado de máquina, inteligência artificial.
Karimiziarani (2023)	Utilizam os dados do Twitter para classificar os tweets em tópicos e categorias humanitárias para ajudar no gerenciamento de desastres com análise de sentimento.	Mineração de dados, processamento de linguagem natural, análise de dados.
Modelo Proposto	**Construir uma base de dados proveniente da Interface de programação de aplicações do Twitter, coletar dados e produzir uma análise das palavras-chave dos tweets coletados.**	**Mineração de dados, análise das palavras-chave coletadas.**

4 Proposta

Este trabalho consiste em filtrar e limpar dados espaço-temporais coletados a partir do Twitter, a fim de identificar quais palavras-chave são mais utilizadas em cenários de desastre, para encontrar a localização de uma pessoa desaparecida. Primeiro, apresentaremos o cenário de desastre tratado no trabalho. Em seguida, discutiremos o processamento dos dados coletados para a construção da base de dados.

4.1 Cenário de Desastre

O terremoto ocorreu no Havaí como resultado de fissuras na lateral do sul do vulcão Kilauea, na zona leste do rifte, no qual consiste em grandes fraturas tectônicas. Para Frozza, Mello e Costa (2018) a abertura do solo, a uma profundidade de 50 metros, foi contínua e esteve associada à atividade sísmica. As áreas de erupção foram na Zona Leste do Rift e no topo do Kilauea. Segundo a USGS, os moradores e visitantes perto das fissuras dos fluxos de lava e da área de colapso da cúpula devem se manter atentos aos avisos do Departamento de Defesa Civil e do Parque Nacional do *Havaí*[1].

A Figura 1 foi obtida a partir do site do County of *Hawaii*[2] no qual possui dados geológicos disponibilizados pela USGS. Tal figura mostra as áreas mais afetadas onde escorreu a lava e o fluxo histórico.

Conforme observado na Figura 1, a mancha vermelha representa o fluxo de lava que cobriu o leste do Havaí, especialmente nas cidades de Kapoho e Leilani States. Ademais, tais áreas afetadas, possuem o histórico de serem atingidas em outras ocasiões de desastre, conforme visto nas manchas em cinza representando dados de fluxos históricos.

[1] https://volcanoes.usgs.gov/volcanoes/kilauea/status.html

[2] https://hawaiicountygis.maps.arcgis.com/apps/webappviewer/index.html?id=3428cd9282ff431c865eb32761793078

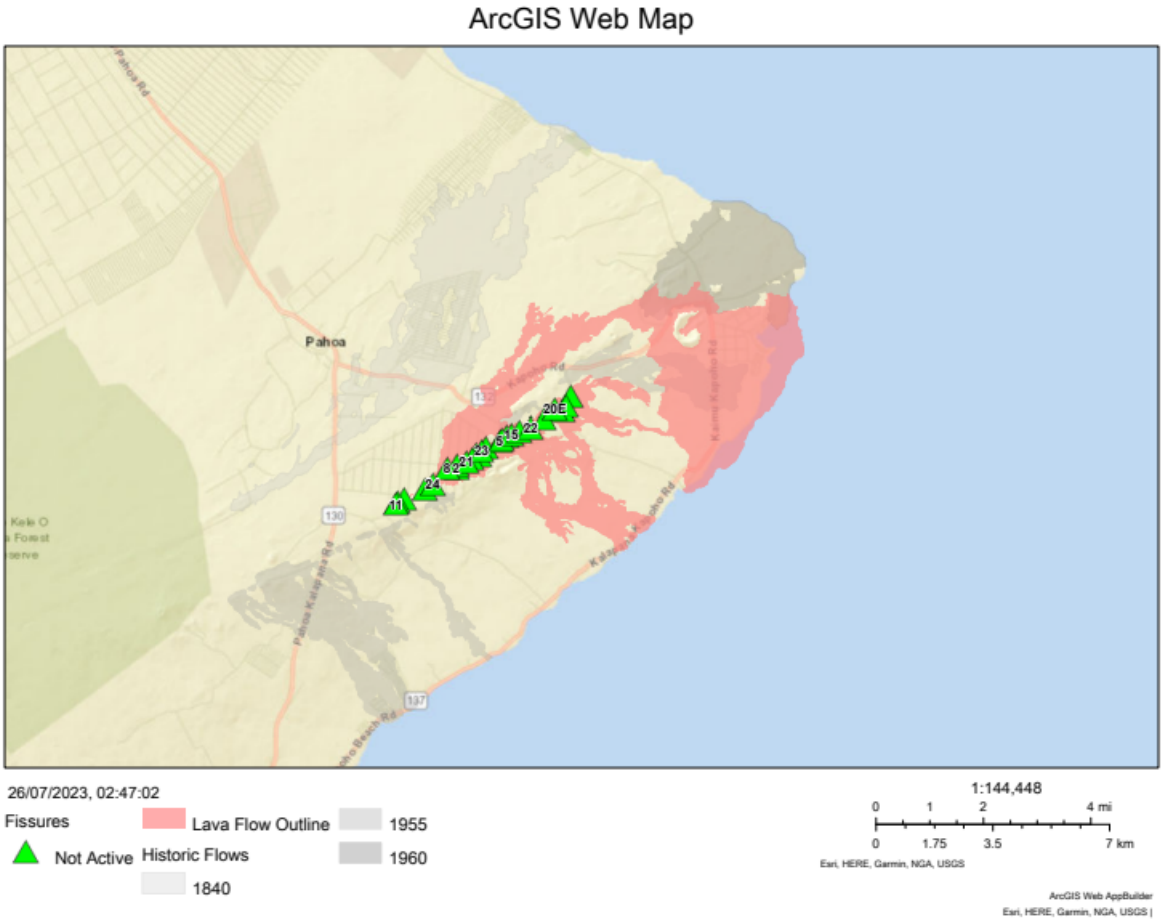

Figura 1 – Mapa do fluxo de lava, em vermelho corresponde ao ano de 2018 e em cinza os fluxos históricos.

4.2 Mineração dos Dados

O estágio inicial desta pesquisa consistiu em realizar a *mineração* de dados no formato de Notação de Objetos JavaScript (JSON) utilizando a linguagem Python. Posteriormente, foi realizado o armazenamento das informações em um banco de dados e iniciado o estágio de obter os padrões dos dados espaço-temporais, instaurando o processo de mineração.

O segundo estágio consistiu em organizar os dados de cada usuário. Cada *agrupamento de usuário* compõe um conjunto de documentos que possuem dados relevantes, por exemplo, nome, ID, horário, data, latitude/longitude. Foi percorrido o campo horário incluso em cada documento e analisado a mudança de localização. Posteriormente, na *análise de desvios*, foram identificados conjuntos de dados que não obedeceram ao padrão de comportamento dos demais dados, para serem tratados ou descartados, utilizando a abordagem estatística do filtro de Kalman.

Em seguida, foi realizado um processo de agrupamento utilizando uma técnica de análise de dados chamada "agrupamento de palavras-chave"ou "clusterização de palavras-

chave"para analisar as palavras-chave mais descritas nos *tweets* coletados. Por fim, os resultados obtidos através de mapas e gráficos foram apresentados. Foi aplicado a *visualização* para apresentar os dados de forma compreensível. A Figura 2 representa um fluxograma de etapas que foram seguidas nesta metodologia.

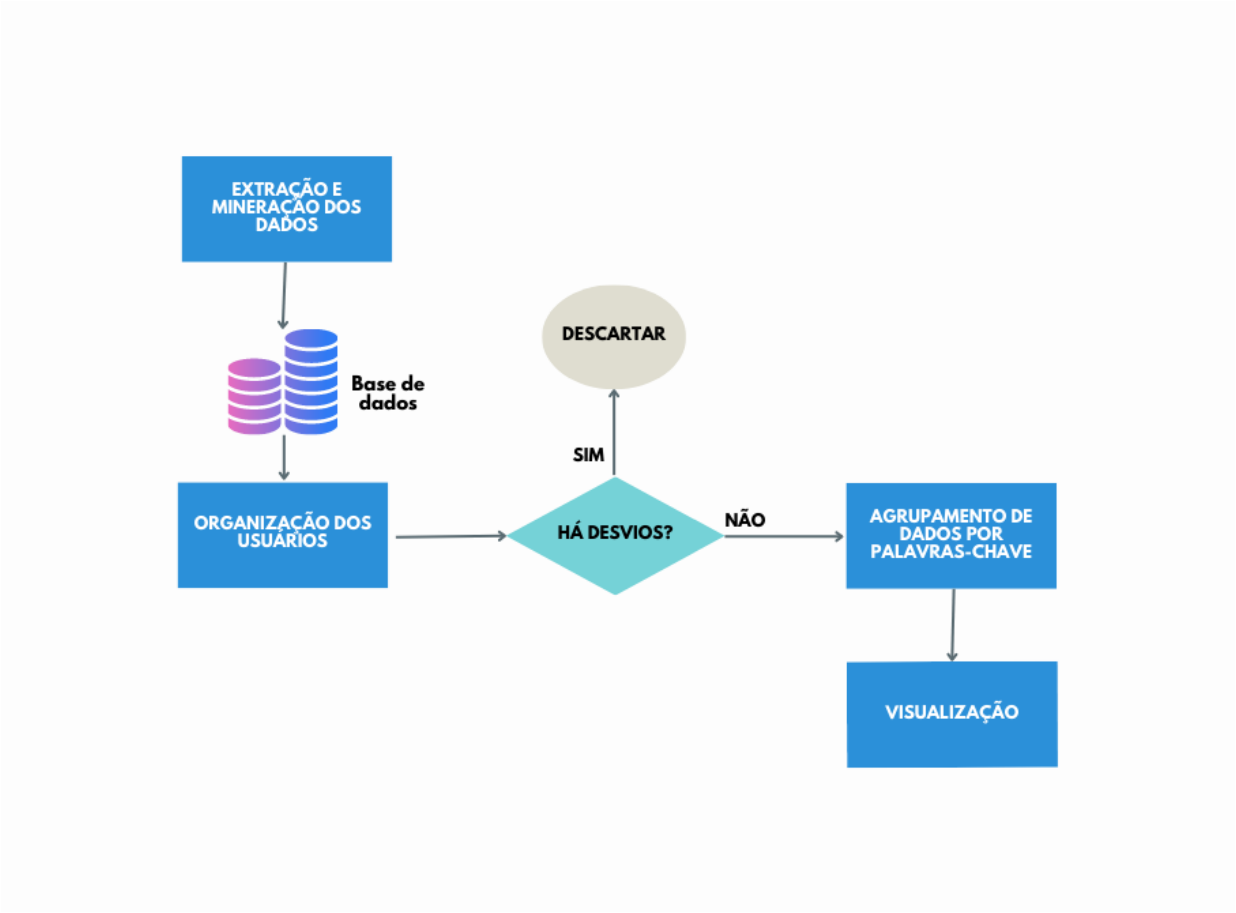

Figura 2 – Metodologia seguida para alcançar os resultados obtidos.

4.2.1 Descrição das Etapas da Metodologia

As etapas principais foram as seguintes:

1. Mineração de dados no formato JSON: O foco foi a obtenção dos dados relevantes de interesse para a pesquisa. O Twitter oferece uma API de fácil acesso e com quantidade significativa de dados disponíveis, porém é importante ressaltar que os usuários que utilizam Twitter com geolocalização ativa é baixa, porém ainda assim bastante significativa. A fim de capturar os dados, é necessário definirmos alguns parâmetros importantes como: latitude, longitude, distância. Levando em conta esses parâmetros, podemos construir um filtro, onde somente *tweets* provenientes da área especificada do desastre possam ser coletados.

2. Armazenamento em banco de dados: Os dados coletados foram armazenados no banco de dados CouchDB, possibilitando o gerenciamento e manipulação dos registros para análises futuras.

3. Identificação dos padrões espaço-temporais: Após o armazenamento, iniciou-se o estágio de obtenção dos padrões espaço-temporais dos dados. A localização foi filtrada com base nas coordenadas do desastre, nessa etapa, os dados foram submetidos a uma análise minuciosa para a obtenção dos padrões espaço-temporais relevantes. Primeiramente, foram extraídas informações sobre a localização e o horário do evento registrado nos *tweets*. A análise de mudança de localização permitiu identificar áreas geográficas mais afetadas pelo desastre, possibilitando a delimitação das regiões mais impactadas.

4. Organização dos usuários: Os dados foram organizados para cada usuário, formando conjuntos de documentos relevantes relacionados a cada indivíduo. Cada agrupamento de usuário continha informações como nome, ID, horário, data e coordenadas de localização. A imagem 3 ilustra como os dados foram organizados para cada usuário. Cada linha representa um usuário diferente, com as seguintes informações disponíveis: nome, ID, horário em que os dados foram registrados, data em que os dados foram registrados e as coordenadas de localização correspondentes ao usuário em questão. É importante ressaltar que as coordenadas de localização foram representadas por valores de latitude e longitude para indicar a posição geográfica de cada usuário.

	date	id_str	user_name	text	time	latitude	longitude
0	2018-05-06	2.379071e+09	tiedyehobo	[majic, sands, beach, kona, hawaiiawesome, tun...	20:41:33	19.594333	-155.971728
1	2018-05-06	2.929292e+07	queeraspoetry	[good, center, spiritual, living, hawai]	20:33:41	19.560340	-154.986600
2	2018-05-06	8.149151e+17	Hazards_Network	[monday, january, earthquake, magnitude, shook...	20:30:04	19.078167	-155.187667
3	2018-05-06	4.142280e+08	NaomiCooper808	[vacay, ladat, waimea, hawaii, county, hawai]	20:09:43	20.020278	-155.667778
4	2018-05-06	2.011251e+07	vioart527	[last, sunset, hawaii, next, time]	19:46:08	19.650000	-155.994000
...	...	...	...	...	...	...	...
6018	2018-04-29	1.446640e+07	MauiBenjamin	[lahaina, sunday, lahaina, hawai]	21:43:25	20.886100	-156.675000
6019	2018-04-29	3.048545e+09	googuns_lulz	[becbefbfdaadeaaaebadbedbbeccdfcdcfffecbbecdc]	21:31:00	19.277470	-153.229383
6020	2018-04-29	2.927180e+08	barkingbbq	[last, day, brunch, favorite, place, maui, vac...	21:27:53	20.944502	-156.691695
6021	2018-04-29	8.743629e+17	lifted_karma	[absolutely, love, switch, blaze, vape, batter...	21:27:01	20.906560	-156.243140
6022	2018-04-29	2.337400e+09	KauCoffeeMill1	[happening, ever, onolicious, recipes, submitt...	21:25:12	19.238094	-155.478760

6023 rows × 7 columns

Figura 3 – Organização dos dados dos usuários.

5. Filtragem dos dados dos tweets coletados: Na análise e processamento de dados coletados a partir de tweets, é comum realizar uma etapa de filtragem para remover elementos que podem ser considerados como "ruídos"e que não contribuem significativamente para a análise ou para a identificação de padrões e informações relevantes.

Entre os ruídos comuns em tweets estão os emojis, números, pontuações e as chamadas "stopwords". O processo de tokenização foi feito através de um token, que pode ser uma palavra, uma frase ou até mesmo um caractere individual. O objetivo da tokenização é dividir o texto em unidades menores para que possa ser tratado e processado pelo modelo de aprendizado de máquina. Nesse método, o texto foi dividido em palavras individuais, usando espaços e pontuações como delimitadores. Por exemplo, a frase "Está acontecendo um desastre no Havaí!"seria tokenizada em ["Está", "acontecendo", "um", "desastre", "no", "Havaí"]. A função que remove os ruídos do dataset tem o objetivo de limpar e pré-processar os dados, a fim de torná-los mais adequados para a tarefa da análise de dados. Ruídos são informações indesejadas ou irrelevantes que podem atrapalhar o desempenho do modelo e causar resultados imprecisos. As etapas típicas de remoção de ruídos incluíram: remoção de caracteres especiais, remoção de stopwords, remoção de emojis e remoção de números. Na Figura 4, mostra como foi realizada essa filtragem.

```
[ ]  #função para remover os numeros
     def remove_numbers(text):
       return re.sub(r'\d+', '', text)

[ ]  #função para remover os emojis
     def remove_emojis(text):
       return emoji.replace_emoji(text)

[ ]  #função para remover pontuação
     def remove_dots(text):
       return re.compile(f'[{re.escape(string.punctuation)}]').sub(' ', text)

[ ]  #função para remover stopwords
     nltk.download('stopwords')

     [nltk_data] Downloading package stopwords to /root/nltk_data...
     [nltk_data]   Unzipping corpora/stopwords.zip.
     True

[ ]  def remove_stopwords(text):
       stop_words = set(stopwords.words('english'))
       make_tokenize = TweetTokenizer()
       tokens = make_tokenize.tokenize( remove_dots( remove_numbers(remove_emojis(text) ) ))
       return [w.lower() for w in tokens if not w in stop_words]
```

Figura 4 – Funções para remover números, emojis, pontuações e stopwords.

6. Técnica de Stemming: A técnica de stemming é um processo linguístico usado no processamento de linguagem natural (NLP) para reduzir palavras à sua forma básica ou radical, removendo sufixos e prefixos. O objetivo do stemming é trazer as palavras relacionadas a um mesmo radical para uma forma comum, o que pode facilitar a análise de texto e a recuperação de informações. Por exemplo, as palavras "correr", "correndo", "corrida"e "corre". Ao aplicar a técnica de stemming, todas essas

palavras seriam reduzidas ao radical "corr", tornando mais fácil para um algoritmo de processamento de linguagem entender que essas palavras estão relacionadas.

Quando é realizada a busca ou recuperação de informações, a técnica de stemming ajuda a aumentar a probabilidade de encontrar todas as variações de uma palavra-chave, mesmo que elas tenham sufixos diferentes. Isso melhora a precisão da recuperação de informações relevantes. Na figura 5, mostra como foi feita essa etapa.

Figura 5 – Tweets antes e depois da técnica de stemming.

7. Análise de palavras-chave nos *tweets* coletados: O próximo passo é extrair as palavras-chave mais frequentes utilizadas nos *tweets* coletados. Isso foi feito através da utilização de algoritmos de processamento de linguagem natural para identificar palavras-chave importantes.

   O processo de agrupamento utiliza uma técnica de análise de dados chamada "agrupamento de palavras-chave"ou "clusterização de palavras-chave". Essa técnica é uma forma de aprendizado não supervisionado, onde o objetivo é identificar padrões ou grupos naturais nos dados. A função "intersection"refere-se à operação de encontrar os elementos comuns entre dois ou mais conjuntos. Essa função é utilizada em bibliotecas de análise de dados para identificar os elementos que aparecem em mais de um conjunto.

   A Figura 6 seleciona palavras-chave no dataset e percorre os textos em cada linha do DataFrame e identifica quais palavras-chave estão presentes em cada texto, atribuindo uma coluna "interest"com valor 1 se houver pelo menos uma palavra-chave

```python
#selecionando as palavras-chave no dataset
df['interest'] = 0
df['keywords'] = ''
for i in df.index:
  result = list(set(df['text'][i]).intersection(words))
  if len(result) > 0:
    df['interest'][i] = 1
    df['keywords'][i] = result
```

Figura 6 – Função para ler e selecionar as palavras-chave no dataset.

encontrada e armazenando as palavras-chave encontradas em outra coluna chamada "keywords". A Figura 7 mostra as *keywords* encontradas.

```
abandon, active, active_fault, administration, administrators, adversity, afraid, aftershocks, aghast, aid, airfall,
aliment,
alimentchow, allowance, amplitude, andesite, annihilated, anxious, apartment, appeal, apprehensive, ash, ashes,
ashfall, assistance, asthenosphere, asylum, bail_out, basalt, begging, bench, block, bomb, broken, calamity, caldera,
careful, cataclysm, catastrophe, celsius, central_vent, central_volcano, charity, chow, cinder, cinder_cone, cinder
kipuka, clearance, clinic, comestible, concerned, concerns, condo, conduit, cone, contribution, convulsion, coulee,
crater, craving, crisis, crowd, dacite, danger, death, deformation, demolished, destroyed, devastated, disasters,
difficulty, direction, directors, disaster, disintegrated, displace, distress, distressed, disturbed, donations,
dwelling,
earthquake, earthquake fault, ejecta, emergency, emission, emptying, endowment, entreaty, epicentre, erupt,
erupting, eruption, eruptions, erupts, evacuee, evacuate, evacuation, evacuee, exigency, extremity, fahrenheit,
fallout,
fault, feed, fidgety, fire, fissures, flow, fog, foodstuff, forsake, fumarole, gas, heavy, help, holocene, hut,
imploring,
insecurity, instability, institution, jittery, kilauea, killed, kipuka, lahar, larder, lava, laze, leadership, leave,
leilani,
lodging, lost, lua, luapele, macroseism, magma, magnitude, mainshock, meal, mess, microseism, migration,
misadventure, mischance, misfortune, mishap, monogenetic, mount, move_out, movement, nervous, nervy,
nourishment, nutriment, organization, orison, pahalacloud, pahoehoe, pali, path, paths, pele, peril, perilousness,
petition, phreatomagmatic, pleading, plight, plug, prayers, probability, problem, pull_out, pumice, pyroclastic, quake,
quaker, quit, ravaged, razed, refreshment, refuge, reliefsubsidy, removal, remove, repose, request, restless, rhyolite,
risk, riskiness, road, route, ruined, safe, safety, sanctuary, scared, scoria, seismicity, seism, seismicity, seismism,
seismograms, seismograph, seismology, service, shake, shaky, shattered, shelter, shelters, shifting, shock, shook,
seismic, sismo, situation, slag, slip, smashed, smoke, solfatara, spooked, steaming, store, subsidy, subsistence,
supervision, support, surface, surge, survive, sustenance, tears, tectonic, tectonics, temblor, tephra, terremoto,
tilt,
tragedy, transport, trembler, tremor, trouble, tuff, uneasy, uptight, urgent, vapor, vents, vents_steaming, volcanic,
volcano, volcanoes, volcanowatch, vulcan, vulcanian, volcanoes, vuleao, vulcdo, wasted, withdraw,
```

Figura 7 – Palavras-chave encontradas no dataset.

8. Apresentação dos resultados através de mapas e gráficos: Para tornar os resultados compreensíveis e visualmente atrativos, foram utilizados mapas e gráficos na apresentação das informações obtidas ao longo do estudo e serão apresentadas nos resultados.

   Ao seguir essa metodologia, foram alcançados os objetivos propostos na pesquisa, fornecendo insights sobre padrões de comportamento espaço-temporais, identificação de palavras-chave relevantes e apresentação visual dos dados coletados que estão disponíveis no *Projeto*[3].

[3] https://colab.research.google.com/drive/11IvLtipJrjB5VaAhHrjtajuIRrBOqchi?usp=sharing

5 Resultados

Está seção irá discorrer sobre os resultados obtidos da mineração dos dados do dataset utilizado com o intuito de descobrir quais as palavras-chave mais utilizadas.

5.1 Resultados da Classificação Geral

A seleção dos dados baseou-se na localização geográfica dos *tweets*, especificamente na região do Havaí, com as coordenadas de latitude 19.420120 e longitude -155.253137, dentro de um raio de 200 Km. A coleta foi restrita aos *tweets* postados durante o período de terremoto e erupção do vulcão Kilauea, que ocorreu aproximadamente de 3 de maio a 11 de junho de 2018. A Figura 8 representa a área onde os dados foram coletados. Durante a aquisição dos dados, a coleta ocorreu diariamente durante 12 horas por dia, abrangendo o período de 27 de abril de 2018 a 23 de maio do mesmo ano. Apenas os *tweets* de usuários que possuíam a informação de localização habilitada foram considerados. No total, foram coletados 6023 *tweets*.

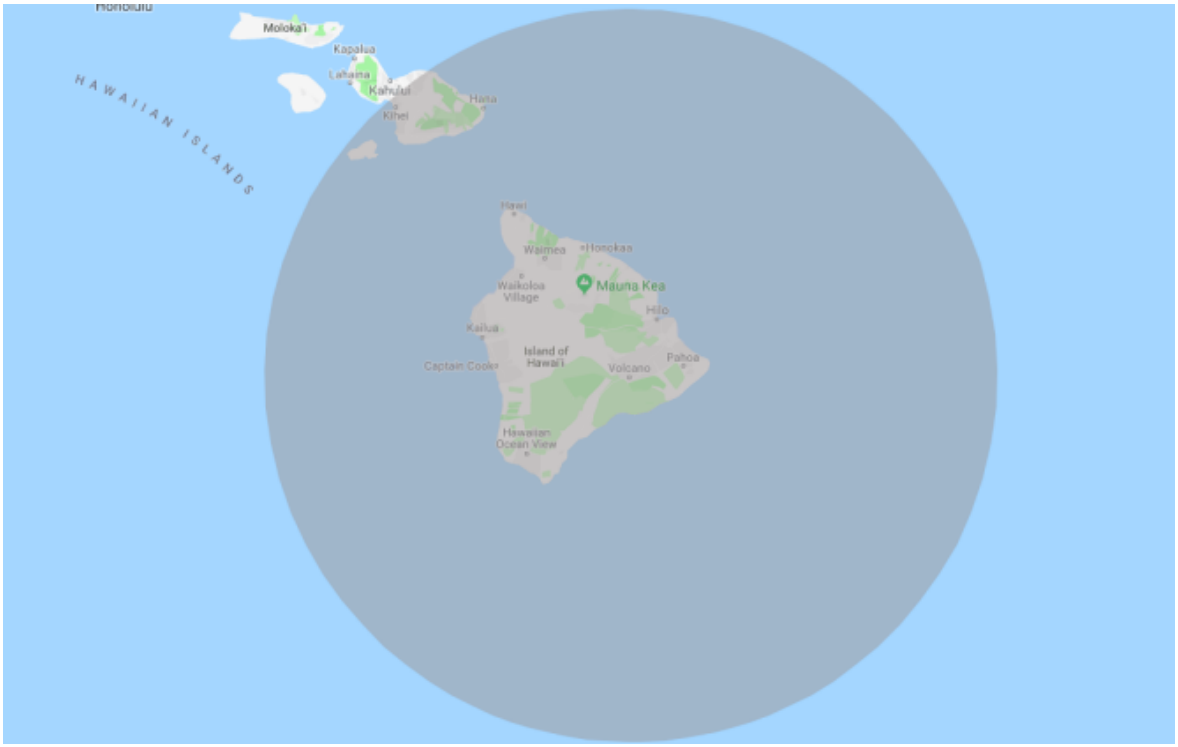

Figura 8 – Área de abrangência dos dados.

Os dados obtidos, estão devidamente organizados e, todos os *tweets* resultantes possuem relação direta com o desastre em questão. Tendo em vista que o Havaí é uma pequena ilha pertencente aos Estados Unidos, a ilha possui uma população pequena, contudo há uma quantidade considerável de usuários de mídias sociais. O conjunto de dados possui

os seguintes campos de interesse: ID (número de identificação do *tweet*), Data, Latitude, Longitude e Horário. As localizações foram disponibilizadas por latitude e longitude em um formato específico definido pelo Twitter. Existem dois objetos JSON usados para descrever o local associado a um *tweet*: coordenadas e *local*[1]. A Figura 9 apresenta um exemplo de *tweet* em formato JavaScript Object Notation(JSON).


```
{
 "created_at":"Thu Apr 06 15:24:15 +0000 2017",
 "id": 850006245121695744,
 "id_str": "850006245121695744",
 "text": "1/ Today we're sharing our vision for the future of the Twitter API platform!nhttps://t.co/XweGngmxlP",
 "user": {},
 "entities": {}
}
```


Figura 9 – Exemplo de como os tweets são disponibilizados.

Na fase de limpeza dos dados, foram aplicadas técnicas de Processamento de Linguagem Natural (PLN) para a normalização do texto do *tweet*, foram usadas técnicas de tokenização, stemming e remoção de ruídos. Diante do exposto, foi aplicada a tokenização no qual consiste em decompor a mensagem inteira em termos/palavras isoladas.

Em seguida nós utilizamos a stemming mantendo somente o sentido principal da palavra. Essa técnica foi feita utilizando a bilioteca do Python NLTK (Natural Language Tool Kit), com a finalidade de criar algoritmos que funcionam com linguagem natural.

E, por fim, removemos o que consideramos ruídos. Executamos um filtro para remoção de conteúdo duplicado e possíveis spams. Durante o processo de remoção de conteúdo duplicado foram identificados perfis com conteúdo similar para todas as suas postagens, perfis de divulgação e impulsionamento de conteúdo e marcas, perfis de postagens dinâmicas sobre hora, tempo e empregos, postagens que não são interessantes para este trabalho. Diante do exposto, verificamos quais as palavras mais frequentes nos *tweets* obtidos e foi gerada o gráfico de palavras da Figura 10.

[1] https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/geo-objects.html

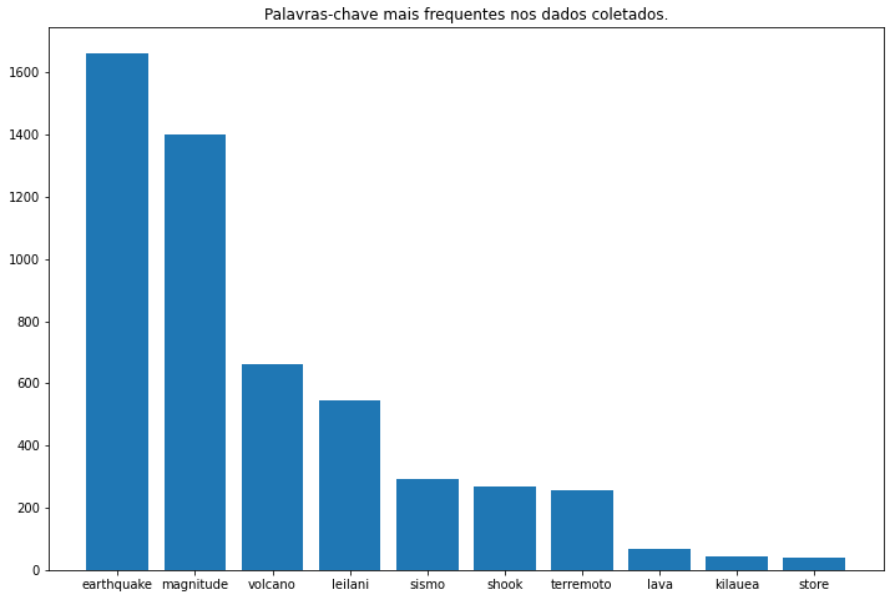

Figura 10 – Palavras-chave mais utilizadas nos tweets durante o evento de desastre no Havaí.

O mapa de calor apresentado na Figura 11 ilustra visualmente a distribuição geográfica das pessoas que publicaram no Twitter durante o período relacionado ao desastre em estudo. O mapa destaca as áreas com maior concentração de usuários da plataforma que fizeram postagens nesse período específico. A representação em forma de calor indica a densidade de pessoas nessas localidades, ou seja, quanto mais intenso for o tom de cor no mapa, maior será o número de usuários que fizeram publicações na região.

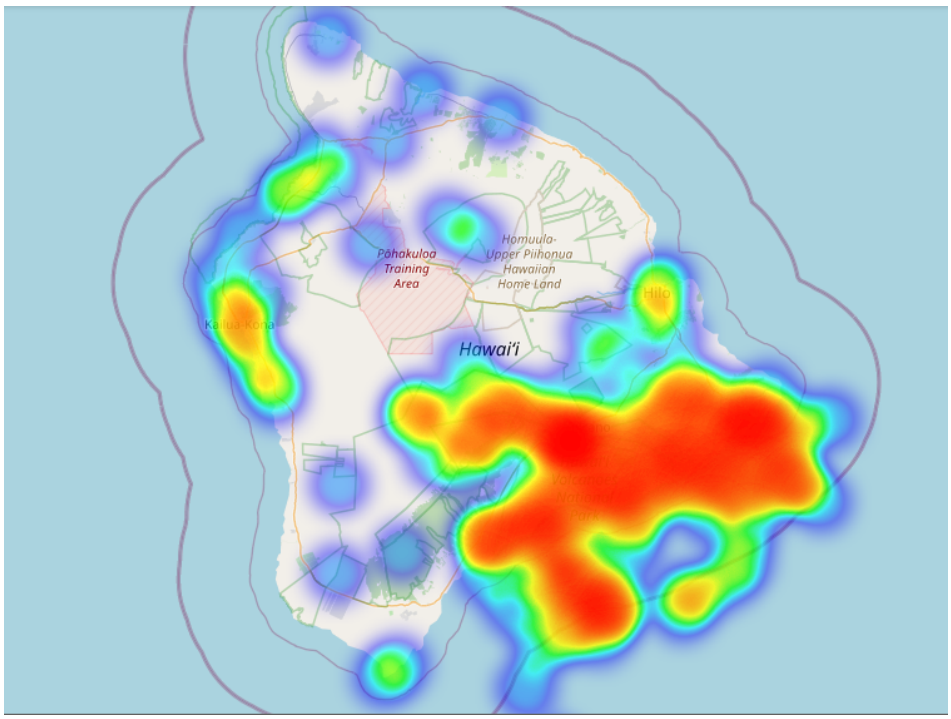

Figura 11 – Locais com maior quantidade de tweets no Havaí.

Conforme apresentado na Figura 12, além de ocorrer uma concentração de *tweets* durante o desastre que ocorreu no Havaí, também foi observada uma significativa concentração de *tweets* na ilha de Maui. O mapa de calor ilustra os locais com maior densidade de *tweets* publicados, indicando que a ilha de Maui também foi afetada pela disseminação de informações e discussões relacionadas ao desastre. Isso sugere que os usuários do Twitter na ilha de Maui estavam ativamente compartilhando e discutindo os eventos em tempo real, demonstrando um alto nível de engajamento e participação online. Essas percepções geográficas podem ser valiosas para entender como as comunidades locais estão reagindo e se envolvendo em situações de crise, fornecendo uma visão mais abrangente dos impactos e percepções do desastre no Havaí, incluindo a ilha de Maui.

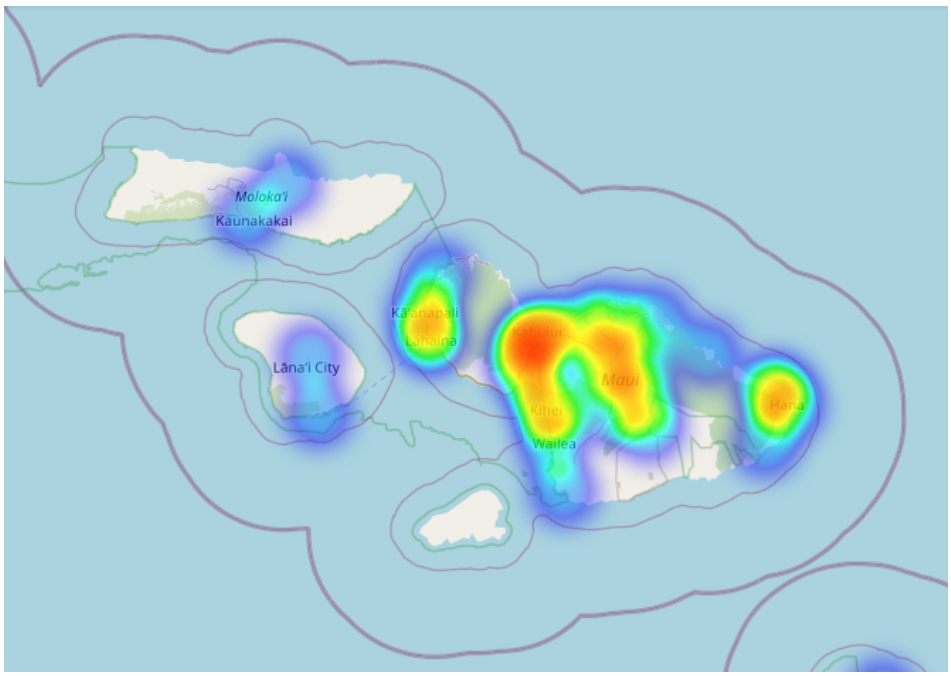

Figura 12 – Locais com maior quantidade de tweets na ilha de Maui.

Para melhor entendimento das palavras-chave usadas para encontrar a localização dos indivíduos, foi usada apenas a palavra-chave "earthquake"(terremoto), que foi a mais mencionada (1662 *tweets*) para criar um mapa de calor do terremoto em Kilauea e oferecer uma visão valiosa sobre a distribuição espacial e temporal desses eventos sísmicos durante o desastre. Essa ferramenta de visualização é uma abordagem eficaz para entender melhor a atividade sísmica durante o evento, permitindo que pesquisadores e autoridades tomem medidas mais informadas para a prevenção e resposta a desastres futuros.

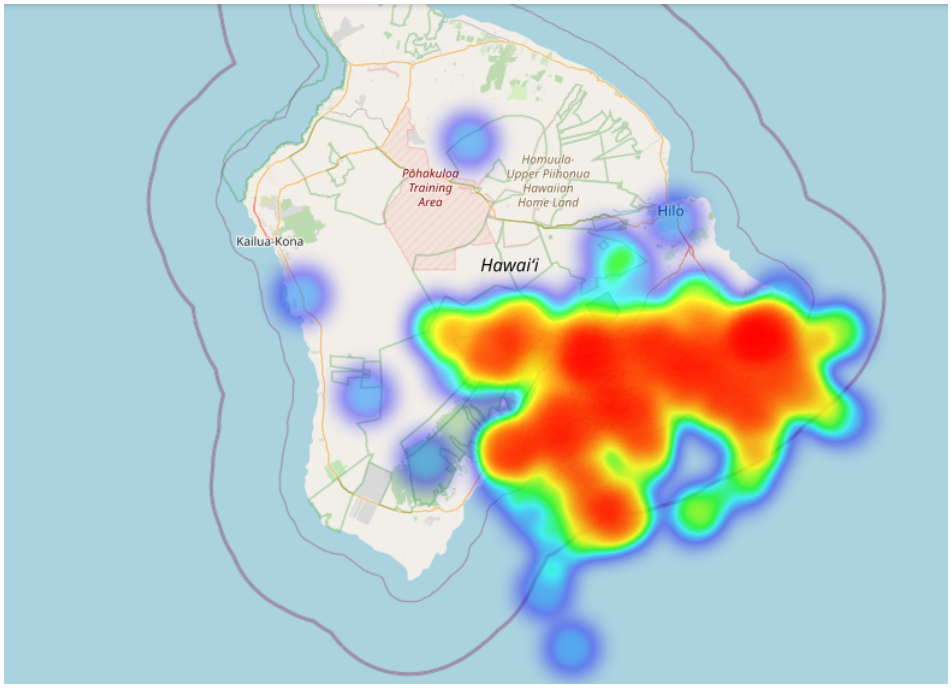

Figura 13 – Mapa de calor usando somente a palavra-chave mais citada nos tweets.

Como visto na imagem 13, a palavra-chave "earthquake", com sua menção em 1662 tweets, foi mais encontrada ao leste do Havaí, onde o desastre ocorreu de fato. Além disso, a análise dos tweets revelou que a maioria das menções à palavra-chave "earthquake"ocorreu dentro de um período de 24 horas após o desastre. Isso demonstra como as redes sociais desempenham um papel fundamental na disseminação rápida de informações durante situações de crise e emergência. Na figura 14, vemos o mesmo mapa mostrando a região dos tweets.

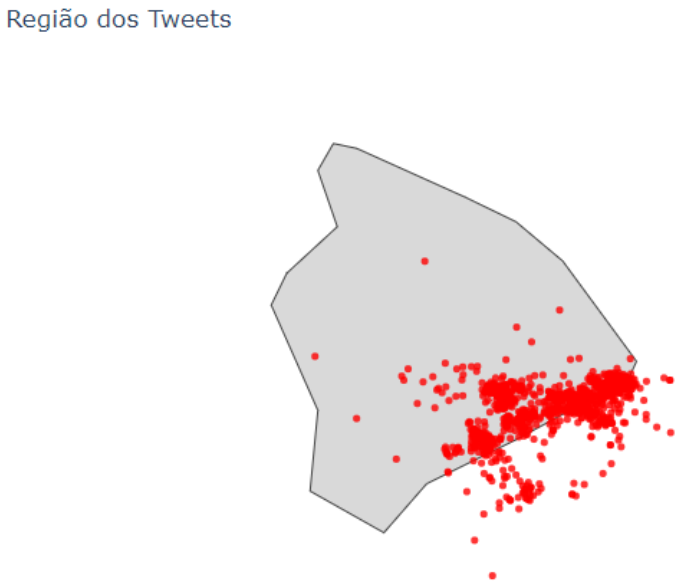

Figura 14 – Região dos tweets usando uma palavra-chave.

Os tweets relacionados ao terremoto no leste do Havaí incluíam diversos temas, desde relatos de moradores locais sobre a intensidade do tremor até mensagens de solidariedade e preocupação de pessoas de outras regiões do mundo. Além disso, hashtags como #HawaiiEarthquake, #PrayForHawaii e #NaturalDisaster foram amplamente utilizadas para categorizar e acompanhar as discussões sobre o evento.

A geolocalização dos *tweets* também permitiu identificar áreas mais afetadas pelo terremoto, fornecendo dados valiosos para equipes de resgate e autoridades locais. Além do leste do Havaí, algumas menções também foram registradas em regiões vizinhas, evidenciando o impacto sísmico em uma área mais ampla.

5.1.1 Discussão

A análise de palavras-chave associada à localização de pessoas nos desastres naturais pode ser uma ferramenta útil para ajudar a localizar as pessoas de forma mais rápida, especialmente quando combinada com a localização fornecida pelo *tweet*. Essa abordagem pode ajudar das seguintes formas:

1. Detecção de palavras-chave relevantes: O primeiro passo é identificar as palavras-chave relacionadas aos desastres naturais, como "terremoto", "inundação", "incêndio florestal"e assim por diante. Essas palavras-chave podem ser pré-definidas com base nos tipos comuns de desastres na região em questão.

2. Análise do *tweet*: Quando uma pessoa está em perigo devido a um desastre natural, ela pode usar as redes sociais para compartilhar informações sobre sua situação. A análise do conteúdo do *tweet* pode ajudar a identificar palavras-chave relevantes relacionadas ao desastre, como "terremoto em andamento", "preso em um prédio desabado", "preciso de ajuda"etc.

3. Extração de informações de localização: Muitos *tweets* possuem informações de localização, seja diretamente na mensagem ou através de recursos como o GPS do dispositivo. Essas informações podem incluir coordenadas geográficas, nomes de cidades ou pontos de referência próximos.

4. Cruzamento de dados: As palavras-chave relacionadas ao desastre extraídas do *tweet* podem ser combinadas com as informações de localização para criar um contexto mais preciso. Por exemplo, se alguém mencionar um terremoto em uma cidade específica, a combinação dessas informações pode ajudar a identificar áreas prováveis onde a pessoa pode estar localizada.

5. Priorização de resgate: Com base nas informações coletadas, é possível priorizar e direcionar os esforços de resgate para as áreas mais afetadas ou onde há maior proba-

bilidade de pessoas estarem em perigo. Isso pode acelerar o processo de localização e ajudar a salvar vidas.

6 Conclusão

A base de dados e formação de diretrizes disponibilizadas, permite que os cientistas, analistas de dados, defesa civil e comunidades afetadas por desastres, possam verificar e examinar situações anormais, facilitando encontrar a localização das pessoas de forma prática. Neste trabalho, realizamos a mineração de dados referentes ao desastre natural do Havaí, onde foram empregadas análises de comportamento de resposta em eventos de desastres usando dados de mídia social. Através da base de dados construída, alcançamos informações úteis em relação ao comportamento das pessoas e em relação ao desastre, e correlacionamos tais informações à geografia da ilha do Havaí. A construção da base de dados se deu por meio da coleta de *tweets*, e esta passou por várias fases de limpeza com a finalidade de se tornar uma base de dados útil para a comunidade científica.

É importante observar que a análise de palavras-chave e a localização fornecida pelo *tweet* são apenas uma parte do processo de localização de pessoas em desastres naturais. É fundamental contar com uma infraestrutura robusta de resposta às emergências, equipes de resgate treinadas e coordenadas, além de outras fontes de informação, como sistemas de alerta e monitoramento de desastres, para garantir uma resposta eficaz.

Em perspectivas futuras, será realizada a coleta de informações de outros tipos de desastres a fim de validar as diretrizes apontadas neste trabalho. Ainda, como consequência natural do trabalho, vislumbro a construção de um sistema de localização rápida de pessoas nos desastres usando os *tweets* coletados.

Referências

AHN SON, C. Understanding public engagement on twitter using topic modeling: The 2019 ridgecrest earthquake case. In: *International Journal of Information Management Data Insights*. [S.l.: s.n.], 2021. Citado na página 11.

AVVENUTI, M.; CRESCI, S.; MARCHETTI, A.; MELETTI, C.; TESCONI, M. Ears (earthquake alert and report system): a real time decision support system for earthquake crisis management. In: ACM. *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining*. [S.l.], 2014. p. 1749–1758. Citado 3 vezes nas páginas 15, 18 e 23.

BECKEN, S.; STANTIC, B.; CHEN, J.; ALAEI, A. R.; CONNOLLY, R. M. Monitoring the environment and human sentiment on the great barrier reef: assessing the potential of collective sensing. *Journal of environmental management*, Elsevier, v. 203, p. 87–97, 2017. Citado 2 vezes nas páginas 20 e 23.

BISHOP, G.; WELCH, G. et al. An introduction to the kalman filter. *Proc of SIGGRAPH, Course*, v. 8, n. 27599-23175, p. 41, 2001. Citado na página 16.

CHAE, J.; THOM, D.; JANG, Y.; KIM, S.; ERTL, T.; EBERT, D. S. Public behavior response analysis in disaster events utilizing visual analytics of microblog data. *Computers & Graphics*, Elsevier, v. 38, p. 51–60, 2014. Citado 2 vezes nas páginas 14 e 15.

FROZZA, A. A.; MELLO, R. dos S.; COSTA, F. d. S. da. An approach for schema extraction of json and extended json document collections. In: IEEE. *2018 IEEE International Conference on Information Reuse and Integration (IRI)*. [S.l.], 2018. p. 356–363. Citado na página 24.

KANAKARAJ, M.; GUDDETI, R. M. R. Nlp based sentiment analysis on twitter data using ensemble classifiers. In: IEEE. *2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN)*. [S.l.], 2015. p. 1–5. Citado 2 vezes nas páginas 14 e 15.

KARIMIZIARANI, M. Social response and disaster management: Insights from twitter data assimilation on hurricane ian. In: *Journal Pre-Proof*. [S.l.: s.n.], 2023. Citado 2 vezes nas páginas 22 e 23.

KRYVASHEYEU, Y.; CHEN, H.; OBRADOVICH, N.; MORO, E.; HENTENRYCK, P. V.; FOWLER, J.; CEBRIAN, M. Rapid assessment of disaster damage using social media activity. *Science advances*, American Association for the Advancement of Science, v. 2, n. 3, p. e1500779, 2016. Citado 3 vezes nas páginas 14, 19 e 23.

MACEACHREN, A. M.; JAISWAL, A.; ROBINSON, A. C.; PEZANOWSKI, S.; SAVELYEV, A.; MITRA, P.; ZHANG, X.; BLANFORD, J. Senseplace2: Geotwitter analytics support for situational awareness. In: IEEE. *2011 IEEE conference on visual analytics science and technology (VAST)*. [S.l.], 2011. p. 181–190. Citado na página 16.

MALDONADO, M.; ALULEMA, D.; MOROCHO, D.; PROAÑO, M. System for monitoring natural disasters using natural language processing in the social network twitter. In: IEEE. *2016 IEEE International Carnahan Conference on Security Technology (ICCST)*. [S.l.], 2016. p. 1–6. Citado na página 15.

MARTÍN, Y.; LI, Z.; CUTTER, S. L. Leveraging twitter to gauge evacuation compliance: Spatiotemporal analysis of hurricane matthew. *PLoS one*, Public Library of Science, v. 12, n. 7, p. e0181701, 2017. Citado 4 vezes nas páginas 14, 15, 19 e 23.

POWERS DEVARAJ, A. D. J. S. M. Using artificial intelligence to identify emergency messages on social media during a natural disaster: A deep learning approach. In: *International Journal of Information Management Data Insights*. [S.l.: s.n.], 2023. Citado 3 vezes nas páginas 11, 21 e 23.

SAKAKI, T.; OKAZAKI, M.; MATSUO, Y. Earthquake shakes twitter users: real-time event detection by social sensors. In: ACM. *Proceedings of the 19th international conference on World wide web*. [S.l.], 2010. p. 851–860. Citado na página 16.

SANTANA, D. D. S. *Navegação terrestre usando unidade de medição inercial de baixo desempenho e fusão sensorial com filtro de Kalman adaptativo suavizado*. Tese (Doutorado) — Universidade de São Paulo, 2011. Citado na página 16.

SUFI. A decision support system for extracting artificial intelligence-driven insights from live twitter feeds on natural disasters. In: *Decision Analytics Journal*. [S.l.: s.n.], 2022. Citado 2 vezes nas páginas 21 e 23.

WU, D.; CUI, Y. Disaster early warning and damage assessment analysis using social media data and geo-location information. *Decision Support Systems*, Elsevier, v. 111, p. 48–59, 2018. Citado 2 vezes nas páginas 20 e 23.

ZHOU, Z. Victimfinder: Harvesting rescue requests in disaster response from social media with bert. In: *Computers, Environment and Urban Systems*. [S.l.: s.n.], 2022. Citado 3 vezes nas páginas 11, 20 e 23.

ZVAREVASHE, K.; OLUGBARA, O. O. A framework for sentiment analysis with opinion mining of hotel reviews. In: IEEE. *2018 Conference on Information Communications Technology and Society (ICTAS)*. [S.l.], 2018. p. 1–4. Citado 2 vezes nas páginas 14 e 15.

TERMO DE AUTORIZAÇÃO PARA PUBLICAÇÃO DIGITAL NA BIBLIOTECA "JOSÉ ALBANO DE MACEDO"

Identificação do Tipo de Documento

( ) Tese

( ) Dissertação

(x) Monografia

( ) Artigo

Eu, Joanny Eva Pacheco Monteiro,

autorizo com base na Lei Federal nº 9.610 de 19 de Fevereiro de 1998 e na Lei nº 10.973 de 02 de dezembro de 2004, a biblioteca da Universidade Federal do Piauí a divulgar, gratuitamente, sem ressarcimento de direitos autorais, o texto integral da publicação Análise de Palavras-Chave Associadas à Localização de Pessoas nos Desastres Naturais

de minha autoria, em formato PDF, para fins de leitura e/ou impressão, pela internet a título de divulgação da produção científica gerada pela Universidade.

Picos-PI 14 de agosto de 2023.

Joanny Eva Pacheco Monteiro

Assinatura