*M*ineração de Textos usando *Word Embeddings* com Contexto Geográfico

Antônio Ronaldo da Silva

SERVIÇO DE PÓS-GRADUAÇÃO DA FACOM-UFMS

Data de Depósito:

Assinatura:_____

*M*ineração de Textos usando *Word Embeddings* com Contexto Geográfico

Antônio Ronaldo da Silva

Orientador: *Prof. Dr. Ricardo Marcondes Marcacini*

Dissertação de mestrado apresentada à Faculdade de Computação da Universidade Federal de Mato Grosso do Sul - FACOM-UFMS como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação.

UFMS - Campo Grande
Setembro/2022

Aos meus pais,
José e Maria,

À minha família,

À minha namorada,

Ao Prof. Dr. Ricardo Marcondes Marcacini.

Agradecimentos

A Deus, por sempre estar presente em minha vida e por guiar meus passos.

Aos meus pais José Maria e Maria José, aos meus irmãos, Carlos, Edicarlos, Roberto, Carliane e Rodrigo, a minha namorada, Jakeline, por todo amor, dedicação, paciência e compreensão nos momentos de ausência para a realização do presente estudo e ter me apoiando em todos os momentos, meu eterno muito obrigado!

Meu agradecimento especial ao Professor Ricardo Marcondes Marcacini, meu orientador, por me direcionar durante todo o desenvolvimento desta dissertação de mestrado. Sempre disposto, me motivando e tentando me mostrar os caminhos a seguir. Obrigado pela paciência, por nunca me deixar desistir e pela dedicação nas orientações.

Aos demais professores, técnicos e servidores da FACOM - UFMS.

Ao grupo do LIA, especialmente por me proporcionar amigos como o Allan Menchik, Rafael Torres, Carlos Monteiro e Juan Carvalho.

Aos amigos da pós-graduação: Matheus Santana, Mário Carvalho e Calebe Lemos.

À Faculdade de Computação (FACOM) da Universidade Federal de Mato Grosso do Sul (UFMS).

Agradecimento ao apoio financeiro da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)..

Por fim, agradeço a todos que contribuíram comigo nesse trajeto percorrido.

Abstract

Many essential phenomena are related to a geographic context, such as events extracted from textual bases in economics, public health, and urban violence. Manually analyzing events would be impractical, considering their significant volume and different data sources. Thus, there was a need for intelligent computational methods such as Text Mining that enable the exploration of textual content with geographic information and return patterns that traditional models would not find. The traditional model for analyzing the relationship between terms and regions is to calculate the probability of a term being used in texts associated with a region, in general, through the frequency of terms in regions. However, it is recognized that this approach fails for new terms presented to a model and texts with ambiguous terms. In this context, models based on *Word Embeddings* are recognized for improving the identification of the relationships between a word and the possible associated location. In this sense, this project investigates textual representations based on *Word Embeddings* from BERT models (*Bidirectional Encoder Representations from Transformers*) in a fine-tuning process, in which the georeferenced information of the texts is used as context. We named this proposal the *GeoTransformers Language Model*. One of the differentials of this proposal is to automatically identify macro-regions and micro-regions from the events and use them as a context for fine-tuning a language model. Compared to other models in the literature, the results generated by the *GeoTransformers* model obtained higher values for precision metrics, recall, F1-Score. Moreover, our model was the only one capable of dealing with regions with fewer events.

Resumo

Muitos fenômenos importantes estão relacionados a um contexto geográfico, como eventos extraídos de bases textuais na área da economia, saúde pública, violência urbana e questões sociais. A análise de eventos de maneira manual seria impraticável considerando a sua grande quantidade e as diversas formas nas quais os dados são encontrados. Assim, passou-se a ter a necessidade de processos baseados em métodos computacionais inteligentes como a Mineração de Textos que, por meio das suas etapas, torna capaz a exploração do conteúdo textual com informação geográfica e retorna padrões que não seriam encontrados por modelos tradicionais. O modelo tradicional para analisar a relação entre termos e regiões é o de calcular a probabilidade de um termo ser utilizado em textos associados a uma região, em geral, por meio da frequência de termos em regiões. No entanto, é reconhecido que essa abordagem falha para novos termos apresentados a um modelo, bem como para textos com termos ambíguos. Nesse contexto, modelos baseados em *Word Embeddings* são reconhecidos por melhorar a identificação das relações entre uma palavra e o possível local associado. Nesse sentido, neste projeto são investigadas representações textuais baseadas em *Word Embeddings* do modelo BERT (*Bidirectional Encoder Representations from Transformers*) em um processo de ajuste fino, na qual as informações georreferenciadas dos textos são utilizadas como contexto, culminando na proposta deste trabalho denominada *GeoTransformers Language Model*. Um dos diferenciais da proposta é automaticamente identificar macrorregiões e microrregiões a partir dos eventos e utilizá-las como contexto para ajuste fino de um modelo de linguagem. Os resultados gerados pelo modelo *GeoTransformers*, em comparação com outros modelos da literatura, apresentaram maiores valores para métricas de precisão, revocação, F1-Score. Além disso, o modelo proposto foi o único capaz de lidar com regiões com menor quantidade de eventos e difíceis de classificar.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

BoW Bag of Words

PLN Processamento de Linguagem Natural

PTE *Predictive Text Embeddings*

MLM Modelagem de Linguagem Mascarada

NSP Previsão da Próxima Frase

GIR Recuperação de informação geográfica

NER Reconhecimento de entidade nomeada

CNN Rede Neural Convolucional

GSAM Modelo de Endereço Geoespacial-Semântico

MAIR Modelo Atento para Incorporação de Rótulos

LEAM *Label-Embedding Attentive Model*

BERT *Bidirectional Encoder Representations from Transformers*

NBSVM *Naive Bayes Support Vector Machines*

GDELT *Global Database of Events, Language, and Tone*

KDE Estimativa de Densidade do Kernel

GNN Redes Neurais de Grafos

KNN *k-Nearest Neighbors*

TF-IDF *Term Frequency - Inverse Document Frequency*

TF *Term Frequency*

CAPÍTULO

1

Introdução

1.1 Contextualização e Motivação

A informação georreferenciada é importante para diversos sistemas de apoio à tomada de decisão [Pei et al., 2020]. Em geral, muitos fenômenos importantes estão relacionados a um contexto geográfico, como eventos na área da economia, saúde pública, violência urbana, questões sociais, entre outras. Em especial, neste trabalho, há o interesse em bases textuais com informações geográficas associadas, a partir das quais são extraídos os conjuntos de eventos [Chen and Li, 2020]. Um evento pode ser compreendido como algo que ocorre em determinado tempo e local [Allan, 2012]. A análise de eventos a partir de textos é um tema que vem atraindo a atenção de muitos pesquisadores, que encontraram nessas grandes bases de textos uma forma de mapear e interpretar diferentes fenômenos que ocorrem em nossa sociedade [Hogenboom et al., 2016; Xiang and Wang, 2019].

Os dados de eventos são escritos em linguagem natural. Por isso, são considerados um tipo de dado não estruturado - o que deixa a sua manipulação ainda mais complexa. No entanto, os textos georreferenciados podem oferecer muitas oportunidades para as pesquisas em um vasto mar de informações valiosas, como palavras-chave, tópicos, topônimos, entidades e sentimentos, por meio das ligações entre acontecimentos extraídos dos textos e locais de ocorrências [Hu, 2018].

A análise dos eventos de maneira manual seria impraticável, haja vista a grande quantidade e as diversas formas em que os dados são encontrados. Assim, surgiu a necessidade de métodos computacionais inteligentes que fos-

sem capazes de explorar conteúdo textual, contendo informação geográfica, extraído de portais de notícias, redes sociais, blogs e bases de conhecimento como o *Wikipédia*, de forma ágil e precisa [Purves et al., 2018].

Buscando sanar esse problema, pesquisas computacionais passaram a ser utilizadas por intermédio de técnicas automatizadas para extração de conhecimento em base de dados, tais como a Mineração de Textos Georreferenciados e Aprendizado de Máquina [Hu, 2018; Pei et al., 2020]. Recentemente, vêm sendo apresentados estudos na literatura que buscam a melhor forma para representar textos, possibilitando a captura dos diversos cenários semânticos dos quais uma palavra pode ser encontrada dentro de uma frase (sensíveis ao contexto) [Sinoara, 2021].

Diante dessa perspectiva, estudos promissores envolvendo *Word Embeddings* têm sido utilizado como alternativas para lidar com essas representações [Aggarwal, 2018], pois permitem mensurar as diversas informações contidas nos dados e mapeá-las em um espaço multidimensional que contém propriedades geométricas utilizadas no processo de extração de padrões. Para esse objetivo, alguns modelos como *word2vec* [Mikolov et al., 2013], *FastText* [Bojanowski et al., 2017], *ELMo* [Peters et al., 2018] e *BERT* [Devlin et al., 2018] têm sido utilizados em muitas aplicações e alguns deles serão discutidos em mais detalhes neste trabalho.

Há vários métodos conhecidos que utilizam *Word Embeddings* para extrair relações semânticas entre os textos levando em conta o seu contexto geoespacial. Muitos desses métodos são baseados no *word2vec* [Mikolov et al., 2013], devido a sua eficácia em aprender *Word Embeddings* com bases textuais das palavras. Com o tempo, surgiram várias pesquisas com o objetivo de criar algoritmos com maior eficácia para produzir *Word Embeddings*. Os resultados foram algoritmos como *GloVe* [Pennington et al., 2014] e *FastText* [Joulin et al., 2016]. Esses algoritmos são exemplos que buscam transformar uma representação (textual) complexa em outra representação (de máquina) que possibilite a utilização de inúmeros métodos de aprendizado de máquina.

O aprendizado de *Word Embeddings* é baseado no conceito de modelo de linguagem - que pode ser obtido por meio de redes neurais. Explora-se a probabilidade de um conjunto de palavras aparecerem em um mesmo contexto [Mikolov et al., 2013; Pennington et al., 2014; Joulin et al., 2016; Gudivada, 2018]. Para um modelo de linguagem de sucesso, estima-se a distribuição sequencial de palavras, indo além da codificação da estrutura gramatical e do conhecimento que pode estar contido nos corpora de treinamento. Um modelo de linguagem consegue associar com alta probabilidade a sequência “Rio de Janeiro” do que “Lago de Janeiro” ou diferenciar o sentido semântico de uma palavra. Por exemplo, o vocábulo “banco” pode estar se referindo ao banco de

uma praça ou a uma instituição financeira.

Palavras tendem a seguir padrões geográficos de uso, de modo que é possível observar as relações existentes entre o conteúdo de um texto e o ponto de sua origem geoespacial [Cocos and Callison-Burch, 2017]. Em muitos casos, a correlação é evidente; é comum ouvir conversas sobre praias em cidades litorâneas, tendo em vista que pessoas que estão próximas a praias tendem a ser mais propícias a falarem sobre esse tema [Dassereto et al., 2020]. Já em outros momentos, não é possível identificar com clareza a existência de tal relação.

Em geral, o padrão com que os textos são gerados tende a depender do local de origem de seus usuários no momento da sua publicação [Andogah et al., 2012]. Por exemplo, um portal de notícias sobre acontecimentos do Nordeste Brasileiro tende a ter um léxico de expressões, acontecimentos, pessoas e organizações diferente de um portal de notícias do Sul do Brasil.

O modelo tradicional para analisar a relação entre termos e regiões é calcular a probabilidade de um termo ser utilizado em textos associados a uma região, em geral, por meio da frequência de termos em regiões [Purves et al., 2018]. No entanto, é reconhecido que essa abordagem falha para novos termos apresentados a um modelo, bem como para textos com termos ambíguos [Purves et al., 2018].

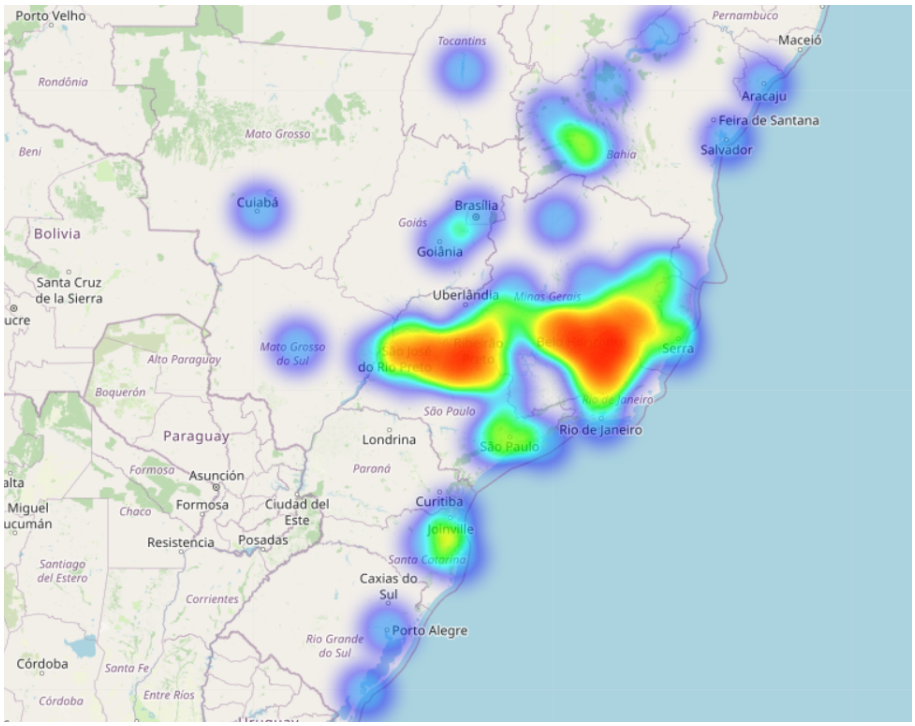
Word Embeddings são rotineiramente treinadas para aprender que palavras semanticamente similares ocorrem dentro do mesmo contexto textual. Com isso, não apenas é possível pesquisar a relevância das palavras, levando em consideração as posições em que são distribuídas ao longo da frase, mas também proporcionam-se estudos que levam em consideração a similaridade semântica entre palavras e o seu contexto geoespacial.

Dessa forma, esse tipo de *Word Embeddings* é útil para analisar os impactos de um determinado evento (extraído de um texto) em uma região distinta, como: desastre natural, conflitos urbanos, políticas de segurança pública, estudo de efeitos climáticos ou uma doença infecciosa. É útil também na construção de indicadores inteligentes em diversos domínios.

Embora existam algumas iniciativas para treinamento de *word embeddings* a partir de textos que contenham informação georreferenciada [Konkol et al., 2017; Cocos and Callison-Burch, 2017; Gong et al., 2020], ainda há desafios de pesquisa que precisam ser superados para o uso efetivo em análise de eventos. Dentre os diversos desafios em aberto, neste trabalho, foi investigado o treinamento de *word embeddings* com contexto geográfico, não apenas para representação de textos, mas também para análises preditivas sobre prováveis localidades de ocorrência considerando o conteúdo de eventos, especialmente em cenários nos quais a informação geográfica do evento é ausente ou muito

abrangente.

Figura 1.1: Mapa de calor que indica as regiões de maior frequência de eventos.



Fonte: Projeto *Websensors Analytics* [Marcacini et al., 2017]

Um método capaz de estimar a região a qual um texto foi escrito (regiões de interesse) a partir de dados não estruturados é potencialmente útil para melhorar o desempenho de sistemas de recuperação de informação e de sistemas de recomendação baseados em contexto geográfico [Weitzel et al., 2010]. No contexto deste trabalho, há maior interesse em utilizar modelos treinados em bases de eventos para gerar automaticamente a cobertura geográfica de um texto e respectivos mapas de calor, conforme ilustrado na Figura 1.1. Nesse exemplo, é ilustrado o mapa de calor do texto "*mortes por febre amarela no Brasil*", onde regiões mais quentes do mapa indicam maior frequência de eventos. Na Figura 1.1, é utilizada uma base de eventos extraída de notícias sobre febre amarela no Brasil no ano de 2017 do projeto *Websensors*[1]. Observe que, enquanto no treinamento do modelo é necessário possuir textos georreferenciados e anotados, no uso do modelo isso é mais flexível, pois apenas um pequeno trecho textual é utilizado de entrada para inferência de cobertura geográfica.

[1]Websensors Analytics: https://websensors.net.br/

1.2 Hipótese e Objetivos

A hipótese que rege esta pesquisa é a de que modelos de *word embeddings* baseados em contexto geográfico são mais eficazes em estimar regiões de interesse do que estratégias atuais que realizam estimativas por meio de probabilidade de ocorrência de termos em determinadas regiões. O objetivo geral deste trabalho de mestrado é desenvolver e avaliar um processo de mineração de textos baseado em *word embeddings* de contexto geográfico, visando a geração automática de mapas de calor a partir de expressões textuais. Para atingir esse objetivo geral, são propostos três objetivos específicos:

1. Propor e avaliar um processo de Mineração de Textos em que as etapas de pré-processamento e extração de padrões envolvam, respectivamente, o treinamento de *word embeddings* considerando o contexto geográfico e modelos para inferir regiões (localidades) de interesse a partir de expressões textuais.

2. Adaptar e treinar modelos de *word embeddings* que consideram o contexto geográfico utilizando como base o modelo BERT [Devlin et al., 2018].

3. Desenvolver uma ferramenta computacional que explore modelos pré-treinados capazes de inferir as regiões às quais pertence uma expressão textual de interesse do usuário, retornando os resultados através de mapas de calor.

1.3 Organização do Texto

Os próximos capítulos deste trabalho estão organizados da seguinte forma:

- **Capítulo 2:** são apresentados os fundamentos deste projeto com conceitos básicos sobre Mineração de Textos, Recuperação de Informação Geográfica e *Word Embeddings*.

- **Capítulo 3:** trabalhos relacionados sobre combinação de informações textuais e informações geográficas são discutidos, indicando quais as lacunas existentes que motivaram o desenvolvimento desta proposta.

- **Capítulo 4:** a metodologia e discussões dos resultados obtidos neste trabalho.

- **Capítulo 5:** as conclusões, dificuldades encontradas e sugestões para possíveis estudos futuros.

CAPÍTULO

2

Fundamentos

Nos últimos anos, grandes repositórios de textos vêm sendo gerados e armazenados por empresas, universidades, governos e outras organizações [Amaral, 2016], tornando-se uma matéria-prima rica e abundante para a geração de conhecimento [Miller and Goodchild, 2015].

A universalização dos dispositivos computacionais com GPS integrado tem contribuído para o crescimento de bases textuais capazes de relacionar localizações geográficas com texto em linguagem natural por meio dos *metadados* encontrados em suas bases [Hu, 2018]. Esse tipo de informação pode ser encontrado em notícias da *internet*, postagens em redes sociais, *sites* de avaliação de viagens e textos da *Wikipédia* [Li et al., 2019]. Devido à grande quantidade de textos diariamente publicados nesses repositórios, técnicas automatizadas de extração de conhecimento baseadas em mineração de textos e recuperação de informações georreferenciadas têm sido o objeto de investigações acadêmicas.

Neste capítulo, são apresentados os principais conceitos relacionados a este trabalho. Na Seção 2.1, são abordadas as etapas do processo de Mineração de Textos. Posteriormente, são descritos os Sistemas de Recuperação de Informações Geográficas, conforme apresentado na Seção 2.2. A representação de textos por *Word Embedding* é um tópico importante neste trabalho e métodos relacionados são discutidos na Seção 2.3.

2.1 Mineração de Textos

A Mineração de Textos (MT) pode ser definida como um conjunto de técnicas para extrair padrões úteis em bases de dados textuais. É uma área

interdisciplinar que engloba recuperação de informações, aprendizado de máquina, estatística e processamento de linguagem natural [Fayyad et al., 1996; Rezende, 2003; Aranha and Passos, 2006].

A MT pode ser dividida em cinco grandes etapas [Rezende, 2003], conforme ilustrado na Figura 2.1. A primeira etapa é a **Identificação do Problema**. Ela visa compreender o domínio da aplicação, estabelecer objetivos e selecionar bases textuais e ferramentas que serão utilizadas. A segunda etapa é formada pelo **Pré-processamento dos textos**. Consiste em eliminar eventuais inconsistências nas bases textuais, empregar a padronização dos dados e gerar uma representação estruturada e concisa dos textos. Em seguida, há a etapa de **Extração de Padrões**. Ela considera a aplicação de técnicas de aprendizado de máquina capazes de extrair padrões do *corpus* textual. A etapa de **Pós-processamento** tem a finalidade de validar ou não o conhecimento obtido na etapa anterior, podendo ainda ser feito o refinamento do conhecimento adquirido. Por fim, há a **Utilização do Conhecimento** válido e sua disponibilização para os envolvidos [Rezende, 2003].

Figura 2.1: Etapas da Mineração de Textos.

Fonte: Rezende [2003]

Nas seções seguintes cada etapa é discutida em mais detalhes, já considerando o contexto de textos georreferenciados.

2.1.1 Identificação do problema

Projetar os objetivos e metas a serem alcançados com um processo de Mineração de Textos Georreferenciados é uma etapa importante, pois afetará a escolha das técnicas das próximas etapas.

Uma atividade relevante é a identificação e seleção das coleções de dados textuais georreferenciados. Algumas bases de textos, como eventos extraídos de notícias, já possuem *metadados* associados, incluindo entidades geográficas e respectivas informações de latitude e longitude. No entanto, a grande maioria das bases possui apenas os textos, sendo necessário o uso de técnicas de pré-processamento para identificação de entidades geográficas conhecidas como *geoparsing* e *geocoding*, discutidas em detalhes na Seção 2.2.

Por fim, há uma outra categoria de coleção textual na qual as entidades geográficas não estão explícitas. Desse modo, é necessário recorrer ao tópico do texto para estimar prováveis localizações. Por exemplo, um texto relatando eventos sobre carnaval tem maior chance de ocorrer no Brasil e no Rio de Janeiro. Essa última categoria é especialmente importante para este trabalho, uma vez que se busca explorar bases de textos que possuem entidades geográficas explícitas para aprender modelos capazes de estimar as regiões de ocorrência de textos e tópicos sem entidades geográficas definidas.

2.1.2 Pré-processamento

Os textos pré-selecionados para análise, na etapa anterior, podem vir de base de dados única ou, geralmente, são provenientes de diversas bases, sendo comum que os dados não estejam preparados para aplicação dos métodos de Extração de Padrões [Rezende, 2003]. Dessa forma, o pré-processamento visa realizar a estruturação dos textos georreferenciados de maneira a torná-los processáveis por algoritmos de aprendizado de máquina.

Em um primeiro momento, a base textual é processada para padronização e limpeza dos textos da coleção. O objetivo é deixar todos os dados em um único formato, removendo os textos duplicados, truncados, bem como caracteres não alfanuméricos [Sansome and Hacker, 2020; Makrehchi and Kamel, 2008]. Em seguida, é realizado o processo de transformação dos textos em um formato estruturado, geralmente usando um modelo de representação no espaço-vetorial. Nesse modelo, cada documento textual é representado por meio de um vetor m-dimensional. Cada dimensão representa um atributo do texto e o valor indica o peso desse atributo para o documento.

Diferentes técnicas de pré-processamento obtêm diferentes modelos de representação no espaço-vetorial. Um modelo clássico na área é o de *Bag-of-Words* (BoW). Esse modelo é baseado na frequência da palavra no documento

(por exemplo, tf, tf-idf, e binário que podem ser usados como esquemas de pesos dos termos). Nessa estrutura, as palavras são consideradas atributos e os pesos indicam a frequência dessas palavras no respectivo documento. Modelos BoW geralmente são descritos no formato Matriz Documento-Termo ou Matriz Atributo-Valor [Naseem et al., 2021; Rezende et al., 2011; Weiss et al., 2010], conforme a Tabela 2.1. Deve-se observar que um atributo especial (última coluna) foi adicionado, determinando a classe ou categoria de um documento. Esse atributo é utilizado como informação rotulada em métodos supervisionados de extração de padrões, como métodos de classificação.

No caso de um modelo BoW, são geralmente aplicadas técnicas para apoiarem na identificação de atributos mais relevantes, como: (i) *stemming*, que consiste na obtenção de radicais, tendo como propósito reduzir as palavras ao seu radical; (ii) remoção de *stopwords*, que consiste na exclusão de palavras repetitivas que não possuem uma semântica relevante no texto, como artigos, preposições e verbos auxiliares; (iii) padronização da caixa de texto, que tem como objetivo padronizar todas as letras de uma palavra em maiúsculo ou minúsculo.

O BoW é representado por uma matriz composta por n linhas e m colunas (ver Tabela 2.1). As n linhas estão relacionadas aos documentos de d_1 a d_n da coleção textual. As colunas indicam os m atributos de t_1 a t_m que discriminam os documentos. Os rótulos ou classes são simbolizados pela coluna Y. Os valores da matriz, a_{11} a a_{nm}, descrevem uma função que destaca o grau de relevância do atributo para um documento. É importante destacar que há outras técnicas envolvidas nessa etapa de pré-processamento, como identificação de n-grams (termos compostos e geração de *Bag-of-n-grams*), lematização e substantivação. No entanto, elas não serão abordadas em profundidade, pois não estão no contexto deste projeto de pesquisa.

Tabela 2.1: Matriz atributo-valor.

	t_1	t_2	$\cdots$	t_m	Y_1
d_1	a_{11}	a_{12}	$\cdots$	a_{1m}	Y_1
d_2	a_{21}	a_{22}	$\cdots$	a_{2m}	Y_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
d_n	a_{n1}	a_{n2}	$\cdots$	a_{nm}	Y_m

Fonte: Autor

No contexto deste trabalho, algumas palavras estão relacionadas a entidades geográficas, como nomes de cidades, estados, países, etc. Identificar essas palavras é uma etapa importante, pois, também deseja-se mapear tais entidades geográficas às informações de latitude e longitude. Essa atividade, em

geral, é vista como parte da etapa de pré-processamento dos textos, embora alguns estudos também categorizem tal atividade como um tipo de extração de padrões. Por envolver uma gama maior de conceitos, a análise de informação geográfica dos textos será detalhada na Seção 2.2.

Mais recentemente, representações textuais baseadas em *word embeddings* têm recebido destaque [Zhong et al., 2020; Liu et al., 2020; Medad et al., 2020]. O objetivo é fornecer representações vetoriais de palavras, nas quais cada palavra é mapeada em um espaço latente e palavras com semântica similar tendem a estar próximas umas das outras. Por exemplo, o *word vector* da palavra "cachorro" tende a estar próximo do *word vector* da palavra "gato", uma vez que essas palavras frequentemente aparecem juntas. Por outro lado, esses dois *word vectors* devem estar distantes do *word vector* da palavra "carro", já que essa palavra refere-se a outro contexto. Diferentes métodos para aprender *word embeddings* têm sido propostos e são discutidos em mais detalhes na Seção 2.3.

Dados os resultados promissores de *word embeddings* para uma representação mais semântica de textos, neste trabalho os *word embeddings* são escolhidos para representação de textos considerando informações geográficas neles. A ideia geral é que os *word vectors* de duas palavras estejam próximos em um espaço latente, caso tais palavras sejam frequentemente utilizadas em um mesmo contexto geográfico.

2.1.3 Extração de padrões

Considerando que os textos estejam representados em um formato estruturado, esta etapa visa aplicar algoritmos de extração de padrões para a descoberta de conhecimento. Tais algoritmos são organizados em tarefas descritivas e preditivas [Rezende, 2003; Fayyad et al., 1996].

Algoritmos para Tarefa Descritiva visam identificar estruturas de agrupamento ou padrões frequentes nos dados, sem considerar informação de rótulo ou classe. Podem-se destacar como exemplo os modelos: (1) regras de associação para busca de padrões que aparecem frequentemente entre os atributos de um grupo de textos; (2) agrupamento (*clustering*) para obter grupos de documentos de acordo com uma medida de similaridade; e (3) sumarização, que envolve métodos cujo objetivo é encontrar uma descrição que seja simples e compacta para um texto ou conjunto de textos. Esses algoritmos são chamados de modelos de aprendizado de máquina não-supervisionados [Fayyad et al., 1996; DA GAMA et al., 2017].

Já os algoritmos para Tarefas Preditivas têm como objetivo generalizar exemplos ou evidências passadas que possuem respostas conhecidas (classe), formando um modelo capaz de identificar a classe de um novo documento tex-

tual. Entre as tarefas preditivas, encontram-se: (1) a classificação de textos, que visa encontrar os relacionamentos existentes entre os atributos e a classe (atributo categórico) dos documentos de um conjunto de treinamento (dados rotulados) para predizer a classe de um exemplo ainda não visto (dados não rotulados); e (2) a regressão que atua de forma similar ao processo de classificação, diferenciando-se apenas no atributo a ser predito, que, nesse caso, é contínuo. Esses algoritmos são conhecidos como modelos de aprendizado de máquina supervisionados [Fayyad et al., 1996; DA GAMA et al., 2017].

O estudo realizado neste trabalho tem como interesse a análise dos dados por intermédio das tarefas do tipo preditiva, na qual as classes são determinadas por regiões de interesse. Assim, serão utilizados algoritmos de classificação de textos. Em especial, serão utilizados métodos de classificação baseado em redes neurais artificiais.

As Redes Neurais Artificiais (RNA) representam um conjunto de técnicas computacionais que são geradas por modelos matemáticos. São algoritmos inteligentes que buscam simular as capacidades cerebrais, semelhantes às redes neurais biológicas [Sivamani et al., 2019].

Na literatura é possível encontrar vários tipos de RNAs, entre eles o modelo de rede neural conhecido como *Perceptron* de Múltiplas camadas (MLP) recebe destaque por ser possível utilizá-lo em diversas aplicações do mundo real [Ebrahimabadi et al., 2015]. O modelo de MLP é formado por conjuntos de várias camadas, sendo a primeira camada a de entrada, as camadas intermediárias são conhecidas como ocultas, que podem ser uma ou mais camadas e, por fim, a camada de saída [Caires and Simonelli, 2021].

As redes neurais artificiais formadas por modelo MLP têm como ponto forte a resolução de problemas a partir de sua capacidade em aprender observando exemplos previamente rotulados. Por isso, são conhecidas como técnicas de aprendizado supervisionado [Sivamani et al., 2019]. A partir do conhecimento adquirido, o modelo pode fazer inferências sobre novos dados. Para melhorar o seu desempenho, são feitos vários ajustes nos pesos da rede ao decorrer da sua execução [Ferneda, 2006].

O algoritmo utilizado para correção dos pesos é usualmente o *backpropagation*. O seu funcionamento consiste, em primeiro lugar, na propagação do sinal para frente pela camada de entrada, passando pela camada oculta, até a camada de saída. Doravante, é feito o cálculo de erro resultado da diferença entre a camada de saída da rede e a saída real desejada. Em sequência, o resultado percorre o caminho inverso, iniciando pela camada de saída, camada oculta e por fim a camada de entrada, para ajustar os pesos da rede conforme as diferenças entre valores preditos e reais. Esse processo se repete até reduzir o erro a níveis que sejam aceitáveis [Ling et al., 2020].

2.1.4 Pós-processamento

O pós-processamento é a etapa que valida os conhecimentos adquiridos pela etapa de Extração de Padrões. Para esse fim, é preciso validar se o conhecimento extraído está de acordo com os objetivos propostos na etapa de Identificação do Problema. Caso o conhecimento gerado seja avaliado como incorreto para os objetivos traçados na primeira etapa, é necessário voltar às etapas anteriores, discuti-las novamente e refazer o processo.

No contexto deste trabalho, a avaliação do conhecimento extraído foi feita a partir da observação das medidas relativas à área de recuperação de informação, como precisão e revocação na predição de dados não vistos. De forma geral, busca-se validar a capacidade do modelo de associar uma informação textual a uma região em específico. Algumas medidas de avaliação são discutidas em detalhes na seção 4.4, durante a discussão de critérios de avaliação da abordagem proposta neste projeto.

2.1.5 Uso do conhecimento

Uma vez que o conhecimento foi validado pela etapa de pós-processamento, ele é direcionado para a etapa de Utilização do Conhecimento. Essa etapa é dependente da aplicação. No âmbito de pesquisas acadêmicas, ela envolve estudos de casos e provas de conceito para ilustrar na prática o desempenho da abordagem proposta.

Um exemplo de uso de conhecimento de interesse neste projeto é a construção de um mapa de calor que estima a probabilidade de ocorrência geográfica de determinados eventos extraídos dos textos. Esse conhecimento é útil para diversos estudos sociais que analisam o comportamento de indivíduos e organizações a partir de textos, como a área de segurança pública, análise de epidemias, discussões políticas, eventos de impacto ambientais, entre outros.

2.2 Recuperação de Informações Geográficas

2.2.1 Geo-parsing

No processo de georreferenciamento, a tarefa de identificação de georreferência é vital, já que ela é o propulsor para as demais atividades. Normalmente, refere-se à tarefa de identificar georreferência como *geoparsing* ou reconhecimento de topônimo. Ela é equivalente à tarefa de Reconhecimento de Entidade Nomeada (REN) subárea da Extração de Informações (EI). Esse processo consiste na atividade de categorizar cada palavra ou agrupamento de palavras para um conjunto de classes ou entidades predefinidas. As cate-

gorias podem ser das mais diversas possíveis, como local, pessoa, organização ou, até mesmo, não categorizado. O processo de Reconhecimento de Entidade Nomeada é composto por várias etapas, entre elas o processo de tokenização, muito utilizado em algoritmos de aprendizado de máquina. Esse processo consiste em decompor o texto em palavras.

Para identificar o nome de lugares é necessário definir a qual região ele pertence, que pode ser o nome da cidade, estado ou outro. Durante o processo de identificação de topônimo, podem aparecer nomes que tenham mais de um significado, como "Bonito", que pode ser qualidade de uma pessoa qualquer ou a cidade do estado de Mato Grosso do Sul, comumente conhecida como ambiguidade semântica. Quando isso ocorrer, é necessário que essa distorção seja corrigida para prosseguir. Uma forma comumente usada para resolver essa situação de ambiguidade semântica é a mistura entre listas com nomes de lugares, pessoas e organizações usando regras ou técnicas de aprendizado de máquina para que seja possível capturar o valor semântico do contexto ao redor do topônimo.

Abordagens para geoparsing

Quando há necessidade em saber se um nome qualquer indica uma georreferência, a primeira coisa que se pensa é em pesquisar em listas de nomes de lugares, endereços, códigos postais e outros. Essa foi, por muito tempo, a forma mais confiável para descobrir se realmente um local é ou não uma georreferência. Um estudo coordenado por Mikheev et al. [1999] mostra que é possível chegar a uma precisão de mais de 90% usando lista simples para locais. Os autores do estudo ainda destacam que a qualidade dos itens da lista teve uma relevância maior do que simplesmente a quantidade de itens na lista. A lista contou com 5.000 locais obtidos da base de dados do *CIA World Fact Book* e avaliados pela *Message Understanding Conference* (MUC-7).

Leidner and Lieberman [2011] vão além da abordagem das listas e apresentam três estratégias para reconhecer possíveis georreferências em um texto de linguagem natural: (1) utilização de lista simples de topônimo; (2) métodos baseados em conhecimento ou regras e (3) aprendizado de máquina.

A abordagem mais simples - utilização de lista simples de topônimo - pode ter uma precisão de 90 a 94% e recall de 75 a 85% segundo Mikheev et al. [1999]. No entanto, essa abordagem traz consigo uma série de desvantagens em relação às demais estratégias: (i) possui uma lista finita de itens, assim por essa entidade não é possível identificar novos topônimos que não tenham sido previamente escolhidos para compor os dicionários geográficos; (ii) quando uma correspondência estiver na entrada da fila, será necessário tomar uma decisão para permitir com que ela seja usada parcialmente ou por completo;

(iii) é comum que palavras tenham mais de um significado, com isso, em listas simples, muitas vezes, não é correto afirmar prontamente que uma palavra está sendo usada em um contexto geográfico de forma correta. Para entender melhor, nesse exemplo pode-se observar o seguinte caso: Seara pode ser uma cidade brasileira com 17.576 habitantes, uma área para plantação ou uma empresa bilionária de alimentos; (iv) Não ter um padrão definido para mapear as georreferências, pode haver, por exemplo, duas palavras escritas diferentes, mas que remetam ao mesmo local geograficamente, como “BR” e “Brasil”.

Ao perceber-se que várias entradas de topônimos não eram vinculadas a um sentido geográfico devido às limitações do uso de listas simples, que por sua vez buscavam uma melhoria na maneira de como identificar possíveis georreferências, começou-se a analisar o contexto circundante de um topônimo. Para obter o sentido circundante de uma palavra, foram criadas regras capazes de identificar gatilhos capazes de auxiliar a reconhecer georreferências com mais precisão. Essa estrutura contextual pode ser identificada como interna e externa.

As estruturas internas ou frasais podem ser identificadas pelo uso de letras maiúsculas, prefixos, sufixos e listas de nomes. Assim, olha-se para a estrutura das palavras para que seja possível compreender o seu contexto e, então, criar os gatilhos de identificação. Essa abordagem é comumente feita por meio da observação das palavras e em seguida definição das regras. Essas estruturas são encontradas com mais frequência em Sistemas de NER mais antigos.

Um passo importante para a classificação correta de topônimos foi a análise de evidências externas, também chamado de contexto. Dessa forma, foram criadas regras capazes de aprender o contexto de algo levando em consideração as palavras e frases que estão ao seu redor. Para tal, é possível criar regras que auxiliem na compreensão sobre a qual classe uma entidade nomeada pertence. Geralmente, utiliza-se essa abordagem em sistemas de NER da atualidade, por meio de métodos de aprendizado de máquina que são capazes de aprender regras novas automaticamente.

Para aprender o contexto do topônimo, é preciso passar por um pipeline de processos para chegar ao objetivo. Um dos processos mais significativos é a fase de treinamento do modelo, momento em que os dados anotados são apresentados várias vezes para um classificador, a fim de criar um modelo capaz de classificar exemplos que ainda não foram apresentados.

Contudo, dois pontos dentro do processo de aprendizagem de máquina que requerem atenção são: (i) os dados de treinamento; (ii) generalização dos classificadores resultantes. No primeiro item, verifica-se se a quantidade de dados será suficiente para que haja um bom conjunto de treino, que seja capaz

de treinar um classificador corretamente e possa ser usado de forma eficaz em conjunto a textos ainda não analisados. Na segunda questão, precisa-se ter certeza de que os novos classificados aprendidos serão capazes de generalizar textos ainda não analisados ou se eles apenas funcionarão para o conjunto de dados já vistos no treinamento.

2.2.2 Geo-coding

A Geocodificação ou resolução de topônimos está associada diretamente à nominação de um único nome de lugar levando em consideração alguma base de conhecimento de localização geográfica [**?**Buscaldi, 2011; Leidner and Lieberman, 2011; Andogah et al., 2012]. Dado o topônimo "Brasília", que é a capital do Brasil, pode-se geocodificá-lo para as coordenadas geográficas longitude -15.7997 e latitude -47.8640. Na identificação do topônimo, podem ocorrer alguns problemas, como quando tem a inserção de diversos locais com o mesmo nome em um dicionário geográfico. Uma forma simples de resolver esse problema consiste na análise do contexto em que o topônimo está inserido. Outra situação vivenciada no processo de geocodificação são nomes diferentes ou com poucas alterações que podem se referir ao mesmo local geográfico, por exemplo, “Jeri” e “Jericoacoara”, ambos se referem à mesma vila no litoral do Brasil.

Abordagens para geocodificação

Um algoritmo de decodificação, também conhecido como desambiguação de referência ou resolução de toponímia, tem como objetivo selecionar um endereço entre várias possibilidades e que vá ao encontro do topônimo ambíguo. Existem métodos que fazem uso de mecanismos geográficos externos ao topônimo para melhorar sua eficácia, tais como a utilização de dicionários geográficos, mapeamento entre coordenadas geográficas e espaciais, heurísticas e o emprego de regras. Frequentemente, utilizam-se heurísticas para facilitar o processo de desambiguação, fazendo uso de informações adicionais existentes ao topônimo para obter o melhor resultado.

Gale et al. [1992] destaca a propriedade de *Word Sense Disambiguation*, com a qual uma palavra assume ter “um sentido único em cada discurso”. Assim, um topônimo terá o mesmo significado ao longo da frase que ele está inserido. Tobler [1970] traz a lei de Tobler, pela qual afirma que elementos que estão mais próximos uns aos outros, tendem a ter uma relação mais próxima do que elementos que estão distantes.

Buscando facilitar o entendimento acerca dos métodos de desambiguação, Buscaldi [2011] descreve três grupos de desambiguação de topônimos: (1) baseados em mapas, consiste na desambiguação que faz uso de técnicas que

computam as distâncias espaciais; (2) baseado no conhecimento, mecanismos que fazem uso de informações advindas de outras fontes externas ao texto para encontrar ponto de desambiguação; (3) baseadas em dados ou supervisionados, referência aos métodos que usam técnicas de aprendizado de máquina na solução do problema.

Olhar o contexto em que um topônimo está inserido facilita determinar o local correto da georreferência. Uma estratégia para isso é analisar as pistas encontradas em seu contexto, bem como nomes de lugares encontrados no mesmo trecho do texto ou no documento como um todo. Essa abordagem é conhecida como relação de contenção.

Buscaldi [2011] mostra que o método baseado em conhecimento faz uso das relações de contenção. Para contexto local ou do documento, realizar a análise apenas do topônimo pode trazer contrariedades, pois, talvez não haja elementos suficientes que possam desfazer a ambiguidade de uma georreferência com precisão. Para contornar essa situação, uma possibilidade seria analisar palavras não geográficas que estejam ligadas ao contexto de um topônimo [Overell and Rüger, 2008].

Speriosu and Baldridge [2013] utilizaram algoritmos de aprendizado supervisionado para treinar um modelo capaz de analisar todas as palavras do documento para que possa realizar a desambiguação. Os autores chegaram a essa conclusão, tendo em vista que termos não geográficos tendem a ocorrer com mais frequência em cenários de contexto específico. Como exemplo, pessoas tendem a falar mais sobre peças de teatro quando estão mais próximas a um teatro, do que próximos ao hospital.

Segundo Speriosu and Baldridge [2013], aprender palavras que estejam ligadas às instâncias específicas do topônimo podem trazer resultados satisfatórios para a geocodificação. A partir das palavras aprendidas, uma ocorrência pode ser desambiguada por meio do seu uso e do contexto circundante. Essa abordagem pode ser classificada como baseada em dados, pois busca encontrar pontos de relação entre as palavras e o contexto em que elas foram inseridas, de forma a capturar as características ao seu redor. Essa abordagem é conhecida como aprendizagem supervisionada ou tarefa de classificação.

Por fim, outro problema encontrado na geocodificação é o de como atribuir características que não podem ser facilmente atribuídas a pontos geográficos de grande complexidade como o Rio Amazonas, entre outros pontos, devido ao seu tamanho ou como lidar com referências históricas a locais (terra da garoa, antigo apelido de São Paulo). Nesses casos, é importante executar etapas adicionais para capturar mais informações sobre o seu contexto.

2.3 Modelos de Linguagem

Os modelos de linguagem são utilizados para os mais diversos fins, desde reconhecimento de fala, reconhecimento de manuscrito, correção de ortografia, à análise, tradução automática e recuperação de informações [Gudivada, 2018]. Um modelo estatístico de linguagem consiste na distribuição de probabilidade seguindo uma sequência de palavras. Para um vocabulário finito $V = \{x_1, x_2, \ldots x_n\}$, a distribuição visa computar o quanto é provável a ocorrência de uma sequência de palavras, em que cada palavra pertença a um texto V^* pertencente a V. Pode-se mostrar a probabilidade do modelo de linguagem como:

$$\sum_{x \in V^*} p(x) = 1,\ e\ p(x) \geq 0\ para\ todo\ x \in V^*, \tag{2.1}$$

Considere um vocabulário $x_1, x_2, ...x_n$, em que x_1 é a primeira palavra, x_2 é a segunda palavra, assim por diante, até x_n ser a última palavra. Considere também a construção de uma frase, de forma que o primeiro termo foi escolhido de maneira consciente e presumindo que seja feita a inserção de uma palavra por vez, de modo que o próximo termo faça sentido com o conteúdo anterior. Assim, é possível calcular a probabilidade das palavras utilizando a regra de encadeamento:

$$p(x_1,\ x_2,\ \ldots,\ x_n) = \prod_i q(x_i \mid x_1,\ x_2,\ \ldots,\ x_i - 1) \tag{2.2}$$

Exemplificando, p ("o carro foi roubado") = q (o) x q (carro | o) x q (foi | o carro) x q (roubado | o carro foi). Para calcular p, estima-se a distribuição de probabilidade de q usando contagens de ocorrências de frase nos dados do corpus. Pode-se dividir e contar, tal como:

$$q(roubado \mid o\ carro\ foi) = \frac{count(o\ carro\ foi\ roubado)}{count(o\ carro\ foi)} \tag{2.3}$$

Embora essa seja uma forma fácil de ser processada pelo algoritmo, não é a melhor, pois existem diversos problemas sobre calcular a probabilidade dessa forma. Existem muitas frases que podem ser apresentadas. Outro problema é que não temos *dataset* que possa fornecer todas as contagens de sequência suficientes para demonstrar todas as possibilidades de todas as frases possíveis de uma linguagem. Para contornar essa situação, é usada a Premissa de Markov. O intuito de usar a Premissa de Markov parte do princípio de que não é necessário olhar para sentenças grandes, uma vez que a suposição de Markov é válida para modelos nos quais os valores seguintes dependem do valor anterior ou pouco antes do valor anterior. O modelo que

bem exemplifica isso é o modelo Markov oculto (HMM), conforme abaixo:

$$p(x_1, x_2, \ldots, x_n) \approx \prod_i p(x_i | x_{i-k},\ x_{(i-k)+1},\ x_{(i-k)+2},\ \ldots,\ x_{i-1}),\ 1 \leq k < i \tag{2.4}$$

Na equação acima, o valor de k escolhido é o que determina o tipo do modelo de linguagem. Atualmente, um dos modelos que se tem destacado é o modelo *n-grama*, de maneira que as "n" palavras mais recentes da sentença são usadas para prever a probabilidade da palavra seguinte. Por exemplo, k = 0 produz o modelo de linguagem *unigrama*. No modelo de *unigrama*, a probabilidade de observar uma determinada palavra não depende do contexto. Quando k = 1 e k = 2, temos os modelos de linguagem bigrama e trigrama, respectivamente. A seguir temos as equações dos modelos de linguagem bigrama (k = 1) e trigrama (k = 2):

$$p(x_1\, x_2 \ldots\, x_n) \approx \prod_i q(x_i | x_{i-1}) \tag{2.5}$$

$$p(x_1\, x_2 \ldots\, x_n) \approx \prod_i q(x_i | x_{i-2},\ x_{i-1}) \tag{2.6}$$

Para o modelo de *bigrama*, a sua probabilidade é estimada levando em consideração a palavra anterior para estimar a próxima palavra. Semelhante ao anterior, para obter o valor estimado do modelo *trigrama*, uma palavra leva em consideração as duas últimas palavras da sentença.

2.3.1 Word embeddings e word vectors

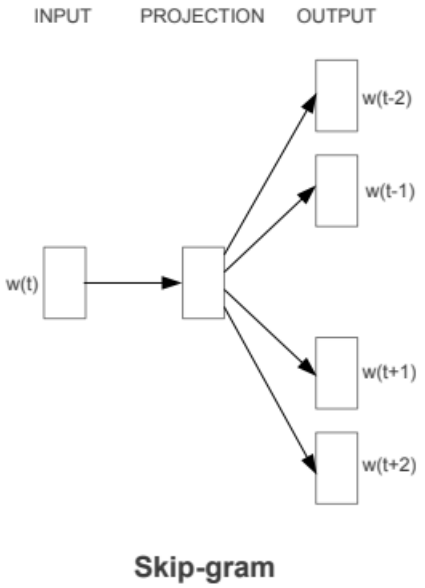
Para melhorar a capacidade de analisar os relacionamentos entre palavras, frases e documentos, foram propostos os *Word Vectors* [Mikolov et al., 2013]. Em razão disso, as tecnologias têm avançado a fim de fornecer às máquinas muito mais informações sobre palavras do que anteriormente era possível ao utilizar apenas representações tradicionais de palavras. São os *Word Vectors* que tornam possíveis tecnologias como reconhecimento de fala e tradução automática, por exemplo. Essa melhoria surgiu pela possibilidade de transformar texto em vetores numéricos, que representam as palavras em um espaço vetorial. Em outras palavras, um *word vector* é uma linha de números com valor real, em que cada ponto captura uma dimensão do significado da palavra e palavras semanticamente semelhantes têm vetores semelhantes. Para esse modelo, busca-se relacionar as palavras que têm significado semelhante e estejam próximas umas às outras em um espaço vetorial, como "pneu"e "motor"devem ter vetores de palavras semelhantes ao da palavra "carro"(devido à semelhança de seus significados), enquanto a palavra cachorro deve estar muito distante. Deste modo, as palavras usadas em um contexto semelhante serão mapeadas para um espaço vetorial próximo.

Uma forma para treinar o modelo de *word embedding*, inicialmente cada palavra é representada por meio de *one-hot encoding*. Para gerar uma *word embedding*, toma-se o vetor de *one-hot encoding* e o treina em uma rede neural para obter vetor denso que realmente consiga capturar o sentido da palavra, conforme descrito na próxima seção.

Rede neural do tipo skip-gram

No modelo de rede neural *Skip-gram* proposto por Mikolov et al. [2013], dado um vocabulário de palavras de tamanho W, o objetivo é aprender uma representação vetorial para cada palavra do vocabulário. Inspiradas na hipótese distributiva de Harris, as representações de palavras são treinadas para prever de forma satisfatória as palavras que aparecem em seu contexto com base em uma palavra-alvo. Na Figura 2.2, temos uma arquitetura do *Skip-gram*.

Figura 2.2: Arquitetura do modelo *Skip-gram*.



Fonte: Mikolov et al. [2013].

Conforme ilustrado na Figura 2.2, o tamanho da janela que tomamos é 2, $w(t)$ representa a palavra-alvo e $w(t-2)$, $w(t-1)$, $w(t+1)$, $w(t+2)$ simbolizam as palavras circundantes dela, o termo de entrada do modelo é denominado palavra-alvo. Primeiro, a palavra de destino é mapeada para um vetor de representação de camada oculta e podemos prever as palavras de contexto de acordo com esse vetor.

De forma mais clara, dado um corpus de treinamento representado como uma sequência de palavras de $w_1,\ldots,w_t$, o modelo *Skip-gram* tem como objetivo, buscar maximizar a probabilidade de um termo alvo w_t dado o seu contexto w_c

conforme a Equação 2.7:

$$\sum_{t=1}^{T} \sum_{c \in C_t} log\ p(w_c|w_t) \tag{2.7}$$

Na Equação 2.8, c é o tamanho da janela de contexto e w_t é o conjunto de índices de palavras que cercam a palavra w_t. A probabilidade de observar uma palavra de contexto w_c dado w_t será parametrizada usando os vetores de palavras mencionados acima. Por enquanto, vamos considerar que recebemos uma função de pontuação s que mapeia pares de (palavra, contexto) para pontuações em R. Uma opção possível para definir a probabilidade de uma palavra de contexto é por meio da função *softmax* conforme é descrito na Equação 2.8:

$$p(w_c|w_t) = \frac{e^{s(w_t,w_c)}}{\sum_{j=1}^{W} e^{s(w_t,j)}} \tag{2.8}$$

No entanto, esse modelo não está adaptado ao caso deste trabalho, pois implica que, dada uma palavra w_t, prevê-se apenas uma palavra de contexto w_c. O problema de prever palavras de contexto pode ser enquadrado como um conjunto de tarefas de classificação binária independentes. Então, o objetivo é prever independentemente a presença (ou ausência) de palavras de contexto. Para a palavra na posição t consideramos todas as palavras de contexto como exemplos positivos e mostramos negativos de forma aleatória no dicionário. Para uma posição de contexto escolhida c, obtemos a seguinte probabilidade logarítmica negativa:

$$log\left(1+e^{-s(w_t,w_c)}\right) + \sum_{n \in N_{t,c}} log\left(1+e^{s(w_t,n)}\right) \tag{2.9}$$

onde $N_{t,c}$ é um conjunto de exemplos negativos amostrados no vocabulário. Ao denotar a função de perda logística $l : x \rightarrow log(1+e^{-x})$, podemos reescrever o objetivo como:

$$\sum_{t=1}^{T} \left[\sum_{c \in C_t} \ell(s(w_t,w_c)) + \sum_{n \in N_{t,c}} \ell(-s(w_t,n)) \right] \tag{2.10}$$

Uma parametrização natural para a função de pontuação s entre uma palavra w_t e uma palavra de contexto w_c é usar vetores de palavras. Vamos definir para cada palavra W no vocabulário, dois vetores U_w e V_w no $\mathbb{R}^d$. Esses dois vetores, geralmente, são chamados de entrada e resultado de vetores na literatura. Em particular, temos vetores U_w e V_w correspondendo, respectivamente, às palavras w_t e w_c. A pontuação pode ser calculada como o produto escalar

entre os vetores de palavras e de contexto, conforme $s\ (w_t,\ W_c) = U_{w_t}^T \cdot V_{w_c}$. O modelo descrito nesta seção é o modelo do *Skip-gram* com amostragem negativa, introduzido por Mikolov et al. [2013].

2.3.2 FastText

Ao usar uma representação vetorial distinta para cada palavra, o modelo do *Skip-gram* normalmente ignora a estrutura morfológica de cada palavra e considera uma palavra como uma única entidade, ignorando a estrutura interna das palavras. Em contrapartida, o modelo *FastText* [Bojanowski et al., 2017], que é uma estrutura para aprender representações de palavras, considera cada palavra como um *n-grama* de caracteres, também conhecido como subpalavras.

Para cada palavra no conjunto W é obtido um conjunto de caracteres *n-gramas*. São adicionados símbolos de limite especiais < e > no início e no final das palavras, para que seja possível distinguir prefixos e sufixos de outras sequências de caracteres. Tomando a palavra verde e n = 3 (*tri-gramas*) como exemplo, ela será representada pelo caractere *n-gramas*: <ve, ver, erd, rde, de> e a sequência especial <verde> que representa a palavra inteira.

Observe que a sequência correspondente à palavra <ver> é diferente do *tri-gramas* ver a partir da palavra verde. O aconselhável é extrair todo o *n-gramas* para n=3 e n=6. Isso ajuda a preservar o significado de palavras mais curtas que podem aparecer como *n-gramas* de outras palavras. Inerentemente, isso também permite capturar significado para sufixos/prefixos.

Dada uma palavra w_t, vamos denotar por $g \subset 1,...,G$ conjunto de n-grama aparecendo em w. Associa-se uma representação vetorial z_g para cada *n-grama*. Representamos uma palavra pela soma das representações vetoriais de seus *n-gramas*. Assim, obtemos a equação de pontuação:

$$s(w,c) = \sum_{g \in G_w} Z_g^T\ V_c \tag{2.11}$$

Este modelo simples permite compartilhar as representações entre palavras, permitindo também aprender uma representação confiável para palavras raras e para palavras não existentes no conjunto de treinamento.

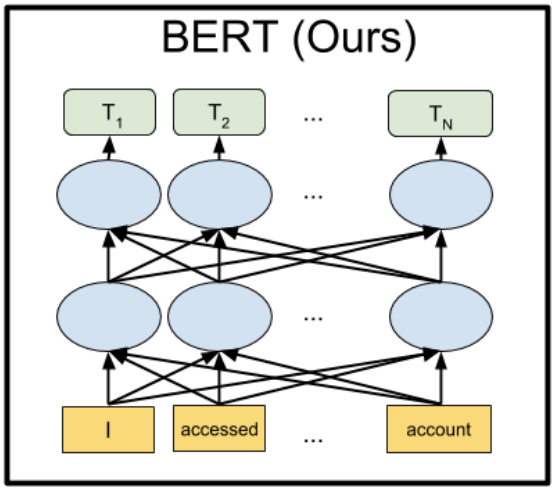
2.3.3 BERT - Bidirectional Encoder Representations from Transformers

Com o intuito de melhorar as atividades do Processamento de Linguagem Natural (NLP), Vaswani et al. [2017] trouxeram o conceito dos *Transformers*, impulsionando o estado da arte, possibilitando a utilização aprimorada da

paralelização e melhorando a modelagem de dependências de longo alcance [Lee et al., 2020]. Entre os modelos que usam essa abordagem, o BERT [Devlin et al., 2018] tem conquistado notoriedade nos meios acadêmico e corporativo nos últimos tempos por conseguir resultados significativos em sua execução [Rogers et al., 2021].

O BERT é um modelo de representação de palavras contextualizadas, formado por um empilhamento de camadas de codificadores dos *Transformers* [Devlin et al., 2018], que aprende as representações de palavras de acordo com o contexto em que as palavras estão inseridas [Lee et al., 2020]. Ao contrário de modelos linguísticos unidirecionais (em outras palavras, da esquerda para a direita e da direita para a esquerda), o BERT usa a técnica de *bi-Transformers* que pode explorar de forma efetiva as informações semânticas profundas de uma frase. O mesmo é constituído pelo modelo de linguagem mascarada de modo a ser capaz de predizer palavras mascaradas aleatoriamente por uma sequência de *tokens* [Devlin et al., 2018]. Na Figura 2.3, é possível verificar uma síntese do modelo BERT.

Figura 2.3: Modelo BERT.



Fonte: Devlin et al. [2018]

Devlin et al. [2018] e Rogers et al. [2021] definem o fluxo de implementação do BERT em duas etapas: pré-treinamento e ajuste fino. No decorrer da etapa de pré-treinamento, são utilizados dados não rotulados para realizar o treinamento do modelo. Este momento pode ser usado em duas tarefas: Modelagem de Linguagem Mascarada (MLM), que é útil para prever os próximos *tokens* de entrada mascarados aleatoriamente; e Previsão de Próxima Sentença (NSP), que pode ser aplicado para predizer se duas sentenças de entrada são adjacentes uma à outra. Tendo em mente a etapa de ajuste fino, os parâmetros são ajustados por meio de dados rotulados das tarefas *downstream* (por exemplo: um classificador), em que uma ou mais camadas estão

completamente conectadas na camada final do decodificador.

Pré-treinamento consiste na execução de duas tarefas não supervisionadas: Modelagem de Linguagem Mascarada (MLM), ou seja, o mascaramento de alguns *tokens* de entrada de forma aleatória para, em seguida, realizar a predição dos mesmos. Pode-se observar que os vetores ocultos finais são equivalentes aos *tokens* mascarados, posteriormente os mesmos são alimentados em um *softmax* de saída do vocabulário. Devlin et al. [2018] obteve bons resultados com 15% de mascaramento dos *tokens* aleatoriamente, tendo como tarefa prever as palavras que estão mascaradas. Previsão de Próxima Sentença (NSP), representa uma tarefa de previsão binária da próxima frase, de forma que se pode prever se duas sentenças de entrada são adjacentes entre si. A importância dessa tarefa se baseia no controle de qualidade e Inferência de Linguagem Natural.

O ajuste Fino é responsável por realizar a conexão entre as entradas e saídas especificadas da tarefa e faz os devidos ajustes em todos os pesos originais, incluindo os pesos de *embeddings* de palavras, blocos do *Transformer* e o classificador. No final, a resultante entre essas conexões, representações de *tokens*, é utilizada para alimentar uma nova camada de saída específica de tarefa e ajustar o modelo por completo. O seu resultado inclui aprender novos parâmetros da camada de saída que podem ser utilizados para realização de tarefas de nível de *tokens*, tais como marcação de sequência, resposta a perguntas e classificação de textos [Devlin et al., 2018; Zhang et al., 2020].

Com isso, é possível compreender que o pré-treinamento em conjunto com o ajuste fino são partes cruciais para implementação do BERT. Por fim, pode-se acrescentar que a primeira etapa é responsável por fornecer o conhecimento de forma autônoma à tarefa, já a segunda, tem como objetivo central mostrar ao modelo quais as representações úteis que tem maior relevância e, assim, possibilitar ao modelo saber quais tipos de representações podem ser desconsideradas.

2.4 Considerações Finais

Neste capítulo, foram apresentados diversos conceitos que estão relacionados ao presente trabalho, como Mineração de Textos Georreferenciados, Recuperação de Informações Geográficas e Modelos de Linguagem. Primeiro, foi apresentada a Mineração de Textos Georreferenciados, processo que tem como objetivo descobrir padrões até então desconhecidos em bases de dados não rotulados ou semiestruturados.

Discorremos sobre as etapas da Mineração de Textos, sendo elas: Identificação do Problema, Pré-processamento, Extração de padrões, Pós-processamento

e Utilização do Conhecimento. Doravante, explicou-se como é o funcionamento da Recuperação de Informações Geográficas por meio dos processos de georreferenciamento *Geo-parsing* e *Geo-coding*, que são capazes de realizar a identificação e desambiguações de topônimos. O segundo ponto conceitual abordado foi o impacto que os Modelos de Linguagem, por meio das técnicas de inteligência artificial, têm causado nas atividades de Processamento de Linguagem Natural. Com os modelos *Bag of Words* e *Word Embedding* é possível converter palavras e frases em vetores numéricos que podem ser utilizados em ferramentas capazes de extrair informações novas a partir dessas representações. Com o melhoramento de *embedding*, métodos como o *FastText*, torna-se capaz a aprendizagem das estruturas das palavras por meio dos *n-gramas*. Por fim, foi apresentado o BERT, que aprende *embeddings* considerando o contexto das frases.

CAPÍTULO

3

Trabalhos Relacionados

3.1 Combinando Similaridade Textual e Espacial

Tendo em vista que os sistemas GIR manipulam seus componentes separadamente, surge a necessidade de combinar seus resultados em uma lista única, de forma a facilitar a busca dos mesmos. Buscando resolver essa questão, Andrade and Silva [2006], Hariharan et al. [2007] e Chen et al. [2013] analisaram a utilização de uma combinação linear, obtida a partir do cálculo da pontuação entre a similaridade textual e a similaridade espacial para um documento de forma individualizada. Para saber o quão próximo um texto está de uma georreferência, esse dado vem por meio da pontuação combinada $comb\ (Q,D)$, onde Q é a consulta e D é o documento:

$$comb(Q,D) = \alpha_1 \cdot textsim(Q_t,D_t) + \alpha_2 \cdot geosim(Q_s,D_s), \tag{3.1}$$

onde $textsim\ (Q_t,D_t)$ calcula a similaridade de um conteúdo (texto), que pode ser calculado usando o modelo espaço vetorial mais a similaridade espacial (geográfica) representada por $geosim\ (Q_t,D_t)$. Cada similaridade tem um peso atribuído a ela, que reflete o seu grau de importância, sendo que o somatório do $\alpha 1$ e $\alpha 2$ deve ser equivalente a 1. Martins et al. [2005] trazem outros métodos com a finalidade de combinar pontuações, entre eles são mencionadas várias combinações lineares.

Palacio et al. [2010] utilizam a abordagem de combinação de listas classificadas de documento. Para isso, eles usam as posições de classificação dos documentos (por exemplo: Round-Robin e Contagem de Borda) ou fazem uso das combinações dos métodos baseados em pontuação e em classificação. Na

classificação por Contagem de Borda, cada lista de classificados é associada a um documento. Assim, a quantidade de associações determina uma pontuação final para a lista de classificados. Frontiera et al. [2008] e De Sabbata and Reichenbacher [2010] aplicam modelos probabilísticos buscando combinar pontuações, a fim de determinar qual documento tem maior potencial para ser escolhido a partir de uma consulta.

Buscando georreferenciar topônimos que não estão listados em dicionários geográficos, DeLozier et al. [2015] implementam o método de modelagem de linguagem. Feita a identificação de um possível topônimo em sistemas NER, eles usam as palavras de contexto (ou janela de contexto, que consiste em um conjunto de palavras antes e depois do possível topônimo) para associá-las a uma probabilidade de ocorrência para o local em uma grade. As combinações de todas as probabilidades associadas às palavras de contexto são responsáveis por identificar qual a localização que tem maior probabilidade de estar associada ao topônimo.

Dois locais podem estar relacionados a mesma região se existir coocorrência de textos entre os locais ou se houver *links* de *web* georreferenciados entre eles [Spitz et al., 2016]. Como a distância geográfica tem um impacto de forma irregular em diferentes locais, os autores utilizaram uma abordagem de modelagem de tópicos a fim de compreender a semântica e a relação dela com a distância do texto.

3.2 Escopo do Documento de Computação (escopo geográfico dos recursos da web)

De acordo com Andrade and Silva [2006], a classificação do escopo geográfico de um documento é dada de duas formas: (i) conjunto de regiões geográficas contidas no documento; (ii) espaço geográfico sobre o qual o documento está se referindo. O processo de resolução do escopo geográfico consiste na sua automação, de maneira que dado um documento, ela possa realizar a atribuição do escopo geográfico de forma independente.

Ding et al. [2000] apresentam um estudo sobre o escopo geográfico voltado para as páginas da *web* (recursos), sendo capaz de estabelecer as relações entre o local de origem das páginas e o público que consome o seu conteúdo. Os autores defendem que um documento (no estudo, páginas da web) tem mais relevância para um público de uma região específica (algo local) do que em relação a um público mais global. Essas relações podem ser observadas por meio da forma como o conteúdo textual é exposto ao longo da página web e levando em consideração a forma como os hiperlinks são distribuídos geograficamente.

Buscando melhorar o entendimento sobre as localizações dos recursos da web, Wang et al. [2005] categorizam seus recursos em três tipos: (i) a localização do provedor, que consiste na localização física na qual o recurso está armazenado; (ii) localização do conteúdo, remete a localização em que o conteúdo da web está se referindo; (iii) local de atendimento, apresenta a localização do escopo geográfico em que o recurso web está direcionado, o seu público-alvo.

3.3 Modelagem de Locais Implicitamente por meio de Modelo de Linguagem

Com a disponibilização de grandes bases de dados associados diretamente a coordenadas geográficas, o aprimoramento de métodos cada vez mais eficazes de aprendizado de máquina e a crescente procura por abordagens que sejam capazes de analisar essas grandes massas de dados em tempo real, resulta-se na criação de novos métodos que, ao invés de utilizarem como ponto principal o georreferenciamento por meio de topônimos contidos no texto, eles são capazes de aprender a localização de um texto de forma mais genérica sem perder a sua precisão. Ao produzir uma grande base de dados compostos por documentos associados a coordenadas geográficas, como dados do *Flickr* com geo-tags, *Wikipédia* ou *Tweets*, é possível extrair um conjunto de informações que estão diretamente relacionadas a determinada localização geográfica. Esse processo é nomeado como modelo de linguagem.

Alguns dos primeiros autores a discutir sobre a utilização de modelos de linguagem no processo de georreferenciamento de texto foram Ahern et al. [2007], que formaram a base de um sistema de navegação baseado em modelo de linguagem capaz de explorar conceitos geográficos específicos. Para realização desse processo os autores utilizaram o método de *k-means* para realizar o agrupamento das coordenadas geográficas, em conjunto com uma versão modificada da técnica *tf-idf*, para que fosse possível localizar e selecionar as palavras-chave que têm maior significado para as *tags* do *Flickr*. Feita a seleção, elas eram armazenadas em uma grade contendo latitude e lon*gitu*de.

Ainda sobre pesquisas feitas utilizando dados geográficos contidos em fotos do *Flickr*, O'Hare and Murdock [2013] trazem uma abordagem similar, utilizando métodos Bayesianos, usando apenas os dados textuais para gerar modelos de linguagem associados a latitude e longitude. No processo para armazenar a foto, também foram utilizados modelos de linguagem para determinar a melhor célula para depositar o conteúdo.

Kinsella et al. [2011] afirmam, em seu estudo, que a utilização de modelos de linguagem são superiores aos modelos de geocodificação tradicionais,

aqueles baseados em dicionário geográfico. Para chegar a esses resultados, os autores observaram o comportamento dos métodos de modelagem de linguagem para obter a localização dos *Tweets* e dos seus usuários. Em seguida, fizeram a utilização da divergência de *Kullback-Leibler* (distância) junto com a probabilidade Bayesiana para combinar os *Tweets* com a sua localização. Já Ozdikis et al. [2019] utilizaram dados do *Twitter* com geotags para criar *word embedding* por meio do algoritmo *FastText*.

Wing and Baldridge [2011] empregaram o uso de modelos de linguagem no processo de indexação de coordenadas geográficas em documentos. Para isso, eles utilizam grades de tamanho regular com o intuito de particionar os modelos de linguagem, para em seguida fazer uso da divergência *Kullback-Leibler* existente entre os modelos de linguagem do documento e o da célula. Roller et al. [2012] trazem em sua proposta a utilização de células de tamanho variado por meio do método de agrupamento e fazem a utilização de conjunto de treinamento para selecionar a localização dos artigos geocodificados da *Wikipédia*.

Com o propósito de melhorar o entendimento sobre a utilização de modelos de linguagem em artigos da *Wikipédia* geocodificados, Laere et al. [2014] fazem uso do método de *k-medoides* para criar modelos mais robustos a partir da combinação de dados da *Flickr*, *Twitter* e *Wikipédia*. Para selecionar os dados de treinamento, foram utilizados aqueles que mais se assemelhavam ao seu conteúdo textual. Os autores descrevem que os melhores resultados foram obtidos nos locais com granularidade fina, como edifícios, já aqueles que representam locais muito extensos ou com granularidade menor, como rios, tiveram resultados menos significativos.

As *Word Embeddings* tendem a abstrair informações relevantes sobre determinado local, região ou sobre o mundo. Por meio do uso de *word embedding*, é possível determinar e compreender o dialeto de uma área geográfica específica, bem como terminar a área geográfica em que um determinado dialeto é falado [Rahimi et al., 2017; Konkol et al., 2017]. Assim, é possível encontrar uma geolocalização do usuário olhando o que ele produz e posta em redes sociais. Para realização do mesmo, Rahimi et al. [2017] recuperaram informações a partir de um dialeto, os seus vizinhos mais próximos no espaço de incorporação, e os compararam aos termos de dialeto associados a esse local. Após as *word embeddings* serem geradas, é possível comparar o seu valor diretamente com dados do mundo real, como as coordenadas de GPS das cidades.

Outra forma é medir o desvio entre a posição dada pela *Word Embeddings* e a sua posição real [Konkol et al., 2017]. Em contraste com o método anterior, Miyazaki et al. [2018] propõe utilizar *bag-of-words* para encontrar a geolocalização de um local. Para isso, é proposto um método de previsão de geolo-

calização do usuário com base em entidades vinculadas e incorporando uma base de conhecimento, que é usada para classificar a localização do usuário. Ambos os trabalhos usam dados do *Tweetter* para pegar suas bases de conhecimento.

Com o intuito de monitorar eventos como desastres naturais e entender o impacto que eles causam, incluindo onde os danos ocorreram e onde as pessoas que precisam de ajuda estão situadas, os pesquisadores Mao et al. [2018] e Hernandez-Suarez et al. [2019] utilizam textos postados em redes sociais por pessoas durante o desastre para acompanhar o ocorrido, mas apresentam em seus trabalhos abordagens e métodos diferentes para resolver essa questão. Mao et al. [2018] fazem uso de duas estruturas, a primeira baseada em modelos probabilísticos que analisam *bag-of-word* para gerar classificadores capazes de definir quando um evento ocorreu, no caso do referido trabalho, a falta de energia. Já o segundo modelo utiliza rede neural de aprendizado profundo, nesse caso o *GloVe*, para extrair expressões que contenham localizações em seus *tweets*.

Os trabalhos de Hernandez-Suarez et al. [2019] e Cardoso et al. [2022], abordam a utilização da detecção de topônimos contidos nos textos. Hernandez-Suarez et al. [2019] utilizam o modelo *Word2Vec* do tipo *Skip-Gram* para *tokenizar* e transformar os *tweets* em *word embeddings* de forma que se notem as relações sintáticas e semânticas estabelecidas por palavras vizinhas. No trabalho de Cardoso et al. [2022], são utilizadas as palavras circundantes ao topônimo como entrada da rede neural para gerar as embeddings de BERT.

A utilização de Redes Neurais Convolucionais (CNNs) para tarefa de aprendizado profundo tem-se mostrado eficaz na resolução de problemas com dependências espaciais [Xu et al., 2020; Blier-Wong et al., 2020]. No trabalho apresentado por Blier-Wong et al. [2020], são utilizadas as informações dos vizinhos mais próximos na criação de *Word embeddings* com contexto geográfico de uma coordenada espacial. As *embeddings* geradas são utilizadas para prever tarefas *downstream* relacionadas a riscos sociodemográficos, buscando suavizar conjuntos de dados de grande escala para uso granular em tarefas relacionadas a riscos.

Tendo em vista a necessidade de se ter uma estimativa precisa da localização do usuário para diversos serviços online, como para a recomendação de produtos por *e-commerce*, Huang and Carley [2019] propõem a utilização de *word embedding* pré-treinadas pelo *GloVe* em uma rede neural de predição de localização hierárquica para predizer a geolocalização de usuários no *Twitter*, combinando informações textuais, metadados e informações de rede. Para este fim, o modelo é utilizado, em um primeiro momento, para prever o país de origem de um usuário e, posteriormente, utiliza o resultado gerado para

orientar uma nova previsão, dessa vez em nível da cidade. A fim de representar ontologias geográficas hierárquicas, Dassereto et al. [2020] fazem uso de *embeddings* hiperbólica para projetar os objetos no espaço sem que perca a relação semântica entre elas. Para representar objetos geográficos, optou-se por utilizar ontologias, pois conseguem manter as relações hierárquicas.

Medad et al. [2020] propõem em seu estudo o reconhecimento e a desambiguação de entidades nomeadas observando suas características espaciais contidas em dados textuais não estruturados. Para isso, os autores utilizaram o método *FastText* com *Word Embeddings* pré-treinadas de entrada para algoritmos de aprendizagem supervisionada poderem identificar se uma entidade nomeada identificada está referindo-se a um objeto espacial por meio do seu contexto. Medad et al. [2020] utilizam a pontuação obtida pela precisão e *F-score* para validar a sua proposta.

Buscando capturar informações linguísticas por meio de rede social que possam representar a geolocalização do usuário, Zhong et al. [2020] faz uso de redes neurais de grafos (GNN) por meio do método de *word2vec*. Dessa maneira é possível fazer com que o modelo foque nas informações mais importantes, sendo capaz de distinguir diferentes aspectos da preferência de publicação do usuário como a linguística.

Contractor et al. [2021] buscam em seu trabalho responder a perguntas por meio de *embedding* do BERT que combinem o raciocínio textual e espacial para obter conhecimento geoespacial. Para este fim, foi implementado um Raciocinador Geoespacial, que é responsável por codificar as perguntas; calcular a distância geoespacial e classificar a relevância espacial.

Xu et al. [2020] e Radke et al. [2022] abordam em seus trabalhos a utilização de modelo de redes neurais profundas para o Processamento de Linguagem Natural, com o intuito de aprender representações de sentenças contextualizadas capazes de predizer preposições em um sentido geoespacial. Em ambos os trabalhos, foi utilizado o modelo de linguagem BERT para aprender as representações semânticas das sentenças. Por fim, os trabalhos fazem uso de técnicas de Ajuste fino para melhorar os seus resultados.

Com base nos trabalhos relacionados, apresentamos a Tabela 3.1, uma síntese comparativa entre os artigos levantados no presente capítulo. Para efeitos de comparação, destacamos os seguintes pontos: (i) Similaridade, que pode ser textual, espacial; (ii) Tarefa: representa a tarefa utilizada no processo de extração do conhecimento, identificada como classificação, *ranking* ou *clustering* ; (iii) Modelo de representação textual utilizado, que está representado na tabela por dois grupos: (a) *Bag-of-words* que engloba os trabalhos que utilizaram a “frequência da palavra por região” para criação do modelo; (b) *Word embeddings*, apresenta os trabalhos que buscaram capturar o valor

semântico no momento de criação do modelo como *GloVe, Word2Vec, FastText e BERT*.

Tabela 3.1: Síntese dos trabalhos relacionados.

Artigos	**Similaridade**	**Tarefa**	***Bag-of-words***	***Word Embedding***
Ding et al. [2000]	Textual	Classificação	Frequência da palavra por região	—
Martins et al. [2005]	Textual e espacial	Classificação e Rank	Frequência da palavra por região - Gazetteers	—
Wang et al. [2005]	Espacial	Classificação	Frequência da palavra por região	—
Andrade and Silva [2006]	Textual e espacial	*Ranking*		—
Hariharan et al. [2007]	Textual e espacial	*Ranking*	Frequência da palavra por região - TF-IDF	—
Ahern et al. [2007]	Textual e espacial	*Clustering*	Frequência da palavra por região	—
Frontiera et al. [2008]	Textual e espacial	Classificação e Rank	regressão logística	—
Palacio et al. [2010]	Textual	Classificação	Frequência da palavra por região - TF-IDF	—
Kinsella et al. [2011]	Textual	Classificação	Frequência da palavra por região	—
Wing and Baldridge [2011]	Textual	Classificação e Rank	Frequência da palavra por região	—
Roller et al. [2012]	Textual	*Clustering*	Frequência da palavra por região	—
Chen et al. [2013]	Textual e espacial	*Ranking*	Frequência da palavra por região	—
O'Hare and Murdock [2013]	Textual	Rank (recuperar)	Frequência da palavra por região	—
Laere et al. [2014]	Textual e espacial	*Clustering*	Frequência da palavra por região - Gazetteers	—
DeLozier et al. [2015]	Textual e espacial	*Clustering*	Frequência da palavra por região	—
Spitz et al. [2016]	Textual	Classificação	Frequência da palavra por região	—
Salvini and Fabrikant [2016]	Textual e espacial	*Clustering*	Frequência da palavra por região	—
Konkol et al. [2017]				
Rahimi et al. [2017]	Textual	Classificação	—	Word2Vec
Miyazaki et al. [2018]	Textual	Classificação	Frequência da palavra por região	—
Mao et al. [2018]	Textual	Classificação	Frequência da palavra por região	GloVe
Huang and Carley [2019]	Textual	Classificação	—	GloVe
Hernandez-Suarez et al. [2019]	Textual	Classificação	—	Word2Vec
Ozdikis et al. [2019]	Textual	Classificação	—	FastText - Bigrama
Zhong et al. [2020]	Textual	Classificação	—	Word2Vec
Blier-Wong et al. [2020]				
Dassereto et al. [2020]				
Liu et al. [2020]	Textual e espacial	Classificação	—	Graph Neural Network - GNN
Medad et al. [2020]	Textual	Classificação	—	FastText - CBOW
Xu et al. [2020]	Textual e espacial	Classificação	—	BERT
Contractor et al. [2021]	Textual e espacial	Classificação	—	BERT
Cardoso et al. [2022]	Textual e espacial	Classificação	—	ELMo e BERT
Radke et al. [2022]	Textual e espacial	Classificação	—	BERT

Fonte: Elaboração própria

CAPÍTULO

4

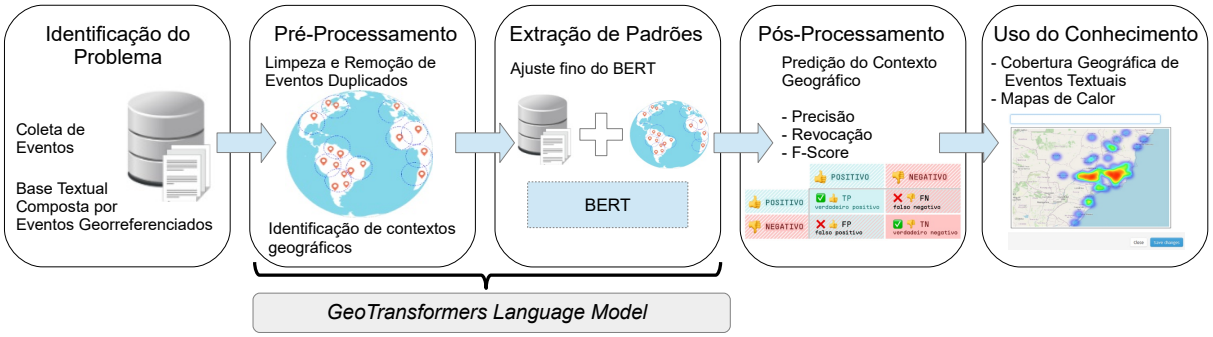
Mineração de Textos com *GeoTransformers Language Model*

Neste capítulo é apresentado o processo de Mineração de Textos proposto neste trabalho que tem, como diferencial, representações textuais baseadas em *Word Embeddings* treinados com contexto geográfico. Para obter essas representações, é proposto o uso do modelo BERT (*Bidirectional Encoder Representations from Transformers*) em um processo de ajuste fino, no qual as informações georreferenciadas dos textos são utilizadas como classes. Esse ajuste fino é denominado *GeoTransformers Language Model* e é uma das principais contribuições deste trabalho de mestrado.

A base metodológica da proposta é um processo de mineração de textos, conforme apresentado por [Rezende, 2003] e discutido no Capítulo 2. Uma visão geral da metodologia é ilustrada na Figura 4.1.

A primeira etapa é a **Identificação do Problema**, a qual busca estabelecer os objetivos, delimitar o problema a ser abordado, selecionar quais fontes de informações geoespaciais serão utilizadas e selecionar as técnicas e modelos. Posteriormente, tem-se o **Pré-processamento** das coleções textuais. Nesse caso, os textos são padronizados, são removidos os ruídos, é feita a delimitação do vocabulário que será utilizado, bem como a identificação do contexto geográfico dos textos georreferenciados. Logo após, realiza-se a **Extração de Padrões**, na qual envolve o ajuste fino de um modelo BERT considerando contextos geográficos identificados na etapa anterior. A união das etapas de pré-processamento e extração de padrões representa o *GeoTransformers Language Model*. Na etapa seguinte, é efetuado o **Pós-processamento**, na qual há a avaliação do modelo obtido na etapa anterior, capaz de estimar a cobertura

Figura 4.1: Processo de Mineração de Textos com *Word Embeddings* de Contexto Geográfico.



Fonte: Elaboração própria.

geográfica de um texto de entrada. Por fim, é realizada a etapa de **Utilização do Conhecimento**, em que são gerados mapas de calor para análise visual dos resultados do modelo obtido.

Cada etapa é apresentada em mais detalhes, no contexto deste projeto, nas subseções a seguir.

4.1 Identificação do Problema

Identificar locais a partir de conteúdo textual pode ser uma tarefa desafiadora. Na frase "*Autoridades do Rio se unem para tentar conter caos e roubos nas praias*", é possível encontrar ambiguidade na palavra "Rio"que pode se referir ao fluxo natural d'água, se referir a abreviação do Estado do Rio de Janeiro ou até mesmo ao Município do Rio de Janeiro. Essas palavras ambíguas prejudicam um sistema computacional baseado em dicionário geográfico para definir a geolocalização do evento.

Nesse contexto, sistemas computacionais baseados em rede neural com *Word Embeddings*, podem melhorar a identificação das relações entre uma palavra e o possível local associado. Portanto, optou-se por modelos de linguagem baseados em *Word Embeddings* que consideram o contexto geográfico da frase, já que esses modelos conseguem abstrair de forma efetiva a representação das palavras, ao contrário dos modelos tradicionais baseados em BoW, conforme mostrado na literatura recente (Tabela 3.1).

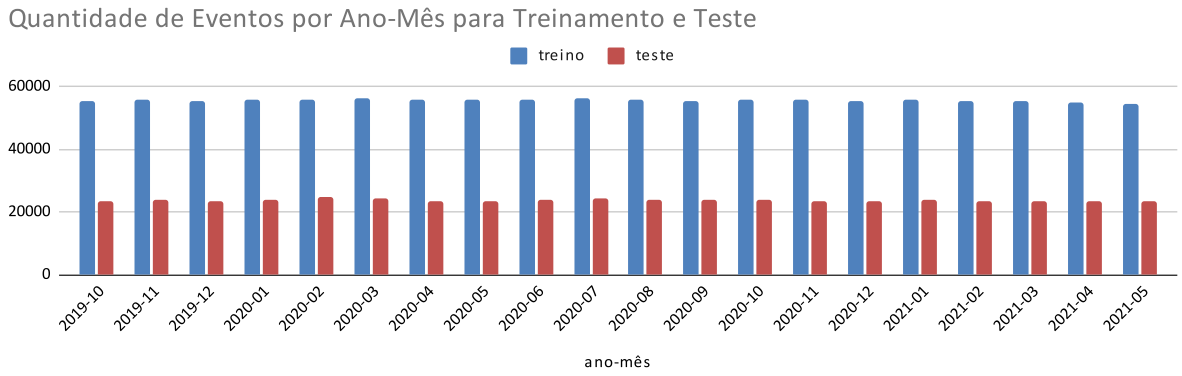
Uma tarefa relevante na identificação do problema é definir uma base de textos que permita o treinamento do modelo na etapa de extração de padrões. Dessa forma, são necessários textos previamente georreferenciados para a etapa de treino. Neste trabalho, foi utilizada a base de eventos GDELT[1] para coletar a base de treinamento e teste. Foram coletados 1.585.750 eventos ge-

[1] https://gdelt.org

orreferenciados no período de Outubro de 2019 até Maio de 2021. Desse total, 1.110.025 eventos (70%) foram utilizados como treinamento e 475.725 para teste (30%), obtidos via amostragem aleatória.

Na Figura 4.2 é apresentada uma distribuição da quantidade de eventos em relação ao período mensal da base textual.

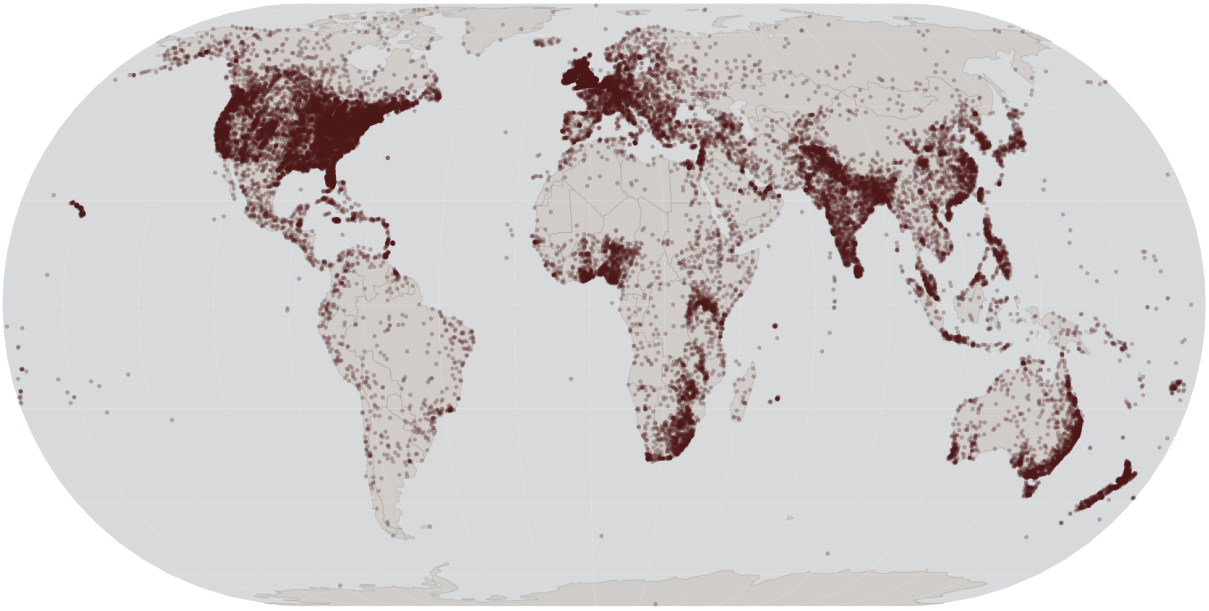
Figura 4.2: Distribuição da quantidade de eventos em relação ao período mensal da base textual.



Fonte: Elaboração própria.

No total, a base textual utilizada contém 38.412 diferentes pontos de latitude e longitude. Na Figura 4.3 é ilustrada a distribuição espacial dos eventos da base textual. Note que há uma maior quantidade de eventos em alguns continentes, representando uma característica própria das fontes de eventos que o GDELT monitora.

Figura 4.3: Distribuição espacial dos eventos textuais da base coletada.



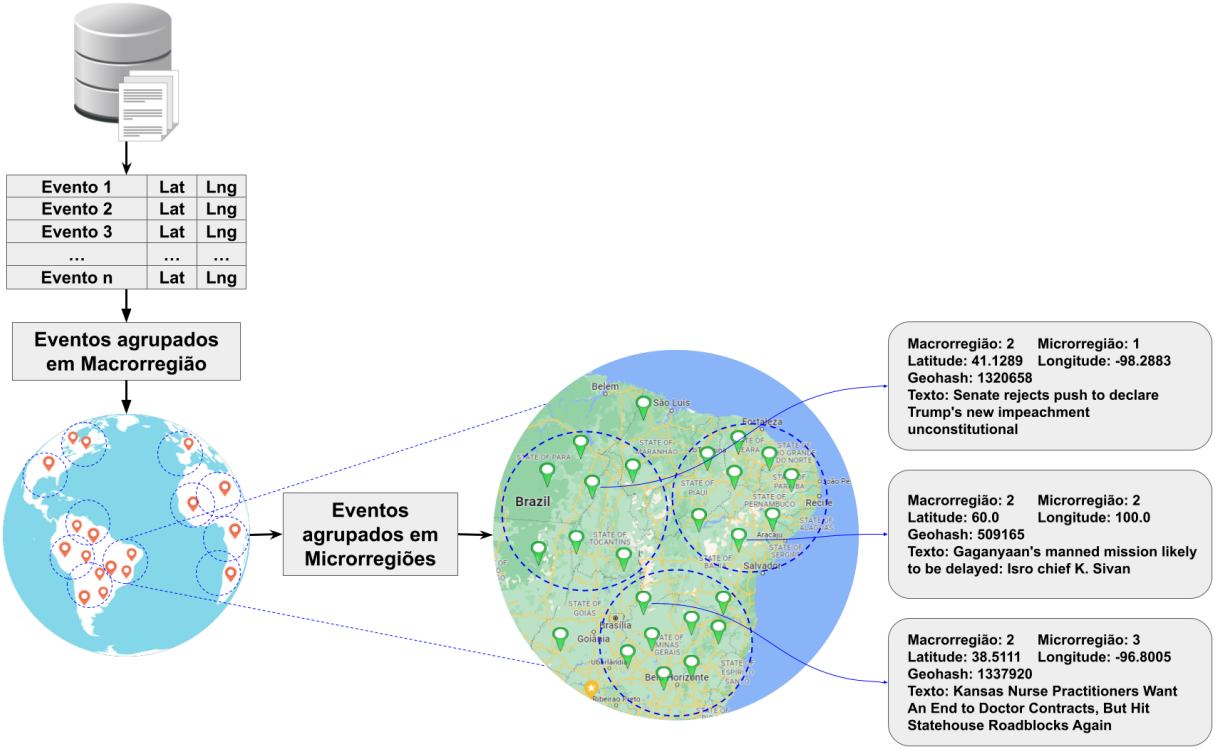
Fonte: Elaboração própria.

4.2 Pré-processamento

Na etapa de pré-processamento foram inicialmente realizadas as atividades de identificação de instâncias duplicadas para remoção, bem como instâncias com pouca informação textual, formadas apenas por 1 ou 2 *tokens*.

Em seguida, a base textual foi preparada para identificação de contextos geográficos. Note que há 38.412 pontos geográficos, sendo uma quantidade muito elevada para utilização direta de cada ponto como contexto geográfico. Além disso, há pontos que ocorrem poucos eventos e outros com muitos eventos, por exemplo, pontos que representam as capitais de estado. Assim, neste trabalho é proposto o uso de métodos de agrupamento para identificar macro e micro-regiões da base textual, como contextos geográficos.

Figura 4.4: Identificação de macro e micro-regiões para determinar contextos geográficos.



Fonte: Elaboração própria.

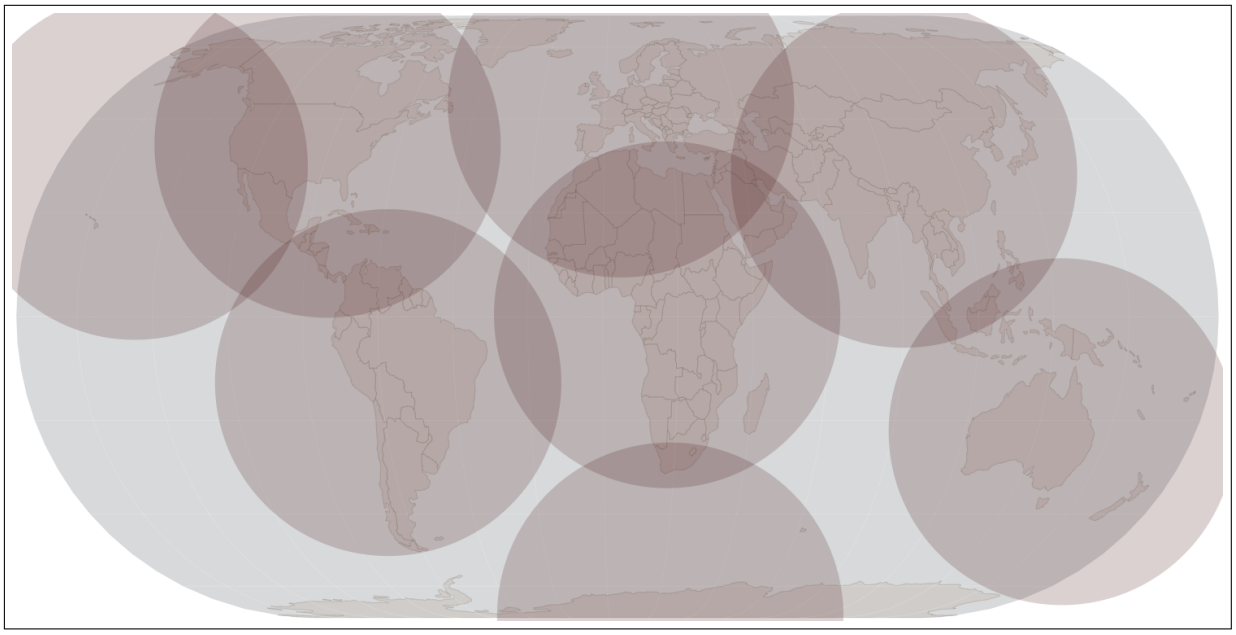
Para macro-regiões, foi utilizado o método k-means com $k = 8$ para identificar oito diferentes macro-regiões do planeta. Esse valor foi selecionado empiricamente, após testes seguidos de análises visuais da distribuição espacial dos grupos. Em seguida, os pontos geográficos de cada macro-região são agrupados novamente. Nesse caso, o valor do número de grupos dentro de uma macro-região é estimado de forma automatizada, determinado pela raiz quadrada da quantidade de pontos da macro-região. Como resultado, são obtidas as micro-regiões. É importante ressaltar que a medida de distância

utilizada para o agrupamento foi a Harversine[2], que é mais apropriada para calcular distâncias entre pontos representados por latitude e longitude.

O processo de geração de contexto geográfico por macro-regiões e micro-regiões é ilustrado na Figura 4.4. Observe que, desta maneira, cada evento da base textual é pré-processado e alocado em seus respectivos contextos geográficos.

Na Figura 4.5 são ilustradas as macro-regiões identificadas pelo método de agrupamento, em que cada macro-região está representada por meio de um círculo. De forma análoga, na Figura 4.6 são ilustradas as micro-regiões identificadas. Por fim, uma visão geral do número de eventos por macro-região e suas respectivas informações geográficas é apresentada na Tabela 4.1.

Figura 4.5: Ilustração das macro-regiões obtidas pelo processo de agrupamento.



Fonte: Elaboração própria.

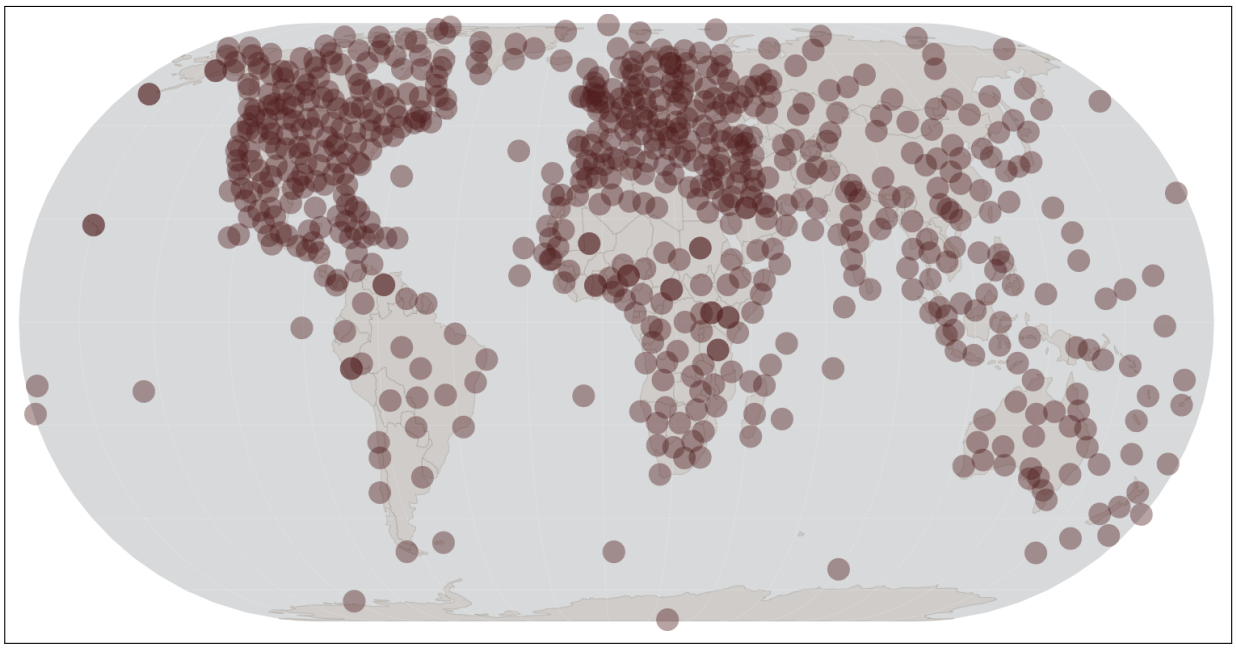
4.3 Extração de Padrões

A etapa de Extração de Padrões visa o treinamento de modelos preditivos a partir de textos georreferenciados. No caso deste projeto, o classificador visa estimar o atributo alvo (classe) que corresponde a uma macro ou micro-região.

Dado um conjunto de treinamento $\mathbf{S} = \{(\mathbf{X}_n, \mathbf{y}_n)\}_{n=1}^{N}$, em que $\mathbf{X} \in \mathcal{X}$ representa um evento textual e $\mathbf{y} \in \mathcal{Y}$ é seu rótulo correspondente, para uma base com N eventos. Especificamente, $\mathbf{y}$ é um vetor de informação geográfica, com

[2] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.haversine_distances.html

Figura 4.6: Ilustração das micro-regiões obtidas pelo processo de agrupamento.



Fonte: Elaboração própria.

Tabela 4.1: Quantidade de Eventos e Informações Geográficas de cada Macro-Região

Macro-Região	**Eventos**	**Latitude**	**Longitude**
R0	213707	31,27 ± 15,09	95,52 ± 22,06
R1	519358	38,48 ± 6,56	-91,06 ± 14,88
R2	259757	49,03 ± 7,81	5,58 ± 14,87
R3	46108	-25,03 ± 14,65	142,65 ± 25,69
R4	49686	0,39 ± 15,81	19,21 ± 16,59
R5	15088	-14,16 ± 16,38	-65,17 ± 8,38
R6	410	-85,23 ± 3,37	32,87 ± 12,67
R7	5911	33,14 ± 24,29	-150,51 ± 42,93

Fonte: Elaboração própria.

as informações de contexto representadas por macro ou micro-regiões. O objetivo é aprender uma função $f : \mathcal{X} \mapsto \mathcal{Y}$, minimizando a Equação 4.1,

$$\min_{f \to \mathcal{F}} \frac{1}{N} \sum_{n=1}^{N} \delta(\mathbf{y}_n, f(\mathbf{X}_n)), \tag{4.1}$$

na qual $\delta : \mathcal{Y}x\mathcal{Y} \mapsto \mathbb{R}$ mede o erro na previsão de $f(\mathbf{X}_n)$ em relação à informação geográfica verdadeira $\mathbf{y}_n$.

Mais especificamente, um evento $\mathbf{X}_n$ é composto por uma sequência de L *tokens*: $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_L\}$. Cada *token* $\mathbf{x}_l$ é representado inicialmente por um vetor *one-hot encoding* no espaço Δ^D, onde D é o tamanho do dicionário. Realizar o aprendizado em Δ^D é computacionalmente caro, sendo então utilizada o modelo de *word embedding* $\Delta^D \mapsto \mathbb{R}^P$, onde P é a dimensionalidade do espaço

de *embedding*. Portanto, a sequência de texto $\mathbf{X}$ é representada por meio da respectiva *embedding* para cada *token* $\mathbf{V} = \{\mathbf{v}_1, \cdots, \mathbf{v}_L\}$, onde $\mathbf{v}_1 \in \mathbb{R}^P$.

Um método de classificação, nessa estrutura, possui três etapas, considerando uma decomposição de função $f = f_0 \circ f_1 \circ f_2$:

- $f_0 : \mathbf{X} \mapsto \mathbf{V}$, cada *token* da sequência de texto é representado no espaço de *embedding* $\mathbf{V}$, de dimensão $\mathbb{R}^P$.

- $f_1 : \mathbf{V} \mapsto z$, agrega as *word embeddings* de cada *token* da sequência em uma representação vetorial z de comprimento fixo, representando todo o evento.

- $f_2 : z \mapsto \mathbf{y}$, um classificador f_2 mapeia a representação de texto z para uma informação geográfica $\mathbf{y}$.

O sucesso desse processo depende muito da eficácia das *word embeddings* em f_0. Eles são frequentemente pré-treinados *off-line* em um grande *corpus* e, em seguida, refinados por meio de f_1 e f_2 para representações de tarefas específicas. Neste projeto, modelos de aprendizado profundo consideram o mapeamento f_1 e f_2 integrados como uma "caixa preta", em uma única arquitetura neural. Essa estratégia explora a transferência de aprendizado via modelos pré-treinados, como o BERT, bem como explorar os conceitos de refinamento de modelos para tarefas específicas.

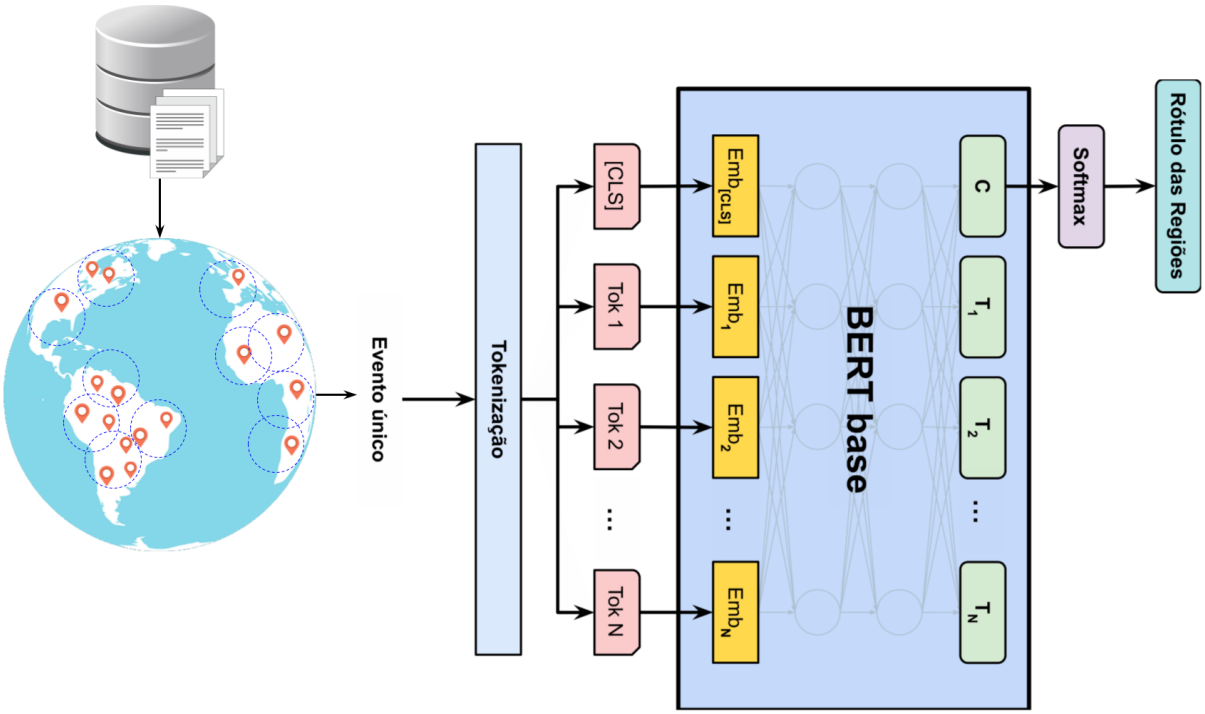
O ajuste fino do modelo *BERT* considerando as macro e micro-regiões representa a proposta de *GeoTransformers Language Model* deste projeto. O treinamento ocorreu em dois momentos distintos, em primeiro temos o treinamento com dados dos eventos rotulados apenas com as informações das macrorregiões; doravante, o modelo passou pelo mesmo processo, mas agora utilizando como entrada o conjunto de dados rotulados com as microrregiões, conforme ilustrado na Figura 4.7.

Para ajuste fino do *BERT*, utilizou-se a biblioteca *Ktrain* em *Python*. Ela consiste em um conjunto de métodos que auxiliam nas atividades de processamento de linguagem natural.

Para o ajuste fino do BERT considerando informação geográfica, optou-se por ajustar todas as 12 camadas de Transformers-Encoders de um modelo BERT-base. Para tal, foi utilizada uma taxa de treinamento reduzida de 0.00002 e 5 épocas. O ajuste fino foi realizado considerando apenas o conjunto de treinamento definido na etapa de identificação do problema.

O modelo resultante foi armazenado de maneira que os pesos da respectiva rede neural pudessem ser carregados novamente para uso posterior, também permitindo realizar a etapa de avaliação no conjunto de testes.

Figura 4.7: Ajuste fino do BERT considerando o contexto geográfico de macro ou micro-regiões como classes.



Fonte: Elaboração própria.

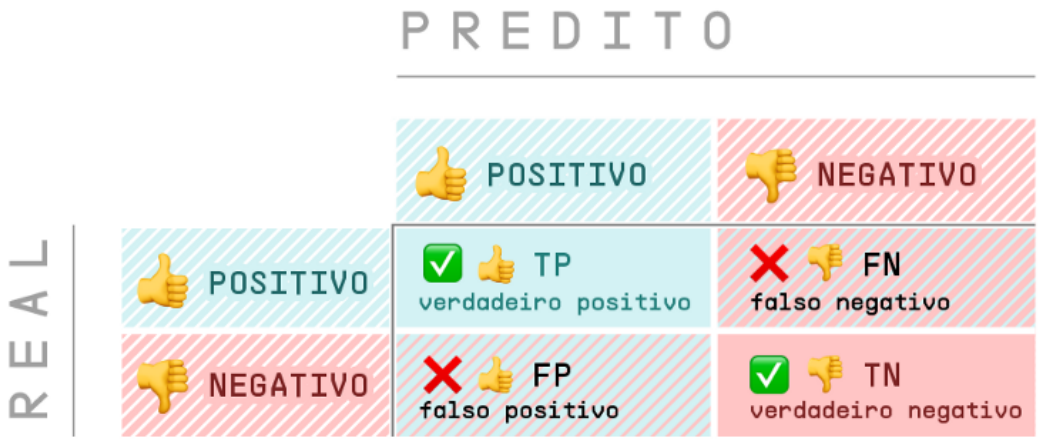
4.4 Pós-processamento

A avaliação dos resultados é baseada na capacidade do modelo em inferir regiões de interesse a partir apenas de um texto de entrada, mesmo que tal texto não possua entidade geográfica associada. Uma forma comumente utilizada para visualizar o desempenho de um determinado modelo é por meio de uma matriz de confusão, conforme ilustrada na Figura 4.8 para um problema de duas classes, mas o seu conceito pode ser estendido para problemas multiclasses.

Para avaliar os resultados obtidos, são usadas as seguintes métricas: precisão (*precision*) que é uma medida de fidelidade; revocação (*recall*), que pode ser compreendida como uma medida de completude e *F-score* representa a média harmônica entre a métrica precisão e revocação. Adiante, detalha-se cada uma dessas métricas.

- **Precisão (P)**: está métrica calcula a eficácia por intermédio da porcentagem entre a quantidade de exemplos que foram classificados de forma correta (TP) e a quantidade total de exemplos que realmente são rotulados como corretos (TP + FN). Pode-se verificar a implementação desta

Figura 4.8: Matriz de confusão de duas classes (Positivo/Negativo).



Fonte: Chen et al. [2020]

métrica na Equação 4.2

$$P = \frac{TP}{TP+FN}. \tag{4.2}$$

- **Revocação (R)**: Evidencia os erros por falso positivo. Ela serve para demonstrar qual a razão entre os exemplos que foram identificados de forma correta (TP) pelo classificador em detrimento a todos os exemplos que foram identificados como corretos (TP + FP), incluindo os exemplos que foram classificados como corretos, mas estavam errados, conforme é descrito na Equação 4.3

$$R = \frac{TP}{TP+FP}. \tag{4.3}$$

- ***F-score* (F)**: Para obter este resultado, são levados em consideração os valores adquiridos anteriormente pela precisão e cobertura. Ela representa a média harmônica entre os resultados obtidos pelas duas métricas que a antecederam. Essa métrica está definida na Equação 4.4

$$F = 2 * \frac{P*R}{P+R}. \tag{4.4}$$

Outro aspecto importante da avaliação é definir modelos de referência para comparar o desempenho obtido pelo GeoTransformers proposto neste trabalho. Assim, foram utilizados os seguintes modelos de referência:

- *Dummy*: É um modelo nulo que realiza predições de regiões de interesse de forma aleatória. Permite calcular um valor mínimo de referência das medidas de avaliação.

- *Multinomial Naive Bayes*: É um modelo probabilístico muito utilizado nos trabalhos relacionados das últimas duas décadas. De forma geral, este modelo considera como base a probabilidade de cada palavra ocorrer em uma região. A partir dessas probabilidades, são inferidas as regiões de interesse para um novo texto. Para esse modelo, foi utilizado pré-processamento dos textos, envolvendo tokenização, remoção de *stopwords*, remoção de tokens com frequência de documentos igual a 1 e ponderação de termos com TF-IDF.

- *FastText+MLP*: É um modelo baseado em *word embeddings* estáticos. Tem sido utilizado nos trabalhos mais recentes relacionados à inferência de contexto geográfico a partir dos textos. Uma rede neural foi inicializada com embeddings pré-treinadas e disponibilizadas no repositório do projeto FastText. Em seguida, foi adicionada uma nova camada densa e uma para classificação por região, como uma Multilayer Perceptron (MLP). As embeddings são ajustadas conforme o erro de classificação. A taxa de treinamento foi de 0.001 e o modelo foi treinado por 10 épocas, parâmetros definidos empiricamente a partir de um conjunto de validação extraído dos dados de treinamento.

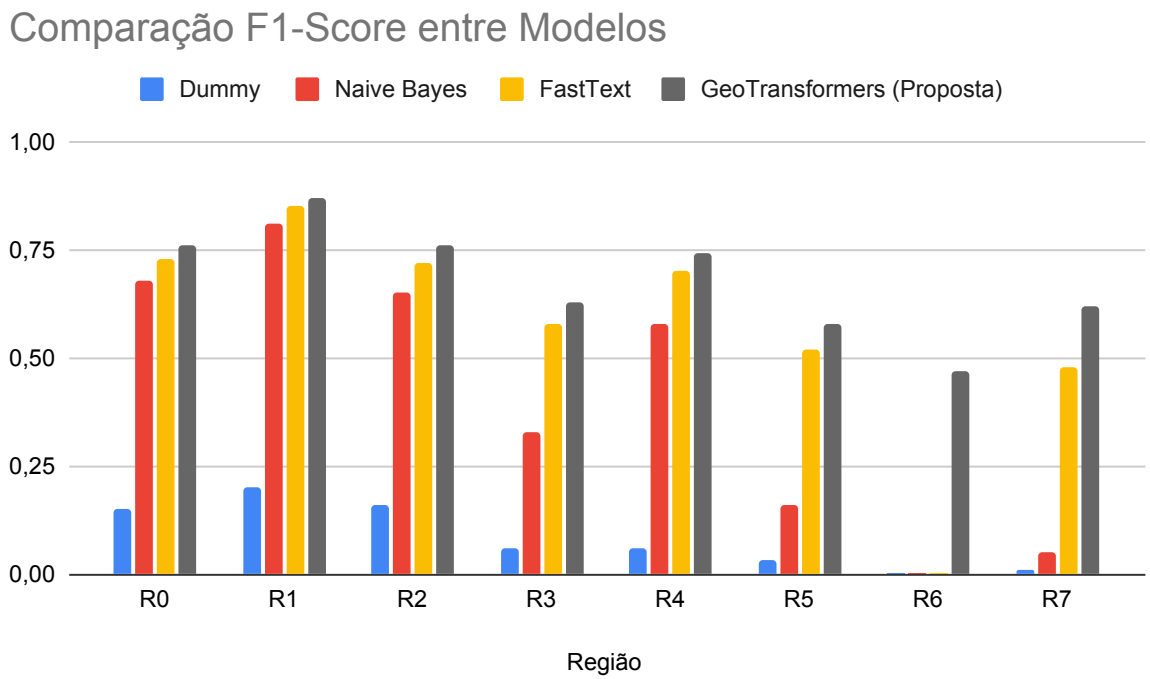
Tabela 4.2: Resultados da Avaliação Experimental

Classificação em Regiões												
	Dummy			*Naive Bayes*			*FastText*			*GeoTransformers*		
Região	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
R0	0.19	0.12	0.15	0.72	0.64	0.68	0.77	0.70	0.73	0.78	0.75	0.76
R1	0.47	0.13	0.20	0.72	0.93	0.81	0.79	0.92	0.85	0.85	0.90	0.87
R2	0.23	0.12	0.16	0.74	0.59	0.65	0.76	0.68	0.72	0.77	0.74	0.76
R3	0.04	0.12	0.06	0.78	0.21	0.33	0.72	0.49	0.58	0.66	0.61	0.63
R4	0.04	0.12	0.06	0.89	0.42	0.58	0.80	0.63	0.70	0.76	0.72	0.74
R5	0.01	0.13	0.03	0.79	0.09	0.16	0.75	0.40	0.52	0.70	0.49	0.58
R6	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.64	0.36	0.47
R7	0.01	0.13	0.01	0.81	0.03	0.05	0.77	0.35	0.48	0.75	0.53	0.62
Média	0.12	0.12	0.08	0.68	0.36	0.41	0.67	0.52	0.57	0.74	0.64	0.68

Fonte: Elaboração própria.

Na Tabela 4.2 é apresentada uma comparação geral dos modelos de referência em relação ao GeoTransformers, que nesse projeto utilizou um ajuste fino do BERT (base) usando as regiões de contexto como classe. É possível observar que o uso de modelos de embeddings contextuais (ou dinâmicas) permite obter maiores valores de Precisão (P), Revocação (R) e F1-Score (F) para todas as regiões. Em especial, vale destacar que o modelo proposto foi o único capaz de lidar com regiões com menor quantidade de eventos e difíceis de classificar, como a Região R6. Para facilitar a visualização do comportamento

Figura 4.9: Comparação F1-Score entre os modelos em relação a predição do contexto geográfico.



Fonte: Elaboração própria.

da abordagem proposta (GeoTransformers), na Figura 4.9 é apresentada uma comparação da medida F1-Score entre os modelos em relação a predição do contexto geográfico.

4.5 Uso do Conhecimento

Como possibilidade de representação do conhecimento gerado pelo modelo, o uso do conhecimento envolveu o desenvolvimento de uma ferramenta computacional que utiliza modelos pré-treinados resultantes deste trabalho de mestrado.

A entrada do modelo será um texto e a saída um mapa de calor com as regiões provavelmente mais associadas ao conteúdo do texto. Note que a abordagem utiliza uma estratégia ponta-a-ponta, ou seja, sem a necessidade de métodos intermediários para reconhecimento de entidades nomeadas (geográficas), *geoparsing* e *geocoding*.

A ferramenta desenvolvida está disponível no GitHub do projeto, no endereço https://github.com/ronaldjijoca/geotransformers/. A seguir, é ilustrado o funcionamento da ferramenta.

O primeiro passo é instanciar um objeto da classe principal, denominada GeoTransformers, conforme apresentado na Figura 4.10. Nessas etapas, é

realizado o *download* dos modelos GeoTransformers e o carregamento das redes neurais.

Figura 4.10: Código para instanciar a ferramenta com o modelo GeoTransformers.

```
GT = GeoTransformers()
GT.download_geomodels()
GT.extract_geomodels()
GT.load_geomodels()
GT.load_geoclusters()
```

Fonte: Elaboração própria.

O segundo passo é apresentar um texto de entrada. Considere o seguinte texto "*Deforestation rate increased by 52 percent during the years Bolsonaro has been in power.*", extraído de https://www.greenpeace.org/usa/news/amazon-deforestation-hits-second-highest-rate-ever-recorded-in-august/. Note que não há uma entidade geográfica explícita no texto, logo o modelo deve utilizar o conhecimento adquirido durante a etapa de treinamento para inferir o contexto geográfico.

A ferramenta retorna duas informações. A primeira informa a confiança de mapeamento do texto para cada macro-região, conforme ilustrado na Figura 4.11. Nesse exemplo, a macro-região R5 (representado por level1_geo_5) possui confiança de 0.94.

Figura 4.11: Confiança de mapeamento de um texto para cada macro-região.

```
s = '''Deforestation rate increased by 52 percent during the years Bolsonaro has been in power.'''
df_results, L = GT.predict(s,n=5)
```

```
df_results
```

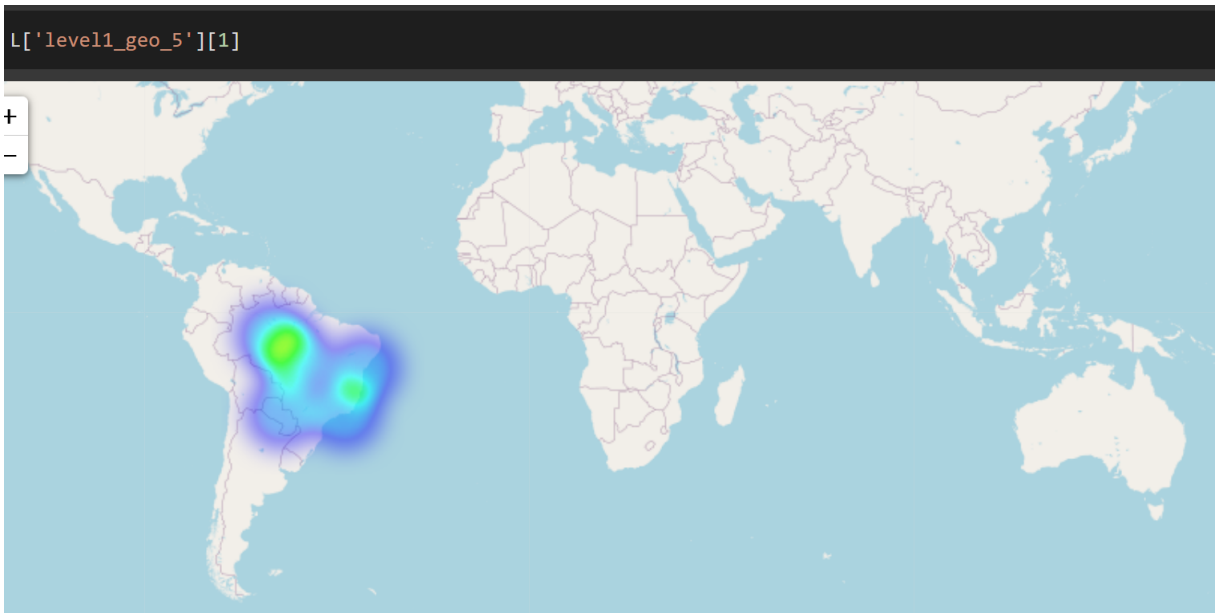
level1_geo	probs
level1_geo_0	0.035855
level1_geo_1	0.005952
level1_geo_2	0.009523
level1_geo_3	0.000283
level1_geo_4	0.000237
level1_geo_5	0.948041
level1_geo_6	0.000089
level1_geo_7	0.000020

Fonte: Elaboração própria.

Em seguida, o usuário pode selecionar a macro-região de interesse para

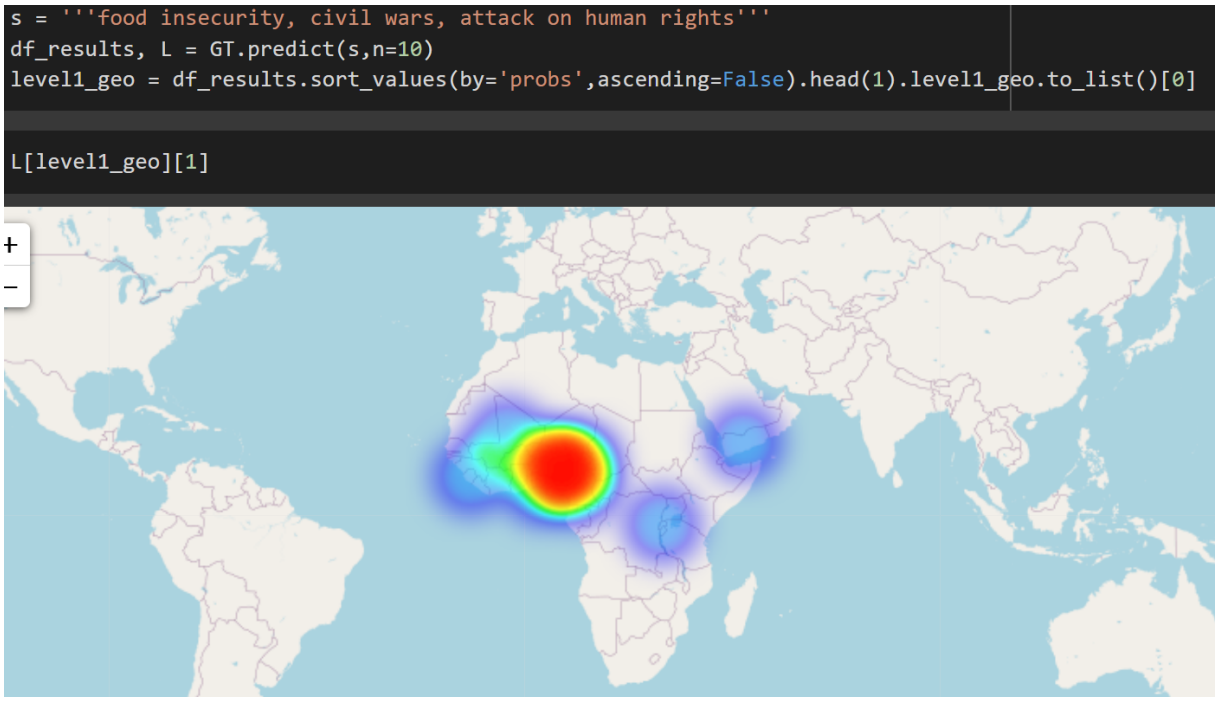
gerar o mapa de calor conforme as n micro-regiões com maior confiança (Figura 4.12). Quanto maior a confiança de mapeamento em uma micro-região, mais realçado estará no mapa.

Figura 4.12: Mapa de calor gerado a partir das n micro-regiões com maior confiança.



Fonte: Elaboração própria.

Figura 4.13: Resultado (Mapa de Calor) obtido para a consulta "*food insecurity, civil wars, attack on human rights*" no modelo GeoTransformers.



Fonte: Elaboração própria.

Por fim, vale destacar que o modelo GeoTransformers desenvolvido neste trabalho pode ser utilizado para explorar a cobertura geográfica de tópicos ou combinações de tópicos. Por exemplo, na Figura 4.13 é apresentada a cobertura geográfica do texto "*food insecurity, civil wars, attack on human rights*".

CAPÍTULO

5

Considerações Finais

Compreender a relação entre o que é escrito (evento) e o local onde foi escrito pode auxiliar os sistemas de tomada de decisões. Analisar grandes quantidades de eventos de forma manual seria inviável, dado a sua grande quantidade e alta demanda de tempo necessária para analisar, para no final obter resultados que não teriam eficácia elevada. Na tentativa de aumentar a abrangência da análise e de reduzir o tempo e esforço, pode-se utilizar técnicas de Mineração de Textos para automatizar esse processo. Assim, neste trabalho de mestrado foi investigado o uso de *Word Embeddings* com contexto geográfico, com o propósito de aprimorar a aprendizagem do modelo de classificação *GeoTransformers Language Model* sobre conjuntos de eventos.

Nas seções seguintes são apresentadas as limitações, as contribuições resultantes desse trabalho e as possibilidades de trabalhos futuros relacionados ao tema proposto.

5.1 Limitações

As principais limitações deste projeto de mestrado são listadas a seguir.

- **Escopo geográfico da base GDELT:** O *GeoTransformers* foi treinado apenas com a base do GDELT que possui muitos eventos em determinadas regiões da América do Norte e europa, e significativamente poucos eventos nas outras regiões, como na América Latina. Assim, menos microrregiões são geradas para locais com poucos eventos.
- **Avaliação por Regiões:** Observamos que os erros do *GeoTransformers* muitas vezes ocorre por associar um evento a uma região vizinha. Na

forma atual, esse erro tem o mesmo peso de um erro de classificação para uma região muito distante da região correta. Assim, ainda é necessário avançar em estratégias de treinamento que utilizam as distâncias entre latitudes e longitudes como medidas de erro.

- **Avaliação experimental em regiões que contêm baixa quantidade de eventos:** Nesse cenário, novos eventos dessas regiões não serão classificados de forma correta. Uma maneira de lidar com essa limitação seria explorar estratégias de aumento de dados.

5.2 Contribuições

Apesar das dificuldades, o trabalho apresenta contribuições que podem auxiliar, tanto a comunidade científica, quanto a comunidade em geral. Assim, é esperado que o desenvolvimento do modelo *GeoTransformers*, bem como da ferramenta proposta possa auxiliar outros trabalhos na compreensão de *Word Embeddings* com contexto geográfico. Dessa forma esse trabalho contribui com:

- Inferir a geolocalização de um evento, a partir de *Word Embeddings* com contexto geográfico;
- Avaliação do modelo *GeoTransformers Language Model*, com a inovação de explorar *Word Embeddings* dinâmicas, em comparação com outros modelos baseados em frequências e *Word Embeddings* estáticas;
- Modelo capaz de classificar textos para regiões com menor quantidade de eventos associados;
- Desenvolvimento de uma ferramenta computacional para gerar os locais com a maior possibilidade do evento ter ocorrido, com técnicas de visualização baseadas em mapas de calor.

5.3 Trabalhos futuros

Com o desenvolvimento da proposta alguns pontos de melhoria e ideias foram identificados, que serviram como sugestões para trabalhos futuros. Alguns deles são:

- Explorar quantidade maior de dados de eventos, em especial para as regiões com pequena quantidade de amostra, especialmente envolvendo estratégias para aumento de dados

- Desenvolver uma versão do *GeoTransformers Language Model* voltado exclusivamente para o português do Brasil;
- Avançar no desenvolvimento do projeto GeoTransformers (https://github.com/ronaldjijoca/geotransformers/), desenvolver uma ferramenta computacional mais simples e intuitiva para que pessoas que não sejam da área de computação possam gerar mapas de calor a partir do texto de eventos.

Todas as possibilidades de estudos citadas acima possivelmente tendem a melhorar a abordagem proposta neste trabalho, enriquecendo o modelo apresentado, bem como a derivar novos trabalhos. Espera-se que essas possibilidades possam contribuir para a construção de novos modelos que usem *Word Embeddings* com contexto geográfico que fomentem a busca por novos campos de pesquisa.

Referências Bibliográficas

Aggarwal, C. C. (2018). *Machine learning for text*. Springer. Citado na página 2.

Ahern, S., Naaman, M., Nair, R., e Yang, J. H.-I. (2007). World explorer: visualizing aggregate data from unstructured text in geo-referenced collections. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, páginas 1–10. Citado nas páginas 29 e 34.

Allan, J. (2012). *Topic detection and tracking: event-based information organization*, volume 12. Springer Science & Business Media. Citado na página 1.

Amaral, F. (2016). *Introdução à Ciência de Dados: mineração de dados e big data*. Alta Books Editora. Citado na página 7.

Andogah, G., Bouma, G., e Nerbonne, J. (2012). Every document has a geographical scope. *Data & Knowledge Engineering*, 81:1–20. Citado nas páginas 3 e 16.

Andrade, L. e Silva, M. J. (2006). Relevance ranking for geographic ir. In *GIR*. Citado nas páginas 27, 28, e 34.

Aranha, C. e Passos, E. (2006). A tecnologia de mineração de textos. *Revista Eletrônica de Sistemas de Informação*, 5(2). Citado na página 8.

Blier-Wong, C., Baillargeon, J.-T., Cossette, H., Lamontagne, L., e Marceau, E. (2020). Encoding neighbor information into geographical embeddings using convolutional neural networks. In *The Thirty-Third International Flairs Conference*. Citado nas páginas 31 e 34.

Bojanowski, P., Grave, E., Joulin, A., e Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146. Citado nas páginas 2 e 22.

Buscaldi, D. (2011). Approaches to disambiguating toponyms. *Sigspatial Special*, 3(2):16–19. Citado nas páginas 16 e 17.

Caires, K. V. L. e Simonelli, G. (2021). Previsão da demanda brasileira de biodiesel utilizando redes neurais artificiais. *Research, Society and Development*, 10(5):e17410513381–e17410513381. Citado na página 12.

Cardoso, A. B., Martins, B., e Estima, J. (2022). A novel deep learning approach using contextual embeddings for toponym resolution. *ISPRS International Journal of Geo-Information*, 11(1):28. Citado nas páginas 31 e 34.

Chen, D., Nigri, E., Oliveira, G., Sepulvene, L., e Alves, T. (2020). Métricas de avaliação em machine learning: Classificação. *Medium*. Citado na página 43.

Chen, L., Cong, G., Jensen, C. S., e Wu, D. (2013). Spatial keyword query processing: An experimental evaluation. *Proceedings of the VLDB Endowment*, 6(3):217–228. Citado nas páginas 27 e 34.

Chen, X. e Li, Q. (2020). Event modeling and mining: a long journey toward explainable events. *The VLDB Journal*, 29(1):459–482. Citado na página 1.

Cocos, A. e Callison-Burch, C. (2017). The language of place: Semantic value from geospatial context. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, páginas 99–104. Citado na página 3.

Contractor, D., Goel, S., e Singla, P. (2021). Joint spatio-textual reasoning for answering tourism questions. In *Proceedings of the Web Conference 2021*, páginas 1978–1989. Citado nas páginas 32 e 34.

DA GAMA, J. M. P., FERREIRA, A. C. P. D. L., CARVALHO, D., FACELI, K., LORENA, A. C., e OLIVEIRA, M. (2017). Extração de conhecimento de dados: data mining. Citado nas páginas 11 e 12.

Dassereto, F., Rocco, L. D., Shaw, S., Guerrini, G., e Bertolotto, M. (2020). How to tune parameters in geographical ontologies embedding. In *Proceedings of the 4th ACM SIGSPATIAL Workshop on Location-Based Recommendations, Geosocial Networks, and Geoadvertising*, páginas 1–9. Citado nas páginas 3, 32, e 34.

De Sabbata, S. e Reichenbacher, T. (2010). A probabilistic model of geographic relevance. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, páginas 1–2. Citado na página 28.

DeLozier, G., Baldridge, J., e London, L. (2015). Gazetteer-independent toponym resolution using geographic word profiles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29. Citado nas páginas 28 e 34.

Devlin, J., Chang, M.-W., Lee, K., e Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. Citado nas páginas 2, 5, 23, e 24.

Ding, J., Gravano, L., e Shivakumar, N. (2000). Computing geographical scopes of web resources. Citado nas páginas 28 e 34.

Ebrahimabadi, A., Azimipour, M., e Bahreini, A. (2015). Prediction of roadheaders' performance using artificial neural network approaches (mlp and kosfm). *Journal of Rock Mechanics and Geotechnical Engineering*, 7(5):573–583. Citado na página 12.

Fayyad, U., Piatetsky-Shapiro, G., e Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37. Citado nas páginas 8, 11, e 12.

Ferneda, E. (2006). Redes neurais e sua aplicação em sistemas de recuperação de informação. *Ciência da Informação*, 35:25–30. Citado na página 12.

Frontiera, P., Larson, R., e Radke, J. (2008). A comparison of geometric approaches to assessing spatial similarity for gir. *International Journal of Geographical Information Science*, 22(3):337–360. Citado nas páginas 28 e 34.

Gale, W. A., Church, K., e Yarowsky, D. (1992). One sense per discourse. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*. Citado na página 16.

Gong, H., Bhat, S., e Viswanath, P. (2020). Enriching word embeddings with temporal and spatial information. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, páginas 1–11. Citado na página 3.

Gudivada, V. N. (2018). Natural language core tasks and applications. In *Handbook of statistics*, volume 38, páginas 403–428. Elsevier. Citado nas páginas 2 e 18.

Hariharan, R., Hore, B., Li, C., e Mehrotra, S. (2007). Processing spatial-keyword (sk) queries in geographic information retrieval (gir) systems. In *19th International Conference on Scientific and Statistical Database Management (SSDBM 2007)*, páginas 16–16. IEEE. Citado nas páginas 27 e 34.

Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Perez-Meana, H., Portillo-Portillo, J., Sanchez, V., e García Villalba, L. J. (2019). Using twitter data to monitor natural disaster social dynamics: A recurrent neural network approach with word embeddings and kernel density estimation. *Sensors*, 19(7):1746. Citado nas páginas 31 e 34.

Hogenboom, F., Frasincar, F., Kaymak, U., De Jong, F., e Caron, E. (2016). A survey of event extraction methods from text for decision support systems. *Decision Support Systems*, 85:12–22. Citado na página 1.

Hu, Y. (2018). Geo-text data and data-driven geospatial semantics. *Geography Compass*, 12(11):e12404. Citado nas páginas 1, 2, e 7.

Huang, B. e Carley, K. M. (2019). A hierarchical location prediction neural network for twitter user geolocation. *arXiv preprint arXiv:1910.12941*. Citado nas páginas 31 e 34.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., e Mikolov, T. (2016). Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*. Citado na página 2.

Kinsella, S., Murdock, V., e O'Hare, N. (2011). "i'm eating a sandwich in glasgow"modeling locations with tweets. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, páginas 61–68. Citado nas páginas 29 e 34.

Konkol, M., Brychcín, T., Nykl, M., e Hercig, T. (2017). Geographical evaluation of word embeddings. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, páginas 224–232. Citado nas páginas 3, 30, e 34.

Laere, O. V., Schockaert, S., Tanasescu, V., Dhoedt, B., e Jones, C. B. (2014). Georeferencing wikipedia documents using data from social media sources. *ACM Transactions on Information Systems (TOIS)*, 32(3):1–32. Citado nas páginas 30 e 34.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., e Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240. Citado na página 23.

Leidner, J. L. e Lieberman, M. D. (2011). Detecting geographical references in the form of place names and associated spatial natural language. *Sigspatial Special*, 3(2):5–11. Citado nas páginas 14 e 16.

Li, X., Xie, Q., e Huang, L. (2019). Identifying the development trends of emerging technologies using patent analysis and web news data mining: the case of perovskite solar cell technology. *IEEE Transactions on Engineering Management*. Citado na página 7.

Ling, Y., Yue, Q., Chai, C., Shan, Q., Hei, D., e Jia, W. (2020). Nuclear accident source term estimation using kernel principal component analysis, particle swarm optimization, and backpropagation neural networks. *Annals of Nuclear Energy*, 136:107031. Citado na página 12.

Liu, Z., Miranda, F., Xiong, W., Yang, J., Wang, Q., e Silva, C. (2020). Learning geo-contextual embeddings for commuting flow prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, páginas 808–816. Citado nas páginas 11 e 34.

Makrehchi, M. e Kamel, M. S. (2008). Automatic extraction of domain-specific stopwords from labeled documents. In *European Conference on Information Retrieval*, páginas 222–233. Springer. Citado na página 9.

Mao, H., Thakur, G., Sparks, K., Sanyal, J., e Bhaduri, B. (2018). Mapping near-real-time power outages from social media. *International Journal of Digital Earth*. Citado nas páginas 31 e 34.

Marcacini, R. M., Rossi, R. G., Nogueira, B. M., Martins, L. V., Cherman, E. A., e Rezende, S. O. (2017). Websensors analytics: Learning to sense the real world using web news events. In *Anais Estendidos do XXIII Simpósio Brasileiro de Sistemas Multimídia e Web*, páginas 169–173. SBC. Citado na página 4.

Martins, B., Silva, M. J., e Chaves, M. S. (2005). Challenges and resources for evaluating geographical ir. In *Proceedings of the 2005 workshop on Geographic Information Retrieval*, páginas 65–69. Citado nas páginas 27 e 34.

Medad, A., Gaio, M., Moncla, L., Mustière, S., e Nir, Y. L. (2020). Comparing supervised learning algorithms for spatial nominal entity recognition. *AGILE: GIScience Series*, 1:1–18. Citado nas páginas 11, 32, e 34.

Mikheev, A., Moens, M., e Grover, C. (1999). Named entity recognition without gazetteers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*. Citado na página 14.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., e Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*. Citado nas páginas 2, 19, 20, e 22.

Miller, H. J. e Goodchild, M. F. (2015). Data-driven geography. *GeoJournal*, 80(4):449–461. Citado na página 7.

Miyazaki, T., Rahimi, A., Cohn, T., e Baldwin, T. (2018). Twitter geolocation using knowledge-based methods. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, páginas 7–16. Citado nas páginas 30 e 34.

Naseem, U., Razzak, I., Khan, S. K., e Prasad, M. (2021). A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5):1–35. Citado na página 10.

Overell, S. e Rüger, S. (2008). Using co-occurrence models for placename disambiguation. *International Journal of Geographical Information Science*, 22(3):265–287. Citado na página 17.

Ozdikis, O., Ramampiaro, H., e Nørvåg, K. (2019). Locality-adapted kernel densities of term co-occurrences for location prediction of tweets. *Information Processing & Management*, 56(4):1280–1299. Citado nas páginas 30 e 34.

O'Hare, N. e Murdock, V. (2013). Modeling locations with social media. *Information retrieval*, 16(1):30–62. Citado nas páginas 29 e 34.

Palacio, D., Cabanac, G., Sallaberry, C., e Hubert, G. (2010). On the evaluation of geographic information retrieval systems. *International Journal on Digital Libraries*, 11(2):91–109. Citado nas páginas 27 e 34.

Pei, T., Song, C., Guo, S., Shu, H., Liu, Y., Du, Y., Ma, T., e Zhou, C. (2020). Big geodata mining: objective, connotations and research issues. *Journal of Geographical Sciences*, 30(2):251–266. Citado nas páginas 1 e 2.

Pennington, J., Socher, R., e Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, páginas 1532–1543. Citado na página 2.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., e Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, páginas 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics. Citado na página 2.

Purves, R. S., Clough, P., Jones, C. B., Hall, M. H., e Murdock, V. (2018). Geographic information retrieval: Progress and challenges in spatial search of text. *Foundations and Trends in Information Retrieval*, 12(2-3):164–318. Citado nas páginas 2 e 3.

Radke, M. A., Gupta, A., Stock, K., e Jones, C. B. (2022). Disambiguating spatial prepositions: The case of geo-spatial sense detection. *Transactions in GIS*. Citado nas páginas 32 e 34.

Rahimi, A., Cohn, T., e Baldwin, T. (2017). A neural model for user geolocation and lexical dialectology. *arXiv preprint arXiv:1704.04008*. Citado nas páginas 30 e 34.

Rezende, S. O. (2003). *Sistemas inteligentes: fundamentos e aplicações*. Editora Manole Ltda. Citado nas páginas 8, 9, 11, e 35.

Rezende, S. O., Marcacini, R. M., e Moura, M. F. (2011). O uso da mineração de textos para extração e organização não supervisionada de conhecimento. *Embrapa Informática Agropecuária-Artigo em periódico indexado (ALICE)*. Citado na página 10.

Rogers, A., Kovaleva, O., e Rumshisky, A. (2021). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866. Citado na página 23.

Roller, S., Speriosu, M., Rallapalli, S., Wing, B., e Baldridge, J. (2012). Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, páginas 1500–1510. Citado nas páginas 30 e 34.

Salvini, M. M. e Fabrikant, S. I. (2016). Spatialization of user-generated content to uncover the multirelational world city network. *Environment and Planning B: Planning and Design*, 43(1):228–248. Citado na página 34.

Sansome, G. e Hacker, A. (2020). Management and curation of multidimensional data in biobank studies. In *Population Biobank Studies: A Practical Guide*, páginas 171–202. Springer. Citado na página 9.

Sinoara, R. A. (2021). Mineraçao de textos e semântica: desafios, abordagens e aplicaçoes. *Revista de Sistemas de Informaçao da FSMA*, (27):41–53. Citado na página 2.

Sivamani, S., Selvakumar, S., Rajendran, K., e Muthusamy, S. (2019). Artificial neural network–genetic algorithm-based optimization of biodiesel production from simarouba glauca. *Biofuels*, 10(3):393–401. Citado na página 12.

Speriosu, M. e Baldridge, J. (2013). Text-driven toponym resolution using indirect supervision. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 1466–1476. Citado na página 17.

Spitz, A., Geiß, J., e Gertz, M. (2016). So far away and yet so close: Augmenting toponym disambiguation and similarity with text-based networks. In *Proceedings of the third international ACM SIGMOD workshop on managing and mining enriched geo-spatial data*, páginas 1–6. Citado nas páginas 28 e 34.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240. Citado na página 16.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., e Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*. Citado na página 22.

Wang, C., Xie, X., Wang, L., Lu, Y., e Ma, W.-Y. (2005). Web resource geographic location classification and detection. In *Special interest tracks and posters of the 14th international conference on World Wide Web*, páginas 1138–1139. Citado nas páginas 29 e 34.

Weiss, S. M., Indurkhya, N., Zhang, T., e Damerau, F. (2010). *Text mining: predictive methods for analyzing unstructured information*. Springer Science & Business Media. Citado na página 10.

Weitzel, L., de Oliveira, J. P. M., Carbonera, J. L., e Torres, P. A. (2010). Expansão de consulta semântica aplicadas a sistemas de recuperação de informação de contexto geográfico. *Cadernos de Informática*, 5(1):7–22. Citado na página 4.

Wing, B. e Baldridge, J. (2011). Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, páginas 955–964. Citado nas páginas 30 e 34.

Xiang, W. e Wang, B. (2019). A survey of event extraction from text. *IEEE Access*, 7:173111–173137. Citado na página 1.

Xu, L., Du, Z., Mao, R., Zhang, F., e Liu, R. (2020). Gsam: A deep neural network model for extracting computational representations of chinese addresses fused with geospatial feature. *Computers, Environment and Urban Systems*, 81:101473. Citado nas páginas 31, 32, e 34.

Zhang, T., Wu, F., Katiyar, A., Weinberger, K. Q., e Artzi, Y. (2020). Revisiting few-sample bert fine-tuning. *arXiv preprint arXiv:2006.05987*. Citado na página 24.

Zhong, T., Wang, T., Wang, J., Wu, J., e Zhou, F. (2020). Multiple-aspect attentional graph neural networks for online social network user localization. *IEEE Access*, 8:95223–95234. Citado nas páginas 11, 32, e 34.