UNIVERSIDADE FEDERAL DE CAMPINA GRANDE

CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA

UNIDADE ACADÊMICA DE SISTEMAS E COMPUTAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

SALATIEL DANTAS SILVA

GERAÇÃO DE *EMBEDDINGS* DE TIPOS DE POI COM BASE EM FEIÇÕES GEOGRÁFICAS

CAMPINA GRANDE – PB

2024

Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Geração de *Embeddings* de Tipos de POI com base em Feições Geográficas

Salatiel Dantas Silva

Proposta de Tese submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Doutor em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Engenharia de Software

Cláudio E. C. Campelo (Orientador)

Campina Grande, Paraíba, Brasil

S586g Silva, Salatiel Dantas.
Geração de *embeddings* de tipos de POI com base em feições geográficas / Salatiel Dantas Silva. – Campina Grande, 2024.
159 f. : il. color.

Tese (Doutorado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2024.
"Orientação: Prof. Dr. Cláudio Elízio Calazans Campelo".
Referências.

1. Engenharia de Software. 2. Pontos de Interesse (POIs) – *Embeddings*. 3. Processamento de Linguagem Natural (PLN). 4. Geo-semântica. I. Campelo, Cláudio Elízio Calazans. II. Título.

CDU 004.41(043)

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECÁRIA SEVERINA SUELI DA SILVA OLIVEIRA CRB-15/225

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
POS-GRADUACAO EM CIENCIA DA COMPUTACAO
Rua Aprígio Veloso, 882, Edifício Telmo Silva de Araújo, Bloco CG1, - Bairro Universitário, Campina Grande/PB, CEP 58429-900
Telefone: 2101-1122 - (83) 2101-1123 - (83) 2101-1124
Site: http://computacao.ufcg.edu.br - E-mail: secretaria-copin@computacao.ufcg.edu.br / copin@copin.ufcg.edu.br

FOLHA DE ASSINATURA PARA TESES E DISSERTAÇÕES

SALATIEL DANTAS SILVA

GERAÇÃO DE EMBEDDINGS DE TIPOS DE POI COM BASE EM FEIÇÕES GEOGRÁFICAS

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação como pré-requisito para obtenção do título de Doutor em Ciência da Computação.

Aprovada em: 20/05/2024

Prof. Dr. CLÁUDIO ELÍZIO CALAZANS CAMPELO, UFCG, Orientador

Prof. Dr. EANES TORRES PEREIRA, UFCG, Examinador Interno

Prof. Dr. DIMAS CASSIMIRO DO NASCIMENTO FILHO, UFAPE, Examinador Interno

Prof. Dr. RENATO FILETO, UFSC, Examinador Externo

Prof. Dr. THALES MIRANDA DE ALMEIDA VIEIRA, UFAL, Examinador Externo

Documento assinado eletronicamente por **CLAUDIO ELIZIO CALAZANS CAMPELO**, **PROFESSOR(A) DO MAGISTERIO SUPERIOR**, em 22/05/2024, às 00:07, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **Dimas Cassimiro do Nascimento Filho**, **Usuário Externo**, em 22/05/2024, às 09:35, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **EANES TORRES PEREIRA**, **PROFESSOR 3 GRAU**, em 23/05/2024, às 18:37, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

A autenticidade deste documento pode ser conferida no site https://sei.ufcg.edu.br/autenticidade, informando o código verificador **4447123** e o código CRC **6220F41D**.

Criado por 05843733416, versão 4 por 05843733416 em 21/05/2024 13:16:25.

Resumo

Pontos de Interesse (POIs) são locais específicos, como restaurantes, *shoppings* e parques, considerados relevantes para os usuários. Representar seus tipos por meio de mecanismos computacionais é crucial para desenvolver soluções que auxiliam em tarefas como planejamento urbano, clusterização e recomendação de POIs. Abordagens recentes têm utilizado vetores de alta dimensão (*vector embeddings*) para representar os tipos de POI com base nas relações contextuais de vizinhança ou palavras associadas aos POIs. Tais representações têm deixado de lado as feições geográficas presentes nas imediações como ruas, edifícios, rios e parques. No entanto, essas feições podem contribuir significativamente para uma melhor representação dos tipos de POI. Nesse contexto, esta pesquisa propõe uma abordagem para gerar *embeddings* de tipos de POI utilizando as feições geográficas presentes no contexto dos POIs. Na abordagem proposta, foi desenvolvido e utilizado o algoritmo GeoContext2Vec, que considera os tipos de POI e as feições geográficas presentes em seu contexto para gerar um conjunto de treinamento, preservando os padrões espaciais de espaço e ocorrência das feições. Tal conjunto é utilizado para treinar os modelos Word2Vec e DistilBert, da área de Processamento de Linguagem Natural (PLN), capazes de gerar os *embeddings* dos tipos. Como principais resultados obtidos, constatou-se que os *embeddings* produzidos com o GeoContext2Vec refletem a similaridade dos tipos de POI conforme estruturas hierárquicas e a opinião de pessoas, com valores de *matching* de aproximadamente 98%, superando estratégias do estado-da-arte. Além disso, os resultados apontam a superioridade dos *embeddings* em uma tarefa de classificação de zonas urbanas, alcançando um valor de F-Score de 90%. Tal resultado demonstra que as feições geográficas são informações relevantes na representação de tipos de POI.

Abstract

Points of Interest (POIs) are specific locations, such as restaurants, shopping centers, and parks, considered relevant to users. Representing their types through computational mechanisms is crucial for developing solutions that assist in tasks such as urban planning, clustering, and POI recommendation. Recent approaches have used high-dimensional vectors (vector embeddings) to represent POI types based on contextual neighborhood relationships or words associated with POIs. Such representations have overlooked the geographical features present in the vicinity, such as streets, buildings, rivers, and parks. However, these features can significantly contribute to a better representation of POI types. In this context, this research proposes an approach to generate embeddings of POI types using the geographic features present in the context of POIs. In the proposed approach, the GeoContext2Vec algorithm has been developed and employed, which considers POI types and the geographical features present in their context to generate a training set, preserving spatial patterns and occurrences of the features. This set is used to train the Word2Vec and DistilBert models, from the Natural Language Processing (NLP) area, capable of generating the embeddings of types. As the main results obtained, it was found that the embeddings produced with GeoContext2Vec reflect the similarity of POI types according to hierarchical structures and people's opinions, with matching values of approximately 98%, surpassing state-of-the-art strategies. Furthermore, the results indicate the superiority of these embeddings in an urban zone classification task, achieving an F-Score value of 90%. This result demonstrates that geographical features are relevant information in the representation of POI types.

Agradecimentos

Primeiramente, gostaria de expressar minha gratidão a Deus por estar presente em minha vida, derramando suas bênçãos sobre mim e minha família, nos dando força para lutar e sempre nos proporcionar vitória.

Aos meus pais, Francisco Damião e Antônia Núbia, e ao meu irmão, Sóstenes Nemuel, pelo constante incentivo, apoio e investimento em todas as etapas da minha jornada. Agradeço também a todos os familiares e amigos pelos momentos que compartilhamos juntos e por acreditarem em nossos sonhos.

À minha esposa, Fernanda Grasiane, pelo apoio, companheirismo e amor dedicados durante toda esta trajetória. Sua dedicação e ajuda foram fundamentais para alcançar meus objetivos.

Aos meus orientadores, Cláudio Campelo e Maxwell Guimarães de Oliveira, pela orientação, amizade e apoio ao longo desses anos. Suas contribuições foram essenciais para o desenvolvimento deste trabalho.

A todos os meus colegas do Laboratório de Computação Inteligente Aplicada (LACINA), especialmente Alexandre Ribeiro, Filipe Gomes, Matheus Lisboa, Helen Cavalcanti, Maria Eduarda, José Manoel, José Davi e Carlos Vinícius, pelos diversos momentos de aprendizado e descontração compartilhados.

Ao Programa de Pós-Graduação em Ciência da Computação da UFCG, pelo corpo docente, direção e administração, cuja competência, comprometimento e ética foram fundamentais para esta conquista.

Por fim, agradeço a todos que, de alguma forma, contribuíram para minha formação acadêmica e pessoal.

Conteúdo

Lista de Siglas

BERT	*Bidirectional Encoder Representations from Transformers*
BHE	*Binary Hit Evaluation*
CBOW	*Continuous Bag-of-Words*
CNN	*Convolutional Neural Network*
ERNIE	*Enhanced Representation through kNowledge IntEgration*
EUA	Estados Unidos da América
GeoContext2Vec	*Geographic Context to Vector*
GloVe	*Global Vectors*
GNN	*Graph Neural Network*
IA	Inteligência Artificial
ITDL	*Information Theoretic Distance Lagged*
LLM	*Large Language Model*
LSTM	*Long Short Term Memory*
MLM	*Masked Language Model*
MRR	*Mean Reciprocal Rank*
NSP	*Next Sentence Prediction*
OP	*Occurence Proportion*
OSM	*Open Street Map*
PCA	*Principal Component Analysis*
PLN	Processamento de Linguagem Natural
POI	*Point of Interest*
RHE	*Ranking Hit Evaluation*
RNN	*Recurrent Neural Network*
SP	*Space Proportion*
SVM	*Support Vector Machine*
t-SNE	*t-distributed Stochastic Neighbor Embedding*

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Pontos de Interesse (POIs) são locais específicos considerados úteis ou relevantes. Para que uma entidade seja classificada como POI, é necessário que ela possua os seguintes atributos: i) um nome; ii) um local indicado por coordenadas geográficas; iii) pelo menos um tipo (categoria), que indica sua natureza ou serviço; iv) um identificador; v) e alguma informação para contato [34].

POIs possuem ciclo de vida bem definido. Eles surgem quando atividades humanas passam a ser realizadas em seu espaço e deixam de existir quando essas atividades cessam. Exemplos de POIs incluem mercearias, restaurantes, escolas, cafés, como também locais que apresentam espaços físicos mais amplos, como parques, arenas, praias ou atrações turísticas [81; 92].

POIs são uma parte importante da vida cotidiana da sociedade, pois as atividades realizadas em seu espaço implicam diretamente em diversos setores da sociedade, como a cultura, segurança, saúde e economia [59]. Por esse motivo, os dados de POI têm ganhado bastante atenção em diversas áreas de pesquisas, visando a criação de soluções computacionais capazes de auxiliar pessoas, empresas ou governos em seu cotidiano. Tais mecanismos permitem que pessoas naveguem em ambientes e descubram comodidades e serviços; as empresas podem analisar o mercado e tomar melhores decisões; os governos podem mapear regiões com precisão e otimizar serviços como transporte público, serviços de emergência, aplicação da lei, entre outros [21; 32; 65; 77; 84; 106; 112; 114].

As soluções computacionais empregadas nas áreas anteriormente citadas, geralmente uti-

lizam os tipos de POI como seus representantes. Esses tipos servem como entrada em algoritmos de classificação, clusterização, recomendação, entre outros. Consequentemente, a qualidade dos resultados dependem do quão precisas são as representações computacionais dos tipos.

Nesse contexto, estudos recentes têm utilizado modelos neurais da área de Processamento de Linguagem Natural (PLN) para gerar representações de tipos de POI, como apresentado em [69; 89; 103]. Esses modelos permitem capturar a relação contextual dos tipos a partir dos dados brutos de POIs, sem a necessidade de pressupostos ou conhecimentos de especialistas, como a árvore ontológica. Como exemplo, Yu, Wanyan e Wang [108] utilizaram modelos de PLN para aprender as sequências de visitação de POIs com o intuito de melhorar a recomendação de POIs. Yang, Bo e Zhang [105] utilizaram modelos de PLN para aprender a relação de vizinhança de POIs com o objetivo de classificar regiões da cidade. Liu *et al*. [42] também utilizaram a relação de vizinhança de POIs para gerar uma representação da vizinhança, denominada "nicho". Portanto, o desenvolvimento de representações computacionais dos tipos de POI mais robustas possibilita que tarefas relacionadas a POIs atinjam melhores resultados. É nesse contexto que se insere este trabalho. Especificamente, buscou-se utilizar novos mecanismos e características para incorporar mais informação nas representações dos tipos de POI.

1.1 Motivação

Pesquisas recentes [11; 13; 24; 33; 89; 109; 103] propõem representar os tipos de POI utilizando modelos neurais, como o Word2Vec. Quando aplicado a dados textuais, esse modelo é capaz de produzir representações vetoriais das palavras com base em seu contexto em um corpus de texto [48]. Nessa técnica, cada palavra é representada por vetores de valores reais, conhecidos como *vector embeddings* ou apenas *embeddings*. A partir dos vetores, é possível calcular a similaridade entre palavras.

No cenário de POIs, o Word2Vec vem sendo empregado para aprender a relação contextual de vizinhança de POIs em diferentes regiões espaciais visando a obtenção dos vetores dos tipos de POI (também conhecidos como *POI type embeddings*) [23; 46; 68; 90]. Como exemplo, a Figura 1.1.a ilustra o contexto de vizinhança de um lava-jato e de

um posto de gasolina. Nos dois contextos, é possível perceber a presença de um bar, estacionamentos e locadoras. Considerando as relações de vizinhança, podemos afirmar que lava-jatos e postos de gasolina compartilham certa similaridade.

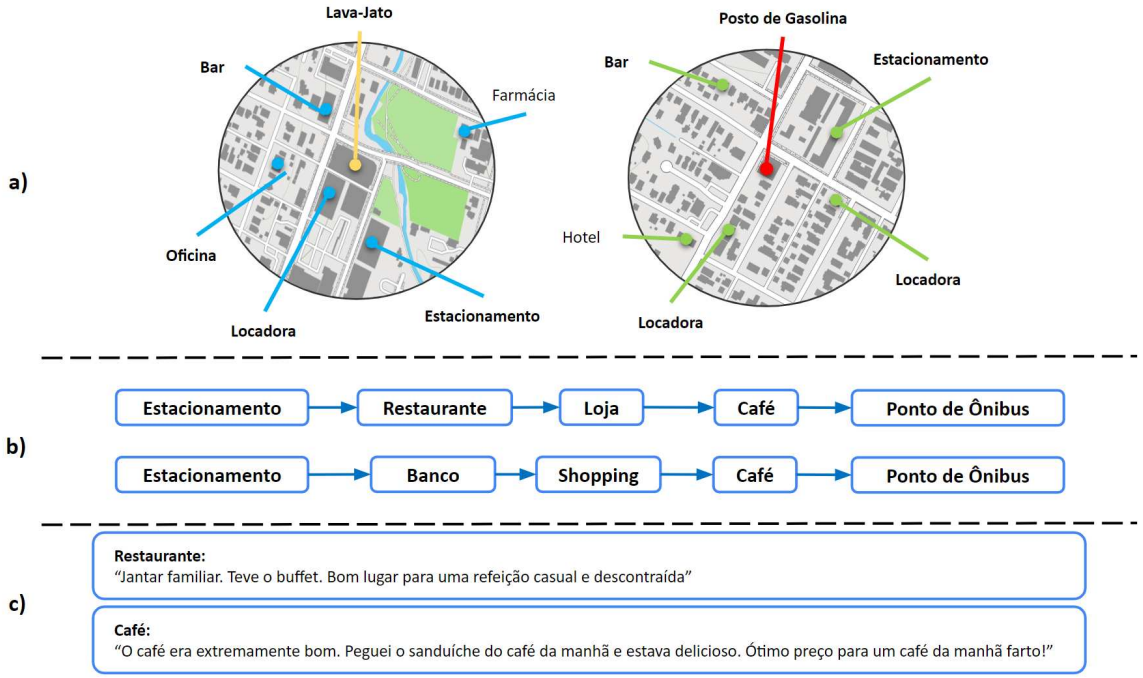


Figura 1.1: a) Coocorrência de POIs na vizinhança; b) Sequência de visitação de POIs; c) Comentários sobre os POIs.

Fonte: Autoria própria

Outra linha de pesquisa sugere que *embeddings* de tipos de POI podem ser gerados utilizando a relação contextual de sequência de visita entre os POIs [11; 46; 49; 88; 90]. Como exemplo, a Figura 1.1.b ilustra duas sequências de visitação que envolvem os tipos loja e *shopping*. No exemplo, é possível perceber que as duas sequências compartilham o POI estacionamento, café e ponto de ônibus. Desse modo, podemos dizer que loja e *shopping* compartilham certa similaridade pois possuem uma sequência de visitação semelhante.

Por fim, uma terceira linha de pesquisa aponta que utilizar textos relacionados aos POIs podem indicar sua similaridade [69; 80; 91]. Como exemplo, a Figura 1.1.c ilustra dois comentários de usuários de um POI do tipo restaurante e de um POI do tipo café. Nos comentários, é possível notar palavras positivas, indicando que o estabelecimento oferece boa refeição, entre outros. Desse modo, POIs relacionados à palavras similares, podem ser considerados semelhantes.

Embora as recentes pesquisas tenham possibilitado a geração de *embeddings* úteis em diversas tarefas, observou-se que estas não consideram as feições geográficas[1] presentes no contexto dos POIs, tais como rios, edifícios, pontes, entre outros. No entanto, assim como POIs podem ser similares com base em suas relações contextuais de vizinhança, sequência de visitação ou palavras relacionadas, levantou-se a hipótese de que tipos de POI podem ser considerados semelhantes com base nas relações contextuais com as feições geográficas do contexto. Por exemplo, parques possuem feições como árvores, grama e lagos (ilustrado na Figura 1.2). Geralmente, as pessoas frequentam parques para realizar atividades físicas ou apreciar a paisagem. Da mesma forma, POIs do tipo café tendem a apresentar feições mais naturais, como lagos, árvores e grama em sua vizinhança (ilustrado na Figura 1.2). Esse tipo de paisagem no contexto dos POIs café decorre da preferência que muitas pessoas têm de apreciar o ambiente enquanto se alimentam.

Apesar da distinção funcional entre os dois POIs, percebe-se a existência de características geográficas similares dentro de seus contextos. Assim, incorporar as relações contextuais das feições geográficas nos *embeddings* dos tipos de POI pode aprimorar sua representação e melhorar a qualidade dos resultados para diversas tarefas, como a recomendação de POIs. Nesse caso, pessoas que gostam de apreciar o ambiente dos parques também podem gostar do ambiente dos POIs café. Algumas abordagens já fazem uso de feições geográficas nas representações de regiões, bairros ou vizinhanças inteiras de POI [23; 46; 68; 90]. Porém, com o melhor do nosso conhecimento, tais abordagens não utilizam essas informações para gerar *embeddings* de tipos de POI, mas sim de representações mais amplas, como regiões ou zonas urbanas.

Outro ponto observado é que as recentes pesquisas utilizam predominantemente o Word2Vec para aprender as relações contextuais (vizinhança de POI, sequência de POI e palavras relacionadas a POI) [11; 24; 41; 103; 111]. No entanto, elas ignoraram avanços recentes em modelos de PLN, como o BERT [18], que representam o estado da arte em várias tarefas [3; 62; 75]. O modelo BERT oferece a vantagem de capturar relacionamentos complexos e nuances de palavras, permitindo a geração de *embeddings* dinâmicos que se adaptam a diferentes contextos de palavras. Como exemplo, considere a Figura 1.3, que

[1]Disponível em https://support.esri.com/pt-br/gis-dictionary/search?q=feição. Acesso em 20 de maio de 2024.

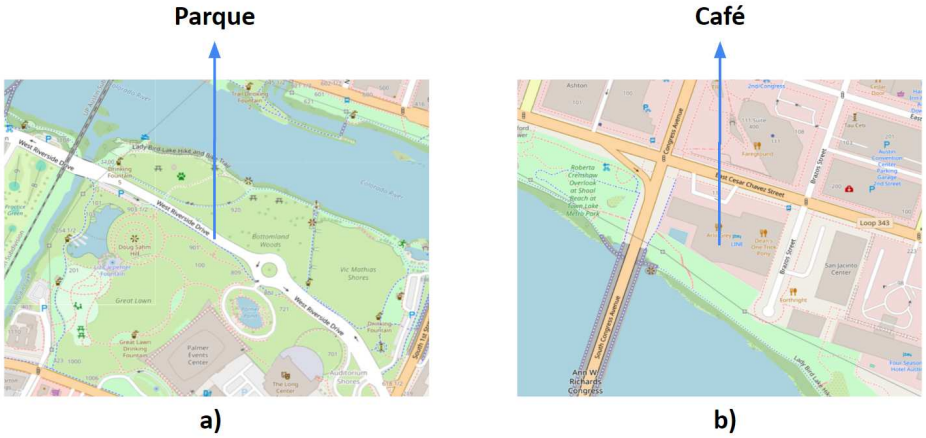


Figura 1.2: Feições geográficas do contexto de um parque (a) e um café (b).
Fonte: Autoria própria

ilustra o contexto geográfico de dois POIs café. O primeiro POI café está em uma área central cercada por prédios e ruas, enquanto o segundo está em uma região mais verdejante perto de um rio. Apesar de seus contextos distintos, métodos tradicionais como o Word2Vec produzem o mesmo *embedding* para ambos os POIs café. No entanto, utilizando um modelo como o BERT, é possível obter dois *embeddings* distintos capazes de refletir essa variação contextual. Esse comportamento pode beneficiar tarefas como a clusterização, permitindo a obtenção de grupos de POIs que levam em conta as variações contextuais.

Figura 1.3: Feições geográficas no contexto do POI café: a) no centro; b) em uma área mais aberta.

Diante deste cenário, este trabalho propõe uma abordagem de geração de *embeddings* de tipos de POI utilizando feições geográficas de seu contexto, como rios, ruas, edificações, entre outros. Além disso, também é proposto o uso dos modelos Word2Vec e BERT para ge-

rar os *embeddings* dos tipos de POI, com o objetivo de comparar o desempenho de modelos tradicionais e recentes de PLN. Com essa abordagem, espera-se obter *embeddings* que capturam as relações contextuais entre os tipos de POI e as feições geográficas. Além disso, tais *embeddings* podem beneficiar diversas tarefas relacionadas, como recomendação de POIs, classificação de zonas urbanas, análise de similaridade de áreas, entre outros.

1.2 Objetivos

Esta pesquisa tem como objetivo desenvolver uma abordagem para a geração de *embeddings* de tipos de POI utilizando feições geográficas do contexto dos POIs. Dessa forma, pretende-se produzir *embeddings* que representem de forma mais precisa os tipos de POIs, levando em consideração as relações contextuais com as feições geográficas.

1.2.1 Objetivos Específicos

Considerando o objetivo principal e o fato de que feições geográficas do contexto dos POIs não são utilizadas para representar seus tipos, este trabalho tem os seguintes objetivos específicos:

- Identificar quais feições geográficas podem ser utilizadas para gerar *embeddings* de tipos de POI;
- Identificar as propriedades relacionadas às feições geográficas e como elas podem ser empregadas para capturar os padrões espaciais;
- Identificar os principais modelos de PLN utilizados para gerar *embeddings* de tipos de POI;
- Desenvolver uma abordagem que utilize as feições geográficas do contexto dos POIs para gerar *embeddings* de seus tipos;
- Avaliar se modelos que utilizam *embeddings* de tipos de POI gerados com feições geográficas são melhores que modelos que utilizam *embeddings* de tipos de POI produzidos com outros dados geográficos.

1.3 Relevância

Na área de Planejamento Urbano, POIs podem ser utilizados para auxiliar nas tomadas de decisões e no entendimento da estrutura das cidades. Planejadores urbanistas podem definir a melhor localização para novas moradias ou serviços considerando os tipos de POI, que fornecem informações sobre os padrões locais. Esses dados ajudam a classificar ou identificar as diversas regiões da cidade e, dessa forma, *embeddings* dos tipos são comumente empregados em tarefas capazes de definir e extrair as funções urbanas. Em geral, a partir dessas representações é possível identificar a função de cada região, como áreas residenciais, comerciais, de maior ou menor fluxo, dentre outros [14; 17; 25; 37; 41; 47; 49; 97; 111].

Na área de sistemas de recomendação, existe a tarefa de prever possíveis POIs nos quais os usuários podem estar interessados. Nessa tarefa, os POIs previamente visitados por usuários são essenciais para execução dos algoritmos de recomendação. Por meio dos *embeddings* dos tipos de POI, é possível definir a similaridade contextual, identificando POIs de possível interesse, e assim prover melhores recomendações [19; 30; 45; 113].

A partir dos exemplos supracitados, é possível notar que POIs estão presentes em diversas aplicações relevantes do cotidiano das pessoas. Nesse sentido, acredita-se que a abordagem presente nesta tese pode beneficiar as aplicações supracitadas e consequentemente seus usuários. No ramo de planejamento urbano os *embeddings* produzidos nesta tese podem dar suporte à classificação de áreas de cidades. Em sistemas de recomendação é possível identificar tipos de POI com contexto geográfico semelhantes aos interesses do usuário. Em sistemas de busca, é possível recuperar resultados que contemplam aspectos geográficos dos POIs candidatos, e assim por diante.

1.4 Contribuições da Pesquisa

As contribuições deste trabalho de pesquisa são as seguintes:

1. **Uma abordagem para geração de *embeddings* de tipos de POI levando em consideração as feições geográficas**: essa abordagem utiliza as propriedades espaciais de

espaço ocupado e *ocorrência* das feições geográficas para permitir que modelos de PLN capturem os padrões espaciais e produzam *embeddings* mais precisos dos tipos de POI. Os resultados dessa contribuição estão detalhados na Seção 4.3;

2. **Os *embeddings* de tipos de POI que incorporam feições geográficas do contexto dos POIs**[2]: os *embeddings* foram treinados com dados da região de Austin no Estados Unidos, que compreende 22.399 POIs distribuídos em uma área geográfica de $704\ km^2$. Estes *embeddings* podem ser utilizados em diversas aplicações, como algoritmos de recomendação, buscas espaciais, planejamento urbano, recuperação de informação geográfica, entre outros. Além disso, a disponibilidade desses *embeddings* permite que outros pesquisadores os utilizem como *baseline*. Os resultados dessa contribuição podem ser encontrados na Seção 4.4.

3. **Uma base de dados pré-processada de feições geográficas** [72]: foi realizado um trabalho de limpeza dos dados, removendo-se informações que não apresentam feições geográficas. Foi realizado um trabalho de associação das feições com POIs de uma segunda base de dados. Essa base permite a recuperação dos tipos de um POI e todas as feições geográficas de seu contexto em um determinado raio em tempo reduzido. Os resultados dessa contribuição estão descritos na Seção 4.2.

4. **Uma revisão bibliográfica sobre a geração de *embeddings* de tipos de POI utilizando várias abordagens e modelos**: nesta revisão, os trabalhos dos últimos anos nessa linha de pesquisa são agrupados em categorias de abordagens. A revisão pode ser encontrada no Capítulo 3.

5. **Os *embeddings* de tipos de POI gerados com as técnicas mais recentes da área de PLN, tais como o BERT**. Esses *embeddings* poderão ser utilizados nas mesmas aplicações que os *embeddings* produzidos com o Word2Vec. Além disso, essa contribuição permitirá que pesquisadores utilizem os novos *embeddings* como *baseline* em pesquisas futuras.

6. **O código fonte do algoritmo GeoContext2Vec**[3]: todo o código da abordagem, capaz

[2]https://drive.google.com/drive/folders/1mjXampBlO0fnfSCNq3F8cdqNZmnFkk1H?usp=sharing
[3]https://github.com/salatielsilva/GeoContext2Vec

de produzir o conjunto de treinamento para geração dos *embeddings* dos tipos de POI, está disponibilizado em um repositório. A partir desse código, outros pesquisadores podem replicar os experimentos e produzir seus próprios *embeddings*.

Esta pesquisa produziu, até então, as seguintes contribuições bibliográficas:

1. Um artigo descrevendo o método de geração de *embeddings* de tipos de POI considerando as feições geográficas, publicado nos anais da conferência *The 38th ACM/SIGAPP Symposium On Applied Computing 2023 - Geographical Information Analytics Track (SAC 23)* [71];

2. Um artigo descrevendo o método de geração de *embeddings* de tipos de POI considerando a distância geográfica entre os POIs em conjunto com sentenças de tamanho variável, publicado na conferência *XXIII Brazilian Symposium on Geoinformatics (GEOINFO) 2022* [70];

3. Um artigo aceito para publicação no *Jounal of Information and Data Management*, que estende os resultados do artigo publicado no GEOINFO. Este artigo contempla mais detalhes técnicos e novos experimentos considerando a distância geográfica entre os POIs e as sentenças de tamanho variável (*In press*);

4. Um artigo submetido ao *International Journal of Geographical Information Science* para avaliação, que compreende os resultados desta pesquisa obtidos até o momento, com foco nas feições geográficas do contexto dos POIs, utilizando modelos tradicionais (Word2Vec) e modelos recentes (BERT) de PLN.

1.5 Metodologia

Para esta tese, foram definidas as seguintes questões de pesquisa (QP):

- **QP**$_1$: Como gerar *embeddings* de tipos de POI utilizando feições geográficas do contexto dos POIs e modelos de PLN?

- **QP**$_2$: *Embeddings* de tipos de POI gerados a partir de relações contextuais com feições geográficas indicam a similaridade dos tipos?

- **QP**$_3$: Modelos que usam *embeddings* de tipos de POI gerados com feições geográfica são melhores que modelos que utilizam *embeddings* de tipos de POI gerados com outros dados geográficos?

- **QP**$_4$: Modelos que utilizam *embeddings* de tipos de POI produzidos com modelos recentes de PLN são melhores que modelos que utilizam *embeddings* de tipos de POI produzidos com modelos clássicos?

Para conduzir a pesquisa e responder as questões elencadas, foram definidas as seguintes atividades:

1. **Investigação da Literatura**: esta atividade tem por objetivo identificar metodologias e técnicas para geração de *embeddings* de tipos de POI. Tal atividade foi realizada durante todo o período da pesquisa, para manter o conhecimento da área sempre atualizado. Para isso, as principais publicações relacionadas à área foram revisadas periodicamente. A execução de tal atividade possibilitou que as questões de pesquisa **QP**$_1$ e **QP**$_4$ fossem contemplada indicando como as feições geográficas podem ser utilizadas, e quais os métodos podem ser aplicados para gerar *embeddings* de tipos de POI;

2. **Aquisição de Dados**: esta atividade envolveu a investigação de bases de dados geográficas para identificar quais informações acerca dos POIs e feições geográficas podem ser obtidas. Tais dados foram utilizados para treinar os modelos de PLN que fornecem os *embeddings* bem como para avaliar a abordagem proposta nesta pesquisa. A realização da referida atividade, permitiu contemplar a questão de pesquisa **QP**$_1$, por meio da qual foi possível definir de maneira detalhada, quais feições geográficas podem ser utilizadas para gerar *embeddings* de tipos de POI;

3. **Implementação da Abordagem**: nesta atividade foi desenvolvida uma abordagem para gerar os *embeddings* de tipos de POI utilizando feições geográficas do contexto e modelos de PLN. Essa atividade partiu dos primeiros resultados após a revisão literária, e consistiu no desenvolvimento, adaptação e melhoria das abordagens do estado-da-arte utilizados na geração de *embeddings* de tipos de POI. A implementação envolveu não apenas a codificação, mas também a realização de experimentos e estudos de caso para identificar problemas e superar os desafios desta pesquisa. A execução desta

tarefa possibilitou contemplar a questão de pesquisa $\mathbf{QP}_1$, pois foi desenvolvido um abordagem para geração de *embeddings* de tipos de POI utilizando feições geográficas;

4. **Validação da Abordagem**: esta tarefa consistiu na realização de experimentos utilizando dados baseados na opinião humana, hierarquia de tipos e uma tarefa de classificação de zonas urbanas. Esses experimentos fornecem evidências sobre a capacidade da abordagem proposta de indicar a similaridade dos tipos de POI com base em suas relações contextuais considerando as feições geográficas. Para isso, foi realizado um comparativo com a opinião humana e com estruturas hierárquicas acerca da similaridade de tipos de POI. Além disso, os *embeddings* produzidos com esta proposta foram empregados em uma tarefa de classificação de zonas, para avaliar o desempenho alcançado. Os resultados dessa tarefa respondem às questões de pesquisa $\mathbf{QP}_2$, $\mathbf{QP}_3$ e $\mathbf{QP}_4$.

5. **Divulgação dos Resultados**: esta tarefa compreendeu a escrita e submissão de artigos científicos com o objetivo de informar e tornar público os resultados para a comunidade através de conferências e periódicos científicos. A tese final foi escrita e publicada, respondendo as questões relacionadas à esta pesquisa.

1.6 Estrutura do Documento

Os capítulos restantes que compõem este documento estão estruturados da seguinte forma:

Capítulo 2: Fundamentação Teórica. Apresentam-se os conceitos gerais necessários para entendimento desta pesquisa, que servem para dar embasamento teórico aos leitores.

Capítulo 3: Trabalhos Relacionados. Discutem-se os trabalhos relacionados na área de geração de *embeddings* de tipos de POI.

Capítulo 4: Geração de *Embeddings* de Tipos de POI. Apresenta-se a abordagem proposta com detalhes sobre como as feições geográficas e os modelos de PLN são utilizados para gerar *embeddings* de tipos de POI.

Capítulo 5: Configuração Experimental. Apresentam-se os dados utilizados, as atividades avaliativas e a configuração de parâmetros da abordagem proposta e dos *baselines*.

Capítulo 6: Resultados. Apresentam-se os resultados obtidos em cada tarefa realizada, bem como uma análise visual dos *embeddings* produzidos.

Capítulo 7: Estudo de Caso. Apresenta-se um estudo de caso utilizando os *embeddings* das feições geográficas em uma tarefa de busca de POIs para demonstrar o benefício das feições.

Capítulo 8: Conclusão. Apresenta-se um resumo das questões de pesquisa respondidas, sumarizando os resultados obtidos e indicando possibilidades futuras de pesquisa.

Capítulo 2

Fundamentação Teórica

Este capítulo tem como objetivo fornecer um resumo dos conceitos básicos necessários para a compreensão desta pesquisa. Inicialmente, apresenta-se o conceito de feição geográfica. Em seguida, são explorados alguns métodos relacionados à representação de palavras por meio de *word embeddings*. Mais adiante, são abordadas algumas medidas utilizadas para calcular a similaridade entre tipos de POI, considerando estruturas hierárquicas e representações vetoriais. Posteriormente, são discutidos alguns métodos empregados para analisar correlação entre variáveis e similaridade de ranques. Por fim, apresentam-se as considerações finais do capítulo.

2.1 Feição Geográfica

De acordo com o Dicionário GIS de Suporte da Esri[1], uma feição geográfica é um objeto do mundo real que pode ser representado em um mapa, tais como estradas, rios, prédios, sinais de trânsito, entre outros. As feições possuem uma geometria, como ponto, linha ou polígono, e podem apresentar atributos relacionados. Por exemplo, os atributos de um rio podem incluir seu nome, comprimento e carga de sedimento em uma estação de medição. A Figura 2.1 ilustra um contexto composto por diversas feições geográficas, como ruas, prédios, estacionamentos, entre outros.

Nesta tese, o foco é mantido em feições geográficas do contexto dos POIs, com ênfase nas

[1]Disponível em https://support.esri.com/pt-br/gis-dictionary/search?q=feição. Acesso em 20 de maio de 2024.

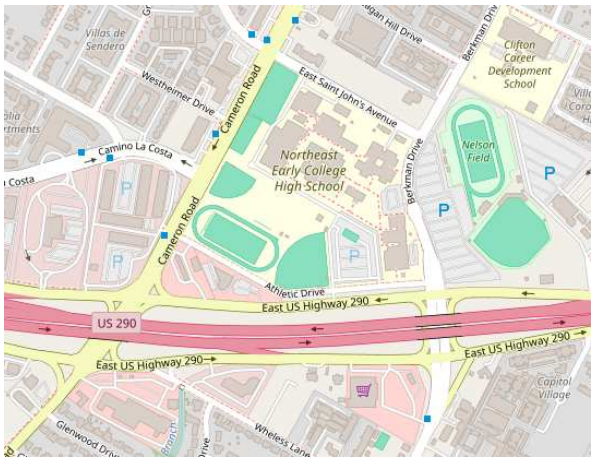


Figura 2.1: Exemplo de feições geográficas em um mapa.

Fonte: Autoria própria

que podem ser visualmente percebidas, como prédios, ruas, postes, calçadas e representados em bases de dados geográficas como o OpenStreetMap(OSM)[2] e Google Maps[3].

2.2 *Word Embeddings*

Word embeddings são representações numéricas de palavras em um espaço vetorial de alta dimensão. Elas são utilizadas para capturar a relação contextual das palavras em textos. Nessa técnica, cada palavra é mapeada para um vetor de valores reais que são estimados por meio de redes neurais e, portanto, a técnica é muitas vezes colocada no campo de *deep learning*. A ideia chave é utilizar uma representação densa para cada palavra em que cada dimensão do vetor representa uma característica contextual da palavra. Isto contrasta com os milhares ou milhões de dimensões necessárias para representações esparsas. Quando produzidos, os *word embeddings* podem ser usados em diversas tarefas de PLN, como classificação de texto, agrupamento de documentos e geração de texto [22].

Uma das principais abordagens para obtenção de *word embeddings* é o uso de redes neurais treinadas para prever a próxima palavra em um contexto, a partir de grandes massas de dados não tratados. Essas representações são aprendidas de forma não supervisionada, o que significa que elas são geradas a partir do próprio corpus de texto, sem a necessidade de

[2]Disponível em https://www.openstreetmap.org. Acesso em 20 de maio de 2024.

[3]Disponível em https://www.google.com/maps. Acesso em 20 de maio de 2024.

etiquetas ou rótulos.

Os *word embeddings* têm mostrado resultados promissores em diversas tarefas de PLN. Entre as principais abordagens para gerar *word embeddings* pode-se destacar o Word2Vec [48] e as recentes abordagens baseadas em *transformers*, tais como o BERT [18]. Tais abordagens são explicadas nas subseções seguintes.

2.2.1 *Word2Vec*

Introduzido em 2013 por Mikolov *et al.* [48], o Word2Vec é um algoritmo de PLN que codifica as palavras em vetores densos (*vector embeddings*). Os vetores gerados são capazes de capturar as relações contextuais das palavras em qualquer corpus analisando a vizinhança de uma palavra central.

O Word2Vec utiliza duas arquiteturas principais: *Continuous Bag-of-Words* (CBOW) e *Skip-Gram*, conforme ilustrado na Figura 2.2. A arquitetura CBOW tenta prever a palavra central a partir das palavras vizinhas, enquanto a arquitetura Skip-Gram tenta prever as palavras vizinhas a partir da palavra central. Nesta tese, o foco será a arquitetura Skip-Gram, por ser a arquitetura mais empregada na área de geração de *embeddings* de tipos de POI (mais detalhes estão no Capítulo 3).

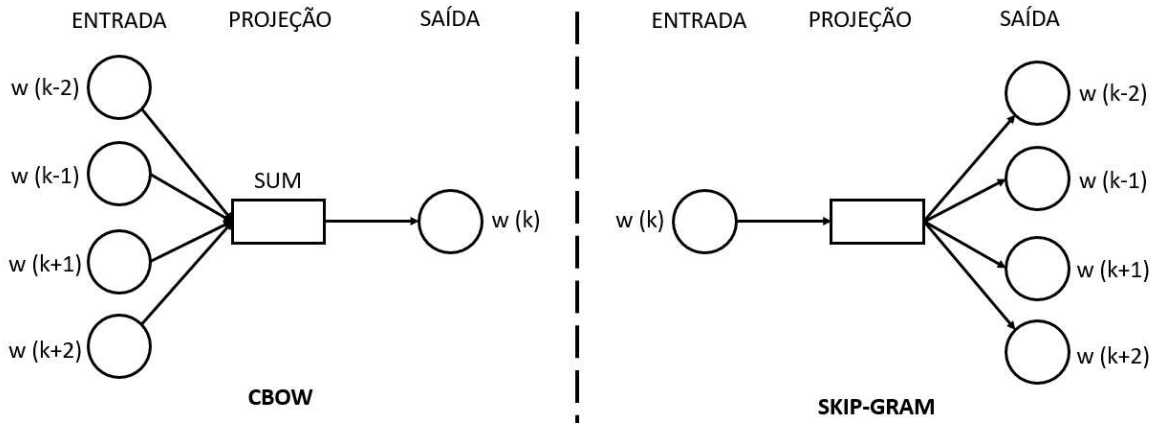


Figura 2.2: Arquiteturas do Word2Vec: CBOW e SKIP-GRAM

Fonte: Autoria própria

Como exemplo, considere a seguinte frase: "Futebol é um esporte muito popular". A Tabela 2.1 sumariza como ocorre a geração do conjunto de treinamento para o Word2Vec a partir dessa frase. Para definir as palavras do contexto, uma janela de tamanho k é deslocada

ao longo da frase (no exemplo $k = 2$). Para cada palavra analisada (em negrito), serão consideradas como palavras de contexto as k palavras predecessoras e as k palavras sucessoras (palavras dentro dos colchetes). A partir do exemplo, é possível notar que a primeira e última palavra da frase, apresentam menos palavras de contexto, e podem influenciar o resultado do modelo.

Frase	Skip-grams
Futebol é um] esporte muito popular	(**Futebol**, é), (**Futebol**, um)
Futebol **é** um esporte] muito popular	(**é**, Futebol), (**é**, um), (**é**, esporte)
[Futebol é **um** esporte muito] popular	(**um**, Futebol), (**um**, é), (**um**, esporte), (**um**, muito)
Futebol [é um **esporte** muito popular]	(**esporte**, é), (**esporte**, um), (**esporte**, muito),(**esporte**, popular)
Futebol é [um esporte **muito** popular	(**muito**, um), (**muito**, esporte), (**muito**, popular)
Futebol é um [esporte muito **popular**	(**popular**, esporte), (**popular**, muito)

Tabela 2.1: Geração de dados de treinamento do Word2Vec Skip-Gram utilizando uma janela de tamanho 2.

Depois que o conjunto de treinamento é gerado, a arquitetura *Skip-Gram* utiliza uma camada oculta (*projeção*) para aprender representações vetoriais densas das palavras, sendo treinada para maximizar a probabilidade de ocorrência das palavras vizinhas dada a palavra central. Durante o treinamento, o *Skip-Gram* usa um par de palavras por vez.

Formalmente, seja um vocabulário V e um conjunto de treinamento W que contém sequências de palavras $w_1, w_2, ..., w_T$, a função objetivo do *Skip-Gram* é maximar a probabilidade de predizer as palavras vizinhas dada a palavra central conforme a Equação 2.1:

$$Pr(w_{n-c}, ..., w_{n+c} \mid w_n) = \frac{1}{T} \sum_{n=1}^{T} \sum_{-c \leq m \leq c} \log p(w_{n+m} \mid w_n) \tag{2.1}$$

em que c é o tamanho de uma janela que define a quantidade de palavras do contexto, T é a quantidade de palavras no conjunto de treinamento, w_n é a palavra central, w_{n+m} $(-c \leq m \leq c)$ é uma palavra de contexto, e o logaritmo empregado tem base natural (base e). O termo

$p(w_{n+m} \mid w_n)$ define a probabilidade da palavra de contexto w_{n+m} ocorrer dada a palavra central w_n. Ela é formulada usando a função softmax conforme demonstra a Equação 2.2:

$$p(w_{n+m} \mid w_n) = \frac{exp(u_{w_{n+m}}^{\intercal} v_{w_n})}{\sum_{m=1}^{W} exp(u_{w_{n+m}}^{\intercal} v_{w_n})} \tag{2.2}$$

em que v_{w_n} e $u_{w_{n+m}}$ são os vetores da palavra central w_n e palavra de contexto w_{n+m}, respectivamente.

Uma das principais vantagens do Word2Vec é que ele é capaz de capturar relações contextuais entre palavras, o que permite que ele seja utilizado para várias tarefas de PLN como resposta a perguntas, recuperação de informações, tradução automática, modelagem de linguagem, entre outros [2; 15; 99; 102]. Além disso, o Word2Vec é computacionalmente eficiente e capaz de lidar com grandes conjuntos de dados de texto.

2.2.2 Transformers

Transformer é uma arquitetura de modelo de aprendizagem profunda que revolucionou o campo de PLN desde sua introdução em 2017 [83]. Aprendizagem profunda é um subconjunto da área de Aprendizado de Máquina (*Machine Learning*), que se baseia em redes neurais artificiais com múltiplas camadas [36]. A arquitetura do *transformer* foi projetada para resolver problemas de natureza *sequence-to-sequence*, ou seja, recebe uma sequência de entrada e a transforma em uma sequência de saída. Diferentemente das arquiteturas tradicionais baseadas em Redes Neurais Recorrentes (RNNs) ou Convolucionais (CNNs), o *transformer* se destaca por sua capacidade de processar sequências de entrada de comprimento variável de forma altamente paralela, resultando em um treinamento mais rápido e eficiente.

Além disso, a arquitetura do *transformer* é composta por dois componentes principais: o codificador e o decodificador. Cada um desses componentes é composto por várias camadas, que por sua vez são compostas por subcamadas de atenção e redes neurais totalmente conectadas (também conhecidas como camadas de *feedforward*).

A atenção é o componente fundamental do *transformer*, pois permite que o modelo foque em partes específicas da sequência de entrada enquanto realiza operações em paralelo. Durante o treinamento, os pesos de atenção são calculados para indicar a importância relativa de

cada palavra em relação a todas as outras palavras na sequência de entrada. Esses pesos são então usados para ponderar as representações das palavras antes de combiná-las para formar uma representação contextualizada da entrada.

2.2.2.1 BERT

O BERT (*Bidirectional Encoder Representations from Transformers*) é um modelo de aprendizado de máquina de aprendizagem profunda baseado em *transformers* que foi proposto pela Google em 2018 [18]. O BERT é um modelo geral de linguagem que se adapta a vários problemas de PLN [88]. Sua arquitetura é composta por uma pilha de codificadores. Conforme ilustra a Figura 2.3, cada codificador é composto por múltiplas camadas de subcomponentes, incluindo camadas de atenção e camadas de redes neurais totalmente conectadas (*feedforward*). Essas camadas são empilhadas para formar uma arquitetura profunda que pode capturar informações contextuais em diferentes níveis de abstração.

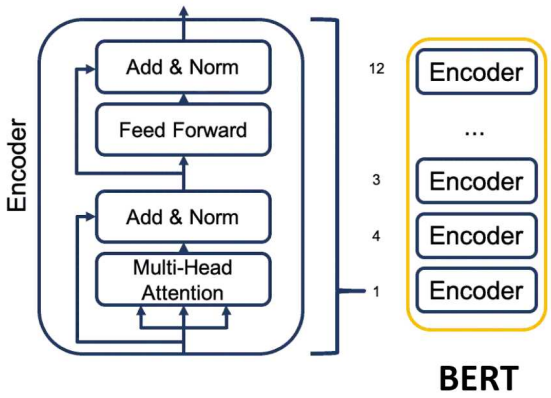


Figura 2.3: Arquitetura do BERT.

Fonte: Autoria própria

Uma das principais diferenças entre BERT e outros modelos é que esse modelo é bidirecional, ou seja, ele considera o contexto tanto à esquerda quanto à direita de cada palavra ao codificar o texto. Isso permite que o modelo capture relações mais complexas entre palavras e melhore a compreensão do texto. Além disso, modelos como o Word2Vec são treinados em uma tarefa de aprendizado não supervisionado, onde a rede aprende a prever palavras próximas em um contexto, enquanto o BERT é treinado em tarefas supervisionadas e semi-supervisionadas, como prever palavras mascaradas em uma sentença ou prever a próxima

sentença em um par de sentenças.

O treinamento do BERT se baseia em duas estratégias denominadas de modelo de linguagem mascarado (MLM do inglês *Masked Language Modeling*) e predição da próxima sentença (NSP do inglês *Next Sentence Prediction*). No modo MLM, antes de alimentar o BERT com sequências de palavras, 15% das palavras de cada sequência são substituídas por um símbolo [MASK], que funciona como uma máscara [18]. O modelo então tenta prever a palavra original entre as palavras mascaradas, com base no contexto fornecido pelas outras palavras não mascaradas. A Figura 2.4 ilustra um texto com uma sequência de palavras mascaradas.

```
[CLS] O Brasil é o país mais [MASK] da América do Sul.
Seu território abrange uma área de aproximadamente 8,5
milhões de [MASK], sendo o quinto maior país do mundo
em área territorial. Possui uma população de mais de
210 milhões de [MASK], tornando-se o sexto país mais
populoso do mundo. A capital do Brasil é [MASK],
enquanto a maior cidade é São Paulo. O Brasil é
conhecido por sua rica [MASK] cultural, belas praias e
diversidade natural. [SEP]
```

Figura 2.4: Texto mascarado para treinamento no BERT.

Fonte: Autoria própria

No exemplo da Figura 2.4, a tag [CLS] é uma marca especial que indica o início de uma sequência de *tokens* em um texto de entrada. A tag [SEP] é utilizada para informar ao modelo que a primeira sequência terminou e que a segunda sequência está prestes a começar.

No modo NSP, o modelo recebe pares de frases como entrada e aprende a prever se a segunda frase do par é subsequente da primeira. Durante o treinamento, 50% das entradas são formadas por pares subsequentes, enquanto que os outros 50% são formados por pares de frases aleatórias do corpus. A suposição é que a sentença aleatória será desconectada da primeira sentença. A Figura 2.5 ilustra um texto de entrada do treinamento do BERT.

O BERT foi treinado em uma grande quantidade de dados textuais, permitindo a generalização para novos conjuntos de dados. Ele também foi treinado em várias tarefas de PLN, incluindo análise de sentimentos, classificação de texto e resposta a perguntas, permitindo que ele seja facilmente adaptado para novas tarefas [3; 62; 75]. Além disso, diferentemente do Word2Vec que é conhecido como modelo *static word embeddings* (representam cada pa-

```
[CLS] O Brasil é o país mais populoso da América do
Sul. [SEP] Seu território abrange uma área de
aproximadamente 8,5 milhões de habitantes, sendo o
quinto maior país do mundo em área territorial. [SEP]
Possui uma população de mais de 210 milhões de
habitantes, tornando-se o sexto país mais populoso do
mundo. [SEP] A capital do Brasil é Brasília, enquanto a
maior cidade é São Paulo. [SEP]
```

Figura 2.5: Texto para treinamento no BERT usando NSP.

Fonte: Autoria própria

lavra por um único *word embedding* independentemente do contexto em que ela ocorre), o BERT considera o contexto em que a palavra ocorre para representá-la, sendo assim um modelo *contextualized word representations*. Em outras palavras, a mesma palavra em diferentes contextos possuirá diferentes *embeddings*.

2.3 Medidas de Similaridade

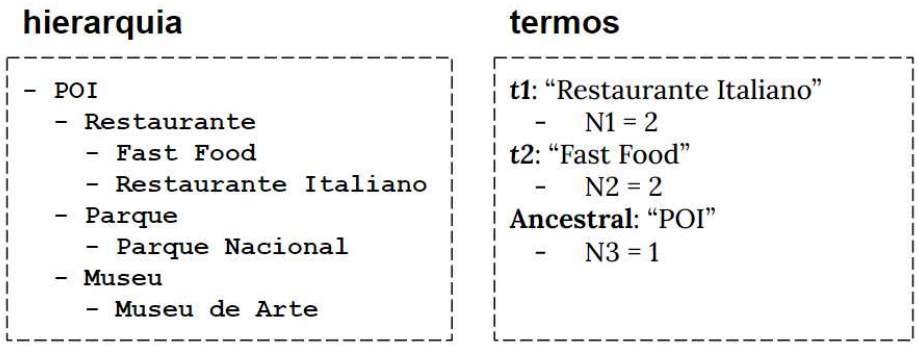
2.3.1 Similaridade Baseada no Caminho e Profundidade

Medidas de similaridade quantificam o quão similares dois termos são com base em informações contidas em uma hierarquia (estrutura que armazena informações do tipo A é um B por exemplo). Um automóvel pode ser considerado mais similar a um barco do que a uma árvore, devido o automóvel e o barco compartilharem o tipo veículo como um ancestral comum em uma hierarquia. Nesse contexto, Wu e Palmer [96] propuseram um método clássico baseado no caminho e na profundidade de uma hierarquia. Esse método determina a proximidade entre dois termos t_1 e t_2 conforme a Equação 2.3.

$$SIM_{WP}(t_1, t_2) = \frac{2 * N_3}{N_1 + N_2} \tag{2.3}$$

onde N_1 e N_2 representam as profundidades dos termos t_1 e t_2 na hierarquia, respectivamente. N_3 é a profundidade do termo mais baixo que é um ancestral comum a ambos os termos t_1 e t_2. Os valores que essa similaridade pode assumir estão no intervalo $0 \leq \textit{similaridade} \leq 1$.

Para ilustrar, considere a hierarquia de tipos de POI mostrada na Figura 2.6. Considerando os tipos `Restaurante Italiano` e `Fast Food`, é possível observar que o termo `Restaurante` é o ancestral comum mais próximo na hierarquia. Ao calcular as profundidades de cada um dos termos e substituí-los na equação, obtém-se uma similaridade de $0,50$.



$$SIM_{WP}(t_1, t_2) = \frac{2 * 1}{2 + 2} = \frac{2}{4} = 0.50$$

Figura 2.6: Exemplo da similaridade de Wu & Palmer.

Fonte: Autoria própria

Leacock & Chodorow [35] também formularam um método de avaliação considerando o caminho entre os termos t_1 e t_2 e a profundidade da taxonomia. Esse método é definido da seguinte forma:

$$SIM_{LC}(t_1, t_2) = -log\left(\frac{N}{2D}\right) \tag{2.4}$$

em que D é o grau máximo da taxonomia e N é o caminho mais curto entre os tipos t_1 e t_2. Considerando o mesmo exemplo ilustrado na Figura 2.6, pode-se calcular $N = 1$, pois os dois tipos são descendentes imediatos do mesmo nó (`Restaurante`). $D = 2$ representa a profundidade do nó mais interno na hierarquia. Portanto, a similaridade é calculada como: $SIM_{LC}(t_1, t_2) = -log\left(\frac{1}{2*2}\right) = -log\left(\frac{1}{4}\right) = 0,602$.

Nesta tese, as métricas mencionadas são empregadas para avaliar se as relações contextuais capturadas pelos *embeddings* se assemelham à relação hierárquica dos tipos de POI. Como exemplo, pretende-se verificar se dois tipos de POI, que são similares conforme a hierarquia, também apresentam similaridade contextual em relação às feições geográficas.

Essas métricas são empregadas em tarefas descritas no Capítulo 6.

2.3.2 Similaridade do Cosseno

A similaridade do cosseno é utilizada para calcular a similaridade entre dois vetores t e e no espaço vetorial conforme a Figura 2.7. Ela é baseada no cálculo do cosseno do ângulo entre os dois vetores, onde um ângulo de 0 graus entre t e e indica que os vetores são completamente similares, ou seja, apontam na mesma direção, e um ângulo de 90 graus indica que os vetores são completamente dissimilares [95]. A Equação 2.5 ilustra o cálculo da similaridade entre t e e:

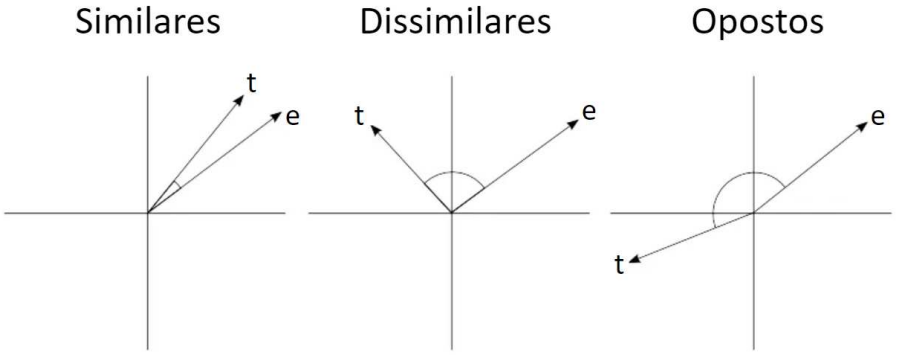


Figura 2.7: Exemplo da similaridade dos vetores t e e com base no ângulo estabelecido entre eles.

Fonte: Adaptado de pyimagesearch[4]

$$\cos(\mathbf{t}, \mathbf{e}) = \frac{\mathbf{t} \cdot \mathbf{e}}{\|\mathbf{t}\|\|\mathbf{e}\|} = \frac{\sum_{i=1}^{n} \mathbf{t}_i \mathbf{e}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{t}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{e}_i)^2}} \tag{2.5}$$

em que t_i e e_i são componentes dos vetores t e e respectivamente. O cálculo resultante varia de -1 significando exatamente o oposto, a 1 significando exatamente o mesmo, com o valor 0 indicando ortogonalidade ou decorrelação. Assim, termos relacionados se aproximarão de 1, termos opostos se aproximarão de -1, e termos não-relacionados terão valor 0.

A similaridade do cosseno também é frequentemente utilizada para calcular a similaridade entre *word embeddings* devido às suas propriedades [24; 41; 105; 106; 111]. Essa medida avalia a semelhança entre vetores de palavras com base na direção, ao invés da distância entre eles no espaço vetorial. Essa abordagem é crucial, uma vez que palavras com

relações contextuais semelhantes tendem a estar na mesma direção ou em direções similares nos espaços vetoriais de *embeddings* de palavras [18; 48; 56].

A robustez da similaridade do cosseno em espaços vetoriais de alta dimensionalidade também é um ponto positivo, visto que os *embeddings* de palavras gerados por modelos de linguagem, como Word2Vec, GloVe ou BERT, frequentemente possuem centenas ou até milhares de dimensões. Mesmo em espaços tão complexos, a similaridade do cosseno mantém sua eficácia como medida de similaridade entre vetores de palavras. Adicionalmente, o cálculo da similaridade do cosseno é computacionalmente eficiente e rápido, o que o torna uma escolha prática para análises em larga escala de conjuntos de dados de *embeddings* de palavras. Portanto, nesta tese foi utilizada a similaridade do cosseno para analisar a similaridade contextual dos embeddings dos tipos de POI.

2.4 Correlação de *Spearman*

O coeficiente de correlação de *Spearman*, também chamado de ρ (rho) de *Spearman*, é um método para avaliar a relação monotônica entre duas variáveis quantitativas. Ele é baseado na ideia de que, se duas variáveis estão relacionadas, então as posições relativas dos seus valores devem ser semelhantes. De maneira simplificada, esse método indica que quando o valor de uma variável aumenta ou diminui, o valor da outra variável aumenta ou diminui [73]. A Figura 2.8 demonstra um exemplo de variáveis e a relação existente entre elas.

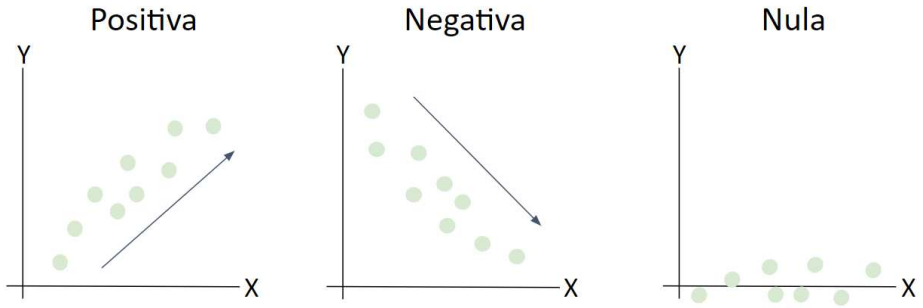


Figura 2.8: Exemplo de correlação de duas variáveis X e Y.

Fonte: Autoria própria

O coeficiente de correlação de *Spearman* é dado pela Equação 2.6:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{2.6}$$

em que d_i é a diferença entre cada observação das variáveis, n é a quantidade de observações. O valor de ρ pode variar entre $+1$ e -1 onde:

- Um valor de $+1$ indica uma associação de classificação perfeita entre as variáveis;
- Um valor de -1 indica uma associação negativa perfeita entre as variáveis;
- Valor de 0 indica que não há associação de classificação.

Nesta tese, a correlação de *Spearman* é utilizada para investigar se existe alguma correlação entre os valores de similaridade do cosseno obtidos dos *embeddings* dos tipos de POI e os valores de similaridade indicados pela opinião humana em uma tarefa de análise de similaridade (a descrição da análise está no Capítulo 6). Optou-se por utilizar a correlação de *Spearman* devido à sua menor sensibilidade aos valores em comparação com a correlação de *Pearson*, que depende mais dos valores fornecidos pela similaridade do cosseno. Além disso, o foco da tese é investigar a existência da correlação, não se atentando à outras possibilidades de correlação (linear, logarítmica, entre outros).

2.5 *Mean Reciprocal Rank*

O *Mean Reciprocal Rank* (MRR) é uma métrica utilizada para avaliar o ranqueamento de respostas de consultas. Ela calcula a média dos inversos das posições das primeiras respostas relevantes encontradas para um conjunto de consultas. Essa métrica é baseada no *Reciprocal Rank* [63], que é o inverso da posição da primeira resposta correta para uma consulta, ou seja, $1/1$ para a primeira resposta correta, $1/2$ para a segunda resposta correta, $1/3$ para a terceira resposta correta e assim por diante. O MRR é a média dos valores do *Reciprocal Rank* para todas as consultas. A Figura 2.9 ilustra um exemplo do cálculo do MRR.

A Equação 2.7 ilustra o cálculo do MRR:

$$MRR = \frac{1}{Q} \sum_{i=1}^{Q} \frac{1}{rank_i} \tag{2.7}$$

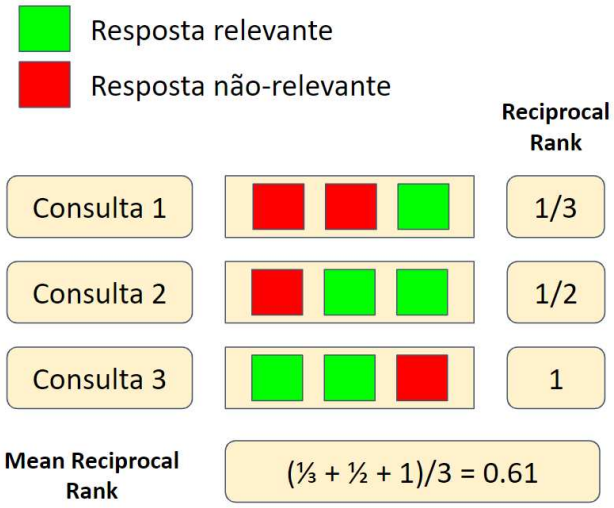


Figura 2.9: Exemplo do cálculo do MRR.

Fonte: Autoria própria

em que o $rank_i$ refere-se à posição da primeira resposta correta para a i-ésima consulta Q.

O MRR é frequentemente utilizado como uma métrica de desempenho para sistemas de recuperação de informação, pois ele leva em consideração não apenas a presença de documentos relevantes nas primeiras posições, mas também a ordem deles. Além disso, é uma métrica fácil de entender e interpretar, e menos suscetível a distorções causadas por consultas com poucos ou muitos documentos relevantes.

Nesta tese, o MRR é utilizado para comparar um *ranking* gerado a partir da similaridade do cosseno dos *embeddings* de tipos de POI com um *ranking* gerado a partir a hierarquia dos tipos e das similaridades SIM_{WP}, SIM_{LC}. Essa análise visa identificar se os tipos de POI que são similares conforme a hierarquia também apresentam similaridade contextual em relação às feições geográficas. Mais detalhes dessa análise estão disponíveis no Capítulo 6.

2.6 Considerações Finais

Este capítulo apresentou a terminologia básica e os principais conceitos necessários para o entendimento desta pesquisa. O escopo desta pesquisa compreende a representação de tipos de POI por meio de *embeddings* considerando as feições geográficas do contexto dos POIs.

Também foi discorrido sobre o termo *word embeddings*, proveniente da área de PLN, que engloba uma variedade de métodos capazes de gerar esse tipo de representação. Este capítulo

também introduziu métodos para calcular essa similaridade, considerando o conhecimento modelado em hierarquias e representações vetoriais. O próximo capítulo abordará uma série de trabalhos relacionados à representação de POIs e seus tipos, utilizando abordagens e conjuntos de dados diversos.

Capítulo 3

Trabalhos Relacionados

A representação computacional dos tipos de POI é fundamental para diversas tarefas, incluindo recomendação, planejamento urbano e recuperação de informações. Vários estudos têm proposto abordagens e recursos para representar POIs e seus tipos, empregando informações de vizinhança ou sequência de visitação, dados textuais georreferenciados e dados geográficos. Para identificar os principais trabalhos nesta área, foram utilizados periódicos GIS (GIS do inglês *Geographic Information Systems*), como o IJGIS[1], Geoinformática[2], Computers, Environment and Urban Systems[3] e o TRANSACTIONS IN GIS[4], além de conferências relevantes, como o SIGSPATIAL[5], COSIT[6] e o GIScience[7]. Além disso, o Google Scholar foi empregado como mecanismo de busca, utilizando a seguinte *string*: "("POI embeddings"AND ("category"OR "type")) OR ("POI type embeddings") OR ("POI category embeddings")".

Nesse capítulo, as Seções 3.1 e 3.2 descrevem trabalhos que utilizam informações de vizinhança ou sequências de POI para gerar *embeddings*. A Seção 3.3 lista trabalhos que utilizam dados textuais georreferenciados para gerar *embeddings* de POIs ou regiões completas. A Seção 3.4 descreve como informações geográficas são utilizadas em representações mais

[1]Disponível em https://www.tandfonline.com/journals/tgis20. Acesso em 20 de maio de 2024.

[2]Disponível em https://link.springer.com/journal/10707. Acesso em 20 de maio de 2024.

[3]Disponível em https://www.sciencedirect.com/journal/computers-environment-and-urban-systems. Acesso em 20 de maio de 2024.

[4]Disponível em https://onlinelibrary.wiley.com/journal/14679671. Acesso em 20 de maio de 2024.

[5]Disponível em https://sigspatial2023.sigspatial.org. Acesso em 20 de maio de 2024.

[6]Disponível em https://geosensor.net/cositseries/. Acesso em 20 de maio de 2024.

[7]Disponível em https://giscience2023.github.io. Acesso em 20 de maio de 2024.

gerais. A Seção 3.5 discorre sobre como os métodos mais recentes de PLN são aplicados para enriquecer a representação de POIs em modelos de linguagem de propósito geral. A Seção 3.6 discute o posicionamento desta pesquisa em relação aos trabalhos relacionados. Finalmente, a Seção 3.7 apresenta as considerações finais do capítulo.

3.1 *Embeddings* Baseados na Vizinhança de POIs

Trabalhos nesse ramo propõem utilizar relações binárias entre todos os POIs dentro de uma vizinhança. Geralmente, são gerados pares da forma ⟨`tipo de POI central, tipo de POI de contexto`⟩ para todas as combinações de tipos do POI central e de todos os POIs de contexto. Esses pares são então utilizados para treinar um modelo de inteligência artificial, tipicamente o Word2Vec, em uma tarefa de predição: prever o tipo de POI de contexto dado o tipo de POI central. Como resultado, tipos de POI com relações binárias semelhantes terão seus *embeddings* situados próximos no espaço latente.

Entre os trabalhos que utilizam a vizinhança dos POIs para gerar *embeddings* dos tipos, Yan *et al*. [103] estão entre os pioneiros. Os autores propõem o algoritmo *Information Theoretic, Distance Lagged* (ITDL) com o objetivo de capturar a relação e similaridade dos tipos. Para isso, a vizinhança contínua é dividida em várias caixas discretas, que são faixas de espaço em torno de um POI. A Figura 3.1 ilustra esta configuração. Cada caixa representa um contexto centrado em um POI (denominado POI central), e é utilizada para codificar a relação contextual entre o POI central e seus vizinhos (denominados POIs de contexto). A partir da relação de vizinhança entre o POI central e os POIs de contexto, relações binárias do tipo ⟨`tipo` l `do POI central, tipo` j `do POI de contexto`⟩ são geradas, em que $l \in L$ e $j \in J$, sendo L e J os conjuntos de todos os tipos do POI central e do POI de contexto.

Para capturar com mais precisão a "importância" dos tipos na vizinhança, dois elementos são considerados: a *popularidade*, calculada a partir da contagem de *check-ins* nos POIs, e a *unicidade* dos tipos, calculada a partir da probabilidade de um tipo de POI ocorrer em determinada caixa discreta. Esses elementos são combinados para calcular um fator numérico que é utilizado para replicar a relação ⟨`tipo POI central, tipo POI contexto`⟩ no conjunto de treinamento do Word2vec. Assim, tipos que são mais raros ou mais populares

Figura 3.1: Vista das caixas discretas na vizinhança de um POI central.

Fonte: Autoria própria

terão a relação binária com os tipos do POI central replicadas no conjunto de treinamento.

O conjunto de relações binárias é então utilizado como entrada no modelo Word2Vec. Durante o treinamento, o Word2Vec estima a probabilidade de ocorrência de um tipo de POI de contexto, dado o tipo de POI central. Assim, o modelo aprende as relações contextuais de cada tipo a partir das coocorrências com outros tipos em uma caixa. Ao final, o modelo permite a obtenção dos *embeddings* para cada tipo de POI, em que tipos com padrões de coocorrência semelhantes terão vetores próximos no espaço vetorial latente.

De maneira semelhante, Liu *et al*. [42] utilizam a coocorrência em uma vizinhança contínua para gerar uma representação abrangente da vizinhança, denominada nicho. Nesse método, o Word2Vec é empregado para aprender a representação latente do nicho. Além disso, esse método produz como saída as probabilidades de cada tipo pertencer a um nicho. As probabilidades do contexto são geradas minimizando a diferença entre as predições do modelo e as relações binárias de entrada do modelo. Nessa abordagem, uma adaptação do Word2Vec é utilizada para representar o nicho considerando a cidade a qual ele pertence. Desse modo, pode-se averiguar quais nichos são semelhantes entre diferentes cidades.

Jin *et al*. [33] apresentam um modelo para geração de *embeddings* de tipos utilizando o método geohash [51]. O geohash é empregado para definir o contexto espacial dos POIs. Em seguida, a coocorrência dos POIs do contexto é utilizada para produzir o conjunto de treinamento de um modelo Word2Vec e consequentemente os *embeddings* dos tipos.

Em seu trabalho Zhu *et al*. [117] utilizam *embeddings* de POI para representar ROIs (Regiões de Interesse). Esses *embeddings* são gerados utilizando o modelo Word2Vec em

conjunto com um corpus de POIs produzidos a partir da vizinhança de POIs em cada ROI. Os *embedding* dos POIs são utilizados para representar cada ROI. Os vetores das ROIs são então utilizados para calcular a similaridade entre as palavras-chave de uma consulta e as ROIs candidatas. Diferentes padrões de consulta são projetados para medir a similaridade e encontrar as k ROIs candidatas mais relevantes para a consulta.

Wang e Moosavi [90] propõem o Urban2Vec, um método semelhante ao ITDL, porém empregando o algoritmo MeanShift [10] para definir a borda da vizinhança dos POIs através de agrupamentos baseados na densidade dos POIs. A partir disso, áreas mais concentradas apresentarão uma vizinhança menor e áreas mais dispersas (subúrbios, áreas rurais) possuirão vizinhanças maiores. Como resultado do agrupamento, os *embeddings* dos tipos incorporam a multiescalabilidade dos POIs em diferentes ambientes. Nesse método, a coocorrência também é aplicada e os *embeddings* são gerados a partir do Word2Vec.

Em sua proposta, Liu *et al*. [41] fornecem um *framework* para visualizar e explorar POIs. Nessa abordagem, *embeddings* de tipos de POI são gerados utilizando o Word2Vec e as coocorrências dos POIs em uma vizinhança. Além disso, utilizando uma técnica de redução de dimensão (isto é, t-SNE), todos os *embeddings* dos tipos são mapeados para o espaço vetorial bidimensional, onde tipos relacionados ou similares estão mais próximos uns dos outros. Finalmente, tomando o espaço vetorial dos tipos como mapa base, a configuração de POIs em cada região é apresentada como um mapa temático. Por meio desse *framework*, a configuração de uma cidade pode ser visualizada e utilizada para compreender e comparar as regiões da cidade.

Zhou *et al*. [115] fornecem um *framework* denominado HREP (*Heterogeneous Region Embedding with Prompt Learning*), que se concentra na geração de *embeddings* de regiões através da utilização de informações heterogêneas, incluindo dados de POI. Os autores utilizam um grafo de região heterogêneo, abrangendo contextos inter-região, como mobilidade humana e vizinhança geográfica, e contextos intra-região, como informações de POI. Os modelos empregados para gerar os *embeddings* baseiam-se em grafos sensíveis a relações e autoatenção para capturar correlações globais entre regiões. Além disso, os *embeddings* das regiões são enriquecidos por um módulo de fusão baseado em atenção e são utilizados em tarefas *downstream* através do aprendizado de *prompt*, especificamente a técnica de *prefix-tuning*.

3.2 *Embeddings* Baseados na Sequência de POIs

Trabalhos nesse ramo se concentram em sequências que conectam POIs. Nesse contexto, o Word2Vec geralmente é empregado para prever o próximo POI na sequência dado um POI atual. Isso permite que o modelo aprenda as relações sequenciais entre POIs e identifique padrões em suas sequências. Consequentemente, POIs que frequentemente aparecem juntos ao longo de sequências semelhantes terão seus *embeddings* situados próximos no espaço vetorial latente.

Lui, Liu e Li [44] utilizam a sequência de POIs visitados pelos usuários para aprender as relações contextuais entre os POIs. Ao considerar a sequência de locais visitados, um modelo Word2Vec visa capturar a influência do contexto de um POI, que inclui o conjunto de POIs visitados antes e depois de um POI específico. Essa abordagem permite uma compreensão mais profunda dos comportamentos de visita dos usuários e ajuda a melhorar a qualidade das recomendações de localização personalizadas.

Zhou e Huang [116] modelam POIs considerando a distribuição de movimentos das pessoas nesses locais. Para isso, um conjunto de coocorrências de movimento são geradas para treinar um modelo Word2Vec. Por exemplo, se houver um movimento de restaurante para local de trabalho, então forma-se um par como $\langle$`restaurante`, `local de trabalho`$\rangle$, em que o local de trabalho é o contexto do restaurante do local de origem. Ao se treinar o modelo, POIs que têm movimentações semelhantes, também apresentam vetores semelhantes no espaço latente. Assim, por meio desse modelo, recomendações de POIs podem ser realizadas considerando o fluxo de movimento entre os POIs.

Wang *et al*. [86] propõem uma abordagem que captura tanto os aspectos estáticos (representação dos POIs) quanto os aspectos dinâmicos (padrões de mobilidade entre os POIs) de comunidades urbanas. O processo envolve a criação de grafos espaciais-temporais periódicos de mobilidade humana e a aplicação de um modelo do tipo auto-encoder para aprender os *embeddings* dos POIs. Esses *embeddings* são então agregados para representar as estruturas de comunidades urbanas a fim de entender a estrutura das comunidades.

Zhai *et al*. [111] propõem uma abordagem para detectar regiões funcionais na escala de uma área de vizinhança através de dados de POI obtidos com modelos baseados no contexto espacial. Os *embeddings* são gerados utilizando relações de sequência entre POIs que

são construídas utilizando o vizinho mais próximo de cada POI. O número de relações do conjunto de treinamento é replicado com base na distância entre o POI central e o POI de contexto. Assim, POIs mais próximos terão um maior número de coocorrências replicadas no conjunto de treinamento. Nessa abordagem, o Word2Vec é utilizado como modelo para geração dos *embeddings* dos tipos. Em seguida, os *embeddings* são extraídos e agrupados para detecção das regiões funcionais.

Wang e Huang [85] propõem o método SPENT (*Similarity-based POI Embedding and Recurrent Neural network with Temporal Influence*), que utiliza uma árvore de similaridade para organizar os POIs em uma hierarquia e aplica o Word2Vec para gerar *embeddings* dos POIs. Em seguida, o SPENT utiliza uma rede neural do tipo LSTM para modelar os sucessivos comportamentos de transição dos usuários entre os POIs, aplicando os vetores dos POIs juntamente com os *check-ins* inseridos pelos usuários. Dessa forma, um mecanismos de recomendação baseado nas transições dos usuários entre POIs é fornecido.

Hu *et al*. [24] também fornecem um *framework* para detectar a estrutura espacial urbana e descobrir regiões funcionais urbanas com base em dados de POIs. O método consiste na construção de um corpus utilizando uma abordagem baseada em pares (POI central, POI de contexto). Um modelo Word2Vec é usado para capturar as relações contextuais dos tipos e consequentemente para extrair os *embeddings*. Na última etapa, um método de agrupamento espacial é utilizado para agrupar os POIs permitindo assim a identificação das regiões e suas funções urbanas.

Xu *et al*. [100] empregam *embeddings* de POI para representar ROIs. Para isso, são utilizadas informações como as preferências dos usuários sobre os POIs e os padrões de mobilidade dos usuários em nível de região. Essas informações são agregadas e processadas por meio de um modelo de rede neural chamado GANR (Graph Attentive Neural Network for Region Recommendation). O modelo GANR é equipado com dois módulos de atenção: o módulo de atenção em nível de POI, que seleciona os POIs informativos de uma região, e o módulo de atenção em nível de região, que aprende as preferências da região. Além disso, o modelo também utiliza o *framework* NGCF (Neural Graph Collaborative Filtering) [87] para aprender as interações entre os usuários e as regiões. Por fim, a ROI é representada por meio dos *embeddings* gerados pelo modelo GANR.

Zhang *et al*. [112] fornecem um *framework* para extrair e identificar regiões funcionais

utilizando *embeddings* gerados a partir do GloVe [56]. Para isso, foram integradas informações de sequência dos POIs no contexto espacial. Inicialmente, zonas centradas em cada POI foram utilizadas para construir um corpus que representa funções urbanas. Em seguida, os *corpus* gerados foram utilizados para treinar um modelo GloVe [56]. Por fim, os *embeddings* são agrupados para identificar as regiões funcionais.

Chen *et al*. [11] propõem o Hier-CEM (*Category Embedding Model*), que gera *embeddings* para cada tipo de POI incorporando também a estrutura hierárquica dos tipos. O Hier-CEM consiste em dois componentes: a incorporação da sequência de tipos em uma sequência de POIs e a incorporação da hierarquia dos tipos no modelo. Na primeira etapa, a relação sequencial entre um tipo e seus vizinhos é capturada. Para melhorar as representações dos tipos, informações hierárquicas são integradas para cada relação sequencial gerada. Assim, usando a hierarquia, são estabelecidas conexões entre o tipo do POI na sequência e os tipos na hierarquia desse POI. As relações construídas são então utilizadas para gerar *embeddings* por meio do Word2Vec.

No contexto de recomendação de POIs, Yu, Wanyan e Wang [108] fornecem um modelo de *embeddings* de POI baseado no Skip-Gram do Word2Vec para capturar a influência contextual dos POIs e aprender a representação vetorial a partir de sequências de visitas. As preferências dos usuários para os POIs de destino são obtidas por meio da similaridade dos vetores aprendidos. Nessa abordagem, dados de *check-in* foram utilizados para modelar as preferências personalizadas dos usuários para um POI.

Bing *et al.* [7] propõem um método chamado CatEM (*POI Category Embedding Method*) para gerar *embedings* de tipos de POI com base em transições sequenciais e relações de vizinhança dos POIs. O método tem como objetivo capturar relações complexas entre os tipos de POIs para criar *embeddings* mais significativos. Ele considera similaridades espaciais entre pares de tipos e localiza de forma adaptativa tipos vizinhos com maior similaridade no espaço vetorial. Os *embeddings* de tipos de POI são gerados aplicando-se a técnica Matriz de Informação Mútua Pontual em conjunto com o algoritmo *Laplacian Eigenmaps* [6] da área de Aprendizado de Máquina.

Em seu trabalho, Huang *et al*. [29] empregam *embeddings* de tipos de POI para estimar funções urbanas. Para isso, é utilizado o algoritmo *random walk* para gerar sequências de visitação, considerando os POIs de cada região da cidade, juntamente com o *Skip-Gram* do

Word2Vec para gerar os *embeddings* dos tipos de POI. Além disso, também é utilizado o algoritmo Laplacian Eigenmaps [6] para enriquecer as relações contextuais dos POIs com os tipos relacionados presentes na hierarquia. Após a obtenção dos *embeddings* dos tipos de POI, as regiões urbanas são convertidas em representações vetoriais, utilizando os *embeddings* de tipos de POI de cada POI presente nas regiões. Em seguida, um modelo *Long Short Term Memory*(LSTM) [53] é empregado para associar a representação vetorial das regiões com os tipos de função urbana correspondentes.

Com o objetivo de classificar as regiões de uma cidade de acordo com sua função, Yang, Bo e Zhang [105] propõem um método baseado em um modelo *DeepWalk* [58] para gerar sequências de nós em um grafo. No grafo, cada nó representa um POI em uma determinada região e as arestas indicam a distância entre os POIs na mesma região. A partir do grafo, conjuntos de sequências de POIs são obtidos e seus tipos são utilizados como entrada no modelo Word2Vec. Desse modo, o modelo aprende a representação de cada tipo com base na sequência de visitação em uma região, considerando a distância entre os POIs. Na etapa seguinte, os *embeddings* dos tipos de POIs no grafo são combinados para representar todas as regiões. Os vetores que representam cada região são utilizados como entrada em um modelo *Support Vector Machine* (SVM) [8] para classificação de cada uma das regiões.

Yao *et al*. [107] propõem um método para detectar mudanças no uso do solo, quantificando, resumindo e representando os aspectos estáticos e dinâmicos das comunidades urbanas. Para isso, são utilizados *embeddings* de tipos de POI que estão presentes em grafos a partir de um algoritmo denominado MT-POI2Vec (multi-temporal POI embedding). A partir do grafo, o algoritmo de *random walk* é utilizado para gerar uma sequência de visitação. Os padrões contextuais de visitação são aprendidos a partir do *Skip-Gram* do Word2Vec. Em seguida, os *embeddings* de tipo de POI são empregados para representar parte das parcelas do solo, que por sua vez alimenta um modelo do tipo Auto-Encoder.

3.3 *Embeddings* Baseados em Dados Textuais Georreferenciados

Trabalhos nessa linha de pesquisa aproveitam a análise de palavras relacionadas aos POIs para definir a relação existente entre eles. Textos georreferenciados associados aos POIs são

utilizados para extrair tópicos ou palavras que são utilizadas para alimentar modelos de IA. Geralmente o objetivo do modelo é prever uma palavra ou tópico relevante dado um tipo específico de POI.

Shoji *et al*. [69] propõem um método para representar a atmosfera da área de um POI usando textos de microblogs georreferenciados. Uma representação vetorial semelhante a representação distribuída do Word2Vec é gerada. Os vetores de características são produzidos considerando *tweets* geolocalizados próximos aos POIs. Assim, o treinamento consiste em predizer as palavras presentes nos *tweets* considerando os POIs próximos. De maneira semelhante ao trabalho de Tsubouchi, Kobayashi e Shimizu [80], POIs que apresentam palavras semelhantes nos *tweets* da região são semelhantes no espaço vetorial.

Em sua proposta, Wei, Anjaria e Samet [91] visam representar POIs considerando dados textuais em *tweets* georreferenciados. Para isso, é proposto o LeGo-CM, que extrai POIs e *tweets* georreferenciados, para construir um grafo de coocorrência que abrange os POIs e elementos textuais dos *tweets* georreferenciados juntamente com o tempo de postagem desses *tweets*. No grafo, os nós representam os POIs e as arestas são ponderadas pelo número de vezes que dois nós coocorrem em *tweets*. Por fim, o LeGo-CM explora um algoritmo de aprendizagem de grafos. Desse modo, os *embeddings* representam POIs considerando os termos presentes nos *tweets*.

Takerngsajsiri, Wakamiya e Aramaki [78] fornecem uma abordagem para comparar áreas entre cidades utilizando dados de *tweets* georreferenciados. Para isso, *tweets* em regiões são utilizados como entrada no Doc2Vec, um modelo capaz de gerar *embeddings* que processa documentos textuais para estimar a similaridade entre áreas em diferentes cidades. Por fim, a similaridade das localizações pode ser visualizada em mapas digitais.

Para capturar a atmosfera ao redor de um POI, Tsubouchi, Kobayashi e Shimizu [80] utilizam consultas realizadas por usuários em ferramentas de busca para capturar as relações contextuais dos POIs. Os autores indicam que, quando pessoas realizam buscas por lugares, geralmente utilizam informações que caracterizam o ambiente do POI como, por exemplo, “parque familiar”. Esse comportamento é utilizado como entrada para treinar um modelo baseado no LSTM para gerar *embeddings* de POIs. A abordagem considera que aspectos do ambiente de um POI estão inseridos no texto da consulta, desse modo, a tarefa consiste em predizer a próxima consulta dentro da estrutura do modelo. Considerando os *embeddings*

dessa abordagem, POIs que possuem descrições de consulta semelhantes estarão próximos no espaço vetorial.

Mousset, Yoann e Lynda [50] abordam o problema de predição de localização de *tweets* georreferenciados. Para isso, os autores propõem o modelo Spatially-aware Geotext Matching (SGM), que combina sinais de correspondência exata palavra-palavra-local com sinais de correspondência global entre *tweets* e POIs. Desse modo, o objetivo do modelo é predizer o POI com base no texto. As interações globais consideram a força da interação entre o *tweet* e o POI, tanto do ponto de vista espacial quanto textual. Desse modo, POIs que possuem relações contextuais com certas palavras, também estão posicionadas próximas no espaço latente.

Para representar ROIs, Paul, Feifei e Jeff [54] utilizam relações espaciais e textuais de POIs de cada ROI. Os POIs são conectados por meio de grafo, onde as palavras associadas aos POIs são utilizadas para criar arestas entre eles. Os *embeddings* dos POIs são obtidos a partir do GloVe pré-treinado, utilizando a descrição dos POIs. Além disso, as informações associadas aos POIs, como comentários de *check-in*, avaliações sociais e microblogs são utilizadas para criar conexões entre os POIs com base nas palavras presentes nesses documentos. Essas relações textuais são utilizadas para gerar os *embeddings* dos POIs e capturar as relações entre eles. Por sua vez, os *embeddings* dos ROIs são gerados a partir da agregação e combinação dos *embeddings* dos POIs.

Rajaonarivo *et al*. [64] propõem um método para gerar automaticamente um grafo de conhecimento a partir de mensagens do *twitter*. O objetivo é utilizar esse grafo para estimar os tipos de POI e responder perguntas feitas por turistas. O método utiliza relações entre palavras para modelar o grafo que é utilizado como entrada em uma Rede Neural Convolucional em Grafos (GCN do inglês *Graph Convolution Network*). Primeiro, são criados *embeddings* dos POIs. Em seguida, é aplicada uma técnica de classificação e medidas de similaridade correspondentes aos *embeddings* dos tipos de POI. Portanto, a partir das relações contextuais entre os tipos de POI e palavras, gera-se *embeddings* dos tipos.

De maneira similar, Wu *et al*. [93] propõem o CAME (category-aware multigraph embedding), uma abordagem que captura informações textuais relacionadas aos POIs com o objetivo de identificar POIs ausentes que o usuário tenha visitado em um determinado momento no passado. O modelo utiliza cinco grafos de informações relacionais para incorporar

dados textuais e espaciais, explorando a influência geográfica e derivando uma lista de pontuação de candidatos para identificação de POIs ausentes. Portanto, o trabalho considera tanto os dados textuais quanto os dados espaciais para relacionar os POIs.

3.4 *Embeddings* Baseados em Dados Geográficos

Nesse ramo, os trabalhos frequentemente se baseiam em características geográficas para criar representações de regiões inteiras. Essas representações se mostram ferramentas valiosas para diversas tarefas, como classificação de áreas urbanas, identificação de regiões de interesse, entre outros.

O estudo de Cocos e Callison [12] explora a relação entre palavras e seus contextos geoespaciais, concentrando-se no valor contextual derivado da localização física de onde determinado texto está associado. A partir do enriquecimento de um corpus de postagens geolocalizadas do *twitter* com dados físicos do *Google Places* e OSM é empregado o Word2Vec para produzir *embeddings* de palavras usando contextos geoespaciais. Esse estudo investiga a relação contextual das palavras codificadas nesses *embeddings* geoespaciais.

Jean et al. [31] e Spruyt [74] fornecem o Tile2Vec e Loc2Vec, respectivamente, para mapear imagens de satélite para o espaço vetorial latente. Cada método consiste em uma CNN, treinada com uma técnica denominada *triplet loss*, que minimiza a distância entre imagens âncora e imagens vizinhas, maximizando assim a distância entre a âncora e os locais distantes. Dessa forma, os dados geográficos presentes nos mapas podem ser representados no espaço vetorial, permitindo identificar a similaridade de diferentes regiões em mapas.

Wang e Rajagopal [90] propõem o Urban2Vec, um *framework* multimodal que incorpora tanto imagens de ruas como dados de POIs para aprender *embeddings* de uma região. Especificamente, é utilizada uma rede neural convolucional [52] (CNN do inglês *Convolutional Neural Network* ou *ConvNet*) para extrair características visuais das imagens das ruas da região, preservando a similaridade geoespacial. Além disso, cada POI é modelado utilizando a técnica de *bag-of-words* contendo informações de categoria, preço, avaliações e comentários. De maneira análoga à similaridade de documentos textuais, estabeleceu-se a similaridade contextual entre vizinhanças, que são tratados como documentos, no espaço vetorial. Os dados das imagens das ruas são então combinados com os dados dos POIs para

formar uma representação mais geral da região.

Huang, Wang e Sheng [28] fornecem uma abordagem para integrar dados multimodais georreferenciados como nós ou arestas de um grafo baseado nas relações existentes entre bairros. Uma representação da vizinhança é gerada com base em imagens de ruas e características, tais como *check-ins*, dos POIs da vizinhança. Também é utilizado a mobilidade humana para caracterizar a relação entre bairros sendo representados como arestas direcionadas no grafo. O grafo construído é utilizado como entrada em um modelo baseado em Redes Neurais de Grafos [67] (GNN do inglês *Graph Neural Network*). Assim, o modelo permite capturar a similaridade entre regiões diferentes.

Han *et al*. [23] propõem um modelo de recomendação de atrações turísticas utilizando fotos georreferenciadas do Flickr[8]. A abordagem integra dados espaciais, temporais e fotos para gerar a recomendação dos lugares. A estratégia utiliza *embeddings* das imagens dos locais, obtidas através de uma rede convolucional pré-treinada, e combina com dados temporais que são modelados com a estratégia de amostragem negativa do Word2Vec. Os *embeddings* resultantes são utilizados como entrada nos métodos *Matrix Factorization* e *Bayesian Personalized Ranking*, que geram a recomendação final.

Em seu trabalho, Xu *et al*. [101] apontam a importância de combinar informações de mapas onde os POIs estão presentes, juntamente com informações dos tipos POIs na previsão do uso da terra. A incorporação dessas informações na representação das zonas urbanas pode melhorar a precisão da previsão do uso da terra. O estudo propõe um modelo que utiliza uma Rede Neural Convolucional (CNN do inglês *Convolutional Neural Network*) e uma abordagem de aprendizado adversarial para enfatizar as partes mais importantes dos mapas e da estrutura categórica dos tipos de POIs.

Li *et al*. [38] abordam a importância da representação das regiões urbanas para a análise e planejamento urbano utilizando técnicas de aprendizado de máquina e dados urbanos. Para lidar com esses desafios, é proposto um *framework* de aprendizado contrastivo chamado RegionDCL, que utiliza dados do OSM, especificamente informações de edifícios, para derivar representações gerais das regiões urbanas. Além disso, também é empregado dados de mobilidade humana na aprendizagem de representações das regiões urbanas. Como resultado, tal estratégia permite a obtenção de *embeddings* das zonas urbanas considerando POIs e prédios

[8]Disponível em www.flickr.com. Acesso em 20 de maio de 2024.

simultaneamente.

Para representar regiões em *embeddings*, Sun *et al*. [76] propõem o modelo HaFusion. Nesse modelo, um conjunto de dados são inseridos para que o modelo aprenda as relações contextuais de todas as informações para gerar a representação da região. Os principais dados utilizados são: mobilidade humana, POIs, características geográficas, como topografia, clima, construções presentes nas regiões, edifícios, casas, entre outros. Esses dados são utilizados por diferentes modelos e métodos mencionados no trabalho para aprender os *embeddings* que são unidos no modelo HaFusion.

3.5 *Embeddings* Baseados em *Transformers*

Nessa linha de pesquisa se concentram trabalhos que buscam especializar modelos de linguagem de grande porte (LLM do inglês *Large Language Models*), como o BERT, para lidar com informações geográficas sobre POIs. Neste sentido, várias abordagens propõem a integração de informações geográficas dentro dos LLMs, abrindo portas para novos e aprimorados modelos de propósito geral.

Liu *et al*. [43] propõem um modelo de pré-treinamento, chamado Geo-BERT, para integrar as informações geográficas nas representações pré-treinadas de POIs. Primeiramente, é simulada a distribuição dos POIs no mundo real a partir de um grafo. Nele, os nós representam POIs de uma vizinhança que são conectados baseados em sua latitude e longitude. Além disso, alguns nós indicando o nível administrativo (ex: rua, bairro, cidade) foram criados para conectar POIs que estão sob as mesmas propriedades. O grafo resultante é treinado utilizando uma GNN. Por fim, dados textuais são integrados com os *embeddings* dos POIs e treinados em um modelo BERT [18].

Bashir e Misic [5] propõem o BERT4Loc, um sistema de recomendação de POIs baseado no BERT [18] com o objetivo de fornecer aos usuários recomendações baseadas na localização. O modelo proposto incorpora dados de localização e preferências do usuário considerando a sequência de visitações. Assim, os vetores gerados por esse modelo podem ser utilizados em abordagens de recomendação, visto que POIs que possuem sequências de visitação semelhantes também apresentam vetores semelhantes.

Li et al. [110] fornecem um modelo de linguagem que contempla aspectos espaciais

para fornecer uma representação de propósito geral de POIs baseado na vizinhança de POIs. O denominado SPABERT, estende o modelo BERT [18] para capturar o contexto espacial linearizado, ao mesmo tempo em que preserva as relações espaciais dos POIs no espaço bidimensional. Esse modelo é pré-treinado como uma tarefa de linguagem mascarada e uma tarefa de predição de entidades mascaradas para aprender assim dependências espaciais. O treinamento fez uso de pseudo sentenças geradas a partir de bases de dados geográficas derivadas do OSM.

Huang et al. [26] apresentam seus esforços para projetar e implementar o ERNIE-GeoL, um modelo baseado no ERNIE [27] (*Enhanced Representation through kNowledge IntEgration*) e pré-treinado com dados geográficos e textuais para melhorar uma gama de tarefas relacionadas ao Baidu Maps[9]. Nessa abordagem um grafo é construído para conter POIs juntamente com textos de consultas usando o banco de dados de POI e *logs* de pesquisas no Baidu Maps. Para integrar tais informações, arestas entre dois nós foram definidas com base na correlação espacial e dados de mobilidade humana. Para gerar cada sequência de entrada para treinamento do ERNIE-GeoL, foi utilizado um algoritmo de caminhada aleatória (*random walk*) para amostrar uma sequência de nós como um documento de entrada. Em seguida, modelos baseados em *transformers* foram utilizados como rede principal para aprender as representações de cada nó. O treinamento foi realizado como uma tarefa de linguagem mascarada e geocodificação. Desse modo, o modelo pode aprender uma representação universal de dados geográficos presentes na linguagem.

Em seu trabalho, Li *et al*. [39] apresentam o GeoLM (*Empowering Language Models for Geospatially Grounded Language Understanding*), um modelo de linguagem que melhora a compreensão de geo-entidades na linguagem natural. Para isso, o GeoLM utiliza menções de geo-entidades como âncoras para conectar informações linguísticas em corpora de texto com informações geoespaciais extraídas de bancos de dados geográficos. O GeoLM conecta os dois tipos de contexto por meio de aprendizado contrastivo e modelagem de linguagem mascarada. Ele também utiliza um mecanismo de incorporação de coordenadas espaciais para codificar relações de distância e direção e capturar contexto geoespacial.

Ding, Ruixue e Chen [20] prpõem um método para resolver o problema de consultas de POIs que considera o contexto geográfico relacionado à consulta. O Modelo de Lingua-

[9]Disponível em https://map.baidu.com. Acesso em 20 de maio de 2024.

gem Geográfico Multimodal (MGeo do inglês *Multi-modal Geographic language*) codifica o contexto, formado por linhas e polígonos que indicam ruas e ROIs, utilizando um modelo BERT. Durante o treinamento, o MGeo utiliza a técnica de mascaramento para prever alguns itens, tais como o ID dos POIs, o tipo de relacionamento do POI com o contexto geográfico (próximo ou coberto) e a posição relativa entre os objetos geográficos do contexto. Os *embeddings* gerados são concatenados com *embeddings* das consultas para associar as consultas e os contextos no mesmo espaço latente. Nesta etapa, o BERT é novamente empregado para prever o contexto geográfico ou informações da consulta. O modelo de linguagem resultante pode ser usado em tarefas específicas, como recomendação ou ranqueamento de POIs em consultas.

Lin *et al*. [40], fornecem o GeoGalactica, um modelo de linguagem projetado especificamente para a área de geociência. Tal modelo foi produzido a partir do Galactica [79], um modelo de linguagem treinado para o domínio da ciência. Para isso, um pré-treinamento adicional do modelo é realizado com um vasto corpus de texto relacionado à geociência e, em seguida, o modelo é refinado usando dados de ajuste supervisionado (Supervised Fine-Tuning - SFT) coletados especificamente para instruir o modelo nas nuances da geociência. O resultado é o GeoGalactica, que consiste em 30 bilhões de parâmetros.

Xie *et al*. [98] propõem o QUERT (Continual Pre-training of Language Model for Query Understanding in Travel Domain Search), um modelo de linguagem pré-treinado cujo objetivo principal é melhorar o desempenho de tarefas relacionadas à busca, como a classificação e a delimitação das consultas. O modelo é capaz de compreender e interpretar as consultas dos usuários de forma contínua. Para isso, o QUERT foi treinado utilizando várias estratégias, incluindo a Predição de Máscara Consciente de Geografia, na qual o modelo é treinado para prever palavras ou frases relacionadas à geografia que foram mascaradas em uma consulta; a Predição de Código Geohash, na qual um sistema de codificação que representa coordenadas geográficas é treinado para prever o código Geohash correspondente a uma consulta, ajudando a capturar informações de localização nas consultas; o Aprendizado de Comportamento de Cliques do Usuário, no qual o modelo é treinado para prever o comportamento de cliques do usuário em relação a uma consulta, possibilitando a compreensão das preferências e intenções dos usuários ao realizar pesquisas de viagens; e a Predição de Ordem de Frase e *Token*, na qual o modelo é treinado para prever a ordem correta das frases

e *tokens* em uma consulta. A partir desses treinamentos, o QUERT permitiu a melhoria de tarefas relacionadas, sendo utilizado como modelo base.

Balsebre, Weiming e Gao [4] propõem um novo *framework* para incorporar conhecimento geoespacial em modelos de linguagem pré-treinados, permitindo que eles forneçam recomendações precisas e minimizem informações incorretas. Para isso, foi descrito um modelo denominado LAMP que utiliza dados sintéticos de consultas relacionadas a POIs. Essas consultas são geradas de forma aleatória, incorporando nomes de lugares ou categorias aleatórias, para que o modelo aprenda que nem sempre há POIs disponíveis nas proximidades e que um POI existente localizado mais distante pode ser uma alternativa mais adequada. Além disso, o treinamento também envolve a tradução da posição do usuário em um endereço utilizando uma API de geocodificação reversa. Como resultado, o modelo consegue recomendar POIs com base em consultas utilizando linguagem natural e também na posição do usuário.

Qi *et al*. [60] propõem o GeoDecoder, um modelo que utiliza tanto informações de texto quanto de imagem para processar dados geoespaciais de forma eficiente e de alta qualidade. Ele é capaz de adquirir conhecimento sobre entidades geográficas e realizar múltiplas tarefas em um único modelo. O GeoDecoder foi treinado em uma grande quantidade de dados de texto e imagem para melhorar seu desempenho. Em seu treinamento, esse modelo é empregado para entender a relação entre diferentes elementos do mapa, como edifícios e estradas. Ele também é treinado para gerar coordenadas precisas para POIs com base em informações contextuais e geográficas. Isso permite que o modelo localize com precisão os POIs no mapa. Desse modo, o GeoDecoder dispõe de uma compreensão abrangente de conhecimentos geográficos e permite que ele resolva problemas relacionados a mapas.

3.6 Posicionamento desta pesquisa em relação aos trabalhos relacionados

Para mapear os trabalhos relacionados, foi definido um conjunto de atributos responsáveis por identificar a informação base utilizada na produção dos *embeddings*, assim como o tipo resultante de *embedding*, e sua natureza (estático ou dinâmico). Isso permite entender quais informações foram utilizadas para produzir os *embeddings*, que tipo de representação é for-

necida e a natureza da representação. A Tabela 3.1 resume os trabalhos relacionados discutidos neste capítulo. A última linha apresenta a caracterização desta pesquisa. As colunas presentes na tabela representam os atributos selecionados para mapear os trabalhos.

- ***Embeddings* de tipos de POI**: trabalhos que focam na representação dos tipos de POIs;
- ***Embeddings* de POI**: trabalhos que focam na representação de POIs como entidade;
- ***Embeddings* de Região/Zona**: trabalhos que focam na representação de regiões ou zonas urbanas;
- **POIs do Contexto**: trabalhos que utilizam os POIs da vizinhança ou de uma sequência no processo de geração de *embeddings*;
- **Palavras Relacionadas**: trabalhos que utilizam palavras relacionadas aos POIs para geração de *embeddings*;
- **Feições Geográficas**: trabalhos que utilizam feições geográficas do contexto dos POIs, tais como rios, ruas, sinais de trânsito, entre outros;
- ***Embeddings* Estáticos**: trabalhos que utilizam modelos que geram *embeddings* estáticos.

De acordo com a Tabela 3.1, é perceptível que muitos estudos têm como objetivo a geração de *embeddings* de tipos de POI. Isso ressalta a relevância dessa representação, amplamente utilizada em tarefas como recomendação de POIs e classificação de áreas urbanas. Também se destaca que a informação base mais comum são os POIs presentes na vizinhança, indicando que as feições geográficas do contexto têm sido subutilizadas na representação dos tipos. Além disso, observa-se que as feições geográficas são frequentemente empregadas para criar representações gerais e não são diretamente utilizadas na geração de *embeddings* dos tipos. Adicionalmente, a maioria dos estudos adota métodos que geram *embeddings* estáticos, geralmente provenientes de modelos clássicos na área de PLN.

Entretanto, o foco desta pesquisa consiste na proposição de uma abordagem para a geração de *embeddings* de tipos de POI utilizando as feições geográficas do contexto dos POIs.

Acredita-se que, da mesma forma que os POIs do contexto produzem *embeddings* que indicam a similaridade dos tipos, as feições geográficas do contexto também podem fornecer tal similaridade, além de serem uma alternativa viável para uso em diversas tarefas. Desse modo, esta pesquisa aborda três dos sete atributos indicados na tabela. Adicionalmente, também é proposta a geração de *embeddings* utilizando modelos mais recentes da área de PLN.

3.7 Considerações Finais

Neste capítulo, foi verificado que *embeddings* de tipos de POI são utilizados em diversas tarefas, como classificação de zonas urbanas, recomendação de POIs, representação de regiões de interesse, entre outros. Foram apresentadas diversas abordagens que geram tal representação e como elas são empregadas nas tarefas relacionadas. Também foi constatado que dados textuais georreferenciados são aplicados para capturar a relação contextual existente entre POIs. Além disso, foram explanados métodos que empregam modelos baseados em *transformers* para gerar modelos de linguagem que incorporam informações geográficas.

A partir dos trabalhos elencados, constatou-se que o uso da informações de vizinhança ou sequência é uma estratégia muito utilizada para gerar *embeddings* dos tipos de POI. Além disso, tais abordagens geralmente fazem uso do Word2Vec, com foco em uma tarefa de predição de tipos de POI de contexto a partir dos tipos do POI central. Nesse ramo, a sequência de visitação de POIs também é utilizada para definir a similaridade contextual dos mesmos. A codificação da relação sequencial geralmente é realizada a partir do Word2Vec ou modelos baseados em redes neurais recorrentes, tais como o LSTM. Nessas abordagens, a tarefa passa a ser predizer os próximos POIs a serem visitados considerando um POI atual.

Constatou-se também que dados textuais georreferenciados são utilizados para capturar aspectos do ambiente ao redor dos POIs. Nessas estratégias, dados de *tweets* e de consultas realizadas em serviços de busca são utilizados para treinar modelos como o Word2Vec, Doc2Vec e LSTM. Assim, POIs são representados considerando os termos presentes nos dados textuais. Além disso, representações maiores, como regiões, também foram geradas com essa estratégia.

Também foram apresentadas abordagens que utilizam dados geográficos para representar

Tabela 3.1: Comparativo de trabalhos relacionados.

Trabalho	Embeddings de tipos de POI	Embeddings de POI	Embeddings de Região ou Zona	POIs do Contexto	Palavras Relacionadas	Feições Geográficas	Embeddings Estáticos
Cocos e Callison (2017) [12]	X				X		X
Yan et al. (2017) [103]	X			X			X
Zhou e Huang (2018) [116]		X		X			X
Shoji et al. (2018) [69]		X			X		X
Wang et al. (2018) [86]		X		X			X
Liu et al. (2019) [42]	X			X			X
Jin et al. (2019) [33]	X			X			X
Takerngsajsiri, Wakamiya e Aramaki (2019) [78]			X		X		X
Wang e Huang (2019) [85]	X			X			X
Wei, Anjraria e Samet (2019) [91]		X			X		X
Zhu et al. (2019) [117]	X			X			X
Zhai et al. (2019) [117]	X			X			X
Hu et al. (2020) [24]	X			X			X
Mousset, Yoann e Lynda (2020) [50]		X		X	X		X
Tsubouchi, Kobayashi e Shimizu (2020) [80]		X			X		X
Wang e Moosavi (2020) [90]	X			X			X
Wang e Rajagopal (2020) [90]			X	X	X	X	X
Xu et al. (2020) [100]			X	X			X
Chen et al. (2021) [11]	X			X			X
Han et al. (2021) [23]				X		X	X
Huang, Wang e Sheng (2021) [28]			X	X		X	X
Liu et al. (2021) [43]		X		X			
Paul, Feifei e Jeff (2021) [54]			X	X	X		X
Zhang et al. (2021) [112]	X			X			X
Bashir e Misic (2022) [5]		X		X			
Bing et al. (2022) [7]	X			X			X
Huang et al. (2022) [29]	X				X		X
Li et al. (2023) [38]			X	X		X	X
Rajaonarivo et al. (2023) [64]	X				X		X
Sun et al. (2023) [76]			X	X		X	X
Wu et al. (2023) [93]	X			X	X		X
Yang, Bo e Zhang (2023) [105]	X			X			X
Yao et al. (2023) [107]	X			X			X
Xie et al. (2023) [98]				X	X		
Xu et al. (2023) [101]			X	X			X
Zhou et al. (2023) [115]			X	X			
Balsebre, Weiming e Gao (2024) [4]		X			X		
Qi et al. (2024) [60]		X			X	X	
Silva (2024) - Este trabalho	**X**					**X**	**X**

regiões e o ambiente dos POIs. Esses métodos utilizam imagens da vizinhança para produzir uma representação geográfica mais ampla. Além disso, a relação entre POIs foi modelada utilizando modelos baseados em grafos. Além de imagens, informações de mapas também foram incorporadas nos *embeddings* das representações gerais.

Por fim, foram apresentados métodos baseados em *transformers* para prover modelos de linguagem que incorporam informações geográficas. Utilizando a vizinhança de POIs, pseudo sentenças são criadas e utilizadas como entradas em modelos como o BERT e o ERNIE. Além disso, POIs foram combinados com dados textuais para que o modelo de linguagem capturasse as informações espaciais dentro da linguagem. Desse modo, os modelos resultantes apresentam aspecto de propósito geral, podendo ser utilizados como base em diversas outras atividades.

O próximo capítulo apresenta a abordagem proposta nesta tese, descrevendo um *pipeline* desenvolvido para a geração de *embeddings* de tipos de POI utilizando feições geográficas.

Capítulo 4

Geração de *Embeddings* de Tipos de POI

Este capítulo apresenta a abordagem proposta para gerar *embeddings* de tipos de POI utilizando feições geográficas do contexto dos POIs. Para isso, foi desenvolvido um *pipeline* que utiliza dados dos POI juntamente com feições geográficas. Essas informações são utilizadas em um algoritmo desenvolvido, denominado *Geographic Context to Vector* (GeoContext2Vec), que realiza a associação dos tipos de POI com as feições geográficas do contexto. Inicialmente, é apresentada a visão geral da abordagem. Em seguida, as bases de dados selecionadas são explicadas. Na terceira seção, o algoritmo GeoContext2Vec é apresentado em detalhes. Na quarta seção, os detalhes sobre como ocorre a geração dos *embeddings* são apresentados. Por fim, a última seção apresenta as considerações finais do capítulo.

4.1 Visão geral da Solução

A geração de *embeddings* é realizada seguindo um *pipeline*, conforme ilustrado na Figura 4.1. Inicialmente, as informações dos POIs e das feições geográficas de seu contexto são utilizadas como entrada no algoritmo GeoContext2Vec, cujo objetivo é associar os tipos dos POI com as feições geográficas em seu contexto mantendo algumas propriedades espaciais. A saída do algoritmo é um novo conjunto de dados que contém relações binárias entre os tipos do POI e as feições geográficas do contexto. Esse algoritmo está descrito na Seção 4.3. No passo seguinte, um modelo neural da área de PLN é treinado utilizando os dados gerados pelo GeoContext2Vec. Ao final do treinamento, é possível obter os *embeddings* para cada tipo de POI. A Seção 4.4 descreve como o treinamento é realizado utilizando modelos de

PLN.

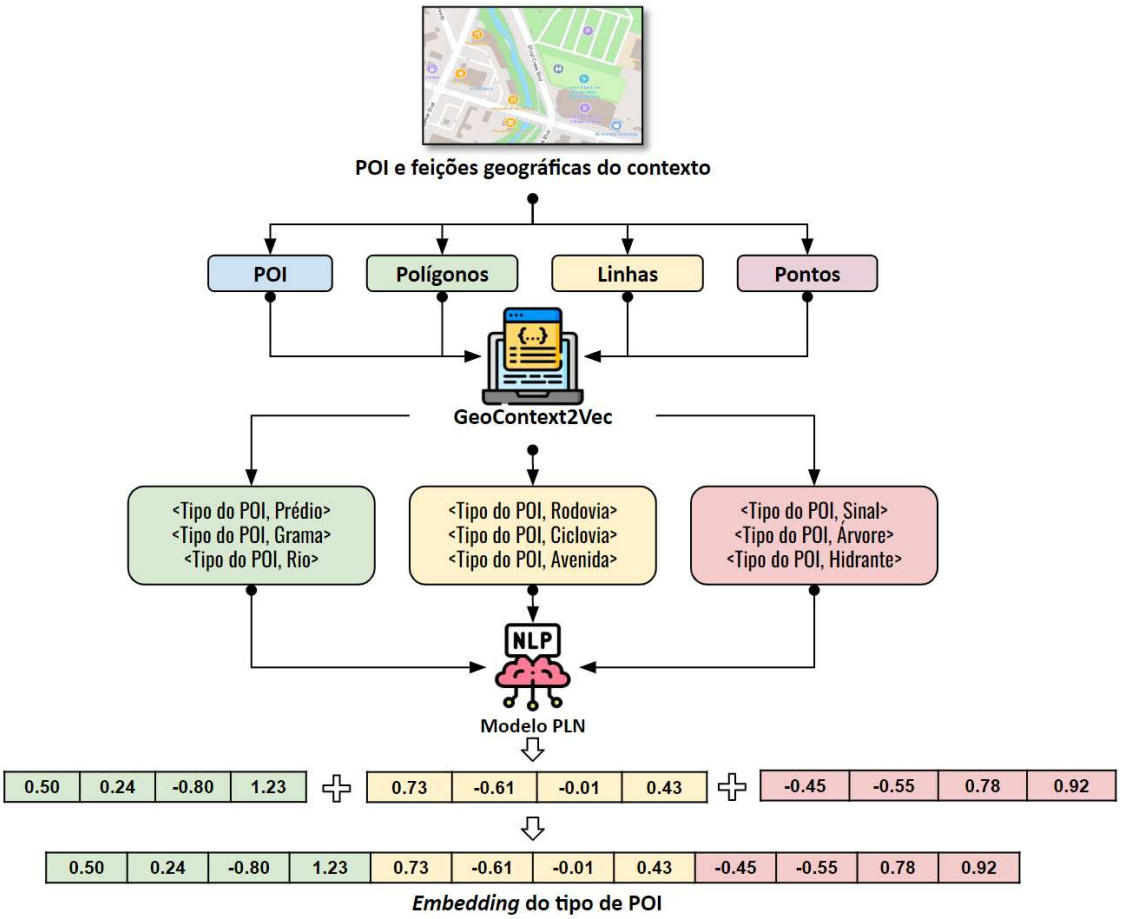


Figura 4.1: *Pipeline* para gerar os *Embeddings* de tipos de POI.

Fonte: Autoria própria

4.2 Bases de Dados

Para obtenção de dados de POIs, foi utilizada uma base de dados proveniente do Yelp, especificamente do *Yelp Challenge*[1], uma inciativa que disponibiliza um subconjunto dos dados contento informações sobre negócios, dados de usuários para uso pessoal, educacional e acadêmico. Essa base foi escolhida devido sua disponibilidade e facilidade para recuperação dos dados, além de não haver restrições para uso acadêmico. Os dados também apresentam boa integridade, pois são constantemente atualizados pela Yelp. Nesta tese, foi utilizada uma instância da base gerada em Fevereiro de 2021.

[1]Disponível em https://www.yelp.com/dataset. Acesso em 20 de maio de 2024.

Cada POI da base de dados é composto por 14 atributos, estando apresentados na Tabela 4.1. Dos atributos presentes na Tabela 4.1, o GeoContext2Vec utiliza os seguintes: *business_id*, *latitude*, *longitude* e *categories*. O *business_id* é utilizado para identificar os POIs e seus respectivos tipos associados. A *latitude* e *longitude* são empregadas para identificar a localização geográfica do POI e definir seu contexto (vizinhança). O atributo *categories* contém os tipos que são o foco desta tese. Por esta tese focar em gerar *embeddings* de tipos de POI utilizando feições geográficas, os demais atributos não são necessários.

Tabela 4.1: Atributos de POIs na base de dados do Yelp.

Atributo	Descrição	Tipo do Atributo
business_id	**identificador único de cada POI**	**String**
address	endereço do POI	String
city	cidade que contém o POI	String
state	estado que contém o POI	String
postal_code	código postal associado à localização do POI	Integer
latitude	**coordenada geográfica**	**Float**
longitude	**coordenada geográfica**	**Float**
stars	avaliação do POI dada pelos usuários	Float
review_count	quantidade de comentários relacionados ao POI	Integer
is_open	*flag* que indica se o POI está aberto ou não	Boolean
attributes	informações gerais como: quantidade de assentos, dentre outros	String
categories	**especifica os tipos associados ao POI**	**String**
hours	horário de funcionamento do POI	Dictionary

Entre as bases de dados geográficas disponíveis, pode-se citar o Baidu[2], Google Maps[3], Gaode[4], OpenStreetMap[5], entre outros. Para obtenção das feições geográficas, foram utilizados dados provenientes do OSM. Essa base é mantida por meio de uma iniciativa de dados aberta, sendo desenvolvida por uma comunidade voluntária de mapeadores que contribuem

[2]Disponível em https://map.baidu.com. Acesso em 20 de maio de 2024.

[3]Disponível em https://www.google.com/maps. Acesso em 20 de maio de 2024.

[4]Disponível em https://gaode.com/. Acesso em 20 de maio de 2024.

[5]Disponível em https://www.openstreetmap.org. Acesso em 20 de maio de 2024.

e mantêm os dados atualizados. A partir do OSM, é possível obter o esquema de mapas através de mecanismos de *tiles*[6], que são arquivos contendo informações renderizadas dos mapas (*raster*[7]) ou vetoriais (*vector*[8]); e através da ferramenta Overpass API[9], que possibilita a obtenção de uma cópia do banco de dados. Os dados provenientes dessa base estão distribuídos em quatro tabelas:

- **planet_osm_polygon**: contém todas as feições geográficas que possuem a propriedade espacial de área, tais como edifícios, rios e parques (Figura 4.2.a);
- **planet_osm_line**: armazena todas as feições geográficas que possuem a propriedade espacial de comprimento e ilustram a estrutura da cidade, como ruas menores e estradas de acesso (Figura 4.2.b);
- **planet_osm_roads**: armazena todas as feições geográficas que também possuem comprimento, mas que ilustram as principais estradas da cidade, como rodovias e avenidas (Figura 4.2.c);
- **planet_osm_point**: engloba todas as feições geográficas que representam aspectos específicos do mapa, como sinais de trânsito, fontes, barreias, árvores, postes, entre outros (Figura 4.2.d).

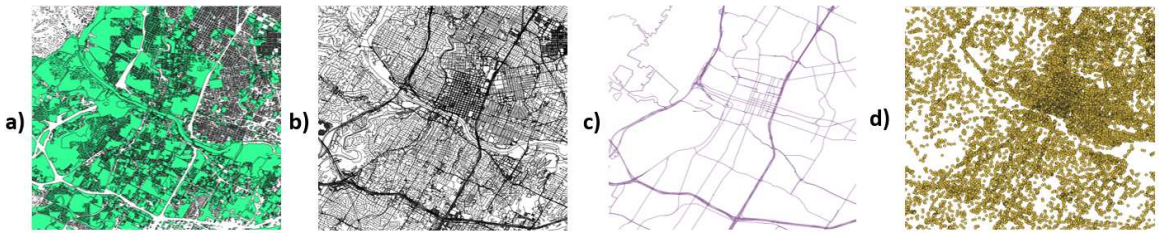


Figura 4.2: Representação em mapa dos dados do OSM das tabelas: a) planet_osm_polygons; b) planet_osm_lines; c) planet_osm_roads; d) planet_osm_points.

Fonte: Autoria própria

Cada tabela possui um conjunto de atributos que descrevem as feições geográficas representada em cada linha. Como exemplo, há um atributo *id* para identificar cada registro e um

[6]Disponível em https://wiki.openstreetmap.org/wiki/Tiles. Acesso em 20 de maio de 2024.
[7]Disponível em https://wiki.openstreetmap.org/wiki/Raster_tiles. Acesso em 20 de maio de 2024.
[8]Disponível em https://wiki.openstreetmap.org/wiki/Vector_tiles. Acesso em 20 de maio de 2024.
[9]Disponível em https://wiki.openstreetmap.org/wiki/Overpass_API. Acesso em 20 de maio de 2024.

também o atributo *area* pra indicar o espaço que determinada feição geográfica ocupa. A lista completa de atributos pode ser verificada na documentação online do OSM[10].

Nesta abordagem, decidiu-se utilizar as tabelas do OSM como provedoras das feições geográficas. Essa decisão foi tomada considerando os seguintes pontos: i) As tabelas do OSM já possuem informações de área e comprimento das feições geográficas, sendo essenciais para capturar os padrões espaciais; ii) A estrutura dos dados é armazenada em bancos relacionais, facilitando o acesso imediato e a manipulação.

4.3 Algoritmo GeoContext2Vec

Com base nos trabalhos relacionados (ver a Seção 3), foi adotada a estratégia de utilizar relações binárias para associar POIs aos dados geográficos do contexto. Para isso, a relação binária ⟨`tipo do POI central, tipo do POI de contexto`⟩, amplamente utilizada nas abordagens de vizinhança de POIs, foi adaptada para uma nova relação ⟨`tipo do POI central, feição geográfica do contexto`⟩. Essa relação é construída considerando todos os tipos associados ao POI que centraliza um contexto e para cada feição geográfica presente no contexto. Como ilustrado na Figura 4.3, um exemplo de contexto em torno de um POI inclui a presença de um rio, um espaço verde, alguns edifícios, entre outros. Supondo que o POI central possua os tipos "restaurante" e "churrascaria", as seguintes relações binárias podem ser criadas: ⟨`Restaurante, Rio`⟩; ⟨`Restaurante, Edifício`⟩; ⟨`Restaurante, Área Verde`⟩; ⟨`Churrascaria, Rio`⟩; ⟨`Churrascaria, Edifício`⟩; ⟨`Churrascaria, Área Verde`⟩.

Utilizar o conjunto de relações binárias simples como entrada em modelos PLN pode não ser suficiente para capturar as propriedades espaciais de forma adequada. Para exemplificar, considere o contexto geográfico da Figura 4.3. É possível notar que o rio e a área verde ocupam boa parte do contexto, enquanto os edifícios ocupam uma parte menor desse espaço. Além disso, há mais ocorrências de edifícios do que rio e área verde. Se a relação binária ⟨`tipo do POI central, feição geográfica do contexto`⟩ for utilizada diretamente nos modelos de PLN, é possível que esses modelos aprendam que edifícios

[10]Disponível em https://wiki.openstreetmap.org/wiki/Tags. Acesso em 20 de maio de 2024.

Figura 4.3: Exemplo de contexto geográfico de um POI (central).

Fonte: Autoria própria

estão mais relacionados ao POI central do que o rio e a área verde, porque há mais ocorrências de edifícios do que rio e área verde. No entanto, como o rio e a área verde ocupam grande parte do contexto, intuitivamente devem ter mais importância, pois são mais perceptíveis. Desse modo, é necessário considerar o espaço que cada feição geográfica ocupa em um contexto para capturar os padrões espaciais.

Para isso, foi definida a proporção de espaço ocupado (SP, do inglês *Space Proportion),* que é calculada considerando o espaço que cada feição geográfica ocupa no contexto[11]. Foram consideradas duas maneiras para calcular tal informação. A primeira delas considera a proporção de espaço de maneira relativa à quantidade de feições presentes em um contexto. Dessa forma, a proporção de espaço ocupado por uma feição é calculada em relação ao espaço total ocupado por todas as feições do contexto. A Equação 4.1 apresenta o cálculo de SP, onde S_{F_j} (S_F do inglês *Space of the Feature*) é o espaço que uma feição F_j ocupa e $\sum_{k=1}^{M} S_{F_k}$ é o espaço total que as M feições ocupam no contexto. Se uma feição ocupar um espaço considerável, o valor de SP será alto, e vice-versa. Essa propriedade é diretamente utilizada na construção das relações binárias do conjunto de treinamento dos *embeddings*. Como SP denota um valor no intervalo $[0, 1]$, é necessário alterar sua escala para números maiores ou iguais a um, para que, para alguns valores de SP, exista pelo menos uma relação binária no conjunto de treinamento. Para isso, foi definida μ, uma constante utilizada para

[11]No caso das feições do OSM, o espaço ocupado se refere à área para polígonos ou comprimento para linhas.

alterar a escala do resultado. Como exemplo, se $\mu = 10$, os valores estarão no intervalo $[0, 10]$, e assim por diante.

$$SP_{F_j} = \left(\frac{S_{F_j}}{\sum_{k=1}^{M} S_{F_k}} \right) \times \mu \tag{4.1}$$

A segunda maneira de calcular a *proporção de espaço ocupado* considera a área total da forma geométrica que representa o contexto de um POI. Portanto, se o contexto for representado por uma geometria circular, deve-se calcular a proporção de espaço ocupado por uma feição em relação à área do círculo. Se o contexto for representado por uma geometria retangular, a proporção de espaço ocupado por uma feição deve considerar a área do retângulo, e assim por diante. A Equação 4.2 demonstra como ocorre o cálculo de SP_abs nesse cenário considerando a área da geometria que representa o contexto (indicado pelo denominador da equação *context_geometry_area*). Nesta tese, o contexto de um POI é representado a partir de uma geometria circular.

$$SP\_abs_{F_j} = \left(\frac{S_{F_j}}{context\_geometry\_area} \right) \times \mu \tag{4.2}$$

Nesta tese, as duas versões de SP foram utilizadas visando investigar como os *embeddings* resultantes se comportam diante da variação da área do contexto. Além disso, outras funções matemáticas, como a logarítmica, não foram empregadas, pois tendem a alterar a distribuição dos resultados, influenciando diretamente na construção do conjunto de treinamento. Entretanto, o objetivo desta pesquisa é utilizar a proporção do espaço preservando a distribuição existente no contexto de POIs.

Além da *proporção de espaço ocupado*, decidiu-se utilizar a *proporção de ocorrência*, que é calculada a partir das ocorrências da feição no contexto. Intuitivamente, se uma feição ocorre repetidas vezes no contexto, ela se relaciona mais com o POI central do que feições que ocorrem poucas vezes. Para calcular a *proporção de ocorrência* (OP do inglês *Occurrence Proportion*) de uma feição, definiu-se a Equação 4.3, em que O_{Fj} (O_F do inglês *occurrence of the feature*) é a quantidade de vezes que a feição Fj ocorre no contexto e $\sum_{k=1}^{M} O_{F_k}$ é quantidade total de ocorrências de todas as M feições no contexto.

$$OP_{F_j} = \left(\frac{O_{F_j}}{\sum_{k=1}^{M} O_{F_k}} \right) \times \mu \tag{4.3}$$

Para utilizar ambas as informações no algoritmo, foi definido um fator de multiplicação da relação binária, que é dado pela Equação 4.4. Nessa equação, β^{lj} é um valor inteiro utilizado para replicar a relação binária $\langle$ `tipo l do POI, feição geográfica j`$\rangle$ no conjunto de treinamento, e ω e $(1-\omega)$ são parâmetros utilizados para definir a proporção de cada informação SP e OP que será considerada nesse processo. Utilizando o fator β^{lj}, é possível criar um conjunto de treinamento que é construído levando em conta tanto a *proporção de ocorrência* quanto a *proporção de espaço ocupado* das feições geográficas no contexto. Dessa forma, o modelo pode capturar os padrões espaciais incorporando-os nos *embeddings*.

$$\beta^{lj} = [\omega SP_{F_j} + (1-\omega)\, OP_{F_j}], \omega \in [0,1] \tag{4.4}$$

Outro aspecto espacial a ser considerado é a distância entre a feição e o POI que centraliza o contexto. No exemplo, é possível notar que o POI está mais próximo de algumas áreas verdes e de alguns prédios do que do rio. Intuitivamente, é esperado que feições que estejam mais próximas do POI sejam mais percebidas do que feições mais distantes. Desse modo, foi incluído a distância para penalizar o cálculo do fator de multiplicação da relação binária, dada pela Equação 4.5.

$$\beta\_dtc^{lj} = [\frac{\omega SP_{F_j} + (1-\omega)\, OP_{F_j}}{1 + distance^j}], \omega \in [0,1] \tag{4.5}$$

em que $distance^j$ indica a distância do ponto da feição j mais próximo do POI que centraliza o contexto.

Nesta tese, as duas versões do fator de replicação β foram utilizadas para investigar como os *embeddings* se comportam diante da aplicação da distância entre o POI e as feições, bem como diante da ausência desse atributo.

O Algoritmo 1 implementa o procedimento considerando as Equações 4.1, 4.3 e 4.4. Tal algoritmo é responsável por criar o conjunto de relações binárias entre tipos de POI e as feições geográficas usando os valores de SP, OP, ω e μ. A entrada do algoritmo é um

conjunto de POIs P, um conjunto de feições geográficas F de seu contexto e os valores dos parâmetro ω e μ. Para cada POI p_i, primeiramente seus tipos são obtidos (linhas 3). Em seguida, as feições geográficas do contexto centrado no POI p_i são recuperadas (linha 4). No passo seguinte, o espaço total (ts) ocupado pelas feições F_{p_i} no contexto é calculado, juntamente com a quantidade total de feições (tf) (linhas $5 - 6$). Para cada a feição geográfica f_j, o espaço ocupado (sf_j) e a contagem das ocorrências (of_j) dessa feição no contexto são calculados (linhas $8 - 9$). Com essas informações, são definidas a *proporção de espaço ocupado* SP_{f_j} e a *proporção de ocorrência* OP_{f_j} (linhas $10 - 11$). Em seguida, o fator de multiplicação da relação binária aug é calculado usando SP_{f_j}, OP_{f_j}, ω e μ (linha 12). Então, para cada tipo t_k do POI p_i, o fator de multiplicação da relação binária aug é utilizado para replicar a relação (t_k, f_j) no conjunto de treinamento (linha 14).

Algorithm 1: ALGORITMO GEOCONTEXT2VEC.

Input: P, F, ω, μ

Output: Lista de relações binárias

1 $tuples\_list \leftarrow empty\_list()$
2 **foreach** $p_i \in P$ **do**
3 $T_{p_i} =$ conjunto de tipos associados ao POI p_i
4 $F_{p_i} =$ feições geográficas do contexto do POI p_i
5 $ts =$ espaço total ocupado por todas as feições geográficas F_{p_i}
6 $tf =$ quantidade de feições geográficas em F_{p_i}
7 **foreach** $f_j \in F_{p_i}$ **do**
8 sf_j = espaço ocupado por f_j
9 of_j = ocorrências de f_j
10 $SP_{f_j} = (sf_j/ts) \times \mu$
11 $OP_{f_j} = (of_j/tf) \times \mu$
12 $aug = ceil(\omega SP + (1 - \omega)OP)$
13 **foreach** $t_k \in T_{p_i}$ **do**
14 adicionar a tupla (t_k, f_j) na $tuples\_list$ aug vezes
15 **return** $tuples\_list$

Os algoritmos que levam em conta a área da geometria utilizada para definir o contexto dos POIs e a distância entre as feições e o POI central são análogos ao Algoritmo 1. Para

esses algoritmos, basta substituir a linha 5 pelo cálculo da área da geometria e incluir, na linha 12, a distância entre a feição e o p_i, dividindo toda a operação para obter o valor de aug.

Além disso, a medida de espaço ocupado pelas feições geográficas muda de acordo com a tabela OSM utilizada. Para dados poligonais (por exemplo, parques, rio), usa-se a área. Para dados do tipo linha (por exemplo, estradas, rodovias), usa-se o comprimento. Para dados do tipo ponto (por exemplo, sinais de trânsito, árvores), usa-se apenas a ocorrência, já que não é possível definir uma unidade de espaço para um ponto. Esta alternância não afeta o resultado dos *embeddings* pois foi utilizada uma tabela OSM de cada vez no algoritmo, gerando assim *embeddings* para cada tipo geográfico que são concatenados no final do processo.

4.4 Representação Latente

Conforme demonstrado no Capítulo 3, o Word2Vec é um dos principais modelos empregados para gerar *embeddings* de tipos de POI. Dado que este modelo é bem estabelecido nessa área, optou-se por utilizá-lo para gerar *embeddings* de tipos de POI considerando feições geográficas. Especificamente, foi escolhida a arquitetura *Skip-Gram* (Figura 2.2, uma vez que a estratégia de *negative sampling* empregada nela contribui para tornar o algoritmo *Skip-Gram* mais rápido em comparação com o CBOW [48].

Baseado em como o Word2Vec é utilizado nos trabalhos relacionados, aplicou-se o *Skip-Gram* com o objetivo de aproximar a distribuição de probabilidade real das feições geográficas ocorrerem no contexto dos POIs a partir do conjunto de treinamento. Utilizou-se uma abordagem que emprega a função de entropia cruzada para medir a diferença entre a probabilidade aprendida e a probabilidade real. Considerando que os dados são discretos, o modelo pode ser simplificado da seguinte maneira:

$$D(\hat{y}, y) = -y_t log(\hat{y_t}) \tag{4.6}$$

Na Equação 4.6, $\hat{y}$ representa distribuição de probabilidade aprendida e y representa a distribuição de probabilidade real. $\hat{y}$ pode ser interpretada como a probabilidade predita de uma feição geográfica ocorrer dado um tipo de POI (denotado pelo t), e y_t pode ser interpretado como a probabilidade real de uma feição geográfica ocorrer dado o tipo de POI.

Pode-se definir $\hat{y_t}$ como:

$$\hat{y_t} = P(g_1, g_2, g_3, ..., g_m \,|\, t) \tag{4.7}$$

na qual, $g_1, g_2, g_3, ..., g_m$ representam feições geográficas e t representa o tipo do POI que centraliza o contexto. Na camada de saída da rede neural, a fim de transformar as saídas em probabilidades e substituir os tipos de POI por representações vetoriais, utilizou-se a função *softmax* [55]. Desta forma, a função objetivo é definida da seguinte forma:

$$Minimize\, J = -log \prod_{t=1}^{M} \frac{exp(u^{\intercal}v)}{\sum_{k=1}^{|M|} exp(u^{\intercal}v)} \tag{4.8}$$

na qual, u e v são os vetores das feições geográficas e dos tipos dos POIs, respectivamente. $|M|$ indica a quantidade das feições geográficas.

As relações binárias geradas no GeoContext2Vec são utilizadas como entrada no Word2Vec. Essas relações podem ser interpretadas como frases de duas palavras, tornando-as adequadas para o treinamento no Word2Vec, que pode ser visto como uma tarefa de predição definida da seguinte forma: a partir do tipo de POI, prever as feições geográficas do contexto desse POI.

Outro aspecto observado nos trabalhos relacionados diz respeito aos primeiros passos no uso de modelos de aprendizagem profunda em tarefas envolvendo POIs, como recomendação ou produção de modelos de linguagem de propósito geral. Diante desse cenário, um dos objetivos deste trabalho é investigar se modelos mais recentes de PLN podem ser empregados no contexto da geração de *embeddings* de tipos de POI. Para este propósito, decidiu-se utilizar o BERT, por ser um dos modelos mais empregados em diversas tarefas da área, apresentando os melhores resultados [62]. Além disso, esse modelo apresenta uma arquitetura baseada em *transformers* sendo capaz de aprender relações mais complexas entre as palavras presentes nos documentos de treinamento.

Para gerar *embeddings* de tipos de POI utilizando o BERT, foi adaptada a abordagem proposta no SpaBERT [110]. No SpaBERT o foco é a representação de entidades geoespaciais (POIs) a partir do BERT. Para isso, as entidades geoespaciais e sua vizinhança são convertidas em documentos utilizados para treinar o BERT por meio da tarefa MLM. A Figura 4.4 ilustra como as entidades geográficas são convertidas para

um documento de entrada do BERT. Dessa forma, o modelo aprende a predizer os *tokens* mascarados, aprendendo também as relações contextuais existentes entre as entidades. Devido a estratégia utilizada no SpaBERT produzir documentos a partir das relações contextuais dos POIs, percebeu-se a viabilidade de empregar uma estratégia similar para criar documentos utilizando POIs e as feições geográficas. Outras abordagens que utilizam o BERT mantêm o foco em associar documentos mais amplos à coordenadas geográficas, sendo menos similares à abordagem proposta nesta tese [20; 40; 98].

Figura 4.4: Transformação das entidades no espaço geográfico (a) para um documento BERT (b); um documento BERT mascarado (c).

Fonte: Autoria própria

No contexto dos tipos de POI e feições geográficas, os documentos são criados utilizando as relações binárias geradas pelo algoritmo GeoContext2Vec. Cada relação, no formato $\langle$`tipo de POI central, tipo de POI de contexto`$\rangle$, se torna uma sentença de duas palavras, e cada documento será composto pelas sentenças que são formadas por uma feição geográfica e todas as relações dessa feição com os tipos de POI do contexto. A Figura 4.5 ilustra a conversão de relações binárias associadas de todos os tipos de POI que se relacionado à feição *River*. Dessa forma, a partir da feição, naturalmente os documentos apresentarão todos os tipos de POI que se relacionam entre si, possibilitando ao modelo aprender as relações contextuais entre esses tipos.

Cada sentença de um documento é separada com o marcador [SEP], que indica o início e o fim de cada sentença. O treinamento é realizado utilizando a tarefa MLM. Desse modo, antes de alimentar o BERT com sequências de palavras, 15% das palavras em cada sequência são substituídas pelo símbolo [MASK] conforme demonstrado em [18]. Então, o BERT tenta prever a palavra original entre as palavras mascaradas. A Figura 4.5 ilustra um exemplo desse

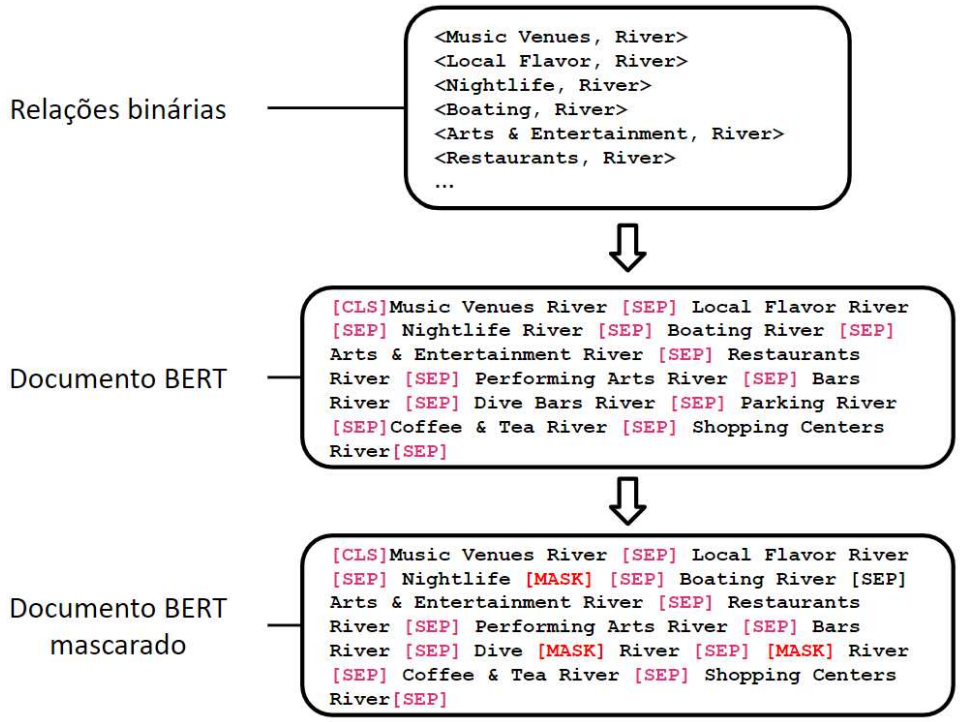


Figura 4.5: Transformação das relações binárias do GeoContext2Vec para um documento BERT mascarado.

Fonte: Autoria própria

processo de geração de um documento com 15% de mascaramento aplicado.

4.5 Considerações Finais

Este capítulo apresentou a abordagem proposta para gerar *embeddings* de tipos de POI considerando as feições geográficas de seu contexto. Para isso, foi apresentado o algoritmo GeoContext2Vec, que associa os tipos de POI às feições geográficas. O algoritmo utiliza as informações de *proporção de espaço ocupado* juntamente com a *proporção de ocorrências* de cada feição no contexto para capturar os padrões espaciais ao redor dos POIs. Também foi discutido como modelos de PLN são empregados para aprender as relações contextuais dos tipos com as feições e fornecer os *embeddings* que podem ser utilizados em tarefas relacionadas a POIs, como classificação de zonas urbanas, recomendação de POIs, entre outros.

O próximo capítulo discute a configuração experimental para validação da abordagem proposta nesta pesquisa.

Capítulo 5

Configuração Experimental

Este capítulo apresenta todo o aparato experimental utilizado para gerar os *embeddings* de tipos de POI utilizando a abordagem proposta e os *baselines*. Ele está estruturado da seguinte forma: a Seção 5.1 apresenta os dados utilizados para gerar os *embeddings*. A Seção 5.2 apresenta as ferramentas utilizadas a fim de permitir a replicação do experimento por terceiros. A Seção 5.3 apresenta os métodos de avaliação utilizados para averiguar a validade da hipótese levantada na pesquisa. A Seção 5.4 apresenta os *baselines* utilizados nesta tese. A Seção 5.5 discute a configuração de parâmetros necessários para executar a abordagem desenvolvida, os *baselines* e os modelos de PLN de maneira correta. Finalmente, a Seção 5.6 apresenta as considerações finais do capítulo.

5.1 Dados Utilizados

Para geração dos *embeddings*, foram coletados POIs do Yelp Challenge[1], versão Fevereiro 2021. Este conjunto de dados contém cerca de 160.585 POIs distribuídos em 836 cidades dos Estados Unidos e Canadá. A região de Austin, nos Estados Unidos, foi escolhida por possuir a maior quantidade de POIs (22.399) no conjunto de dados. Para o contexto geográfico, foram obtidas cópias dos mapas do OSM que compreendem a região de Austin, por meio da ferramenta *Overpass API*[2]. A Figura 5.1 ilustra a região de Austin juntamente com os POIs existentes na cidade, sendo possível verificar a existência de diversos POIs ao longo de toda

[1] https://www.yelp.com/dataset

[2] https://wiki.openstreetmap.org/wiki/Overpass_API

a cidade.

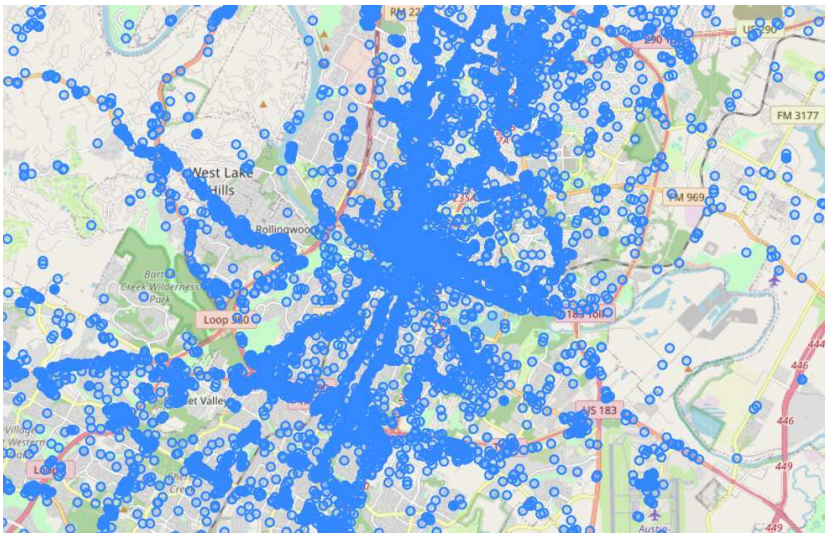


Figura 5.1: Região de Austin (EUA) e os POIs da cidade.

Fonte: Autoria própria

Em relação aos dados geográficos, as tabelas do OSM possuem vários atributos, dos quais nem todos se referem a feições geográficas relevantes para esta proposta. Para identificar as feições importantes, foi realizada uma análise exploratória por meio da documentação do OSM[3], identificando-se assim o tipo de informação associado.

Para decidir quais atributos seriam mantidos, foi considerada a definição de feição geográfica, que indica que uma feição geográfica é um objeto do mundo real que pode ser representado em um mapa e que possui uma forma geométrica. Dessa forma, os atributos que não podem ser representados a partir da forma geométrica não foram considerados, como por exemplo, data de criação, observações gerais, entre outros. A partir dessa análise, construiu-se a Tabela 5.1, que contém os atributos removidos. Vale salientar que, ao todo, existem 107 atributos na estrutura do OSM, dos quais 8 foram removidos, ou seja, $7,5\%$ de remoção.

5.2 Ferramentas

Para desenvolver a abordagem proposta, foram utilizados recursos computacionais já estabelecidos e de fácil acesso, que permitem a manipulação de dados e a geração dos *embeddings*

[3]https://wiki.openstreetmap.org/wiki/Map_features

Tabela 5.1: Atributos do OSM removidos.

Atributo	Descrição
administrative	Subdivisões de áreas/territórios/jurisdições reconhecidas pelos governos ou outras organizações para fins administrativos indicados por valores inteiros.
political	Subdivisões políticas do mapa indicados por valores inteiros
postal_code	Código postal
level	Valores que indicam o tipo de jurisdição do elemento geográfico
start_date	Data de criação do elemento geográfico
place	Identificam se o lugar é um povoado, cidade, vila, subúrbios, bairro, etc.
addresses	Fornece informações postais para um edifício ou instalação.
annotations	Fornecer mais informações sobre valores de atributos, em alguns casos, também para os usuários

de maneira rápida. As seguintes ferramentas foram selecionadas:

- **PostgreSQL**[4]: É um sistema de gerenciamento de banco de dados de fonte aberta que ganhou uma forte reputação por sua arquitetura, confiabilidade, integridade de dados, conjunto robusto de recursos, extensibilidade e a dedicação da comunidade de código aberto. Esse recurso foi utilizado para armazenar os dados dos POIs e os dados geográficos, facilitando a organização e recuperação das informações necessárias;

- **PostGIS**[5]: É uma extensão de banco de dados para dados espaciais que pode ser incorporado ao PostgreSQL. Essa ferramenta adiciona suporte para objetos geográficos, permitindo que consultas baseadas em localização sejam executadas em SQL. Esse recurso foi utilizado para recuperar, de forma rápida e precisa, as vizinhanças de POIs juntamente com as feições geográficas do contexto;

- **Gensim**[6]: É uma biblioteca python gratuita de código aberto, usada para representar documentos como vetores de maneira mais eficiente. Essa biblioteca foi utilizada pois ela implementa o Word2Vec, modelo base dessa abordagem para geração dos *embeddings*;

[4]Disponível em https://www.postgresql.org. Acesso em 20 de maio de 2024.
[5]Disponível em https://postgis.net. Acesso em 20 de maio de 2024.
[6]Disponível em https://radimrehurek.com/gensim/. Acesso em 20 de maio de 2024.

- **Tranformers**[7]: É uma biblioteca python gratuita, usada para treinar modelos baseados em *transformers*, como o BERT, ROBERTA, entre outros. Foi desenvolvida pela *Hugging Face* e fornece uma ampla gama de funcionalidades para pré-processamento de texto, treinamento de modelos, carregamento de modelos pré-treinados e inferência. Nessa tese, essa biblioteca foi utilizada para treinamento de modelos baseados no BERT.

5.3 Tarefas de Avaliação

Esta seção apresenta as tarefas utilizadas para avaliar os *embeddings* gerados com a abordagem proposta. Primeiramente, são descritas as tarefas relacionadas à análise da similaridade contextual capturada nos *embeddings*, comparando-a com a similaridade definida por pessoas, hierarquias de tipos de POI e a visualização no espaço vetorial latente. Em seguida, é apresentada uma tarefa relacionada a POIs, que consiste na classificação de zonas urbanas utilizando os *embeddings* dos tipos.

5.3.1 Análise de Similaridade

Para responder a $\mathbf{QP}_2$ (*Embeddings* de tipos de POI gerados a partir de relações contextuais com feições geográficas indicam a similaridade dos tipos?), foram utilizadas duas tarefas produzidas por Yan *et al*. [103], uma análise de similaridade considerando a hierarquia de tipos e uma análise de similaridade a partir da visualização dos *embeddings* no espaço vetorial latente.

Conforme mencionado no parágrafo anterior, Yan *et al*. [103] propuseram duas tarefas com a participação de 25 voluntários. A primeira tarefa, denominada BHE (*Binary Hit Evaluation*), solicitava que voluntários escolhessem o tipo de POI considerado como mais diferente em uma lista de três tipos. Por exemplo, em uma lista composta pelos tipos `Dentista`, `Educação` e `Ortodontista`, os voluntários deveriam votar no tipo considerado mais diferente. Desse modo, sendo o tipo com maior número de votos considerado o mais diferente. Consequentemente, os pares de tipos menos votados eram considerados

[7]Disponível em https://huggingface.co/docs/transformers/v4.18.0/en/index. Acesso em 20 de maio de 2024.

como os mais similares. Para essa tarefa, foram criadas 77 listas de três tipos utilizando 144 tipos distintos.

Para avaliar se a similaridade contextual entre os tipos de POI correspondia à similaridade percebida pelos voluntários, foi empregada a mesma tarefa, porém utilizando os *embeddings* dos tipos e o cálculo da similaridade do cosseno. Para cada lista de três tipos, calculava-se a similaridade do cosseno entre cada par de tipos. Por exemplo, se a similaridade do cosseno entre os *embeddings* dos tipos `Dentista` e `Ortodontista` fosse a mais alta, o tipo identificado como mais diferente seria `Educação`. Essa análise foi realizada em cada uma das 77 listas. Ao final, contabilizava-se quantos tipos identificados com base nos *embeddings* correspondiam aos tipos mais votados pelos voluntários.

A segunda tarefa, denominada RHE (*Ranking Hit Evaluation*), consistia na identificação do nível de similaridade entre dois tipos. Para isso, os voluntários deveriam indicar numericamente o valor de similaridade entre dois tipos. Por exemplo, dados os tipos `bar` e `clube noturno`, os voluntários deveriam selecionar um valor entre 1 e 7 (quão maior o número, maior a similaridade e vice-versa). No fim, foi calculado a média dos 25 votos para cada par, gerando o valor de similaridade final. Nessa tarefa foram utilizados 70 pares de tipos e 66 tipos distintos.

Uma terceira avaliação foi realizada utilizando a estrutura hierárquica dos tipos (hierarquia do Yelp). Na hierarquia do Yelp[8], os tipos são agrupados em 21 categorias gerais e 1.275 categorias especializadas, distribuídas em quatro níveis. A Tabela 5.2 apresenta a distribuição dos tipos ao longo desses níveis da hierarquia.

Tabela 5.2: Distribuição dos tipos de POI na hierarquia do Yelp.

Nível da Hierarquia	**Tipos de POI por Nível**
01	21
02	857
03	404
04	17

Yan *et al*. [103] também consideraram a hierarquia, e aplicaram os métodos de Wu &

[8]Disponível em https://www.yelp.ca/developers/documentation/v3/category_list. Acesso em 20 de maio de 2024.

Palmer [96] e Leacock & Chodorow [35], conforme demonstrado na Seção 2.3.1. Nesse caso, os termos t_1 e t_2, presentes nos métodos, passam a ser os tipos de POI. Por meio desses métodos, calculou-se a similaridade entre todos os pares de tipos da taxonomia, e com base nos valores de similaridade, foi gerado um ranque para cada tipo, ordenado dos mais similares aos menos similares. Uma vez que tais métricas permitem empates de similaridade entre vários tipos, cada posição do ranque pode conter uma lista dos tipos igualmente similares.

Para investigar se as relações contextuais dos tipos de POI refletem a similaridade dos tipos indicados na hierarquia, foi utilizada a similaridade do cosseno para construir o ranque dos tipos mais similares para cada tipo de POI. A partir desses dois ranques, foi aplicado o MRR para determinar em qual posição o tipo mais similar indicado pela similaridade do cosseno estava no ranque construído com a hierarquia. Por exemplo, se usando a similaridade do cosseno, o tipo mais similar a `Restaurante Chinês` for `Restaurante Francês`, e este tipo estiver na primeira posição da lista de `Restaurante Chinês` construída com a hierarquia, o valor de MRR é 1. Nessa avaliação, foi calculado a média de todos os MRR para cada tipo de POI.

Também foi empregada a visualização de *embeddings* para verificar se as relações contextuais dos tipos de POI refletem a similaridade dos tipos. A visualização é uma prática muito comum no campo do PLN. Seu objetivo é permitir que cientistas de dados ou desenvolvedores compreendam como os *embeddings* se relacionam no espaço vetorial e se existem grupos naturais considerando as relações contextuais das palavras. Para essa tarefa, foi utilizado o t-SNE (*t-distributed Stochastic Neighbor Embedding*), uma técnica de redução de dimensionalidade que visa modelar as similaridades entre pares de pontos nos dados de entrada, tentando preservar as similaridades na visualização final [82].

Um dos parâmetros ajustáveis do t-SNE é a perplexidade, que determina como equilibrar a atenção entre os aspectos locais e globais dos *embeddings*. Esse parâmetro funciona como um palpite sobre o número de vizinhos próximos que cada ponto possui. De acordo com a recomendação do próprio autor do t-SNE, a perplexidade deve estar na faixa entre 5 e 50. Nessa pesquisa, foi escolhido experimentalmente o valor 15.

Outro parâmetro é o número de iterações executadas durante a redução dimensional. Não existe um número fixo de iterações que seja universalmente recomendado. No entanto, geralmente valores entre 1.000 e 5.000 são comuns em muitas implementações do t-SNE.

Nessa pesquisa, optou-se por executar 2.500 iterações.

Para a visualização, foram selecionados os *embeddings* referentes aos tipos de POI do segundo nível da hierarquia. Essa decisão foi tomada devido à considerável quantidade de tipos diferentes nesse nível (857), com o objetivo de investigar possíveis agrupamentos diante dessa diversidade. Adicionalmente, para cada tipo, também foi obtida a informação do tipo pai. Com base nessas informações, realizou-se a redução dos *embeddings* dos tipos do segundo nível, e a visualização apresenta cada um desses tipos com a cor correspondente ao tipo pai. Dessa forma, torna-se possível analisar se os *embeddings* dos tipos de POI filhos de um mesmo pai estão próximos ou não no espaço vetorial.

É importante mencionar que o desempenho do t-SNE depende dos hiperparâmetros selecionados, como a *perplexity* e a *learning rate*. A escolha inadequada desses parâmetros pode levar a resultados insatisfatórios ou difíceis de interpretar. Além disso, a redução de *embeddings* acarreta em uma perda natural das informações incorporadas.

5.3.2 Classificação de Zonas Urbanas

Para responder à $\mathbf{QP}_3$ (Modelos que usam *embeddings* de tipos de POI gerados com feições geográfica são melhores que modelos que utilizam *embeddings* de tipos de POI gerados com outros dados geográficos?), foram aplicados os *embeddings* produzidos nessa abordagem em uma tarefa de classificação de zonas. Estudos anteriores [61; 94; 104; 106] indicaram que uma zona pode ser representada calculando a média ponderada de todos os *embeddings* de tipos de POI dentro da zona. Consequentemente, os *embeddings* das zonas são derivados da média de todos os *embeddings* dos tipos de POI associados a cada zona. Os *embeddings* que representam a j-ésima zona podem ser definidos matematicamente da seguinte forma:

$$\text{Zona}_j = \sum_{i=1}^{N} \frac{\text{POI\_T}_{i,j}}{N} \tag{5.1}$$

em que $\text{POI\_T}_{i,j}$ se refere ao *embedding* do i-ésimo tipo de POI da j-ésima zona, e N é o número de tipos de POI na j-ésima zona.

Na tarefa de classificação de Zonas, o algoritmo de Floresta Aleatória (*Random Forest*) tem sido amplamente adotado [61; 94; 104; 106]. Floresta Aleatória é um algoritmo de aprendizado em conjunto composto por múltiplas árvores de decisão não correlacionadas.

Essa abordagem em conjunto ajuda a mitigar problemas como variáveis sobrepostas e correlacionadas [9]. Esse algoritmo é utilizado para aprender a relação entre os *embeddings* de zonas e suas respectivas categorias.

5.4 *Baselines*

Como base para comparação, foram selecionados o método ITDL proposto por Yan *et al*. [103]. e o método do *Shortest Path* (Caminho Mais Curto), conforme descrito por Yao *et al*. [106]. O ITDL foi selecionado por ser uma referência na área de geração de *embeddings* de tipos de POI, servindo como base para diversos trabalhos. O *Shortest Path* foi selecionado por ser uma estratégia amplamente utilizada para gerar *embeddings* de tipos de POI para a tarefa de classificação de zonas urbanas, tarefa que também foi abordada nesta tese.

O método ITDL baseia-se na relação de vizinhança entre os POIs para gerar *embeddings* de seus tipos, sendo referência em diversos estudos na área. Como detalhado na Seção 3.1, o ITDL utiliza o contexto de vizinhança para estabelecer relações binárias, expressas como $\langle$`tipo de POI central`,`tipo de POI de contexto`$\rangle$. Nesta abordagem, os autores empregam a popularidade, obtidos a partir de dados de *check-in*, e a unicidade dos tipos para capturar a "importância" dos POIs dentro do contexto. Eles partem da hipótese de que POIs com um número significativo de *check-ins* exercem maior influência no contexto, indicando que são locais frequentemente visitados. Além disso, eles consideram a raridade dos tipos de POIs, sugerindo que locais menos comuns em um contexto também podem ter uma influência significativa. Por exemplo, é esperado que tipos de POIs como "estações de polícia" e "shoppings" apareçam menos frequentemente em uma vizinhança, mas exercem muita influência no contexto. O Algoritmo 2 ilustra o funcionamento do método ITDL [103].

Tal algoritmo recebe como entrada uma lista de POIs L, contendo seus nomes N, coordenadas geográficas G e tipos T. Também são fornecidos o número da caixa discretas s, a largura da caixa h, e o parâmetro de balanceamento σ. Este último determina se o algoritmo dará mais peso à popularidade A, calculada com base nos *check-ins*, ou à unicidade U, calculada com base na ocorrência de cada tipo na vizinhança. O objetivo principal do algoritmo é ampliar as relações binárias entre os tipos de POI presentes na mesma vizinhança, especialmente aqueles com mais *check-ins* ou que são mais raros na região. Essa abordagem é

Algorithm 2: Algoritmo ITDL [103].

Input: $L = (N, G, T), s, h, \sigma$

Output: Lista de relações binárias

1 $tuples\_list \leftarrow empty\_list()$
2 **foreach** $l_i \in L$ **do**
3 $T_{l_i} =$ conjunto de tipos associados ao POI l_i
4 **for** $n \leftarrow 0; n < s; n++$ **do**
5 $sc =$ total de *check-ins* dentro da caixa n
6 $so =$ total de tipos de POI dentro da caixa n
7 **foreach** $l_j \in L$ **do**
8 T_{lj} = tipos do POI l_j **if** $nh \leq d(l_i, l_j) < (n+1)h$ **then**
9 **foreach** $t_{ki} \in T_{li}$ **do**
10 **foreach** $t_{kj} \in T_{lj}$ **do**
11 cc = *check-ins* de t_{kj}
12 co = ocorrências de t_{kj}
13 $A = -\log_2(1 - cc/sc)$
14 $U = -\log_2(co/so)$
15 $aug = ceil(\sigma A + (1 - \sigma)U)$
16 adicionar a tupla (t_k, f_j) na $tuples\_list$ aug vezes
17 **return** $tuples\_list$

evidenciada nas linhas 13 e 14 do algoritmo, onde os dados de *check-ins* e ocorrências são utilizados para calcular os valores de A e U. Em seguida, na linha 15, esses dois valores são combinados para determinar o valor de aug, utilizado para replicar as relações binárias entre os tipos de POI em um conjunto de treinamento.

O método *Shortest Path* define relações sequenciais entre POIs para gerar *embeddings* de seus tipos. Essa abordagem é amplamente utilizada em trabalhos na área de classificação de zonas urbanas [106; 112]. De acordo com Yao *et al*. [106], para cada zona urbana, o seguinte algoritmo deve ser executado para construir o caminho mínimo entre POIs:

1. Primeiramente, calcula-se a distância euclidiana para cada par de POIs <P, P> e seguida seleciona-se o par que apresenta a maior distância como os pontos inicial e final

do caminho (denotados por P_s e P_e). Dessa forma, após essa operação, a ordem sequencial do caminho mínimo é P_s, P_e, e os demais POIs estão em uma fila de espera;

2. Em seguida, a tarefa passa a ser a inserção e atualização do caminho mínimo com um POI que ainda não faz parte do caminho. Nesse passo, uma estratégia gulosa é adotada para assegurar que cada POI inserido seja aquele que faz com que o comprimento do caminho seja o menor possível;

3. O passo anterior é repetido até que todos os POIs da fila sejam inseridos no caminho.

Dessa forma, se existirem K zonas urbanas, o método resultará em K caminhos. Durante o treinamento, os POIs são substituídos por seus tipos e são utilizados como entrada no modelo Word2Vec. Conforme mencionado pelos autores [106], é empregada a janela do Word2Vec, que parte do primeiro POI e avança POI a POI no caminho, para definir a vizinhança dos tipos. Com essa estratégia, os *embeddings* de tipos são gerados.

5.5 Configuração de Parâmetros

Esta seção apresenta a configuração necessária para executar os algoritmos GeoContext2Vec, ITDL e *Shortest Path*. Também são apresentadas as configurações utilizadas para geração de *embeddings* utilizando os modelos Word2Vec e BERT. Os algoritmos e modelos foram executados em uma máquina equipada com uma CPU Intel Core $i7 - 12700F$ de $4,90GHz$, acoplada a $32GB$ de memória, placa de vídeo NVIDIA GeForce RTX 4090 acoplada com $24GB$ de memória, operando no sistema operacional Ubuntu.

5.5.1 Parâmetros do GeoContext2Vec

Para executar o GeoContext2Vec, é necessário definir valor de ω, que indica a proporção de SP e OP a ser considerada no algoritmo (ver Algoritmo 1 na seção 4.3), e de μ, que permite mudar a escala dos valores de SP e OP. Para o valor de μ, estabeleceu-se através de experimentos o número 20. Esse valor multiplica diretamente as frações relacionadas à SP e OP, fazendo um mapeamento das proporções conforme a Tabela 5.3.

No algoritmo GeoContext2Vec, os valores de SP e OP são usados em uma função matemática que arredonda os valores para cima. Isso resulta em um aumento de uma unidade

Tabela 5.3: Relação do valor $\mu = 20$ multiplicado pela proporção x.

Intervalo da Proporção	**Intervalo da multiplicação**
$0,00 < x \leq 0,05$	$0 < x * \mu \leq 1$
$0,05 < x \leq 0,10$	$1 < x * \mu \leq 2$
$0,10 < x \leq 0,15$	$2 < x * \mu \leq 3$
$0,15 < x \leq 0,20$	$3 < x * \mu \leq 4$
...	
$0,90 < x \leq 0,95$	$18 < x * \mu \leq 19$
$0,95 < x \leq 1,00$	$19 < x * \mu \leq 20$

a cada intervalo de 5%. A partir da Tabela 5.3, é possível notar que as proporções x das feições geográficas que apresentam valores próximos são mapeadas para o mesmo valor de replicação. Por exemplo, na tabela fornecida, as feições com proporções entre 0% e 5% são mapeadas para o valor 1, o que significa que as relações binárias entre os tipos de POI e essas feições serão inseridas uma vez no conjunto de treinamento. Usar valores maiores para μ pode resultar em um mapeamento mais preciso das proporções. Por exemplo, se $\mu = 100$, cada unidade de porcentagem seria mapeada para um valor inteiro de replicações, ou seja, 1% para 1 unidade, 2% para 2 unidades, e assim por diante. No entanto, isso também aumentaria o tamanho do conjunto de treinamento, o que poderia levar a tempos de treinamento mais longos. Por outro lado, usar valores menores de μ pode resultar em um mapeamento mais amplo, gerando menos replicações no conjunto de treinamento, mas com a perda de informações sobre os padrões espaciais. Observou-se que, com $\mu = 20$, ocorreriam no máximo 20 replicações, com um mapeamento que mantém uma variação de 5%.

Para o parâmetro ω, foram definidos os valores no intervalo $[0, 1]$ com incrementos de $0,1$, pois assim seria possível gerar modelos que consideram exclusivamente a *proporção de ocorrência* das feições geográficas (OP); modelos que consideram exclusivamente a *proporção de espaço ocupado* (SP); e modelos que combinam esses dois componentes (OSP). Essa configuração resulta em 11 modelos, um para cada valor de ω.

Além disso, foram criados modelos considerando contextos de raio no intervalo de $100m$ à $900m$ com um passo de $100m$. O objetivo dessa variação é a de explorar o comportamento dos *embeddings* em contextos de tamanhos diferentes. Considerando que são gerados 11

modelos para cada valor de ω, e que foram definidos nove contextos de tamanhos diferentes, isso resulta na produção de 99 modelos, sendo 11 modelos gerados para cada tamanho de contexto.

Conforme discutido no Capítulo 4, existem duas formas de calcular SP. A primeira utiliza uma proporção relativa, onde as áreas de todas as feições geográficas de um contexto são somadas para encontrar a proporção de cada feição no contexto (descrita na Equação 4.1). A segunda forma utiliza a área da figura geométrica do contexto (descrita na Equação 4.2). Considerando os dados do OSM, apenas a tabela de polígonos apresenta a propriedade de área, o que permite uma comparação direta com a figura geométrica do contexto. As tabelas de linhas apresentam a propriedade de comprimento. Nesse caso, foi calculado o comprimento máximo que cada feição ocupa em cada contexto, e esse valor foi utilizado como denominador na equação. Portanto, foram gerados modelos considerando essas duas formas de calcular SP, denominados *GeoC2Vec_Rel* e *GeoC2Vec_Abs*, respectivamente.

Também foi utilizada a distância para penalizar o fator de multiplicação das relações binárias (descrito na Equação 4.5). Nesse caso, não foi empregada a distância real entre as feições, pois isso resultaria em uma redução acentuada do valor de β para contextos de raio maior. Por exemplo, se uma feição está a $500m$ de distância de um POI e o valor de β é 20 (valor máximo de considerando $\mu = 20$), a fração resultante seria $\frac{20}{500} = 0,04$. Como $\beta < 1$, a relação binária entre essa feição e os tipos do POI do contexto não seria inserida no conjunto de treinamento. Em vez disso, em cada contexto, calcula-se uma distância relativa ao tamanho do contexto, conforme a equação a seguir:

$$distance = \frac{real\_distance}{context\_radius} \tag{5.2}$$

Assim, as feições que estão muito distantes do POI terão o fator β reduzido no máximo pela metade. O modelo resultante dessa abordagem foi denominado *GeoC2Vec_Dtc*.

A execução do algoritmo GeoContext2Vec consome em média 4 minutos para produzir os quatro conjuntos de treinamento para um raio fixo (relações entre os tipos de POI e dados poligonais, lineares e pontuais). Esse tempo foi obtido utilizando recursos de paralelismo do python e mecanismos de recuperação mais eficientes do PostgreSQL como *views* materializadas.

5.5.2 Parâmetros do ITDL

Para executar o algoritmo ITDL, é necessário definir a quantidade de caixas discretas (s), a largura da caixa (h) e os valores de σ. O parâmetro h foi definido como $h = 100m$, conforme especificado por Yan *et al.* [103]. Esse valor faz com que a vizinhança apresente um formato anelar, com espessura de $100m$. Assim, se uma vizinhança começa em $0m$, terminará em $100m$ no formato circular, caso a vizinhança se inicia a $100m$ de distância, terminará a $200m$, e assim sucessivamente.

O parâmetro σ varia no intervalo de $[0, 1]$ com incrementos de $0,1$. Com isso, foram gerados modelos que consideram exclusivamente a unicidade (U) de cada tipo de POI, modelos que consideram exclusivamente a popularidade (A) dos tipos, e modelos que consideram combinações desses dois elementos (UA). Essa configuração resulta em 11 modelos, um para cada valor de σ.

Para a quantidade de caixas discretas, definiu-se s no intervalor $[0, 8]$, com passo de 1. Dessa forma, são criadas caixas em torno de POIs que atingem o raio de $100m$, $200m$, até $900m$ de distância do POI central (mesma distância utilizada no GeoContext2Vec). Esse valor foi definido para que fosse possível comparar as duas abordagens quando o tamanho do contexto cresce nas mesmas proporções. Semelhantemente ao GeoContext2Vec, nesse cenário são produzidos 99 modelos, sendo 11 modelos gerados para cada valor de caixa discreta.

5.5.3 Parâmetros do *Shortest Path*

Para executar o algoritmo *Shortest Path*, é necessário obter zonas urbanas que contenham POIs e também determinar quantos vizinhos serão considerados ao gerar os *embeddings* dos tipos. Como os POIs utilizados nesta tese são da região de Austin, foram obtidos dados das zonas urbanas de Austin a partir de fontes governamentais oficiais[9]. Esses dados incluem 406 zonas, e para cada zoneamento, foram mapeados os POIs do Yelp. Em seguida, foi executado o algoritmo *Shortest Path*, resultando em 406 caminhos (O Apêndide A trás mais detalhes sobre a configuração das zonas urbanas).

No trabalho de Yao *et al.* [106], uma vizinhança de 10 POIs foi definida para cada POI

[9]Disponível em https://data.austintexas.gov. Acesso em 20 de maio de 2024.

do caminho, o que resulta em cinco POIs predecessores e cinco sucessores utilizados como contexto para geração dos *embeddings*. Para relacionar os *embeddings* gerados por essa abordagem com os *embeddings* gerados pelo GeoContext2Vec, foi estabelecida uma relação entre os tamanhos do contexto do GeoContext2Vec e a quantidade de vizinhos do *Shortest Path*. Assim, foi decidido que, para um contexto de $100m$, seria utilizado os *embeddings* produzidos com uma vizinhança de $k = 10$. Para um contexto de $200m$, seria utilizado os *embeddings* produzidos com uma vizinhança de $k = 20$, e assim por diante, até $k = 90$ (correspondendo ao tamanho do contexto de $900m$). Com essa definição, foram criados 9 modelos com a abordagem *Shortest Path*.

5.5.4 Configuração do Word2Vec

O treinamento no Word2Vec requer a definição do tamanho da janela e da dimensão dos vetores. Para o tamanho dos vetores, Yan *et al*. [103] demostraram que vetores de dimensão 70 geram os melhores *embeddings* de tipos de POI.

O tamanho da janela foi adaptado conforme a abordagem utilizada. Para o GeoContext2Vec e ITDL, foi utilizado o tamanho 1, pois o conjunto de treinamento é composto por relações binárias que podem ser vistas como uma sentença de duas palavras. Ou seja, cada palavra central (neste caso, tipo de POI central) tem apenas uma palavra de contexto (neste caso, tipo de POI de contexto ou feição geográfica do contexto) e vice-versa (para mais detalhes sobre a janela do Word2Vec, veja a Seção 2.2.1).

Para a abordagem *Shortest Path*, o tamanho da janela reflete a quantidade de vizinhos a serem considerados como contexto. Por exemplo, se a quantidade de vizinhos é 10, deve-se utilizar uma janela de tamanho 5. Como mencionado anteriormente, são considerados K vizinhos no intervalo de $[10, 90]$ com passo de 10. Esses valores foram então utilizados na janela do Word2Vec para definir o contexto de cada POI do caminho.

Conforme descrito na Seção 4.2, as feições geográficas do OSM estão distribuídas em quatro tabelas distintas. Portanto, o GeoContext2Vec foi executado considerando cada tabela separadamente. Uma vez que os dados de treinamento são gerados, ocorre a etapa de treinamento dos modelos. Em outras palavras, foram criados modelos especializados para cada configuração, como ruas, prédios, parques, rios, sinais de trânsito, fontes, árvores, postes, entre outros. Essa abordagem permite que cada modelo capture as relações contextuais

entre os tipos de POI e as feições de maneira mais precisa, reduzindo a perda de informação ao considerar um modelo para cada tabela. Após o treinamento, para cada tipo de POI, os *embeddings* gerados pelos quatro modelos são concatenados.

O tempo de treinamento com o Word2Vec consome em média $17,26$ minutos para os quatro conjuntos de treinamento de um raio fixo que são compostos pelos dados das relações dos tipos de POI com as feições poligonais, lineares e pontuais.

5.5.5 Configuração do BERT

Com o intuito de responder à $\mathbf{QP}_4$ (Modelos que utilizam *embeddings* de tipos de POI produzidos com modelos recentes de PLN são melhores que modelos que utilizam *embeddings* de tipos de POI produzidos com modelos clássicos?). Também foi realizada a geração dos *embeddings* de tipos de POI utilizando o BERT e os conjuntos de treinamento do GeoContext2Vec, ITDL e *Shortest Path*. Para isso, foi empregado o DistilBert base[10], um modelo *Transformer* menor, mais rápido e leve, treinado por meio da destilação do BERT *base*, que possui vetores de dimensão 768.

Como mencionado na Seção 4.4, o treinamento foi conduzido transformando as relações binárias em documentos e mascarando 15% dos *tokens* em cada documento, conforme [18]. Para realizar o treinamento, foi alocado 80% dos documentos para treino e 20% para validação. O número de épocas foi definido como 3, consistente com os autores do BERT [18]. Além disso, foram utilizados os pesos da versão "distilbert-base-uncased" para aproveitar o conhecimento prévio do modelo sobre os tipos de POI.

Assim como no Word2Vec, foi treinado um modelo DistilBert para cada conjunto de treinamento de cada tabela OSM. No fim, os *embeddings* dos tipos de modelo foram concatenados para formar o *embedding* final de dimensão 3072.

O treinamento com o DistilBERT consome em média 880 minutos para os quatro conjuntos treinamentos (dados poligonais, lineares e pontuais) produzidos pelo GeoContext2Vec para um raio fixo.

[10]Disponível em https://huggingface.co/docs/transformers/model_doc/distilbert. Acesso em 20 de maio de 2024.

5.6 Considerações Finais

Este capítulo detalhou todos os arranjos experimentais utilizados para gerar os *embeddings* dos tipos de POI utilizando o GeoContext2Vec e as abordagens *baselines*. Foram discutidas as fontes de dados e as ferramentas empregadas para obter informações sobre os POIs e as feições geográficas. Também foram apresentadas as tarefas de avaliação e como elas fornecem informações sobre os *embeddings*.

Também foi discutido o conjunto de parâmetros de cada abordagem, assim como a quantidade de modelos gerados com base nas combinações desses parâmetros. Por fim, foi demonstrado como os modelos Word2Vec e BERT foram utilizados para obter os *embeddings* de tipos de POIs.

O capítulo seguinte descreve os resultados desta pesquisa.

Capítulo 6

Resultados

Este capítulo apresenta os resultados obtidos nesta pesquisa. Ele está estruturado da seguinte forma: as Seções 6.1, 6.2 e 6.3 descrevem as análises realizadas para verificar se os *embeddings* do GeoContext2Vec refletem a similaridade dos tipos. A Seção 6.4 apresenta a visualização dos *embeddings* do GeoContext2Vec e debate se existe a formação de grupos de similaridades considerando as relações contextuais com as feições geográficas. A Seção 6.5 apresenta os resultados de uma tarefa de classificação de zonas urbanas utilizando os *embeddings* do GeoContext2Vec e os compara com abordagens *baseline*. Por fim, a Seção 6.6 apresenta as considerações finais do capítulo, demonstrando quais questões de pesquisa foram contempladas a partir dos resultados obtidos.

6.1 Análise de Similaridade com BHE

Esta seção apresenta os resultados da utilização dos *embeddings* do GeoContext2Vec e dos *baselines* na tarefa BHE. Foi explorado se as relações contextuais entre os tipos de POI e as feições geográficas refletem a similaridade dos tipos. Também foi investigado como as propriedades espaciais influenciam na capacidade dos *embeddings* de capturar as diferenças contextuais dos tipos de POI. Além disso, foi investigado se os modelos mais recentes permitem obter *embeddings* mais precisos do que os modelos clássicos de PLN.

6.1.1 BHE com Word2Vec

Conforme descrito no capítulo anterior, os *embeddings* de tipos de POI foram utilizados na tarefa BHE para verificar se as relações contextuais dos tipos de POI com as feições geográficas refletem a similaridade dos tipos. Nesse caso, foram empregados todos os *embeddings* dos três modelos GeoContext2Vec (*GeoC2Vec_Rel*, *GeoC2Vec_Abs* e *GeoC2Vec_Dtc*). Os resultados obtidos nessa tarefa estão presentes nas Figuras 6.1 e 6.2.

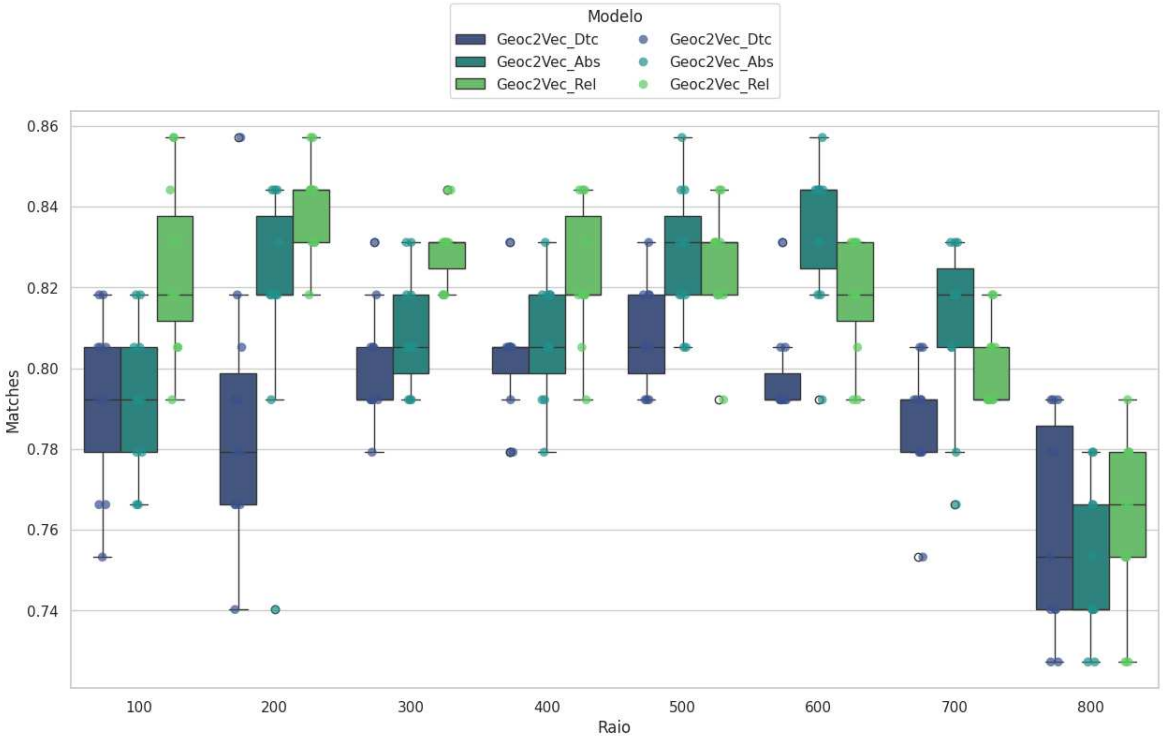


Figura 6.1: Resultados dos três modelos GeoContext2Vec na tarefa BHE por valor de raio.

Observando a distribuição dos resultados para cada valor de raio (Figura 6.1), nota-se que os tipos de POI mais diferentes indicados pelos *embeddings* coincide em uma faixa de aproximadamente 74% a 86% de *matching* com a votação dos voluntários. Isso sugere que as relações contextuais dos tipos de POI com as feições refletem, em certo grau, a similaridade dos tipos. Esse resultado aponta que tipos de POI similares, de acordo com a opinião humana, também apresentam similaridade nas feições geográficas de seus contextos. Conforme detalhes que constam no Apêndice A, 96% do conjunto de teste dessa tarefa utiliza tipos irmãos na hierarquia. Em outras palavras, os tipos de POI irmãos desse conjunto também compartilham similaridade contextual em relação às feições geográficas.

Analisando a diferença nas distribuições entre os três modelos, observa-se que os *embeddings* do modelo *GeoC2Vec_Rel*, que utiliza a área relativa do contexto, tendem a apresentar valores menos dispersos e, para metade dos raios (de $100m$ a $400m$), são os mais altos. Por outro lado, os *embeddings* do *GeoC2Vec_Abs* têm distribuições mais baixas para raios menores. Esse comportamento reflete a relação entre a área absoluta do contexto e a área das feições.

Em contextos menores, é natural que a densidade de feições seja menor e que a área das feições seja limitada pelo próprio contexto. Considerando a divisão que ocorre entre a área de cada feição e a área absoluta do contexto, é provável que os valores resultantes sejam mais próximos de 0, resultando em poucas replicações no conjunto de treinamento. Consequentemente, os *embeddings* capturam menos diferenças contextuais entre os tipos de POI.

Por outro lado, o modelo relativo sempre considera a soma da área ocupada por todas as feições do contexto, relacionando essa área com as áreas de cada feição do contexto. Nesse caso, a divisão que ocorre entre cada feição e a área gera valores que estão mais uniformemente distribuídos entre 0 e 1. Isso permite que as replicações ocorram de forma a evidenciar a diferença entre as feições que ocupam mais espaço daquelas que ocupam menos.

O comportamento mencionado anteriormente é mais evidente na distribuição associada ao raio de $100m$. Entretanto, à medida que o raio aumenta, os contextos dos POIs se tornam naturalmente mais densos e a diferença entre a área absoluta e a área das feições diminui, uma vez que as feições passam a ser completamente inseridas nos contextos. Isso faz com que as replicações se tornem mais representativas, permitindo que os *embeddings* capturem melhor as diferenças contextuais entre os tipos de POI.

Outro fator que influencia os resultados do modelo absoluto está relacionado aos dados lineares do OSM (como ruas e avenidas). Para essas feições, não é possível obter suas áreas, mas apenas seus comprimentos. Conforme mencionado na Seção 5.5.1, foi utilizado o comprimento máximo de cada feição observado nos contextos. Naturalmente, o resultado da divisão dependerá da diferença entre o comprimento de uma feição em análise e o comprimento máximo daquela feição. Essa divisão pode gerar valores próximos de 0, resultando em um número menor de replicações no conjunto de treinamento e, consequentemente, impossibilitando os *embeddings* de capturar com maior precisão as relações contextuais dos POIs.

Além disso, como não é possível obter as áreas das ruas, existe sempre uma penalização para o cálculo das feições poligonais, que considera a área da circunferência ($\pi * r^2$), incluindo naturalmente a área das ruas e avenidas. No entanto, essa limitação é inerente ao OSM e não à abordagem utilizada.

Observando os resultados do modelo baseado na distância, percebe-se que as distribuições para todos os valores de raio são inferiores aos dos demais modelos, com exceção do raio de $800m$. Isso sugere que a estratégia de reduzir o valor de β com base na distância resultou na perda de informações sobre as relações contextuais dos tipos de POI com as feições geográficas.

Esse padrão pode ser atribuído ao decrescimento linear do valor da distância. Nota-se que o maior número de replicações possíveis ocorre quando uma feição está no centro do contexto, resultando em uma distância relativa de 0, e o denominador da equação terá valor 1. Para o raio de $100m$, os resultados são semelhantes aos do modelo absoluto, pois as feições não estão completamente inseridas no contexto e são penalizadas quando se afastam da origem.

No entanto, à medida que o contexto aumenta, apenas as feições muito próximas do centro são menos penalizadas, enquanto aquelas mais próximas da borda têm maiores reduções de β. Esse padrão é observado na maioria das distribuições até aproximadamente $700m$, onde os valores permanecem em torno de 78% a 81% de *matching* (próximos da distribuição de $100m$). Esses resultados indicam que a distância não contribuiu significativamente para que os *embeddings* incorporassem as similaridades contextuais dos tipos de POI. Além disso, observa-se que a área das feições já desempenha um papel relacionado à distância. Se uma feição possui uma grande área de ocupação no contexto, é provável que esteja próxima ao centro e, portanto, receba um maior número de replicações no conjunto de treinamento.

Analisando as distribuições considerando os valores do parâmetro ω (Figura 6.2), observa-se que, para o modelo *GeoC2Vec_Rel* os resultados aumentam à medida que o valor de ω aumenta. Isso sugere que os *embeddings* produzidos ao considerarem em maior proporção o espaço ocupado pelas feições são capazes de capturar com mais precisão as diferenças contextuais entre os tipos de POI, mesmo que haja variação do tamanho do raio. Por outro lado, as distribuições dos valores que dão mais ênfase à propriedade de ocorrência das feições apresentam uma dispersão maior, sugerindo que, para diferentes tamanhos de

raio, os valores variam mais. A mediana de valor mais alto é observada quando $\omega = 0,9$, ou seja, quando se considera 90% da proporção do espaço ocupado e 10% das ocorrências das feições.

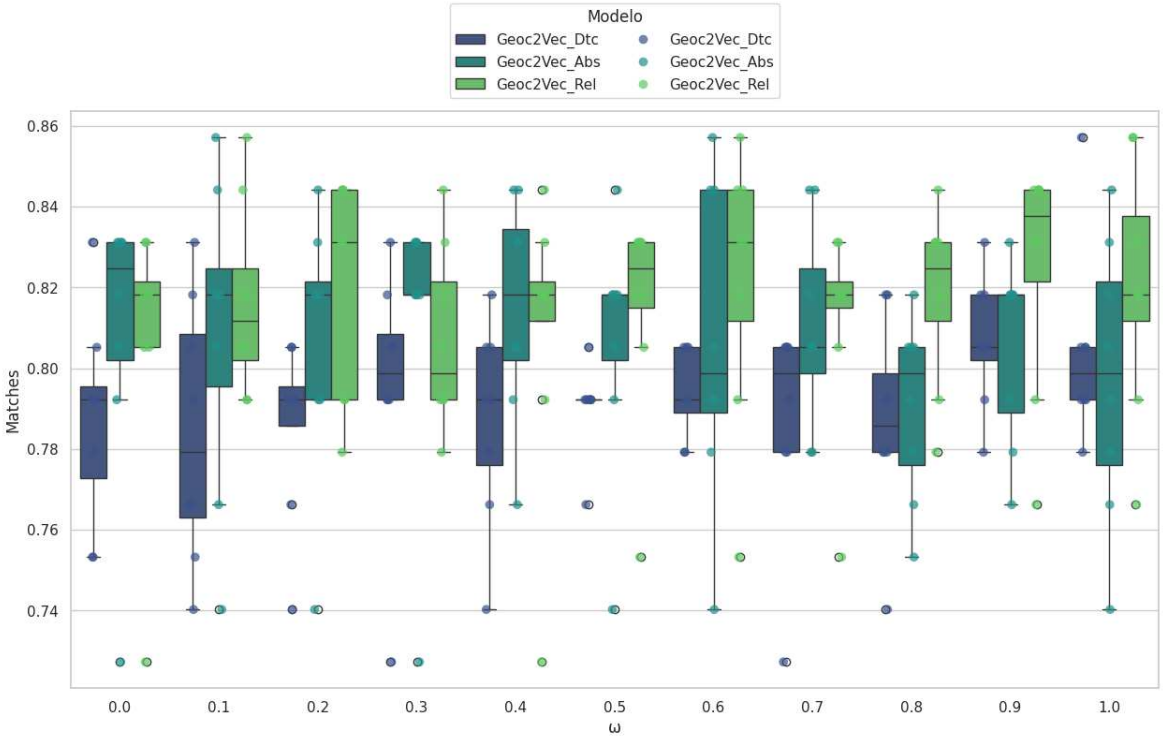


Figura 6.2: Resultados dos três modelos GeoContext2Vec na tarefa BHE por valor de ω.

Fonte: Autoria própria

Observando os resultados do modelo *GeoC2Vec_Abs*, nota-se que as distribuições tendem a diminuir e se tornar mais esparsas quando se considera em maior proporção o espaço ocupado. Isso está em acordo com a discussão anterior, onde utilizar uma área absoluta na divisão das áreas de cada feição resulta em valores menores, gerando menos replicações no conjunto de teste e, consequentemente, fazendo com que os *embeddings* percam nuances das diferenças contextuais entre os tipos de POI. Além disso, quando se utiliza 60% ou 100% do espaço ocupado pelas feições pode-se observar a maior variação dos valores. Isso ocorre, porque para valores de raio menores, a diferença entre a área das feições e a área absoluta do contexto possivelmente é maior, gerando essa variação nos resultados. Além disso, é possível notar que as distribuições, ao considerarem valores mais baixos de ω, sofrem menos com o problema da divisão pelo valor absoluto da área do contexto.

Por fim, no modelo *GeoC2Vec_Dtc*, observa-se que as distribuições tendem a ser um

pouco mais altas para valores elevados de ω, o que mais uma vez destaca a importância do espaço ocupado pelas feições geográficas na diferenciação contextual dos POIs. Além disso, nota-se que os intervalos nos quais as distribuições desse modelo se sobrepõem estão na faixa de 78% a 80%. Isso aponta para o fato de penalizar o valor de β com base na distância. Em outras palavras, independentemente de considerar a ocorrência das feições ou o espaço ocupado, os resultados são semelhantes devido às replicações das relações binárias serem sempre maiores para aquelas próximas à origem do contexto.

A partir das distribuições exibidas nas Figuras 6.1 e 6.2, observa-se interseções entre os resultados de todos os modelos. Para identificar se existe diferença significativa entre as distribuições, foi realizado um teste estatístico. Na análise estatística, para comparações múltiplas, podem ser utilizados testes como ANOVA [16] e *Friedman* [16]. A aplicação do teste ANOVA requer que alguns pressupostos sejam atendidos, como homoscedasticidade e normalidade dos dados, por ser um teste paramétrico. Caso esses pressupostos sejam violados, testes não paramétricos, como o de *Friedman*, podem ser empregados.

Antes de realizar o teste estatístico de comparação de diferentes abordagens, é necessário verificar a normalidade dos dados coletados para determinar qual tipo de teste deve ser aplicado. Para atestar a normalidade dos dados, realiza-se um teste de hipótese conforme as seguintes hipóteses:

- H_0: A amostra provém de uma população normalmente distribuída.
- H_1: A amostra não provém de uma população normalmente distribuída.

Segundo Razali *et al*. [66], o método *Shapiro-Wilk* apresenta o melhor desempenho para diversos tipos de distribuição e tamanhos de amostra. Seguindo um nível de significância de 5% ($\alpha = 0,05$), o resultado do teste *Shapiro-Wilk* retornou *p-valores* inferiores a $0,05$ para cada abordagem, fazendo com que a hipótese H_0 seja refutada com 95% de confiança para a amostra específica. Isso demonstra que a amostra não apresenta distribuição normal. Nesse caso, o teste de *Friedman* [118] pode ser utilizado.

O resultado do teste de *Friedman*, com nível de significância de 5% ($\alpha = 0,05$), retornou um *p-valores* inferiores a $0,05$. Constatou-se, portanto, que existe uma diferença estatisticamente significativa entre as distribuições dos três modelos. Para identificar quais

distribuições diferem significativamente entre si, foi empregado um teste *post hoc* de comparações múltiplas. Nesse caso, foi utilizado o teste de Conover com a correção de Bonferroni (p_{bon}) [57]. Os resultados do teste de Conover estão resumidos na Tabela 6.1.

Tabela 6.1: Resultado do teste de Conover para as distribuições da tarefa BHE.

Modelos		p_{bon}
GeoC2Vec_Rel	*GeoC2Vec_Abs*	0,185
GeoC2Vec_Rel	*GeoC2Vec_Dtc*	< 0,001
GeoC2Vec_Abs	*GeoC2Vec_Dtc*	< 0,001

O valor de p_{bon} para a comparação entre *GeoC2Vec_Rel* e *GeoC2Vec_Abs* é maior que $0,05$. Isso indica que não existe diferença estatisticamente significativa entre essas distribuições. Em outras palavras, os *embeddings* produzidos com o modelo relativo e o modelo absoluto apresentam resultados estatisticamente equivalentes. Esse comportamento pode ser atribuído à maneira similar de calcular a *proporção de espaço ocupado*, utilizando a área das feições sem penalização. Apesar de o modelo *GeoC2Vec_Abs* ter resultados um pouco inferiores ao modelo *GeoC2Vec_Rel* quando o raio é menor, os resultados se equiparam para contextos de raio maior.

Considerando os valores de p_{bon} entre o modelo *GeoC2Vec_Dtc* e os demais, percebe-se que eles são menores que $0,05$. Isso indica que as distribuições apresentam diferenças estatisticamente significativas. Apesar dessa diferença, essa abordagem ainda produziu resultados com cerca de 81% de *matching* com a opinião das pessoas.

A tarefa BHE também foi realizada utilizando os *embeddings* gerados pelos *baselines*. Assim como o GeoContext2Vec, o ITDL utiliza um parâmetro σ para ponderar as propriedades de unicidade e popularidade dos POIs. A Figura 6.3 ilustra a distribuição dos valores ao longo dos raios para o ITDL.

Analisando as distribuições, observa-se que, para um raio de $100m$, há uma convergência dos resultados obtidos com os *embeddings* em torno de 80% a 84% de *matching* com a opinião dos participantes da tarefa. Isso sugere que as relações contextuais de vizinhança dos POIs refletem a similaridade de seus tipos. Além disso, à medida que o valor do raio aumenta, os resultados diminuem quase linearmente. Isso sugere que as diferenças contextuais

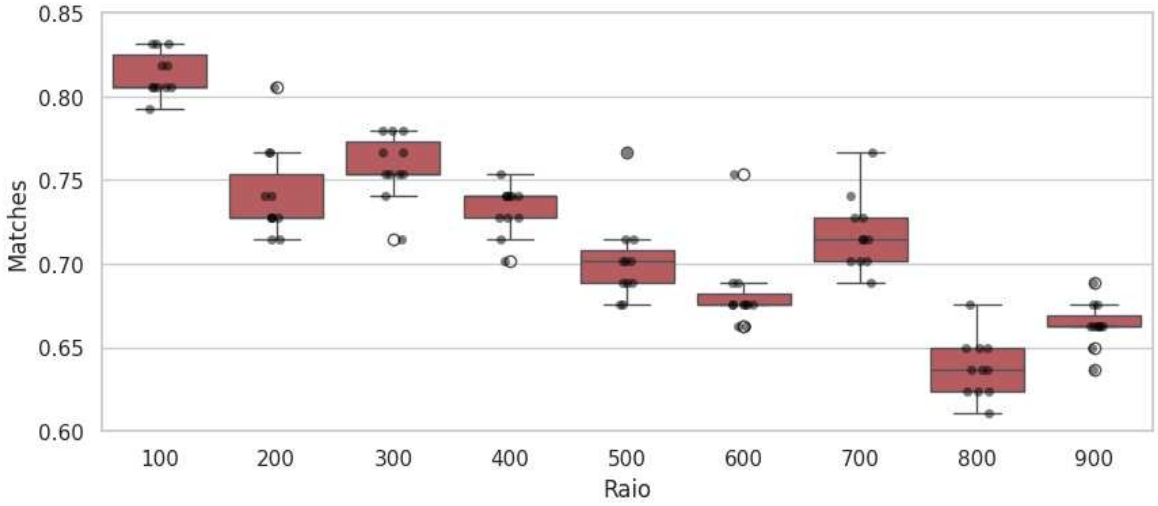


Figura 6.3: Resultados do ITDL na tarefa BHE por valor de raio.

Fonte: Autoria própria

dos POIs são mais evidentes quando se utiliza poucos vizinhos, uma vez que em $100m$ não há uma densidade muito alta de POIs no contexto.

Observando as distribuições em função de σ (Figura 6.4), pode-se perceber que os resultados aumentam à medida que σ aumenta. As distribuições ao longo de σ apresentam valores próximos devido à queda linear dos valores de *matching* quando o raio do contexto cresce. Apesar disso, é possível perceber que considerar mais a popularidade (dada pelos *check-ins*, faz com que os *embeddings* dos tipos de POI capturem com mais precisão as diferenças contextuais. Em outras palavras, os tipos mais similares tendem a apresentar proporções de *check-ins* similares. Nesse caso, com $\sigma = 1,0$, foi alcançado uma mediana em torno de 75%.

Após analisar as distribuições, foi selecionado o *GeoC2Vec_Rel*, pois apresentou o melhor desempenho. Além disso, para o parâmetro ω, observou-se que o melhor resultado ocorre quando se considera 90% do espaço ocupado pelas feições para a produção do conjunto de treinamento ($\omega = 0,9$). No caso do ITDL, foi escolhido o valor $\sigma = 1,0$. Também foi realizada a tarefa utilizando os *embeddings* do *Shortest Path*. Os resultados dessas abordagens estão na Figura 6.5.

Observando os resultados apresentados na Figura 6.5, é possível notar que os *embeddings* produzidos com o GeoContext2Vec alcançaram os melhores resultados de *matching*, conseguindo diferenciar com mais precisão os tipos de POI. Esse resultado demonstra que as relações contextuais dos tipos de POI com as feições geográficas refletem a similaridade

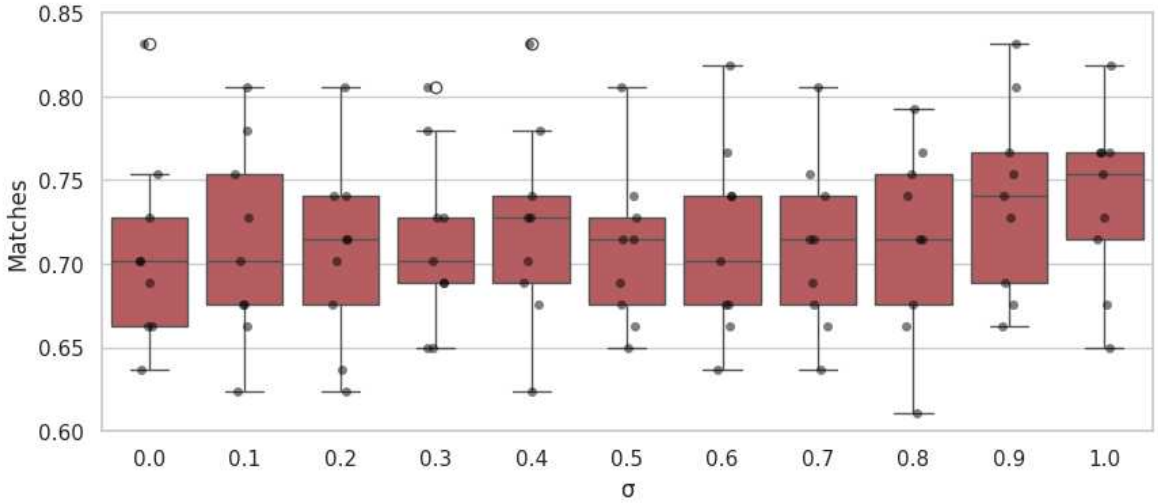


Figura 6.4: Resultados do ITDL na tarefa BHE por valor de σ.

Fonte: Autoria própria

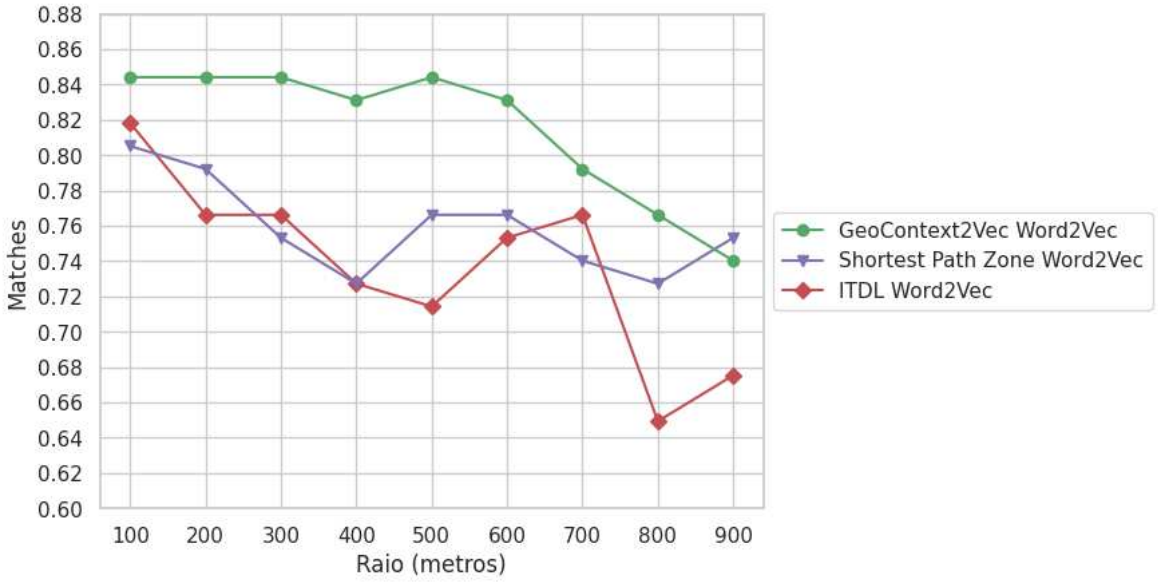


Figura 6.5: Resultados da tarefa BHE por valor de raio para todos os modelos.

Fonte: Autoria própria

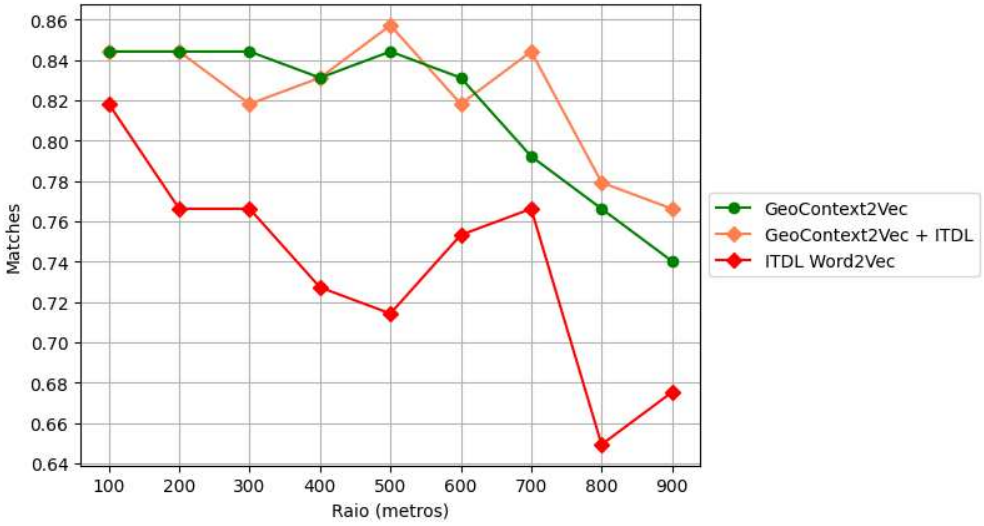
dos tipos. Além disso, mesmo com o aumento do raio, é possível perceber certa estabilidade nos resultados do GeoContext2Vec. Também é notável que os resultados obtidos com os *embeddings* do *Shortest Path* são similares aos resultados do ITDL. Isso demonstra que, mesmo considerando apenas a vizinhança mais próxima e descartando outras informações, como *check-in*, ainda existe alguma relação entre as diferenças dos tipos de POI e as relações contextuais de vizinhança dos POIs.

Os resultados do GeoContext2Vec demonstraram que os *embeddings* gerados a partir das relações contextuais com as feições geográficas possibilitam uma melhor diferenciação dos tipos de POI, com *matching* de 84% com a opinião dos voluntários. A partir disso, foi realizada a concatenação desses *embeddings* com os *embeddings* dos baselines, visando analisar se a combinação considerando diferentes aspectos permite a diferenciação dos tipos de POI com melhor precisão. Como os *embeddings* apresentam tamanhos diferentes (70 para os *baselines* e 280 para o GeoContext2Vec), foi utilizado o método PCA (Principal Component Analysis), que é uma técnica de redução de dimensionalidade amplamente utilizada no contexto de *embeddings* [1]. Nesse caso, reduziu-se os *embeddings* do GeoContext2Vec de 280 dimensões para 70 dimensões. Dessa forma, o cálculo da similaridade do cosseno não é influenciado por conta de dimensões de tamanho diferentes. É importante frisar que o uso de técnicas de redução de dimensionalidade acarretam em perdas naturais de informações dos *embeddings*, podendo fazer com que os resultados dessa tarefa sejam menos precisos. Os resultados estão ilustrados nas Figuras 6.6a e 6.6b.

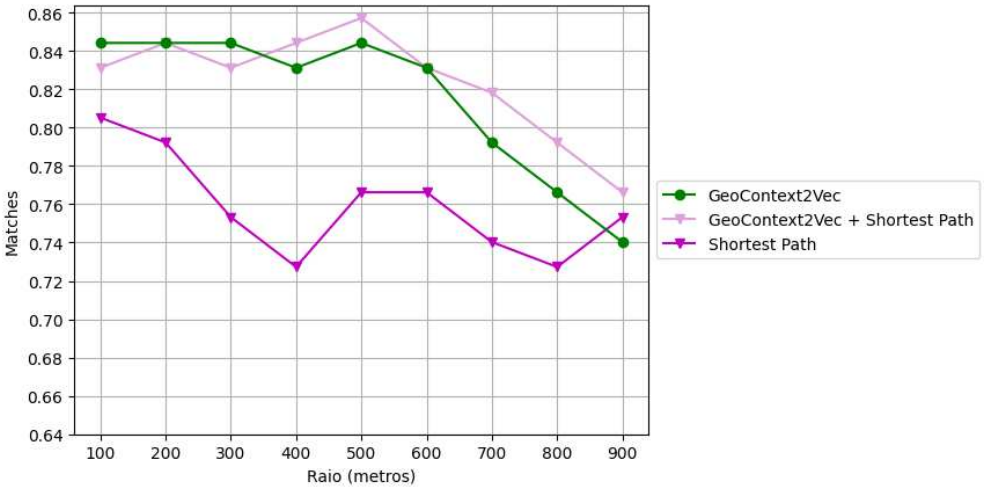
Observando os resultados das combinações do GeoContext2Vec com os dois *baselines*, é evidente que, para todos os valores de raio, os resultados combinados superaram os *baselines*. Isso indica que as feições geográficas, quando combinadas com a vizinhança de POIs, permitem uma melhor diferenciação dos tipos do que apenas utilizar a vizinhança de POIs. Além disso, percebe-se que para valores menores de raio os resultados são próximos ou iguais ao GeoContext2Vec. Isso aponta que as relações contextuais dos POIs com as feições geográficas variam menos do que as relações de vizinhança dos POIs. Porém, quando o raio aumenta, é possível perceber que a combinação das informações permite uma diferenciação de tipos ainda melhor que apenas o GeoContext2Vec. Esse resultado sugere que, em raios maiores, as diferenças contextuais das feições geográficas do contexto dos POIs é menos evidente. Porém, a vizinhança de POIs aponta a melhor a diferença entre os POIs. Considerando o resultado máximo, a combinação dos *embeddings* alcançou aproximadamente 86% de *matching*, sendo o melhor resultado entre todos os modelos.

6.1.2 BHE com DistilBert

Um dos objetivos desta pesquisa é investigar se *embeddings* produzidos com modelos mais recentes, como o BERT, conseguem capturar com maior precisão as relações contextuais dos



(a) GeoContext2Vec concatenado com ITDL.



(b) GeoContext2Vec concatenado com *Shortest Path*.

Figura 6.6: Combinação dos *embeddings* do GeoContext2Vec com os *embeddings* dos *baselines*.

tipos de POI. Nesse caso, considerando os resultados obtidos com o Word2Vec, foram selecionados os conjuntos de treinamento dos melhores modelos do GeoContext2Vec, ITDL e *Shortest Path* para treinamento no DistilBert conforme descrito na Seção 4.4. O treinamento de todos os possíveis conjuntos de dados (99 conjuntos para o GeoContext2Vec e ITDL) demanda uma quantidade de tempo considerável, devido o treinamento com o DistilBert ser mais custoso que o treinamento com o Word2Vec. A Figura 6.7 ilustra os resultados obtidos utilizando os *embeddings* dessas abordagens na tarefa BHE em comparação com os *embeddings* do Word2Vec.

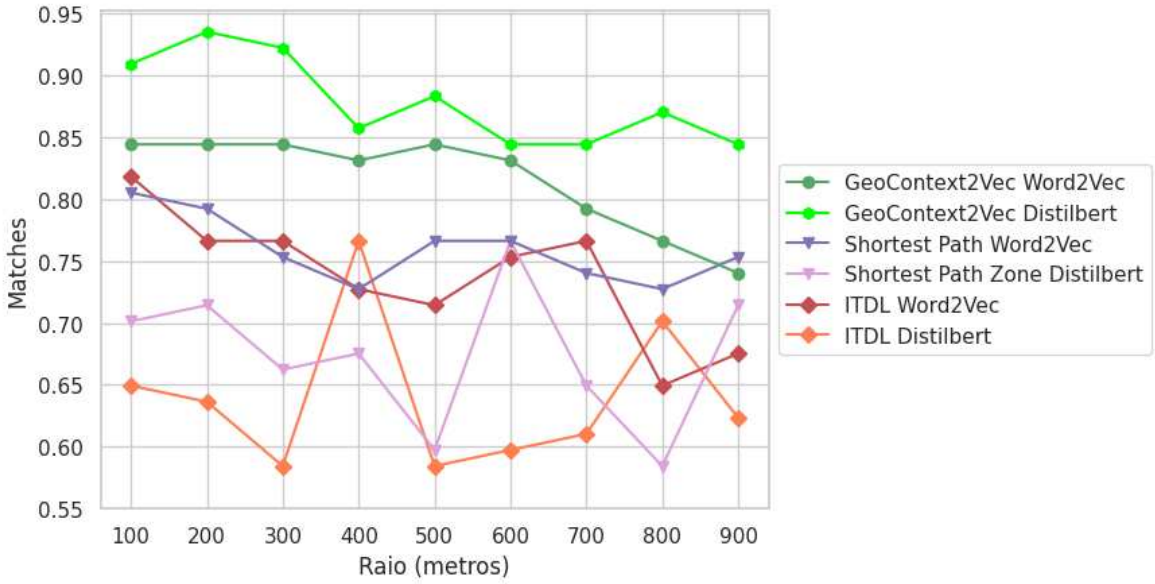


Figura 6.7: Resultados da tarefa BHE por valor de raio para todos os modelos utilizando Word2Vec e DistilBert.

Fonte: Autoria própria

De maneira geral, é possível notar que os *embeddings* produdizos com o DistilBert e GeoContex2Vec atingiram os melhores resultados para todos os valores de raio. Seu valor máximo está em torno de 94% de *matching* com a opinião dos voluntários. Além disso, seu valor mínimo está em torno de 85%, sendo superior ao todos os demais métodos. A partir desse resultado, pode-se afirmar que os *embeddings* produzidos por esse modelo, se mostram mais robustos e capturam com maior precisão, as diferenças contextuais dos tipos de POI que utilizam feições geográficas.

Em contrapartida, os *embeddings* produzidos com os conjuntos de treinamento do

ITDL e *Shortest Path* demonstraram desempenho abaixo dos *embeddings* produzidos com o Word2Vec. Nesse caso, os resultados chegam a no máximo empatar (raio de $400m$ e $600m$) com os valores do Word2Vec. Além disso, os *embeddings* do ITDL configuram resultados ainda menores que os *embedings* do *Shortest Path*.

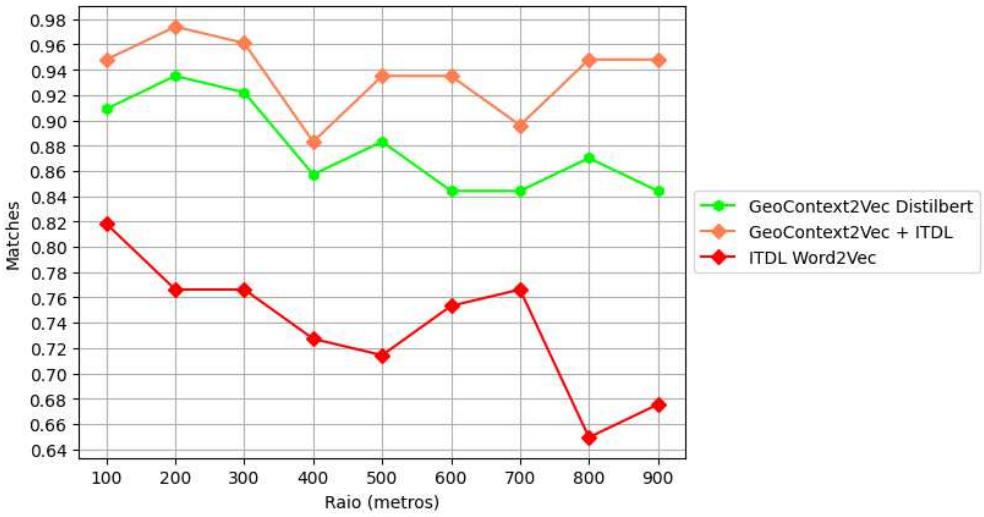
Acredita-se que esse resultado decorre da maneira como os documentos de treinamento são estruturados, e não do modelo em si. Os documentos do GeoContext2Vec associam uma feição do contexto a todos os tipos de POI relacionados a ela em um único documento. Assim, é provável que a feição tenha atuado como uma ponte, exercendo a mesma "função no texto" para o modelo associar todos os tipos de POI do documento entre si. Em cada documento, a feição sempre é a segunda palavra da frase. No entanto, nos documentos que utilizam apenas tipos de POI, os *embeddings* não capturaram com precisão as diferenças contextuais da mesma maneira que o Word2Vec. Acredita-se que, como os tipos podem ser tanto a primeira palavra quanto a segunda, o modelo confundiu sua "função no texto" no documento[1]. Outro ponto a ser mencionado é que o DistilBert é pré-treinado em textos anteriores e possui conhecimento prévio dos tipos de POI de forma inerente. No entanto, correlacionar diretamente esses tipos não melhorou significativamente a capacidade do modelo de discernir suas relações contextuais.

Por outro lado, é improvável que os textos anteriores usados para treinar o DistilBert associem explicitamente os tipos de POI com feições geográficas. Portanto, aproveitar o conhecimento prévio dos tipos de POI incorporando informações sobre feições geográficas resultou em um desempenho superior para o DistilBert treinado com o conjunto de treinamento do GeoContext2Vec. Portanto, acredita-se que são necessários estudos mais aprofundados sobre como um documento de treinamento deve ser gerado nesse caso.

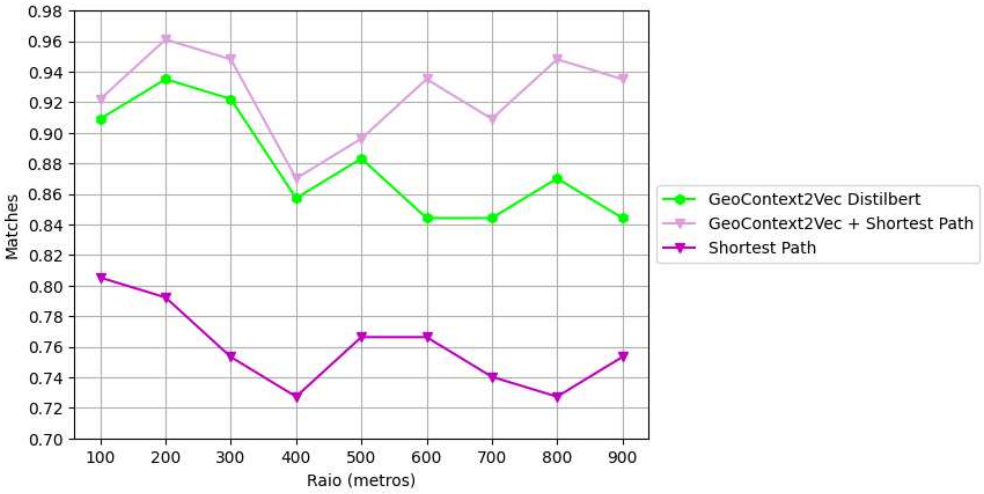
Assim como nos *embeddings* do Word2Vec, foi realizado a concatenação entre os *embeddings* do GeoContext2Vec DistilBert e do ITDL e *Shortest Path* Word2Vec, visando analisar se ocorre melhoria na diferenciação dos tipos de POI. Também foi empregado o PCA para reduzir a dimensão dos *embeddings* do DistilBert de 3072 para 70. Os resultados são apresentados na Figura 6.8.

A partir dos resultados, é evidente que houve um aumento nos valores de *matching* em

[1] Além de pares de palavras por sentença, na geração dos documentos foi investigado manter apenas uma palavra por frase utilizando o separador [SEP]. Entretanto, os resultados foram similares.



(a) GeoContext2Vec DistilBert concatenado com ITDL Word2Vec.



(b) GeoContext2Vec DistilBert concatenado com *Shortest Path* Word2Vec.

Figura 6.8: Combinação dos *embeddings* do GeoContext2Vec com os *embeddings* dos *baselines* na tarefa BHE.

relação a todos os modelos. Os *embeddings* combinados entre o GeoContext2Vec e o ITDL alcançaram uma conformidade máxima aproximada de 98% (no raio de $200m$) de *matching* com a opinião dos voluntários. A combinação entre o GeoContext2Vec e o *Shortest Path* alcançou uma conformidade máxima aproximada de 96% (no raio de $200m$). Esses resultados demonstram que POIs que apresentam tipos com funções semelhantes, apresentam relações contextuais com as feições geográficas e vizinhança de POIs similares. Ou seja, existe um indicativo de que o contexto de tipos com funções similares possua padrões de feições geográficas e de vizinhos que se assemelham.

6.2 Análise de Similaridade com RHE

Esta seção apresenta os resultados da utilização dos *embeddings* do GeoContext2Vec e dos *baselines* na tarefa RHE, explorando se os valores de similaridade entre os tipos de POI fornecidos pelo cálculo do cosseno refletem os valores de similaridade indicados por pessoas. Também foi investigado como as propriedades espaciais afetam os valores de similaridade. Além disso, foi analisado se os modelos mais recentes apresentam valores de similaridade mais próximos ao indicado por pessoas do que os modelos clássicos de PLN.

6.2.1 RHE com Word2Vec

Para esta tarefa, foram calculados a similaridade do cosseno entre os pares de tipos de POI do conjunto de teste. A partir dos valores definidos pelos participantes, calculou-se a correlação de *Spearman*, para identificar se os valores de similaridade obtidos com os *embeddings* refletem os valores de similaridades conforme a opinião humana. Os resultados obtidos nessa tarefa, para os três modelos GeoContext2Vec (*GeoC2Vec_Rel*, *GeoC2Vec_Abs* e *GeoC2Vec_Dtc*) e os diferentes valores de raio estão presentes na Figura 6.9.

A partir dos resultados, observa-se uma correlação em torno de 57% entre os valores de similaridade do cosseno e as valores atribuídos pelas pessoas. Isso aponta que, quando os valores de similaridade atribuídos pelas pessoas aumentam ou diminuem, os valores calculados pelo cosseno também aumentam ou diminuem com uma correlação forte. Com base no conjunto de teste construído (ver Apêndice A), pode-se afirmar que quando os tipos de POI têm funções semelhantes, os valores de similaridade são mais altos, enquanto que, quando os

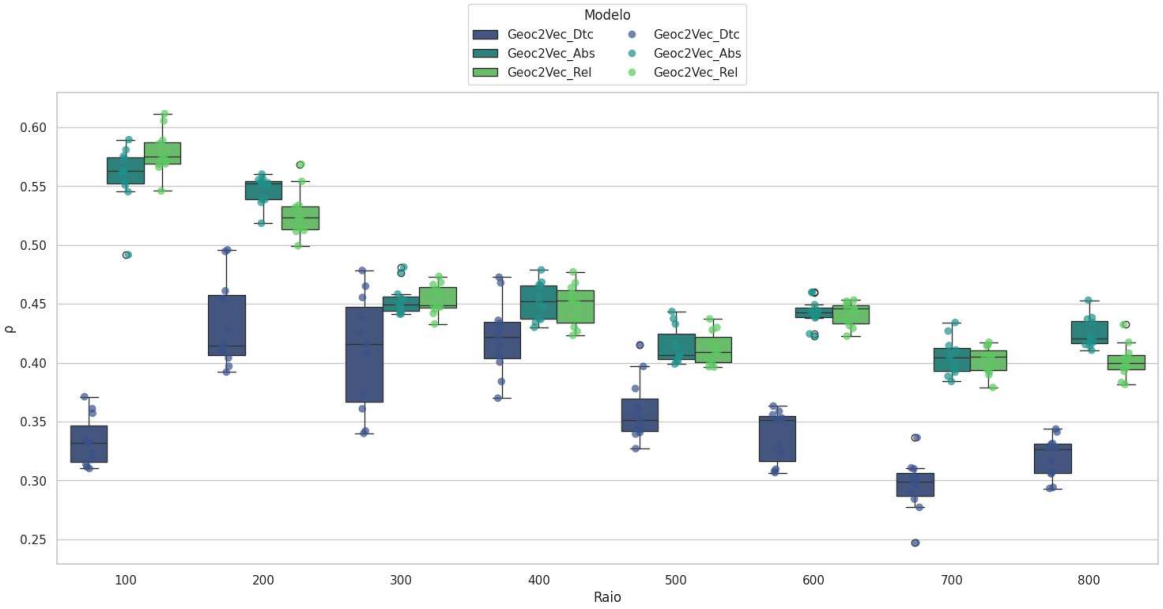


Figura 6.9: Resultados dos três modelos GeoContext2Vec na tarefa RHE por valor de raio para todos os modelos.

Fonte: Autoria própria

tipos possuem funções diferentes, os valores são mais baixos. Uma análise dos valores dos modelos *GeoC2Vec_Rel* e *GeoC2Vec_Abs* sugere que POIs que possuem tipos com funções semelhantes também compartilham padrões geográficos similares. Por exemplo, é comum que estabelecimentos que vendem comida disponham de estacionamento para clientes, enquanto postos de gasolina tendem a estar localizados perto de grandes vias e cruzamentos.

Ao analisar os valores de correlação em relação ao aumento do raio, é observado um decréscimo. Intuitivamente, à medida que o raio aumenta, os contextos dos tipos de POI se tornam mais semelhantes devido ao compartilhamento de feições. Como resultado, a similaridade do cosseno tende a apresentar valores cada vez mais próximos para os diferentes tipos de POI e a correlação com a opinião humana diminui.

Ao analisar as distribuições da correlação por valor de ω, nota-se que os resultados dos modelos *GeoC2Vec_Rel* e *GeoC2Vec_Abs* apresentam extensões praticamente idênticas (Figura 6.10). No entanto, elas são um pouco menos dispersas para valores mais altos de ω. Isso sugere que as ocorrências das feições geográficas nos contextos dos POIs que possuem tipos com funções relacionadas são mais padronizadas do que as áreas ocupadas por essas feições. Por exemplo, é comum que lugares que vendem comida tenham estacionamentos em

seus contextos, mas o tamanho desses estacionamentos pode variar de acordo com o porte do estabelecimento. Essa variação é refletida nas distribuições.

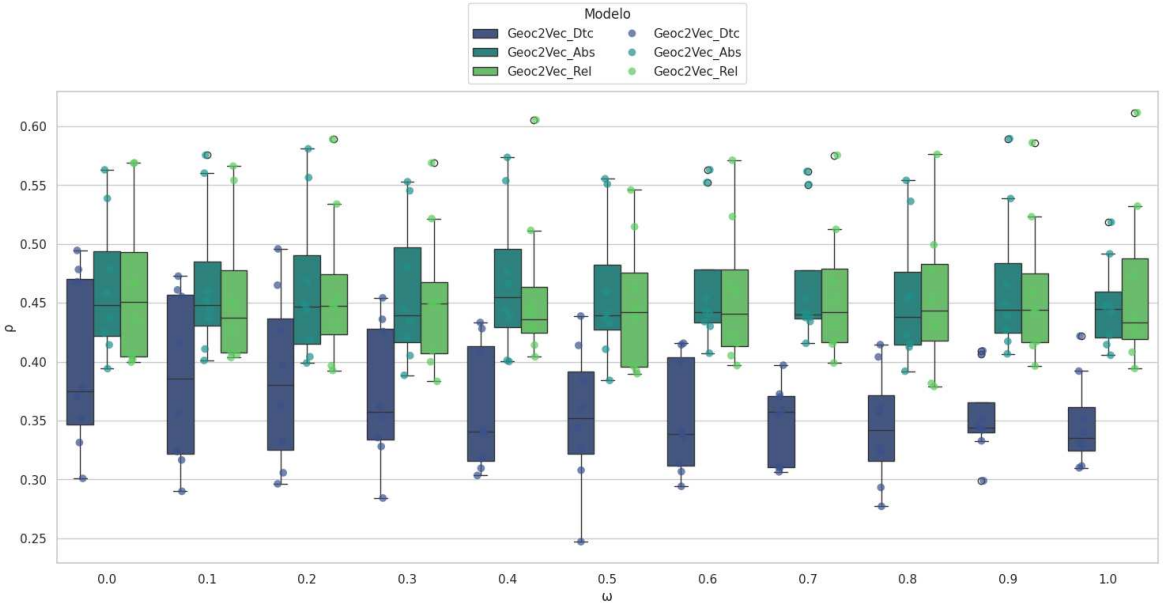


Figura 6.10: Resultados dos três modelos GeoContext2Vec na tarefa RHE por valor de ω para todos os modelos.

Fonte: Autoria própria

No que diz respeito ao modelo *GeoC2Vec_Dtc*, observa-se que a distribuição tende a diminuir à medida que o valor de ω aumenta. Isso denota que a distância tem um impacto maior no atributo SP do que no atributo OP. Em outras palavras, penalizar mais as ocorrências tem menos impacto na correlação do que penalizar o espaço ocupado por uma feição com base na distância. No entanto, como mencionado anteriormente, a redução da quantidade de relações binárias causada pela distância faz com que os *embeddings* incorporem menos as diferenças contextuais dos POIs.

Considerando as interseções nas distribuições entre os três modelos demonstrados nas Figuras 6.9 e 6.10, foi realizado um teste estatístico para averiguar a existência de diferenças estatisticamente significativas. Para esse teste, a mesma hipótese da tarefa BHE foi admitida, ou seja, que a amostra provém de uma população normalmente distribuída (H_0). Nesse caso, o teste *Shapiro-Wilk* retornou p-valores abaixo de $0,05$ para cada abordagem. Portanto, a hipótese nula é refutada com 95% de confiança para a amostra utilizada, indicando que um teste não paramétrico pode ser aplicado, como o teste de *Friedman*.

Tabela 6.2: Resultado do teste de Conover para as distribuições da tarefa RHE.

Modelos		p_{bon}
GeoC2Vec_Rel	*GeoC2Vec_Abs*	0,359
GeoC2Vec_Rel	*GeoC2Vec_Dtc*	< 0,001
GeoC2Vec_Abs	*GeoC2Vec_Dtc*	< 0,001

O resultado do teste de *Friedman*, com nível de significância de 5% ($\alpha = 0,05$), retornou um *p*-valor menor que $0,05$. Isso demonstra que existe uma diferença estatisticamente significativa entre as distribuições dos três modelos. Para identificar quais distribuições diferem significativamente entre si, foi empregado o teste de Conover com a correção de Bonferroni. Os resultados desse teste estão resumidos na Tabela 6.2.

A partir do teste de Conover, observa-se que o p_{bon} dos modelos *GeoC2Vec_Abs* e *GeoC2Vec_Rel* é maior que $0,05$, demonstrando que não existe diferença estatisticamente significativa entre as distribuições desses dois modelos. No entanto, os *embeddings* do modelo *GeoC2Vec_Dtc* mostram uma correlação significativamente inferior a $0,05$. Isso está alinhado com as discussões apresentadas na tarefa BHE, uma vez que a decisão de penalizar as feições mais distantes da origem resulta em menos replicações das demais feições, levando os *embeddings* a não capturarem as diferenças contextuais dos tipos de POI.

Considerando os *baselines*, também foram utilizados os embeddings do ITDL e do *Shortest Path* na tarefa RHE. A Figura 6.11 ilustra os resultados obtidos com o ITDL para todos os valores de raio. Nesse caso, é possível perceber uma correlação máxima em torno de 65% para o melhor resultado ($100m$). Assim como no GeoContext2Vec, as relações contextuais de vizinhança refletem os níveis de similaridade entre tipos que possuem a mesma função. Em outras palavras, POIs que possuem tipos que apresentam funções semelhantes tendem a possuir vizinhanças semelhantes. Além disso, os valores de correlação diminuem à medida que o raio aumenta, pois ao considerar contextos maiores, os POIs passam a ter muitas interseções.

Analisando os resultados agrupados por valor de σ, nota-se que as distribuições mais altas estão associadas aos valores mais elevados de σ (Figura 6.12). Isso sugere que, além de POIs que têm tipos com funções semelhantes compartilharem vizinhanças semelhantes, a quantidade de *check-ins* presentes nos POIs vizinhos também é similar.

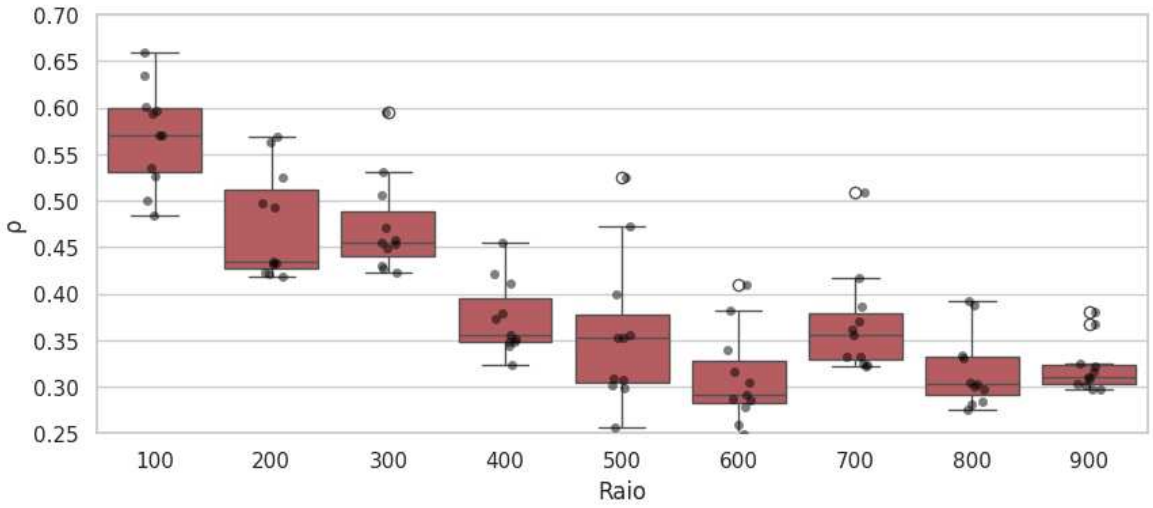


Figura 6.11: Resultados da tarefa RHE por valor de raio para o ITDL.

Fonte: Autoria própria

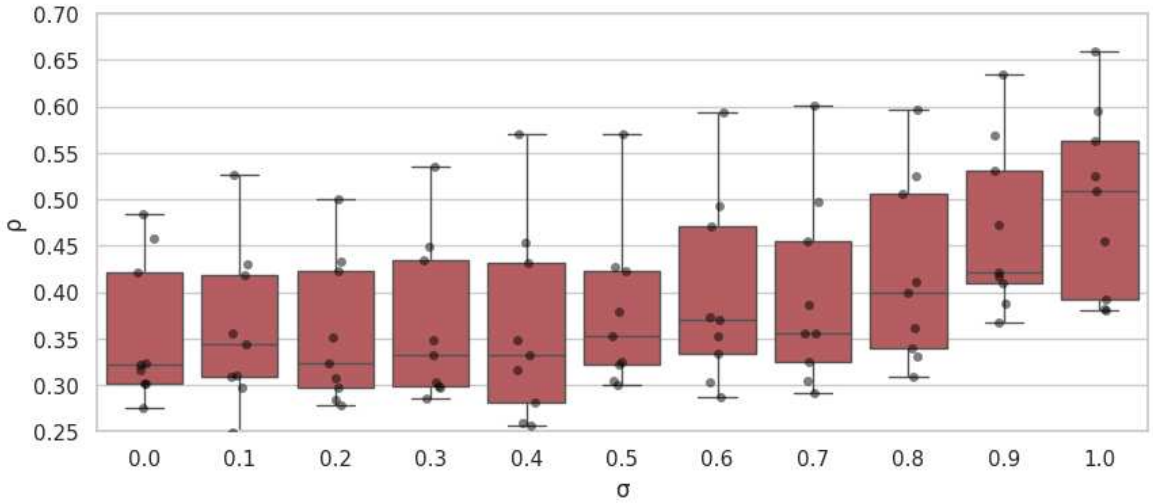


Figura 6.12: Resultados da tarefa RHE por valor de σ para o ITDL.

Fonte: Autoria própria

Definindo as melhores configurações para os modelos GeoContext2Vec ($\omega = 0,9$) e ITDL ($\sigma = 1,0$), os embeddings do Shortest Path também foram aplicados à mesma tarefa. Os resultados dos embeddings dos três modelos estão na Figura 6.13. A partir dos gráficos apresentados na figura, observa-se que os modelos que consideram a vizinhança apresentam uma correlação mais alta do que os modelos que consideram as feições do contexto. Nesse caso, o ITDL apresentou a maior correlação com um valor de $\rho = 66\%$, o Shortest Path com $\rho = 64\%$, e o GeoContext2Vec com $\rho = 59\%$.

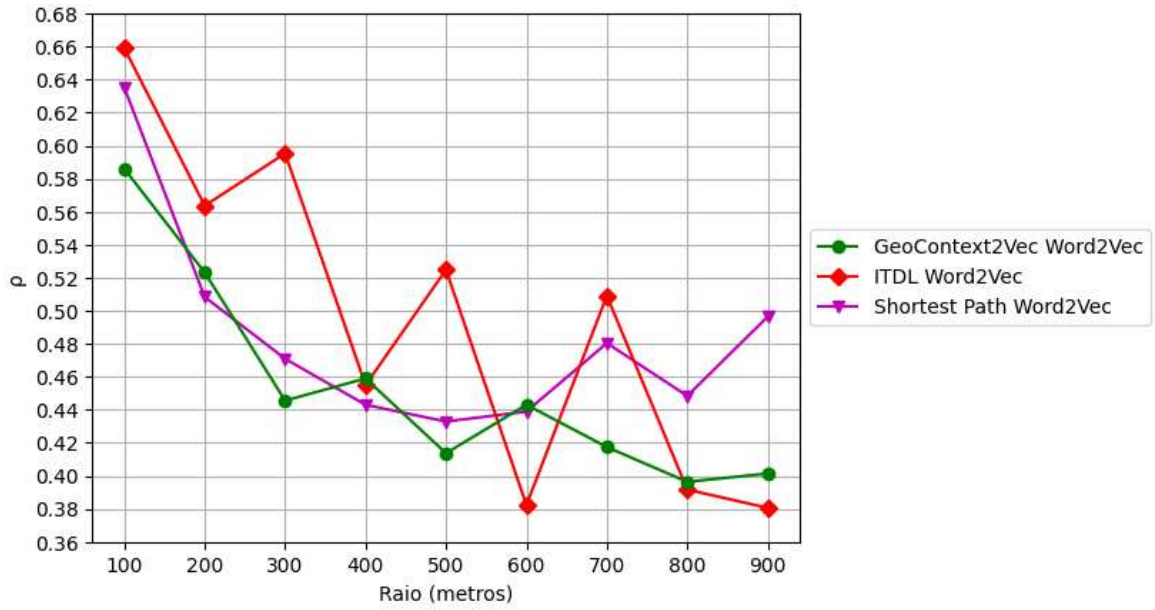


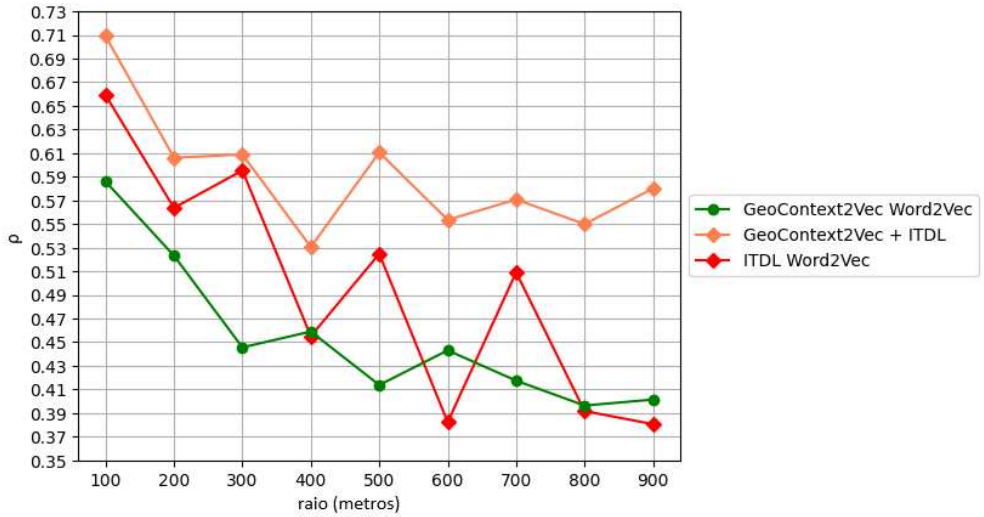
Figura 6.13: Resultados da tarefa RHE por valor de raio para o todos os modelos.

Fonte: Autoria própria

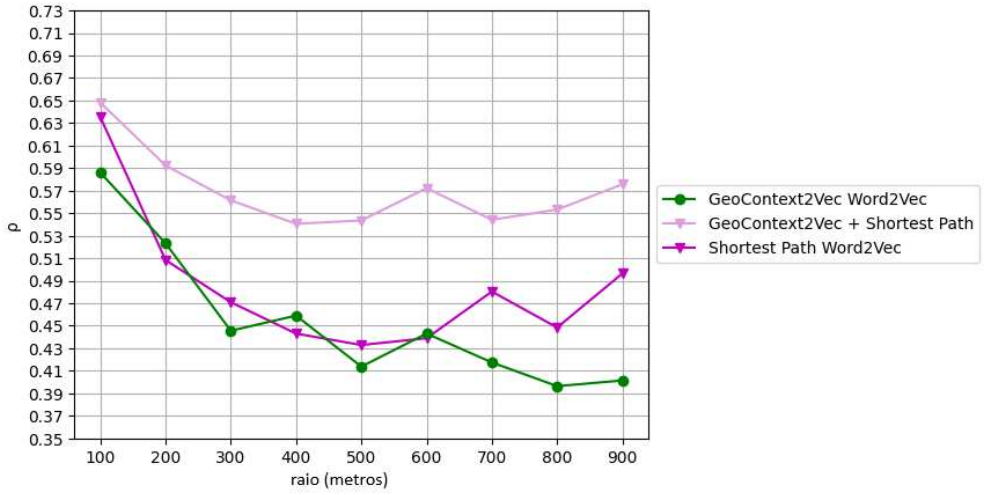
Mesmo com valores ligeiramente inferiores de correlação, esse resultado demonstra que os POIs compartilham mais as feições geográficas do que a vizinhança com outros POIs. Intuitivamente, é comum encontrar prédios, ruas, sinais e outras feições no contexto da maioria dos POIs. Por outro lado, é mais provável que uma vizinhança de POIs mude mais a depender do local. Apesar disso, os resultados do GeoContext2Vec apresentaram valores competitivos com o *Shortest Path*, sugerindo que as feições do contexto dos POIs também refletem a similaridade de seus tipos.

Também foi investigado o comportamento da correlação quando os *embeddings* produzidos com o GeoContext2Vec são concatenados com os *embeddings* dos *baselines*. A Figura 6.14 ilustra os resultados.

É observável que os *embeddings* combinados do GeoContext2Vec e ITDL apresentam melhoria em todos os valores de raio em relação aos demais modelos. Isso sugere que, POIs que possuem tipos com função similar, possuem ao mesmo tempo, padrões de feições geográficas e de POIs vizinhos. A combinação do GeoContext2Vec com o ITDL alcançou uma correlação máxima de 70% e a combinação com o *Shortest Path* alcançou uma correlação máxima de 65%. Além disso, conforme o raio aumenta, a correlação cai, pois intuitivamente, à medida que o contexto cresce, todos os POIs passam a compartilhar as feições geográficas



(a) GeoContext2Vec concatenado com ITDL.



(b) GeoContext2Vec concatenado com *Shortest Path*.

Figura 6.14: Combinação dos *embeddings* do GeoContext2Vec com os *embeddings* dos *baselines* na tarefa RHE.

e outros POIs da vizinhança, torando-se assim menos diferentes.

6.2.2 RHE com DistilBert

Conforme mencionado na Subseção 6.1.2 foram gerados *embeddings* utilizando o DistilBert. Nesta tarefa, foram empregados apenas os *embeddings* produzidos com o GeoContext2Vec, pois os *embeddings* gerados com os conjunto de treinamento dos *baselines* não apresentaram resultado satisfatório na tarefa BHE. A Figura 6.15 ilustra os resultados alcançados.

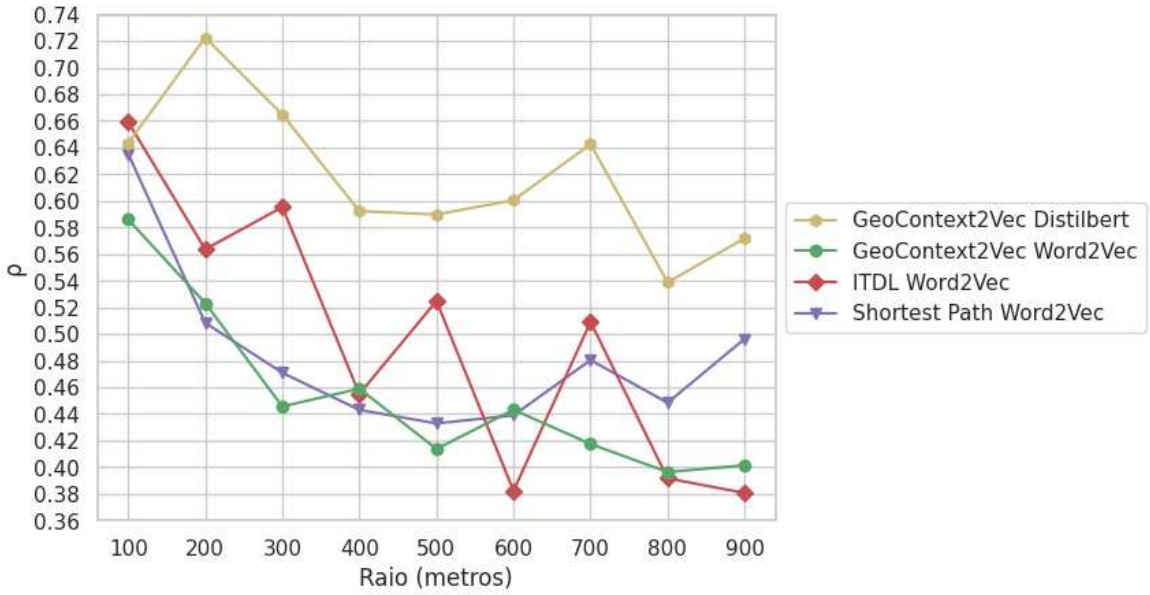
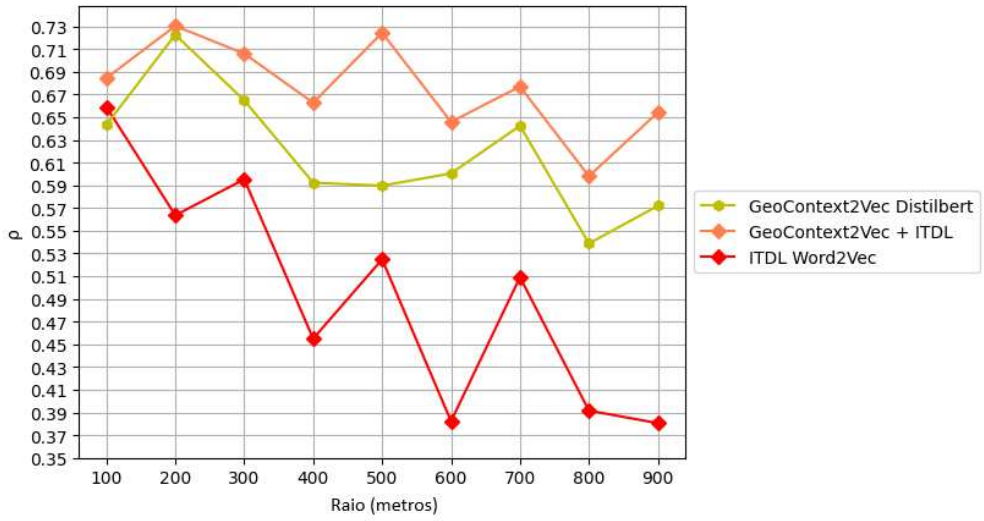


Figura 6.15: Resultados da tarefa RHE por valor de raio para o todos os modelos utilizando Word2Vec e DistilBert.

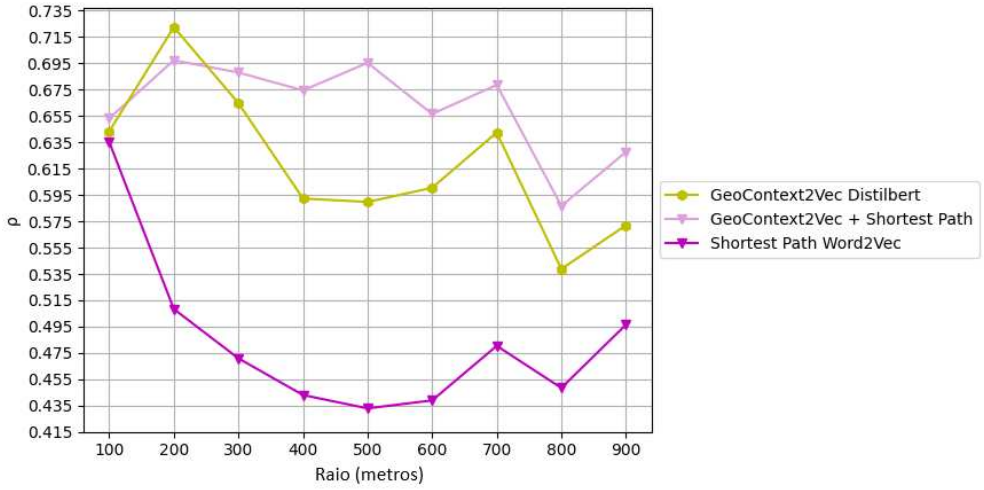
Fonte: Autoria própria

Os resultados obtidos demonstram que os *embeddings* produzidos com o GeoContext2Vec DistilBert superam os demais modelos para todos os valores de raio, exceto para o raio de $100m$. O melhor resultado alcançou a marca de 72% de correlação, superando o ITDL que atingiu no máximo 66% de correlação. Essa superioridade dos *embeddings* pode ser atribuída à robustez do DistilBert em capturar as nuances das relações contextuais entre os tipos de POI e as feições geográficas. Mesmo que muitas vezes os contextos compartilhem as mesmas feições, as replicações geradas considerando as propriedades de espaço ocupado e ocorrência são melhor percebidas pelo DistilBert do que pelo Word2Vec.

Também foram analisados os resultados das combinações dos *embeddings* do GeoContext2Vec DistilBert com os *baselines*. Os resultados são apresentados na Figura 6.16.



(a) GeoContext2Vec DistilBert concatenado com ITDL.



(b) GeoContext2Vec DistilBert concatenado com *Shortest Path*.

Figura 6.16: Combinação dos *embeddings* do GeoContext2Vec Distlibert com os *embeddings* dos *baselines* na tarefa RHE.

Observando a correlação das combinações, é possível afirmar que os resultados foram superiores aos *baselines* para maioria dos valores de raio (exceto para o *Shortest Path* em $200m$). Esses são resultados são novamente atribuídos à capacidade do DistilBert em perceber as nuances contextuais das feições geográficas e dos tipos de POI. Além disso, conforme discutido anteriormente, os resultados dos modelos combinados demonstram que POIs que

possuem tipos com funções semelhantes, tendem a apresentar padrões geográficos e de vizinhança semelhantes. Desse modo, o valor de similaridade entre tipos com funções equivalentes é mais alto do que o valor de similaridade entre tipos com funções diferentes. Nesse caso, o resultado máximo obtido partiu do GeoContext2Vec combinado com o ITDL que alcançou $\rho = 73\%$.

6.3 Análise de Similaridade Hierárquica

Ainda com o intuito de identificar se as relações contextuais dos tipos de POI com as feições geográficas refletem a similaridade dos tipos, foi realizada uma análise de similaridade considerando a hierarquia dos POIs (fornecida pelo Yelp) e a similaridade dos tipos obtidos por meio dos *embeddings*. Conforme demonstrado na Seção 5.3, foram gerados ranques utilizando os métodos de Wu e Palmer [96] e Leacock & Chodorow [35]. Os ranques produzidos utilizando os dois métodos foram iguais, pois eles se baseiam na distância dos termos dentro da hierarquia. Nesse caso, os resultados que serão debatidos valem para ambos. A seguir são discutidos os resultados obtidos com o Word2Vec e com o DistilBert.

6.3.1 MRR com Word2Vec

Considerando o ranque dado pela similaridade do cosseno entre os *embeddings* dos modelos, foi calculado o MRR para cada tipo de POI em relação aos ranques da hierarquia considerando o primeiro lugar do ranque. A Figura 6.17 ilustra os resultados de MRR para todos os modelos GeoContext2Vec (*GeoC2Vec_Rel*, *GeoC2Vec_Abs* e *GeoC2Vec_Dtc*).

As distribuições obtidas demonstram que, para todos os tipos de POI que possuem *embeddings*, o tipo hierarquicamente mais similar fica em média próximo da segunda posição do ranque (40% a 48%). Esse resultado sugere que tipos de POI hierarquicamente irmãos também apresentam similaridade considerando as relações contextuais com as feições geográficas do contexto de seus POIs.

Analisando-se cada modelo GeoContext2Vec, os resultados convergem com os das tarefas BHE e RHE. Ou seja, o modelo *GeoC2Vec_Rel* apresenta as distribuições mais altas para raios pequenos, e apresenta uma distribuição equivalente ao *GeoC2Vec_Abs* para valores de raio maiores. Como discutido anteriormente, esse modelo sofre menos com as diferenças

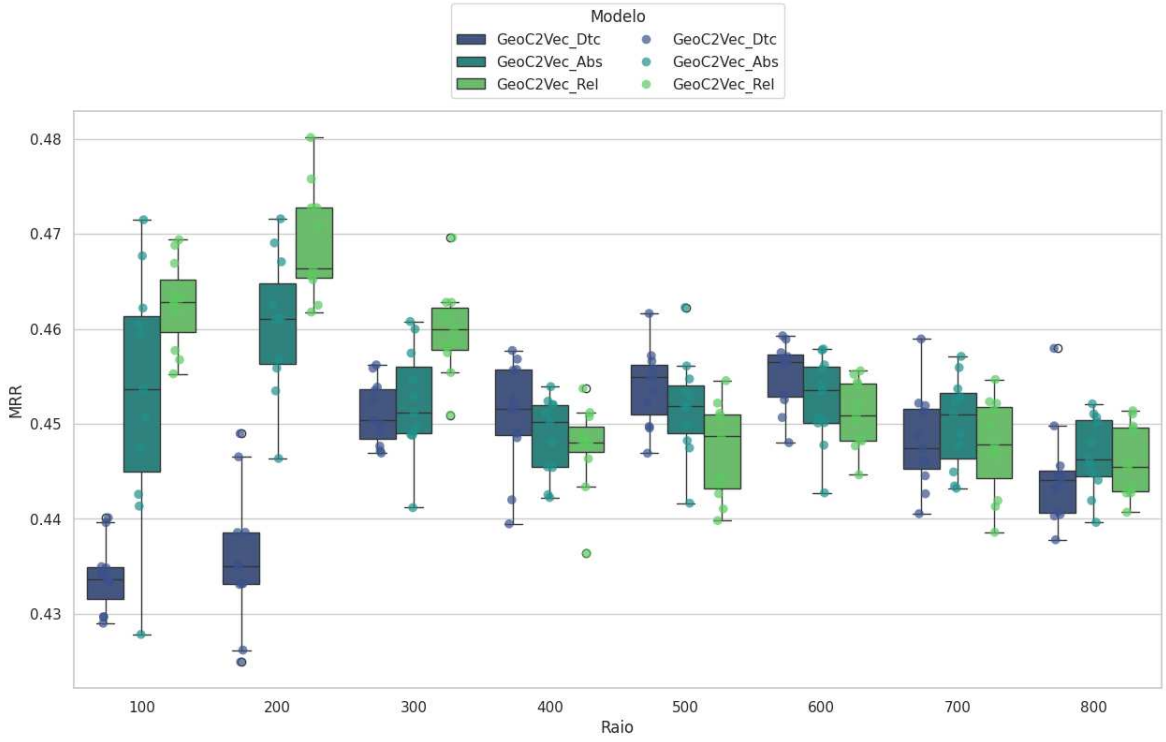


Figura 6.17: Resultados de MRR dos três modelos GeoContext2Vec por valor de raio.

Fonte: Autoria própria

entre as áreas das feições geográficas e a área absoluta do contexto geográfico. Entretanto, mesmo que o modelo *GeoC2Vec_Abs* sofra com essa diferença, ainda sim apresenta resultados similares ao *GeoC2Vec_Abs*.

Em relação ao *GeoC2Vec_Dtc*, o comportamento se assemelha aos resultados do BHE e RHE. Quando o raio do contexto é pequeno, os *embeddings* desse modelo que são penalizados pela distância, não conseguem perceber com mais eficiência as diferenças contextuais dos tipos. Esse comportamento vai se amenizando à medida que o raio cresce, pois as feições mais próximas são menos penalizadas, e permitem ao modelo perceber mais diferenças contextuais dos tipos. Esse modelo alcançou distribuições mais altas que os demais modelos nos raios de $400m$ à $700m$. Mesmo assim, existem muitas interseções entre as distribuições dos três modelos para raios a partir de $400m$.

Analisando as distribuições para cada valor de ω (Figura 6.18), percebe-se que o modelo *GeoC2Vec_Abs* apresenta distribuições mais concentradas que os demais modelos, e a distribuição tende a reduzir à medida que ω aumenta. Isso indica que utilizar mais o espaço ocupado pelas feições faz com que os *embeddings* produzidos com esse modelo não apre-

sentem um tipo irmão como o mais similar próximo ao topo de seu ranque. Esse resultado reflete os pontos já discutidos sobre como esse modelo relaciona a área das feições com a área absoluta da geometria do contexto dos POIs, e considerar ω maior faz com que o espaço seja mais considerado e consequentemente produza relações binários menos uniformes.

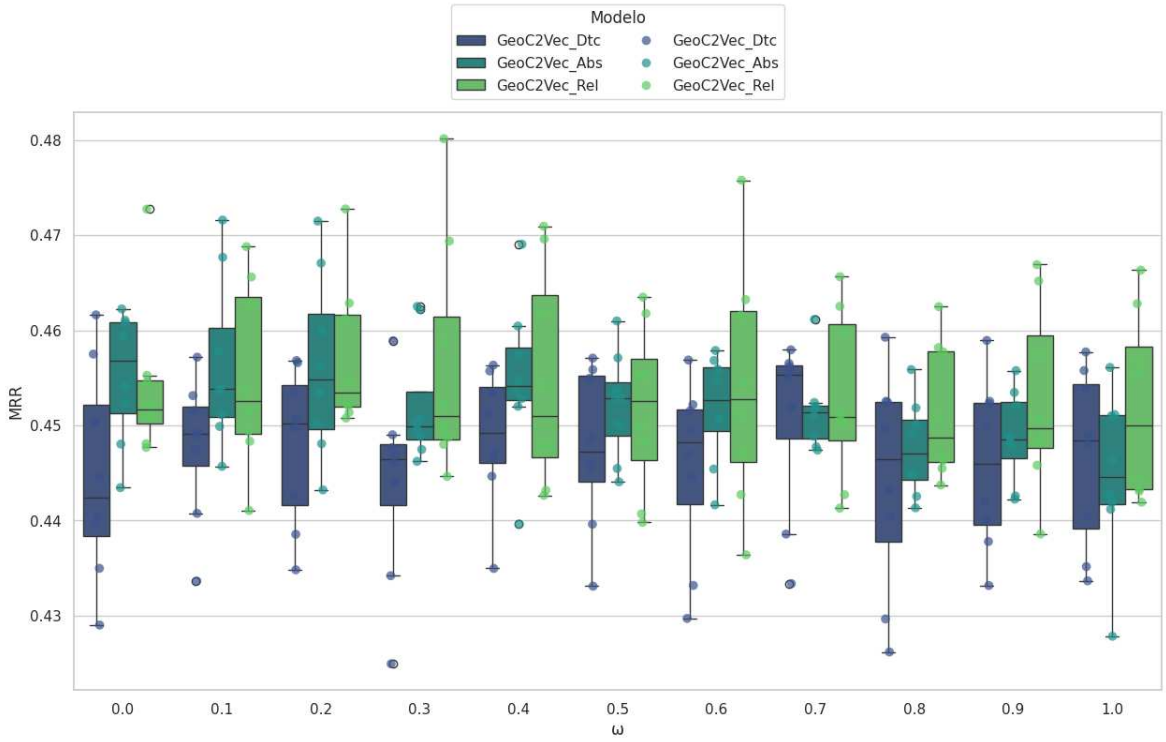


Figura 6.18: Resultados de MRR para os três modelos GeoContext2Vec por valor de ω.

Fonte: Autoria própria

Um comportamento similar é observado na distribuição do modelo *GeoC2Vec_Rel*. Ou seja, as distribuições diminuem um pouco à medida que ω cresce. Esse resultado aponta que os tipos irmãos na hierarquia tendem a apresentar padrões de ocorrência mais similares do que padrões de espaço das feições. Esse comportamento também foi observado na tarefa RHE. Entretanto, existe uma interseção entre todas as distribuições independentemente do valor de ω.

Os resultados do modelo *GeoC2Vec_Dtc* apresentaram comportamento menos uniforme para os diferentes valores de ω. Acredita-se que essa variação ocorra devido à os valores das distribuições para os diferentes raios utilizados. Entretanto, é possível observar que todas as distribuições desse modelo estão mais abaixo. Os resultados do teste estatístico foram os mesmos para os valores dos modelos distribuídos por raio.

Considerando as interseções existentes entre as distribuições dos três modelos, foi verificado se existe diferença estatística significativa entre elas. Inicialmente testou-se a normalidade dos dados a partir do teste *Shapiro-Wilk* para averiguar a hipótese H_0. Nesse caso, o teste *Shapiro-Wilk* retornou $p - valores$ abaixo de $0,05$ para uma das abordagem. Logo, a hipótese nula é refutada com 95% de confiança para a amostra utilizada, indicando que deve-se empregar um teste não paramétrico (como o teste de *Friedman*).

O resultado do teste de *Friedman*, com nível de significância de 5% ($\alpha = 0,05$), retornou um *p*-valor igual a $0,119$. Isso demonstra que não existe uma diferença estatisticamente significativa entre as distribuições dos três modelos. Esse resultado é principalmente evidenciado para valores de raio maiores que $400m$.

Aplicando os *embeddings* do ITDL na mesma tarefa, percebe-se que esse modelo obteve uma média de MRR de aproximadamente 50% (conforme ilustrado na Figura 6.19). Isso ponta que, em média, os tipos hierarquicamente similares ocupam a segunda posição conforme a similaridade do cosseno. Ou seja, as relações contextuais de vizinhança de POI também indicam a similaridade hierárquica dos POIs. Além disso, a variação da distribuição de acordo com os raios não apresentou tanta divergência, sinalizando que tal variação esteja mais associada às ocorrências e *check-ins* do que a vizinhança em si.

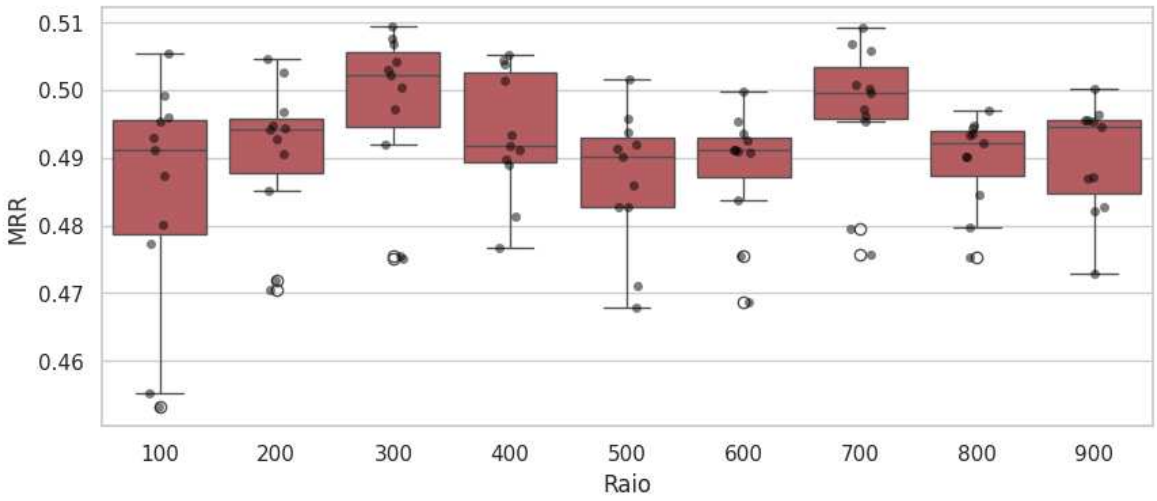


Figura 6.19: Resultados de MRR por valor de raio para o ITDL.

Fonte: Autoria própria

Analisando as distribuições obtidas para cada valor de σ (Figura 6.20), percebe-se que os valores caem à medida que σ aumenta. Isto indica que, a informação de popularidade

(*check-ins*), faz com que a similaridade do cosseno aponte outros tipos como mais similares e estes não são necessariamente os mais similares pela hierarquia. Esse resultado reflete a própria configuração de vizinhança dos POIs. Como exemplo, é possível que shoppings e restaurantes sejam vizinhos e apresentem muitos *check-ins* registrados. Entretanto, hierarquicamente falando, esses tipos não são os mais similares pois estão em ramos diferentes da hierarquia.

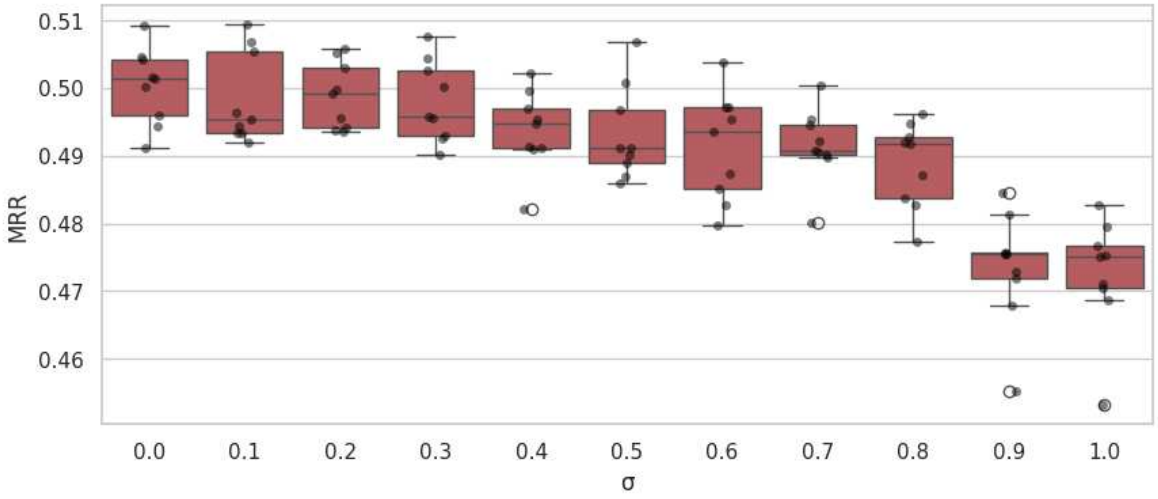


Figura 6.20: Resultados de MRR por valor de σ para o ITDL.

Fonte: Autoria própria

Empregando os *embeddings* das melhores distribuições (GeoContext2Vec com $\omega = 0,2$ e ITDL com $\sigma = 0,0$), juntamente com os *embeddings* produzidos com o *Shortest Path*, observa-se que o ITDL apresenta o melhor resultado (Figura 6.21). Este método não se limita apenas à vizinhança de POIs, mas também utiliza informações de unicidade (raridade do tipo) e popularidade para replicar as relações binárias no conjunto de treinamento. Isso mostra que as relações contextuais de vizinhança dos POIs também refletem a similaridade presente na hierarquia. Especificamente, a raridade (dada por $\sigma = 0,0$) dos tipos indicam com maior precisão a similaridade presente na hierarquia. Além disso, percebe-se que o resultado do *Shortest Path*, baseado exclusivamente na vizinhança, tem o pior desempenho apontando o benefício das estratégias utilizadas no ITDL. Os resultados do GeoContext2Vec permanecem competitivos, estando acima dos resultados do *Shortest Path* e sugerem que as relações contextuais dos tipos de POI com as feições geográficas refletem em certo grau a similaridade dos tipos considerando a hierarquia.

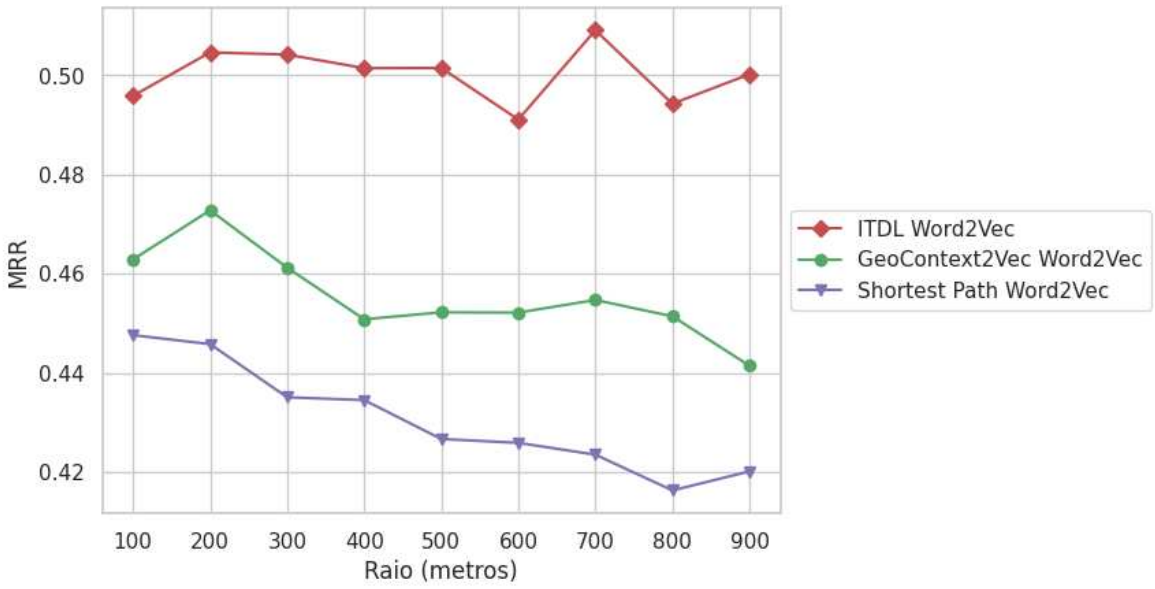
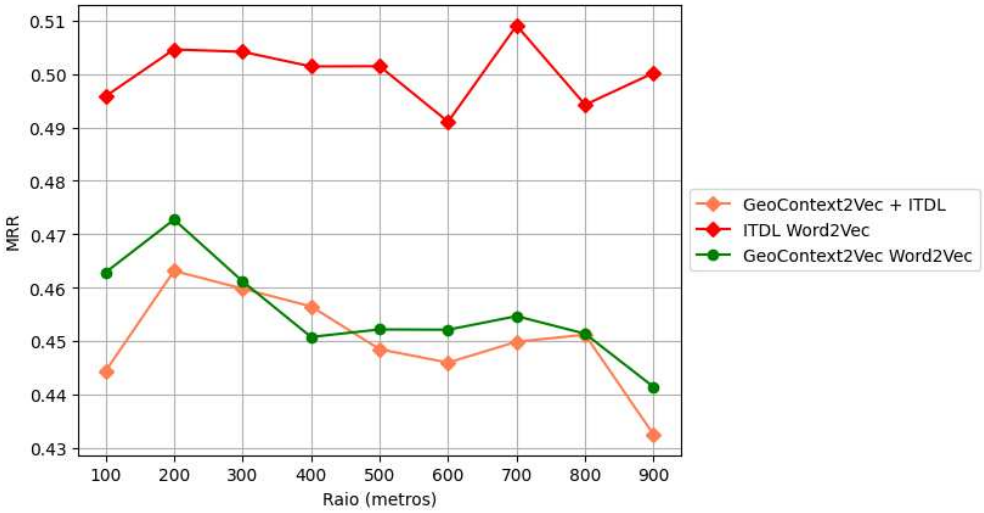


Figura 6.21: Resultados de MRR por valor de raio para todos os modelos.

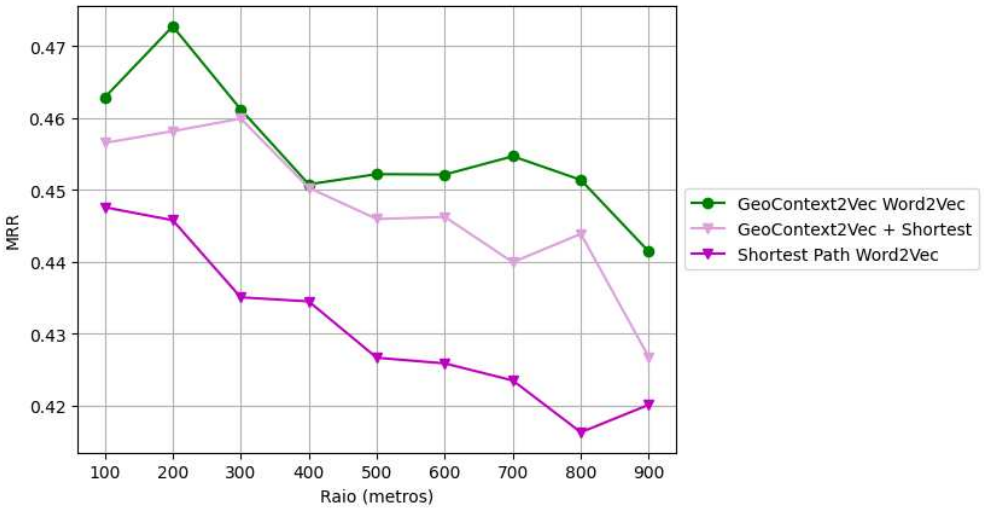
Fonte: Autoria própria

Assim como nas tarefas, BHE e RHE, realizou-se a concatenação dos *embeddings* produzidos com o GeoContext2Vec para analisar se existe melhoria nos resultados de MRR dessa tarefa. Os resultados dessa combinação são ilustrados na Figura 6.22.

Os resultados da combinação do GeoContext2Vec com o *Shortest Path* mostraram uma melhoria em relação ao desempenho do *Shortest Path* isolado. No entanto, essa combinação não superou o GeoContext2Vec, alcançando no máximo resultados equivalentes para alguns valores de raio. Isso sugere que a vizinhança de POIs por si só não reflete tão bem a similaridade hierárquica. No entanto, ao combinar essa informação com as feições geográficas do contexto, as similaridades contextuais entre os tipos de POI tornaram-se mais alinhadas com a hierarquia. Por outro lado, os resultados da combinação do GeoContext2Vec com o ITDL não apresentaram melhoria, com desempenho inferior aos demais modelos, especialmente para raios menores. Esse resultado sugere que a propriedade de unicidade do ITDL revela que os POIs hierarquicamente relacionados tendem a ser únicos no contexto, independentemente das feições geográficas.



(a) GeoContext2Vec concatenado com ITDL.



(b) GeoContext2Vec concatenado com *Shortest Path*.

Figura 6.22: Resultado MRR da combinação dos *embeddings* do GeoContext2Vec com os *embeddings* dos *baselines*.

6.3.2 MRR com DistilBert

Aplicando os *embeddings* do GeoContext2Vec obtidos com o DistilBert (Figura 6.23), observa-se que os valores de MRR são superiores a todos os outros métodos, alcançando um máximo de 61%. Isso sugere que o tipo de POI mais similar, considerando a similaridade do cosseno, varia entre a primeira e segunda posição no ranking hierárquico. Esse resultado é atribuído ao conhecimento prévio do modelo relação aos tipos de POI e também à capacidade do DistilBert de capturar nuances contextuais relacionadas às feições geográficas do contexto dos POIs, tornando esse modelo muito mais eficaz em discernir as diferenças e similaridades contextuais entre os tipos de POI.

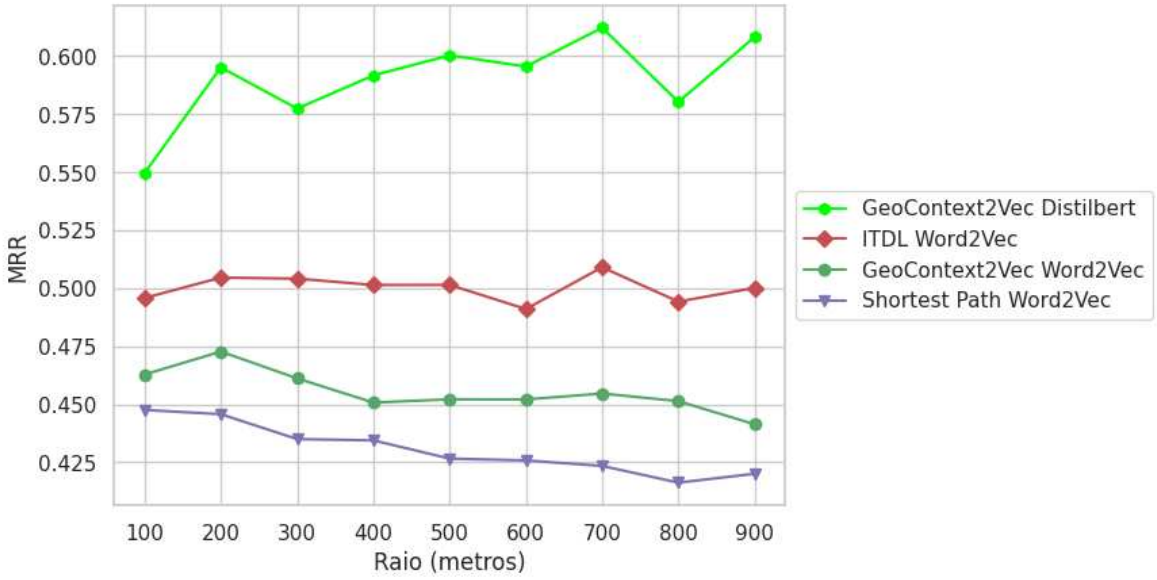
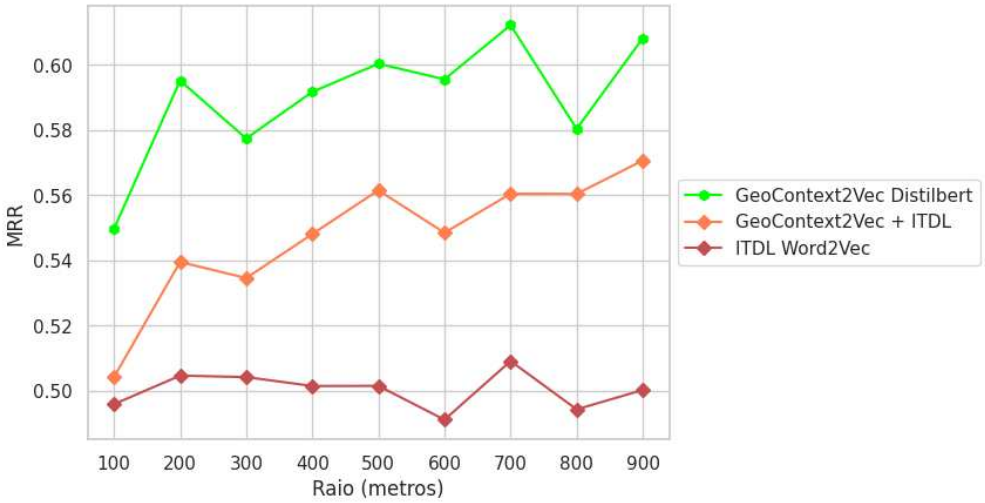


Figura 6.23: Resultados de MRR por valor de raio para todos os modelos e GeoContext2Vec DistilBert.

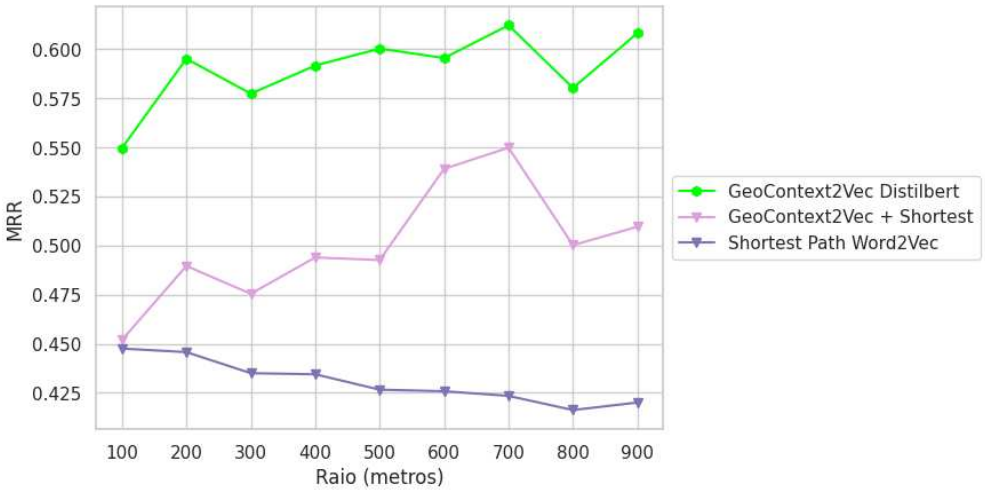
Fonte: Autoria própria

Concatenando os *embeddings* do GeoContext2Vec DistilBert com os *embeddings* dos demais modelos (Figura 6.24), observa-se que os resultados dos modelos concatenados foram superiores aos resultados dos dois *baselines*. No entanto, em nenhum caso, os resultados superaram os *embeddings* do GeoContext2Vec isolado. Esse resultado reforça a superioridade do DistilBert em gerar *embeddings* mais robustos, capazes de capturar a relação contextual dos tipos de POI com as feições geográficas de maneira mais precisa. Além disso, tal resultado demonstra também que o conhecimento prévio do modelo, juntamente com as relações

contexuais dos tipos com as feições geográficas, refletem a similaridade de tipos presente na hierarquia.



(a) GeoContext2Vec DistilBert concatenado com ITDL.



(b) GeoContext2Vec DistilBert concatenado com *Shortest Path*.

Figura 6.24: Resultado MRR da combinação dos *embeddings* do GeoContext2Vec DistilBert com os *embeddings* dos *baselines*.

6.4 Visualização dos *Embeddings*

Entre os modelos investigados nesta pesquisa, optou-se por gerar a visualização dos *embeddings* do ITDL, por terem os melhores resultados nas tarefas anteriores utilizando a vizinhança de POIs, e do GeoContext2Vec DistilBert, por ter apresentado os melhores resultados

em todas as tarefas. A Figura 6.25 ilustra o resultado da redução para os *embeddings* do GeoContext2Vec DistilBert.

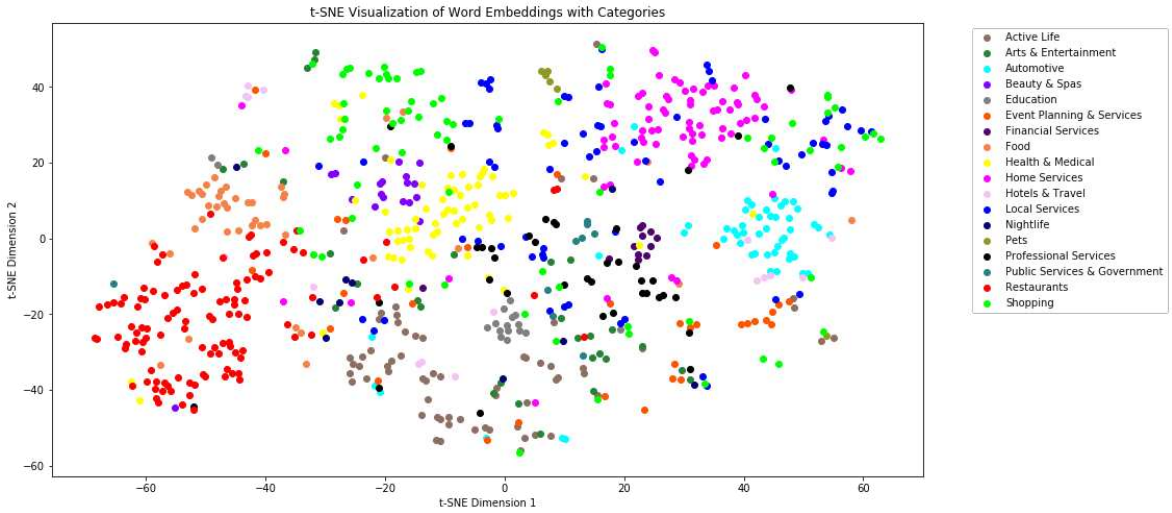


Figura 6.25: Visualização 2D dos *embeddings* do GeoContext2Vec DistilBert.

Fonte: Autoria própria

Observando-se os *embeddings* no espaço bidimensional, é possível notar que existem grupos bem definidos para várias categorias. No lado esquerdo da figura, percebe-se pontos vermelhos e laranja claro próximos. Esses pontos são *embeddings* de tipos de POI que são filhos das categorias `Restaurants` e `Food`. As duas categorias fazem referência à lugares que vendem alimentos. Entre os filhos da primeira categoria, estão os vários tipos de restaurantes registrados pelo Yelp, como `Chinese`, `French`, entre outros[2]. A segunda categoria apresenta tipos como `Coffee & Tea`, `Churros`, entre outros. Essa proximidade na visualização sugere que POIs que possuem esses tipos filhos dessas categorias são similares em relação às feições geográficas do contexto. Ou seja, diferentes tipos de restaurante tendem a apresentar os mesmos padrões de feições no contexto, assim como locais que vendem certos tipos de comida. Além disso, a proximidade entre os *embeddings* das duas categorias ocorre porque, em alguns casos, o mesmo POI pode assumir tipos das duas categorias simultaneamente, pois restaurantes vendem comida.

Um pouco mais ao centro, percebe-se um grupo de pontos amarelos e um pequeno grupo de pontos lilás. Esses dois grupos são formados pelos *embeddings* dos tipos filhos das categorias `Beauty & Spas` e `Health & Medical`. A primeira categoria agrupa tipos

[2]Disponível em https://docs.developer.yelp.com/docs/resources-categories. Acesso em 20 de maio de 2024.

relacionados a serviços de beleza como `Hair Salons` e `Barbers`. Porém, ela também possui tipos que associam beleza à saúde, como `Medical Spas`, entre outros. A segunda categoria abrange todos os tipos relacionados a serviços médicos e de saúde, comportando tipos como `Dentists`, `Acupuncture`, `Saunas`, entre outros. A proximidade dos pontos dessas categorias revela que POIs que possuem tipos filhos dessas categorias apresentam similaridade em relação às feições geográficas de seus contextos. Além disso, não há tantos POIs compartilhando os tipos das duas categorias, como ocorre com os tipos filhos de `Restaurants` e `Food`. A proximidade dos pontos aponta que, de fato, há uma similaridade contextual dos POIs que possuem esses tipos.

A parte superior da figura revela um grupo de pontos verdes. Essa cor faz referência à categoria `Shopping`. Essa categoria agrupa tipos de POIs que indicam diversos serviços de venda, como `Drugstores`, `Electronics`, `Fashion`. Para essa categoria, percebe-se que os pontos são um poucos mais esparsos e constam e mais locais da Figura 6.25. Esse comportamento sinaliza que alguns tipos filhos de `Shopping` apresentam certa similaridade contextual das feições geográficas. Entretanto, devido tal categoria englobar muitos tipos que vendem produtos diferentes, percebe-se que é natural tal diferença. Como exemplo, é provável que mercados públicos, representado pelo tipo `Public Markets`, possuam um contexto de feições diferente de locais que vendem quadrinhos, representado pelo tipo `Comic Books`. Mesmo existindo essa variação de produtos ou serviços ofertados pelos tipos, percebe-se a formação de um grupo.

Na parte central e inferior da figura, há um grupo de pontos em um tom marrom claro, que corresponde à categoria `Active Life`. Essa categoria engloba tipos de POIs onde as pessoas praticam esportes, como `Golf`, `Baseball Fields`, `Gyms`, entre outros. Assim como na categoria `Shopping`, essa categoria abrange tipos relacionados a diversos esportes. Cada tipo representa esportes praticados em espaços diferentes; por exemplo, `Golf` é praticado em áreas abertas com gramado e muito espaço, enquanto os esportes de ginásio (`Gyms`) são praticados em ambientes fechados, muitas vezes em edifícios de grande proporção. Devido a essa diversidade de espaços esportivos, muitos pontos da categoria `Active Life` estão próximos, mas não formam um grupo denso.

Próximo aos pontos `Active Life`, encontram-se os pontos da categoria `Education` (representados na tonalidade cinza). Dentro dessa categoria, estão inclusos os tipos

`Colleges & Universities`, que representam universidades, e `Middle Schools & High Schools`, que representam escolas para crianças e adolescentes. Observa-se que os pontos desse grupo estão bastante próximos entre si, o que demonstra que POIs que apresentam tipos pertencentes a essa categoria possuem uma similaridade contextual maior. Ou seja, ambientes de educação apresentam padrão bem definido em relações às feições contextuais.

Um pouco acima e à direita dos pontos de `Education`, encontra-se um grupo de pontos roxos, que representam a categoria `Financial Services`. Ao redor desse grupo, há um conjunto de pontos em verde-azulado, representados pela categoria `Public Services & Government`, e outro conjunto de pontos pretos, representando a categoria `Professional Services`. A categoria `Financial Services` inclui tipos de POI relacionados a serviços bancários, seguros, investimentos, entre outros. A categoria `Public Services & Government`, por sua vez, engloba tipos como tribunais, embaixadas, escritórios de impostos, entre outros. A categoria `Professional Services` abrange tipos relacionados a advocacia, seguros, arquitetura, gráficas, entre outros.

A proximidade desses tipos sinaliza que seus POIs compartilham muitas feições geográficas do contexto. Nesse caso, observa-se uma vizinhança entre POIs de categorias diferentes, pois é natural que perto de bancos e serviços de impostos existam escritórios de advocacia ou serviços de seguros. Além disso, muitas vezes o contexto geográfico dos bancos se assemelha ao contexto geográfico de locais que fornecem serviços públicos, pois necessitam de prédios maiores para acomodar muitas pessoas, fornecendo estacionamentos. Em relação aos pontos da categoria `Professional Services`, nota-se que eles estão mais dispersos. Isso ocorre porque essa categoria também inclui tipos de POI de serviços mais variados, como reparo de barcos e provedores de internet.

Um pouco mais ao centro e à direita, encontra-se um grupo de pontos na cor ciano. Esses pontos representam a categoria `Automotive`, que inclui tipos como postos de gasolina, lava-jatos, oficinas mecânicas, lojas de veículos e peças automotivas. A visualização revela que POIs associados a essa categoria compartilham similaridades nas feições geográficas do contexto. Além disso, observa-se que esses pontos estão muito próximos, indicando uma alta similaridade.

Esse resultado é em parte atribuído ao compartilhamento de tipos por alguns POIs. Por

exemplo, é comum que algumas lojas de veículos ofereçam serviços de manutenção, e postos de gasolina ofereçam serviços de lavagem. Além disso, mesmo quando os POIs não compartilham tipos, mas pertencem à mesma categoria, tendem a estar localizados próximos às principais rodovias para facilitar o acesso aos veículos. Além disso, os edifícios desses POIs costumam ser maiores para acomodar mais veículos.

Na parte superior à direita, encontra-se um grupo de pontos na cor rosa, associados à categoria `Home Services`. Essa categoria inclui tipos relacionados à instalação de carpetes, jardinagem, limpeza de casas, limpeza de piscinas, instalação de sistemas de segurança, entre outros. A visualização demonstra que, apesar da diversidade de serviços, os POIs que oferecem esses tipos compartilham feições geográficas muito similares, como demonstrado pela proximidade dos pontos.

Além disso, há a presença de alguns pontos azuis, pertencentes à categoria `Local Services`, que engloba tipos como limpeza de carpetes, reparo de móveis, serviços de lavanderia, entre outros. Devido à variedade de serviços oferecidos pela categoria `Local Services`, que não se restringem apenas a elementos domésticos, observa-se que seus pontos estão espalhados por quase toda a figura.

Investigando os *embeddings* do ITDL no espaço vetorial bidimensional, é possível identificar alguns grupos de pontos. Relacionando-se com as observações feitas sobre a visualização gerada com os *embeddings* do GeoContext2Vec, percebe-se na visualização que as categorias `Restaurants`, `Food`, `Health & Medical`, `Beauty & Spas`, `Home Services` e `Shopping` também apresentam grupos de pontos próximos.

Esse resultado sugere que, assim como os POIs são similares em relação às feições geográficas do contexto, eles também são similares em relação à sua vizinhança. Em outras palavras, é comum que POIs que oferecem tipos dessas categorias tenham vizinhanças semelhantes. Além disso, considerando que o ITDL utiliza as informações de *check-in* e unicidade do tipo, também é possível constatar que os tipos da mesma categoria, tendem a apresentar os mesmos padrões dessas duas informações.

Em parte, esse resultado decorre da tendência comum de certos POIs estarem geograficamente próximos. Por exemplo, em áreas onde há um restaurante, é frequente encontrar outros tipos de restaurantes oferecendo alimentos geralmente consumidos em conjunto, como almoço e sobremesa, ou cafés. Da mesma forma, locais que possuem postos de gasolina

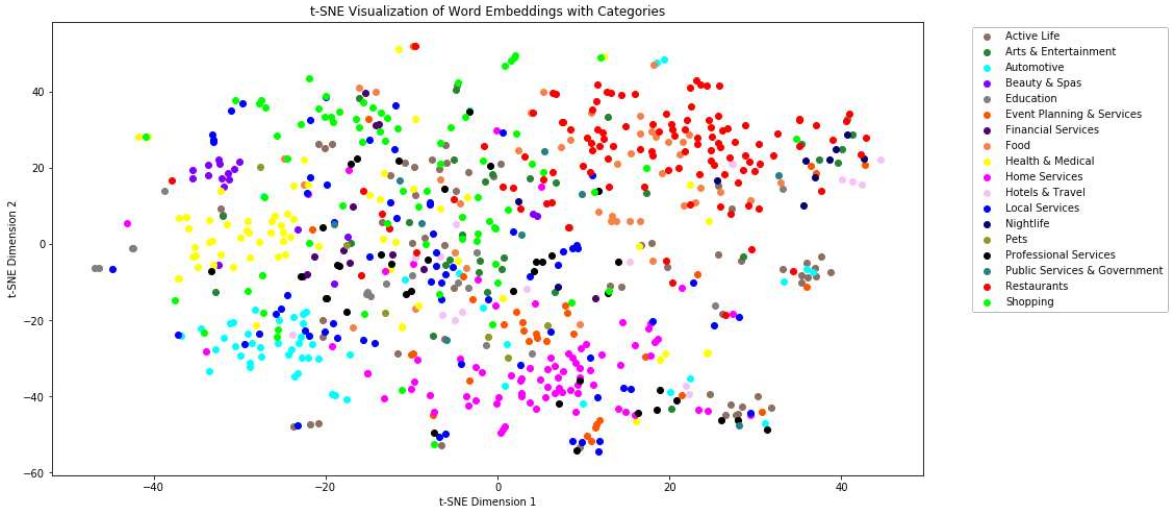


Figura 6.26: Visualização 2D dos *embeddings* do ITDL Word2Vec.

Fonte: Autoria própria

tendem a ter lava-jatos ou lojas de peças automotivas em sua vizinhança. Portanto, há uma relação de vizinhança bidirecional entre alguns POIs.

Comparando as visualizações geradas pelas duas abordagens, observa-se que os pontos dos *embeddings* produzidos pelo GeoContext2Vec exibem grupos mais distintos e com menos variação no centro da figura. Isso sugere que as feições geográficas do contexto dos POIs possuem um padrão mais bem definido do que a vizinhança desses POIs. Esse resultado pode ser associado ao desempenho do GeoContext2Vec na tarefa BHE, onde demonstrou-se que os *embeddings* dessa abordagem apresentaram maior correlação com as diferenças indicadas pelas pessoas.

Por fim, foi gerada a visualização da combinação dos *embeddings* do GeoContext2Vec e do ITDL (Figura 6.27). A visualização revela um agrupamento muito evidente para os pontos da maioria das categorias. Esse resultado demonstra que a maioria dos tipos de POI que estão na mesma categoria está associada a POIs que são similares considerando o contexto geográfico e a vizinhança de POIs. Desse modo, as duas configurações juntas possibilitaram o melhor agrupamento dos *embeddings*. Esse resultado também está em linha com o encontrado na tarefa BHE, onde a combinação dos *embeddings* alcançou o maior resultado em diferenciar os tipos de POI.

Na visualização, percebe-se que há mais distinção entre os pontos das categorias `Restaurants` e `Food`. Nesse caso, os pontos de cada categoria estão mais próximos

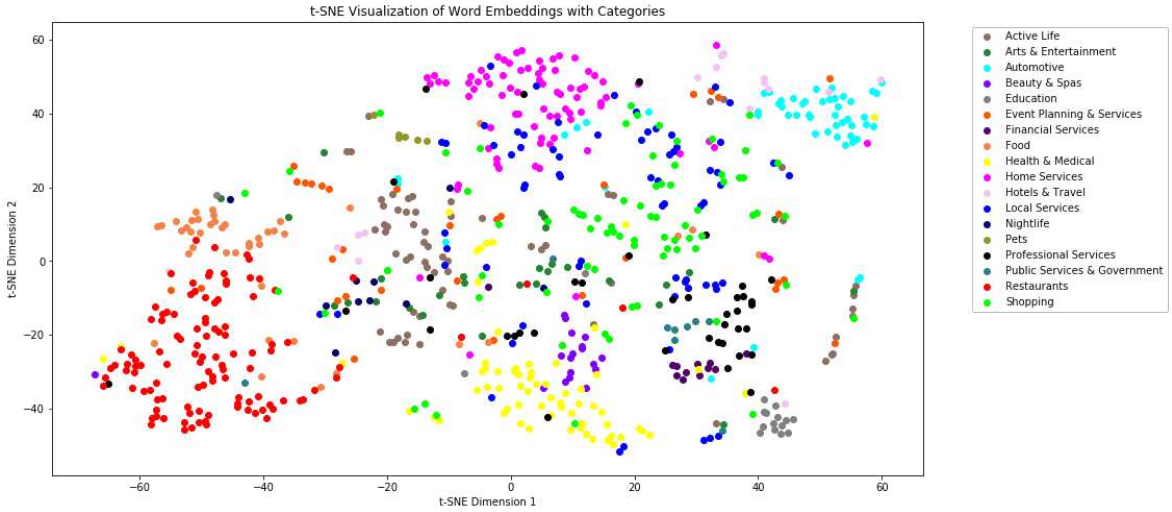


Figura 6.27: Visualização 2D dos *embeddings* combinados entre o GeoContext2Vec e ITDL.

Fonte: Autoria própria

entre si, mas ainda mantêm proximidade entre as duas categorias. O grupo `Beauty & Spas` tornou-se ainda mais próximo do grupo `Health & Medical`, sugerindo que esses dois grupos apresentam similaridade tanto em feições geográficas como em vizinhança de POIs. Percebe-se também que o grupo `Home Services` apresenta muitos pontos do grupo `Local Services` próximos. Conforme mencionado anteriormente, os dois grupos apresentam tipos que dispõem de serviços diversos sendo diretamente ou indiretamente ligados a questões domésticas. O resultado demonstra que essas duas categorias apresentam similaridade de suas feições e vizinhança.

Os pontos do grupo `Shopping` estão menos esparsos ao longo de toda a figura e concentram-se próximos aos pontos das categorias `Local Services` e `Professional Services`. Isso demonstra que locais que fornecem diferentes serviços têm uma vizinhança de POIs e feições do contexto semelhantes a muitos tipos da categoria `Shopping`. A categoria `Education` permanece agrupada, assim como na visualização do GeoContext2Vec, demonstrando que os POIs dessa categoria também apresentam padrões similares de vizinhança e feições geográficas. Por fim, os pontos da categoria `Automotive` se mantiveram agrupados e um pouco mais próximos entre si.

Considerando os resultados das visualizações do *GeoContext2Vec* isolado e combinado com o *ITDL*, pode-se afirmar que no espaço vetorial, tipos de POI que são hierarquicamente similares também estão próximos. Ou seja, esses tipos também apresentam similaridades

com base em suas relações contextuais com as feições geográficas.

6.5 Classificação de Zonas Urbanas

Para realizar essa tarefa foram utilizadas as 406 zonas da cidade de Austin. Essas zonas são divididas nas categorias de nível superior denominadas Comercial (230 exemplos), Residencial (105 exemplos), Industrial (45 exemplos) e Propósito Especial (26 exemplos)[3]. Os experimentos se concentraram nas categorias Residencial e Comercial, pois são as mais representativas no conjunto de dados. Essa decisão de utilizar apenas essas duas categorias está relacionada à quantidade de interseção de tipos de POI entre as zonas, conforme detalhes que constam no Apêndice B. Além disso, o objetivo desse estudo é demonstrar a eficiência dos *embeddings* da abordagem proposta sem aprofundar-se muito na área de classificação de zonas.

Conforme descrito na Seção 5.3.2, os *embeddings* de cada tipo de POI pertencente a cada zona foram utilizados para gerar os *embeddings* das zonas. Especificamente, foram empregados os tipos do terceiro nível da hierarquia do Yelp, pois isso possibilita diferenciar melhor cada zona, uma vez que tipos que estão em níveis mais altos da hierarquia tornam as zonas mais gerais. Para obtenção dos *embeddings*, foram utilizados os modelos GeoContext2Vec, ITDL e *Shortest Path*, bem como suas versões combinadas. Especificamente, foram selecionados os modelos que apresentaram os melhores resultados na tarefa BHE.

A classificação foi realizada utilizando o Algoritmo *Random Forest* em conjunto com a técnica de k-fold estratificada com $k = 5$. Para lidar com problemas de desbalanceamento dos dados de treino, foi utilizado o algoritmo *RandomUndersampling* da classe *imbalanced-learn*[4], que permite reamostrar todas as classes, exceto a minoritária. Para cada resultado do k-fold, computou-se a média e o desvio padrão das métricas de acurácia, precisão, *recall* e F1-Score. Os resultados estão condensados na Tabela 6.3. É importante frisar que não foi realizado nenhum trabalho de busca dos melhores parâmetros do algoritmo *Random Forest* para obtenção da melhor combinação, pois o objetivo é apenas demonstrar a utilidade dos

[3]Disponível em https://www.austintexas.gov/sites/default/files/files/Planning/zoning_guide.pdf. Acesso em 20 de maio de 2024.

[4]Disponível em https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.ht
Acesso em 20 de maio de 2024.

embeddings e não a obtenção dos melhores resultados possíveis.

Levando em consideração os resultados obtidos, percebe-se que o modelo *Shortest Path + GeoC2Vec W2V* se destaca dos demais, alcançando valores em torno de 90% para as métricas de acurácia, *recall* e F1-score. Mesmo que esse modelo não tenha apresentado o melhor resultado de precisão, seu valor está muito próximo do melhor resultado (89%). O método *Shortest Path* foi concebido para fornecer *embeddings* para uma tarefa de classificação de zonas urbanas. É possível notar que ele obteve melhores resultados que o ITDL, que foi projetado para gerar *embeddings* independentemente de tarefas específicas. Percebe-se, portanto, que adicionar os *embeddings* produzidos a partir das relações contextuais das feições geográficas com os tipos de POI permitiu que a classificação fosse ainda mais precisa do que apenas utilizar *embeddings* baseados na vizinhança dos POIs. Além disso, observa-se que os resultados do *Shortest Path* combinados com o GeoC2Vec DistilBert também produziram resultados melhores do que apenas o *Shortest Path*. Nesse caso, esse modelo apresentou valores de métrica em torno de 90%.

Observando-se os resultados do ITDL, percebe-se que individualmente esse modelo apresentou o menor desempenho. Os valores de suas métricas ficaram em torno de 85%. No entanto, quando foram adicionados os *embeddings* do GeoContext2Vec Word2Vec e DistilBert, observa-se que os valores aumentaram em todas as métricas. Os resultados do *ITDL+GeoC2Vec W2V* apresentam valores em torno de 90%, enquanto o *ITDL+GeoC2Vec Distil* mostra valores em torno de 88%. Esses resultados sugerem que os *embeddings* do GeoContext2Vec possibilitaram ao *Random Forest* diferenciar melhor áreas comerciais de áreas residenciais.

Verificando os resultados dos *embeddings* do GeoContext2Vec Word2Vec e DistilBert isoladamente, percebe-se que os valores são próximos dos valores dos modelos *Shortest Path* combinados. Para os *embeddings* gerados com o Word2Vec e DistilBert, observa-se uma média de 90% para todas as métricas. Isso sugere que não houve uma diferença nos *embeddings* produzidos com os dois métodos. Entretanto, os modelos combinados com os *embeddings* do Word2Vec alcançaram melhores resultados. Acredita-se que a superioridade do GeoContext2Vec Word2Vec pode ser atribuída ao fato de que os *embeddings* desse modelo foram gerados exclusivamente com feições geográficas. Por sua vez, o GeoContext2Vec DistilBert possui conhecimento prévio sobre os tipos de POI obtidos dos pesos pré-carregados do

***Embeddings* do Modelo**	**Accuracy (± std)**	**Precision (± std)**	**Recall (± std)**	**F1 (± std)**
Shortest Path ($100m$)	0,8806 (± 0,0283)	0,8571 (± 0,0310)	0,8820 (± 0,0295)	0,8666 (± 0,0307)
Shortest Path + GeoC2Vec W2V ($500m$)	**0,9104** (± **0,0400**)	0,8895 (± 0,0442)	**0,9167** (± **0,0419**)	**0,9002** (± **0,0441**)
Shortest Path + GeoC2Vec Distil ($200m$)	0,9045 (± 0,0179)	0,8833 (± 0,0184)	0,9046 (± 0,0266)	0,8920 (± 0,0209)
ITDL ($100m$)	0,8657 (± 0,0353)	0,8418 (± 0,0364)	0,8634 (± 0,0471)	0,8488 (± 0,0401)
ITDL + GeoC2Vec W2V ($500m$)	0,9075 (± 0,0346)	0,8861 (± 0,0379)	0,9119 (± 0,0396)	0,8963 (± 0,0387)
ITDL + GeoC2Vec Distil ($200m$)	0,8925 (± 0,0257)	0,8703 (± 0,0282)	0,8907 (± 0,0308)	0,8787 (± 0,0291)
GeoC2Vec W2V ($100m$)	**0,9104** (± **0,0340**)	**0,8927** (± **0,0415**)	0,9115 (± 0,0305)	0,8995 (± 0,0368)
GeoC2Vec Distil ($200m$)	0,9045 (± 0,0322)	0,8825 (± 0,0348)	0,9123 (± 0,0334)	0,8938 (± 0,0351)

Tabela 6.3: Resultado da classificação para as categorias Comércio e Residencial.

distilbert-base-uncased. Desse modo, é possível que esse conhecimento tenha confundido um pouco o modelo de classificação nessa tarefa.

Um segundo experimento de classificação foi realizado considerando classes mais específicas das zonas Comercial e Residencial. Nesse caso, utilizou-se as classes Comércio (167 exemplos), Familiar (97 exemplos) e de Escritórios (55 exemplos). A Tabela 6.4 sumariza os resultados obtidos.

Modelo	**Accuracy (± std)**	**Precision (± std)**	**Recall (± std)**	**F1 (± std)**
Shortest Path ($100m$)	0,8656 (± 0,0415)	0,8455 (± 0,0487)	0,8469 (± 0,0353)	0,8422 (± 0,0415)
Shortest Path + GeoC2Vec W2V ($500m$)	0,8781 (± 0,0182)	0,8584 (± 0,0223)	0,8667 (± 0,0259)	0,8590 (± 0,0214)
Shortest Path + GeoC2Vec Distil ($200m$)	**0,8813 (± 0,0290)**	0,8585 (± 0,0309)	**0,8730 (± 0,0301)**	**0,8624 (± 0,0287)**
ITDL ($100m$)	0,8500 (± 0,0449)	0,8180 (± 0,0508)	0,8363 (± 0,0539)	0,8245 (± 0,0529)
ITDL + GeoC2Vec W2V ($500m$)	0,8750 (± 0,0221)	0,8515 (± 0,0209)	0,8684 (± 0,0220)	0,8567 (± 0,0210)
ITDL + GeoC2Vec Distil ($200m$)	0,8750 (± 0,0328)	0,8494 (± 0,0398)	0,8654 (± 0,0381)	0,8545 (± 0,0395)
GeoC2Vec W2V ($100m$)	0,8781 (± 0,0334)	**0,8589 (± 0,0237)**	0,8710 (± 0,0417)	0,8585 (± 0,0341)
GeoC2Vec Distil ($200m$)	0,8656 (± 0,0364)	0,8467 (± 0,0477)	0,8489 (± 0,0388)	0,8428 (± 0,0402)

Tabela 6.4: Resultado da classificação para as categorias Comercial, Familiar e Escritórios.

Analisando os resultados, percebe-se um comportamento similar aos resultados da Tabela 6.3. É possível notar que os valores das métricas alcançados pelo *Shortest Path* estão um pouco acima dos resultados alcançados pelo ITDL, sugerindo que a abordagem do caminho mínimo gera melhores resultados nessa tarefa. Nesse caso, o primeiro método apresenta valores em torno de 85%, enquanto que o segundo apresenta valores em torno de 83%.

Percebe-se também que os valores obtidos a partir da combinação dos *embeddings* do GeoContext2Vec Word2Vec e DistilBert com os *baselines* são um pouco maiores que os valores dos *baselines* isolados. Esse resultado corrobora com o resultado da primeira análise, demonstrando que os *embeddings* produzidos a partir das relações contextuais com as feições forneceram uma representação que permite ao algoritmo *Random Forest* diferenciar melhor cada zona urbana do conjunto de teste. Nesse caso, o modelo *Shortest Path + GeoC2Vec W2V* apresenta valores de métrica em torno de 86%, o *Shortest Path + GeoC2Vec Distil* apresenta valores de métrica em torno de 87%, *ITDL + GeoC2Vec W2V* apresenta valores de métrica em torno de 85%, e o *ITDL + GeoC2Vec Distil* apresenta valores de métrica em torno de 85%.

Investigando os resultados dos *embeddings* do GeoContext2Vec isoladamente, percebe-se que os valores do Word2Vec são ligeiramente maiores. No entanto, os *embeddings* do DistilBert produziram o melhor resultado combinado com o *Shortest Path*. Desse modo, não é possível afirmar qual dos dois modelos obteve o melhor desempenho. Apenas é possível dizer que os *embeddings* obtidos pelos dois modelos combinados com os *baselines* possibilitam resultados melhores. Dessa forma, pode-se concluir que *embeddings* produzidos a partir das relações contextuais dos tipos de POI com feições geográficas permitiram que uma tarefa de classificação de zonas alcançasse melhores resultados.

6.6 Considerações Finais

Este capítulo detalhou os resultados alcançados durante a pesquisa realizada para esta tese. Inicialmente, foram discutidos os resultados de três tarefas que visavam analisar se os *embeddings* produzidos com as relações contextuais dos tipos de POI com as feições geográficas indicavam a similaridade dos tipos. Para isso, foram adotadas três tarefas, denominadas BHE, RHE e Similaridade Hierárquica. Os resultados alcançados demonstraram que tais

embeddings refletem a similaridade dos tipos tanto considerando a opinião humana quanto hierarquias pré-concebidas. Esse resultado permite responder à questão de pesquisa $\mathbf{QP}_2$. Além disso, foi demonstrado que os *embeddings* produzidos com DistilBert produziram resultados de similaridade superiores aos produzidos com o Word2Vec, respondendo à questão de pesquisa $\mathbf{QP}_4$.

Também foram investigadas visualizações dos *embeddings* considerando a abordagem proposta e uma abordagem *baseline*. As distribuições dos *embeddings* no espaço vetorial latente demonstraram a formação de grupos de tipos que configuraram a similaridade das relações contextuais dos tipos com as feições. Além disso, os grupos mostraram-se mais evidentes do que os gerados com abordagens baseadas na vizinhança de POIs. Esse resultado também corrobora com a $\mathbf{QP}_2$, pois a proximidade dos pontos devido à sua similaridade contextual com as feições também refletiu a similaridade dos tipos.

Por fim, foi realizada uma tarefa de classificação de zonas urbanas para investigar se os *embeddings* do GeoContext2Vec permitiam que a tarefa obtivesse resultados melhores em comparação com *baselines*. Foi visto nos resultados que os *embeddings* do GeoContext2Vec isolados alcançaram resultados melhores do que os *embeddings* do ITDL e *Shortest Path*. Além disso, a combinação dos *embeddings* do GeoContext2Vec com os *embeddings* do ITDL e *Shortest Path* produziram os melhores resultados, permitindo assim responder às questões de pesquisa $\mathbf{QP}_3$ e $\mathbf{QP}_4$.

O capítulo seguinte descreve um estudo de caso utilizando *embeddings* das feições geográficas em uma tarefa de busca de POIs.

Capítulo 7

Estudo de Caso

Este capítulo apresenta um estudo de caso que utiliza as feições geográficas do contexto dos POIs em uma tarefa de busca de POIs. A busca por POIs é comumente realizada em aplicações de navegação e turismo, baseando-se em informações como o nome do POI, endereço ou seus tipos. No entanto, basear-se apenas nessas informações pode não ser suficiente para encontrar POIs com características específicas.

Neste sentido, o objetivo desse estudo é demonstrar que, assim como os *embeddings* de tipos de POI podem ser utilizados em tarefas como classificação de zonas urbanas, os *embeddings* das feições também podem ser utilizados em tarefas como busca de POI. Além disso, tal estudo de caso demonstra que o GeoContext2Vec é uma alternativa aos métodos que se baseiam em imagens para realização da busca, pois utiliza *embeddings* das feições obtidos a partir de modelos de PLN. As seções a seguir descrevem os detalhes desse estudo de caso.

7.1 Cenário

Com o objetivo de demonstrar uma aplicação prática, foi planejado um cenário onde se pretende encontrar um local para abertura de um novo negócio ou inserção de algum serviço e que não apresentasse outro tipo igual, ou seja, os locais sugeridos não devem conter nenhum outro POI que compartilhe o mesmo tipo. Além do aspecto de concorrência no contexto, essa decisão de não utilizar POIs que compartilham o mesmo tipo visa evitar resultados "óbvios", pois se um POI de determinado tipo apresenta uma configuração de feições, outros POIs do

mesmo tipo possuem alta probabilidade de apresentarem um padrão muito similar.

Decidiu-se ainda utilizar os contextos de POIs já existentes para reduzir o espaço de busca, pois uma busca minuciosa em toda a cidade pode ter um custo muito elevado. Além disso, o foco desse estudo de caso não é encontrar a solução ótima em toda a cidade, mas sim demonstrar que as feições do GeoContext2Vec fornecem resultados concisos nesta tarefa.

Para realização da busca estabeleceu-se os seguintes critérios:

- Um POI e um tipo devem ser selecionados como âncora para serem utilizados como referência da busca;
- Os demais POIs serão candidatos, exceto aqueles que compartilham o mesmo tipo do POI âncora e aqueles que apresentam algum POI em seu contexto que compartilhe o mesmo tipo do POI âncora;
- A similaridade do cosseno será empregada para calcular a similaridade entre o contexto âncora e os contextos candidatos, e seus valores serão utilizados no ranqueamento;
- Se POIs candidatos apresentarem intersecção de contexto de no mínimo 20%, o candidato de maior similaridade será mantido e o outro será descartado.

7.2 Representação do Contexto do POI

Para realizar a busca é necessário definir o tamanho do contexto e como as feições geográficas pertencentes ao contexto serão utilizadas para representar todo o contexto. Assim como na tarefa de classificação de zonas urbanas, que utiliza a média dos *embeddings* dos tipos na representação do *embedding* da zona, definiu-se que a representação do contexto dos POIs se dará a partir da média de todos os *embeddings* das feições presentes no contexto. Entretanto, diferentemente do POI que tende a ser um elemento pontual, as feições possuem área ou comprimento, e essa propriedade é importante para diferenciar seus contextos.

Desse modo, antes de realizar o cálculo da média dos *embeddings* das feições, utilizou-se o cálculo do logaritmo de base 2 com o valor da área ou comprimento, para obtenção de um número inteiro que é utilizado para replicar uma determinada feição em uma lista de feições do contexto. Após isso, é empregado o cálculo da média sobre os *embeddings* das feições conforme a Equação 7.1.

$$\text{Contexto}_j = \sum_{i=1}^{M} \frac{\text{feição\_F}_{i,j}}{M} \tag{7.1}$$

em que feição_F$_{i,j}$ se refere ao *embedding* da i-ésima feição da lista de feições do contexto j, e M é o número de feições presentes na lista.

Considerando que as feições são divididas de acordo com suas geometrias (polígonos, linhas e pontos), manteve-se essa divisão no cálculo do *embedding* do contexto. Ou seja, o *embedding* do contexto é calculado considerando-se as feições poligonais como um grupo, as lineares como outro grupo e as pontuais como outro grupo. O resultado dos *embeddings* individuais é concatenado para representar o *embedding* final do contexto. Essa separação permite que as informações presentes em cada dimensão dos *embeddings* sejam combinadas com valores de *embeddings* estimados sob os mesmos critérios. Ou seja, valores dos *embeddings* dos polígonos serão combinados entre si, sem a intervenção dos valores dos *embeddings* de pontos ou linhas.

7.3 Resultados da Busca

Para execução do cenário, foram selecionados os *embeddings* do GeoContext2Vec com Word2Vec e GeoContext2Vec com DistilBert, visando analisar a diferença entre os resultados obtidos com essas duas abordagens. Para o tamanho do contexto decidiu-se utilizar $200m$, pois desse modo é possível a obtenção de uma quantidade razoável de feições, além de que raios maiores acarretaria em um custo maior para execução da busca, podendo comprometer o cronograma desta pesquisa.

O primeiro POI selecionado neste estudo possui o tipo `Coffee & Tea`, que vende cafés e chás (indicado abaixo da palavra "Ponto Âncora" na Figura 7.1). Este POI está localizado próximo ao rio da cidade e apresenta áreas verdes, inclusive com árvores (indicadas pela tonalidade verde escura na Figura 7.1). No contexto deste POI, há tracejados azuis e vermelhos, indicando caminhos onde é possível andar de bicicleta e fazer caminhadas. Além disso, há uma estrutura semelhante a um *pier* e alguns prédios mais acima. A área em tom vermelho claro indica uma região de varejo, enquanto as ruas, representadas em tonalidade branca, denotam as vias normais da cidade. Também percebe-se a presença de uma piscina

(geometria oval em azul). Do outro lado do rio, o tracejado em tom marrom indica uma estrada rústica. E o tom mais escuro do solo, denota um terreno abandonado (não utilizado).

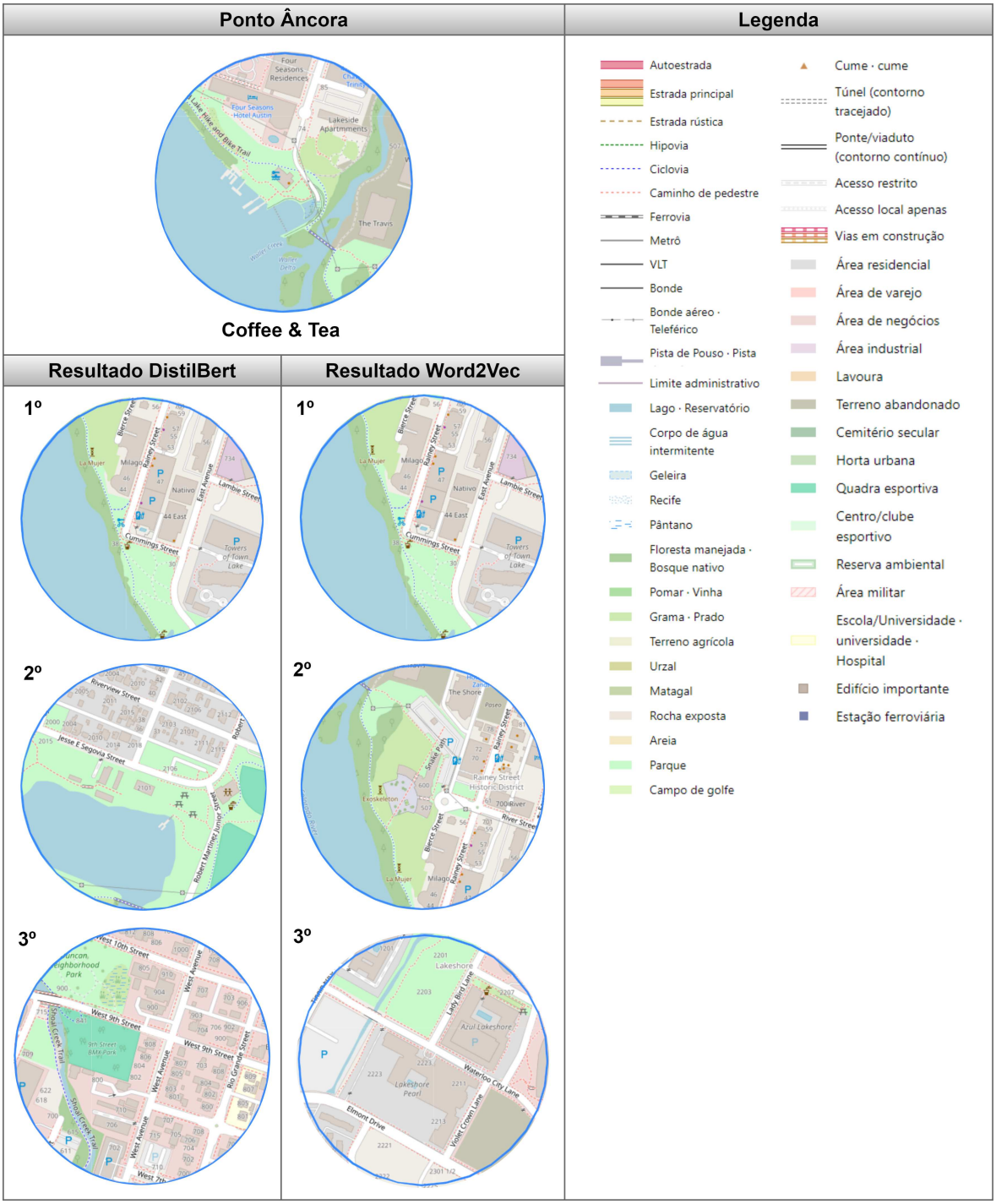


Figura 7.1: Resultado da busca para locais similares ao contexto de `Coffee & Tea`

Fonte: Autoria própria

Considerando o resultado da busca, o primeiro local mais similar utilizando os *embed-*

dings do DistilBert e Word2Vec foram iguais. Nessa resposta (indicado pelo ordinal 1º na Figura 7.1), é possível perceber a presença do rio em uma grande proporção, assim como no local âncora. Também existe uma área verde, composta por gramas (em tonalidade verde claro) e árvores (tonalidade verde escuro). Além disso, também há tracejados representando locais em que se pode caminhar ou pedalar, e uma área avermelhada, apontando um local de varejo. O local é composto parcialmente por prédios e apresenta todas as suas ruas do tipo convencionais, assim como no ponto âncora.

O segundo local fornecido pelos *embeddings* dos dois modelos foi diferente (indicados pelo ordinal 2º na Figura 7.1). Para o DistilBert, o local retornado apresenta um grande lago, um pequeno *Pier*, áreas verdes, uma ponte mais abaixo, assim como tracejados indicando locais para caminhada ou vias de bicicleta. Além disso, há um prédio no centro, um pouco mais isolado do restante dos prédios do local. Para o Word2Vec, percebe-se a presença do rio, das áreas verdes, dos tracejados e de um prédio maior no centro. Além disso, percebe-se que os prédios ocupam um pouco mais de espaço do que no ponto âncora. Em ambos os casos, a configuração de ruas também obedece ao padrão do ponto âncora.

A terceira resposta também difere para os *embeddings* dos dois modelos (indicados pelo ordinal 3º na Figura 7.1). Para o DistilBert, foi retornado uma grande região de área de varejo (em vermelho). Há vários prédios menores, indicando possivelmente uma área mais comercial. Também é possível perceber áreas verdes de grama e árvores, e um canal de água. Além disso, é possível ver duas pontes sobre o pequeno rio e os tracejados utilizados para tráfego a pé ou de bicicleta. O resultado do Word2Vec apresenta prédios mais largos, uma piscina, assim como no local âncora, os tracejados referentes ao tráfego a pé, mas não há para bicicleta (tracejado azul). Além disso, percebe-se a presença de uma área verde clara, mas não há árvores (verde mais escuro). Também é possível notar uma região avermelhada, indicando uma área de varejo, e uma região em um marrom mais escuro, denotando um lugar não utilizado.

Após essa análise, percebe-se que os resultados providos pelos *embeddings* dos dois modelos se assemelham em vários aspectos com a configuração do ponto âncora. Cada resultado apresenta áreas verdes e áreas de água, bem como vias para caminhada ou passeio de bicicleta. A configuração das ruas também é similar entre os resultados. Considerando a diferença entre os dois modelos, percebe-se que os resultados produzidos pelo DistilBert

trouxeram elementos que são um pouco mais similares ao ponto âncora do que os resultados produzidos pelo Word2Vec.

Uma segunda busca foi realizada utilizando um POI do tipo `Chinese`, referente à um restaurante chinês (indicado abaixo da palavra "Ponto Âncora" na Figura 7.2). Esse local fica numa área mais densa, onde existe a presença de alguns estacionamentos, pontos de ônibus, sinais de transito e duas vias principais. Além disso existem cruzamentos entre as vias principais e as vias convencionais. Também é possível perceber a presença de regiões avermelhadas, denotando uma área de varejo, e de áreas amareladas, denotando áreas de escolas ou hospitais.

Nesse caso, novamente, o local mais similar fornecido utilizando os *embeddings* do DistilBert e Word2Vec foi igual (indicados pelo ordinal 1º na Figura 7.2). Nessa área, é possível notar a presença de duas vias principais, apresentando sinais de trânsito, cruzamentos e pontos de ônibus. A área também inclui uma região avermelhada, denotando uma área comercial. Além disso, há espaços para estacionamento e vias convencionais. O local também exibe linhas vermelhas tracejadas, indicando que o tráfego de pedestres é permitido nas vias. Além disso, percebe-se uma pequena área verde na parte superior da região. Toda essa configuração também pode ser encontrada no ponto âncora.

Para o segundo resultado, os *embeddings* dos dois modelos apontaram locais diferentes (indicados pelo ordinal 2º na Figura 7.2). O resultado dos *embeddings* do DistilBert apresenta uma via principal maior e mais cruzamentos. Além disso, percebe-se a presença de pontos de ônibus, estacionamentos e vias para pedestres. A configuração desse local ainda inclui alguns edifícios maiores e mais áreas verdes. Há também áreas avermelhadas indicando uma região de varejo. O resultado obtido com os *embeddings* do Word2Vec também denota vias principais, mas apenas com um cruzamento. Nesse cenário, há prédios menores, dois pontos de ônibus, uma área maior para estacionamento, vias para pedestres e áreas verdes. Esse lugar também apresenta regiões vermelhas relacionadas ao varejo. Adicionalmente, existe um canal de água, porém essa feição não está presente no ponto âncora. Apesar disso, é usual que em áreas verdes existam corpos de água nas proximidades.

O terceiro lugar também apresentou divergências nos resultados fornecidos pelos *embeddings* dos dois modelos (indicados pelo ordinal 3º na Figura 7.2). No caso do DistilBert, observa-se a presença das vias principais e um cruzamento entre elas. Também são iden-

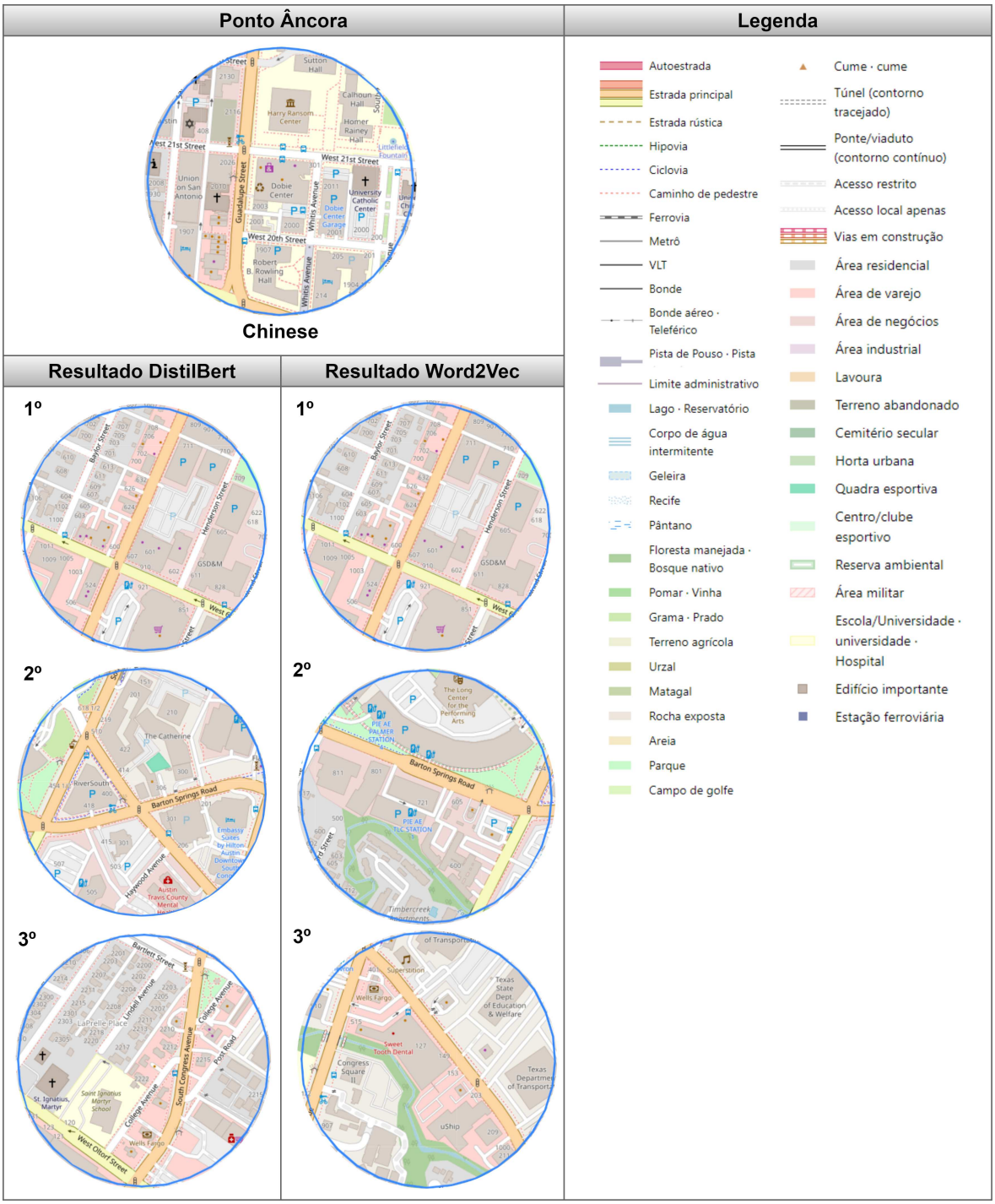


Figura 7.2: Resultado da busca para locais similares ao contexto de `Chinese`

Fonte: Autoria própria

tificados sinais de trânsito e um ponto de ônibus. Nota-se colorações vermelha e amarela, indicando áreas de varejo e escolas, respectivamente. Além disso, há uma região verde e vias para pedestres. Entretanto, essa localização não apresenta estacionamentos, elemento pre-

sente no ponto âncora. Ao analisar o resultado retornado com os *embeddings* do Word2Vec, nota-se a presença da via principal e um cruzamento na parte superior da região. Também são identificados um sinal de trânsito e dois pontos de ônibus. Na localização, há ainda uma área avermelhada, indicando varejo, e áreas verdes com a presença de corpos de água. Percebe-se ainda a existência de uma ponte. Ambos os resultados apresentam diversos elementos do ponto âncora e, por serem menos similares entre si, naturalmente compartilham menos configurações com esse ponto.

Analisando ambos os resultados dos *embeddings* dos dois modelos, pode-se notar que muitas das feições presentes no ponto âncora são respeitadas nos lugares sugeridos. Isso demonstra a eficiência dos *embeddings* das feições geográficas quando empregados para representar as regiões. Em relação aos dois modelos, percebe-se que existe uma diferença nos resultados, verificando-se novamente que os *embeddings* do Distilbert representam as regiões respeitando mais elementos em comum com o ponto âncora. Apesar disso, os resultados de ambos os modelos são concisos.

Aproveitando-se das possibilidades de se trabalhar com representações vetoriais, aplicou-se a operação de média entre duas regiões com feições geográficas diferentes, visando analisar se o resultado da busca demonstrava lugares que detêm aspectos dos dois lugares. Para isso, combinou-se os dois POIs dos exemplos anteriores (`Coffee & Tea` e `Chinese`). Como já mencionado, o POI que tem tipo `Coffee & Tea` fica em um lugar mais aberto com mais feições geográficas naturais, enquanto o outro (`Chinese`) fica em uma região mais central da cidade, demonstrando muitos aspectos urbanos. A Figura 7.3 ilustra o resultado da busca para a combinação desses dois lugares.

Assim como nos resultados dos dois exemplos anteriores, os *embeddings* do DistilBert e Word2Vec produziram o mesmo resultado para o local mais similar (indicados pelo ordinal 1º na Figura 7.3). Examinando-se esse local, percebe-se a presença da via principal, dos cruzamentos, dos sinais de trânsito, dos estacionamentos, das áreas verdes, dos pontos de ônibus, dos trechos para pedestres e para ciclistas. Na parte superior, existem duas pontes. Entre os elementos presentes nas duas regiões âncora, não foi encontrado corpo d'água. Entretanto, é possível perceber uma grande variedade dos elementos presentes nos dois locais.

No segundo resultado, os *embeddings* dos modelos geraram respostas diferentes (indicados pelo ordinal 2º na Figura 7.3). Com o DistilBert, o resultado apresentou várias vias

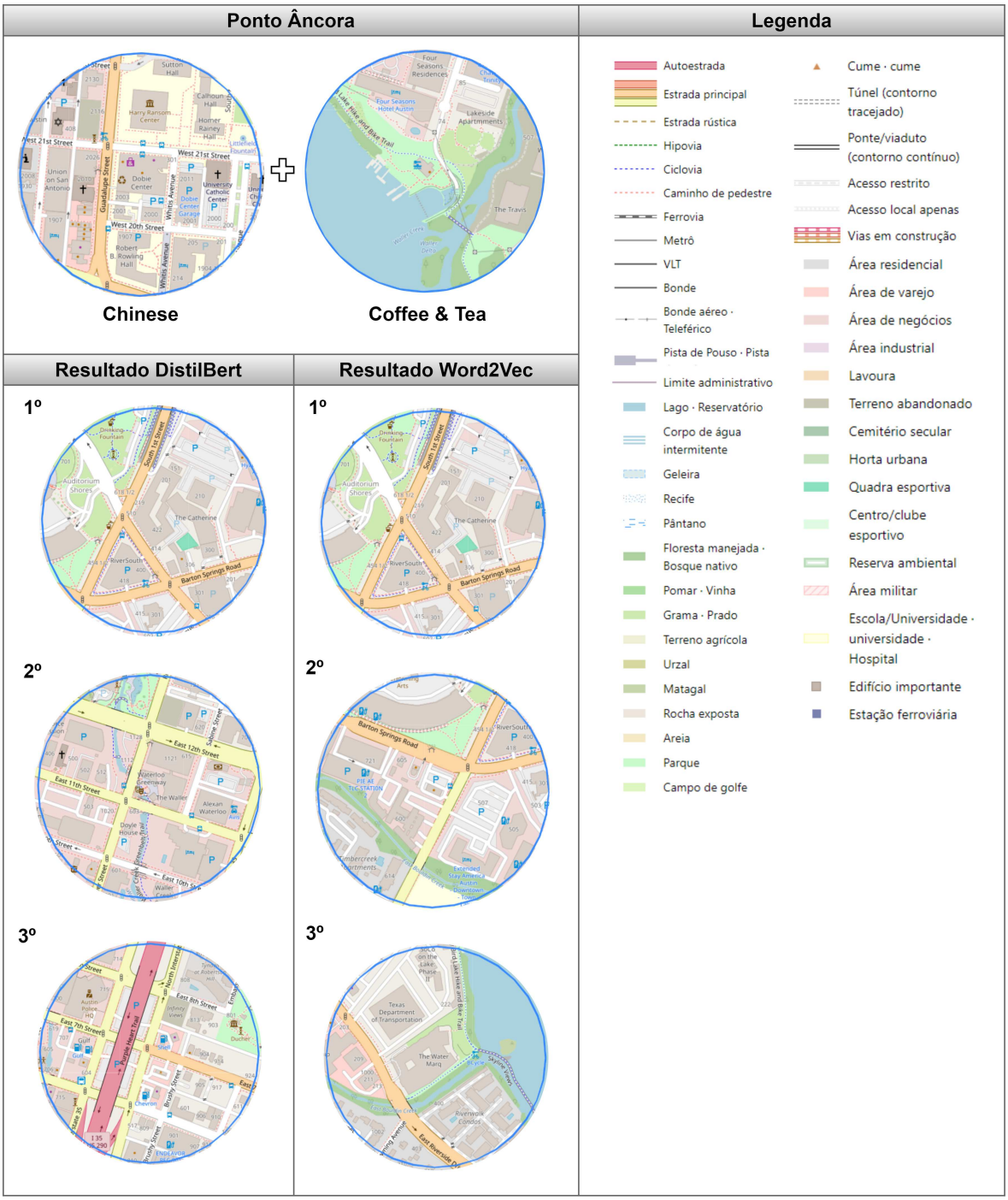


Figura 7.3: Resultado da busca para locais similares ao contexto de `Coffee & Tea` e `Chinese` combinados

Fonte: Autoria própria

principais, cruzamentos, pontos de ônibus, sinais de trânsito, vias de pedestres e ciclistas. Também há uma região de áreas verdes e um canal de água que atravessa a região. Esse resultado também denota muitos elementos presentes nos dois pontos âncora. Porém, percebe-se

que o padrão mais denso se sobressaiu. No resultado produzido com o Word2Vec, percebe-se duas vias principais, um cruzamento, um sinal de trânsito, pontos de ônibus, locais de estacionamento. Percebe-se também a presença das vias de pedestres e ciclistas, bem como a presença de áreas verdes de grama (cor verde mais clara) e de árvores (cor verde mais escura), conforme existe no ponto âncora `Coffee & Tea`. Além disso, também existe um canal de água que atravessa a região. Para esse resultado, percebe-se que os *embeddings* preservaram um pouco mais as feições do contexto do POI de tipo âncora `Coffee & Tea`.

O último resultado também divergiu entre os dois modelos (indicados pelo ordinal 3º na Figura 7.3). Para os *embeddings* do Distilbert, percebe-se a predominância dos elementos mais urbanos, presentes no ponto âncora `Chinese`. Ainda assim, das feições naturais do segundo ponto, existe no lado esquerdo da região um pequeno corpo de água, e no lado direito uma área verde. Um comportamento quase inverso ocorreu no resultado gerado com os *embeddings* do Word2Vec. Nesse caso, percebe-se que o local recomendado apresenta bem as feições de água e áreas verdes, assim como o ponto âncora `Coffee & Tea`. Ainda existem alguns elementos do outro ponto, como a rua principal, e alguns edifícios. No entanto, vários elementos desse ponto não foram identificados, como pontos de ônibus, sinais de trânsito e cruzamentos.

De maneira geral, é possível perceber que os resultados apresentados pelos dois modelos incorporaram feições presentes nos dois pontos âncora. Entre os dois modelos, percebeu-se que os *embeddings* dos locais gerados a partir do DistilBert preservaram mais a variedade de feições, tendendo naturalmente para o ponto âncora urbano, enquanto que os *embeddings* dos locais gerados a partir do Word2Vec preservaram as feições predominantes em área, como o caso do rio e das áreas verdes.

7.4 Considerações Finais

Este capítulo detalhou os resultados de um estudo de caso que visava demonstrar a usabilidade dos *embeddings* das feições geográficas. A obtenção de tais *embeddings* é um resultado complementar à abordagem proposta, pois como o foco é a geração de *embeddings* de tipos de POI utilizando feições geográficas do contexto dos POIs e modelos de PLN, tem-se como vantagem a obtenção dos *embeddings* das próprias feições.

Foram demonstrados exemplos práticos sobre como representar uma região utilizando esses *embeddings* e como tais regiões podem ser utilizadas em uma tarefa de busca. Os resultados alcançados demonstraram coesão nos locais retornados, sendo que cada local apresenta aspectos presentes nos pontos âncora, utilizados como referência. Conforme debatido, esse resultado demonstra que os *embeddings* das feições podem ser uma alternativa diante de outras abordagens baseadas em imagens.

O capítulo seguinte descreve a conclusão, apresentando as limitações da abordagem proposta nesta tese, indicando também direções futuras de pesquisa.

Capítulo 8

Conclusão e Trabalhos Futuros

8.1 Conclusões

Representar os tipos de POI por meio de *embeddings* tem apresentado resultados promissores em várias abordagens na literatura, como foi possível verificar ao longo desta pesquisa. No entanto, muitos desses trabalhos concentram-se nas relações contextuais de vizinhança de POIs, negligenciando as feições geográficas presentes no contexto dos POIs. É importante observar que o contexto geográfico oferece um conjunto de feições úteis que podem ser empregadas na geração de *embeddings* desses tipos e assim enriquecer a representação deles com informações geográficas. Portanto, esta tese formulou a seguinte questão de pesquisa: Como gerar *embeddings* de tipos de POI utilizando feições geográficas do contexto dos POIs e modelos de PLN? ($\mathbf{QP}_1$).

Para responder a essa questão, foi proposta uma abordagem que define um conjunto de passos necessários para a geração dos *embeddings*. Inicialmente, foi fundamental adquirir dados relacionados aos POIs e ao seu contexto geográfico. Em seguida, foi definido um algoritmo chamado GeoContext2Vec, que associa os tipos às feições geográficas. Além disso, esse algoritmo utiliza informações como *proporção de espaço ocupado* e *proporção de ocorrência* da feição para capturar os padrões espaciais do contexto. Na etapa seguinte, foi necessário treinar um modelo de PLN utilizando como entrada os dados gerados pelo algoritmo GeoContext2Vec. Assim, com o modelo treinado, os *embeddings* poderiam ser obtidos e utilizados em diversas tarefas, como agrupamento, planejamento urbano, recomendação de POIs, entre outras.

Nesta tese, foram utilizados o Word2Vec, uma técnica amplamente empregada em PLN para a criação de *embeddings* de tipos de POI, e o DistilBert. Um dos objetivos da pesquisa era investigar se os *embeddings* gerados por técnicas mais recentes de PLN permitiriam que modelos baseados em *embeddings* de tipos de POI alcançassem resultados superiores ($\mathbf{QP}_4$). Além disso, para averiguar se os *embeddings* produzidos com esta abordagem indicam a similaridade dos tipos, assim como outras abordagens, foi formulada a seguinte questão de pesquisa: Os *embeddings* de tipos de POI gerados a partir de relações contextuais com feições geográficas indicam a similaridade dos tipos? ($\mathbf{QP}_2$).

Para abordar as questões $\mathbf{QP}_2$ e $\mathbf{QP}_4$, os *embeddings* do GeoContext2Vec foram utilizados em quatro tarefas que analisam a similaridade entre os tipos de POI. Duas dessas tarefas foram realizadas com conjuntos de dados obtidos por meio da coleta de opiniões de voluntários humanos. A terceira tarefa utilizou uma hierarquia pré-concebida, que naturalmente carrega consigo informações sobre a similaridade entre os tipos. E a última tarefa consistiu na visualização dos *embeddings* para identificar se tipos similares estavam próximos ou formavam grupos no espaço vetorial latente.

Os resultados mostraram que os vetores obtidos com o GeoContext2Vec diferenciam os tipos de POI com maior precisão. O *matching* foi de 84% com a opinião dos voluntários ao utilizar o Word2Vec e de 93% ao utilizar o DistilBert. Além disso, os *embeddings* desses modelos produziram valores de similaridade mais próximos das opiniões humanas, com um valor de correlação de ρ competitivo, alcançando 59% com o Word2Vec e 72% com o DistilBert.

Os resultados também demonstraram que a combinação dos *embeddings* da abordagem proposta com outras abordagens baseadas em vizinhança possibilitou valores ainda mais altos de diferenciação entre os tipos. Isso foi demonstrado a partir do *matching* de 85% e uma correlação ρ de 71% utilizando o Word2Vec, e um *matching* aproximado de 98% e uma correlação ρ de 73% utilizando o DistilBert.

Na tarefa de similaridade utilizando a hierarquia dos tipos, os resultados revelaram que os *embeddings* produzidos com o Word2Vec apresentaram desempenho inferior a um dos *baselines* (ITDL), embora ainda tenham alcançado valores competitivos com o *Shortest Path*, com um MRR de $0,47$. No entanto, os *embeddings* gerados com o DistilBert superaram todos os outros modelos, atingindo um MRR máximo de aproximadamente $0,60$.

Na visualização dos *embeddings* do GeoContext2Vec, constatou-se a presença de diversos grupos formados por tipos irmãos (considerando uma hierarquia de tipos). Isso também corrobora com a afirmação de que as relações contextuais dos tipos de POI com as feições geográficas refletem a similaridade dos tipos.

Também foi realizada uma tarefa de classificação de zonas urbanas para investigar se modelos que utilizam *embeddings* produzidos com o GeoContext2Vec alcançavam melhores resultados do que modelos que utilizam *embeddings* baseados em outros dados geográficos (**QP**$_3$ e **QP**$_4$).

Os resultados obtidos demonstram que os *embeddings* produzidos com as feições e os modelos Word2Vec e DistilBert permitiram a obtenção de valores superiores aos valores obtidos com *embeddings* do ITDL e *Shortest Path*, alcançando um F1-Score de 89% em uma classificação mais geral e 85% em uma classificação mais específica (apenas com o Word2Vec).

Além disso, os resultados indicaram que quando os *embeddings* do GeoContext2Vec são combinados com os *embeddings* do ITDL e *Shortest Path*, os valores mais altos são alcançados em ambas as classificações, com 90% de F1-Score para a classificação mais geral e 86% para a classificação mais específica.

8.1.1 Limitações

Como limitação, pode-se citar que nos experimentos realizadas utilizou-se um raio fixo para diferentes tipos de POI. No entanto, cada tipo pode exigir valores de raio diferentes para que os padrões geográficos sejam capturados de maneira mais adequada. Por exemplo, o tipo parque geralmente ocupa áreas maiores, requerendo assim um raio maior. Utilizar um raio pequeno pode levar à captura de pouca informação geográfica, enquanto que para outros tipos, como postos de gasolina, um valor de raio pequeno pode permitir a captura de informações geográficas suficientes.

O conjunto de teste construído com a participação de voluntários providos por Yan *et al.* [103] também pode apresentar risco para a validade deste trabalho, pois não é possível definir que critérios os participantes consideraram para definir a similaridade dos tipos de POI. Além disso, fatores culturais e sociais podem influenciar o pensamento humano sobre o conceito de similaridade.

Uma segunda limitação é em relação ao uso de dados do OSM. Embora seja um banco de dados aberto, para cidades maiores, as estruturas do mapa são consideravelmente detalhadas, enquanto em cidades menores, geralmente existem apenas informações de ruas e avenidas. No entanto, essa abordagem pode ser usada com outros bancos de dados geográficos, sendo necessário apenas ajustar a a forma como se obtém as feições geográficas do contexto dos POIs.

Outra limitação diz respeito aos resultados obtidos com o Distilbert ao utilizar o conjunto de treino dos *baselines*. Apesar de os resultados não terem sido promissores, acredita-se que esse problema não esteja relacionado ao modelo em si, mas sim à forma como o documento de treinamento do modelo é gerado. Dessa forma, acredita-se que a $\mathbf{QP}_3$ foi respondida parcialmente, uma vez que os resultados do DistilBert foram melhores apenas ao utilizar o conjunto de treino do GeoContext2Vec.

8.2 Trabalhos Futuros

Como perspectivas para estudos futuros e oportunidades de pesquisa diretamente associadas a este trabalho, destacam-se:

1. **Investigar como produzir contextos de tamanhos adaptados aos tipos associados aos POIs**: conforme mencionado anteriormente, é possível que POIs de determinado tipo demandem tamanhos de contexto diferentes para que os *embeddings* desse tipo incorporem informações mais significativas;

2. **Investigar se outros tipos de relação entre os tipos e as feições pode trazer benefícios na geração dos *embeddings***: nesta tese, utilizou-se a relação contextual de forma binária, ou seja, relacionado cada tipo de POI à uma feição. Entretanto, pode ser possível utilizar relações topológicas entre os POIs e as feições, para capturar padrões diferentes das relações, como por exemplo, se um POI está de frente de um rio, se está na esquina, e assim sucessivamente;

3. **Investigar maneiras de converter o contexto de um POI em um texto mais representativo**: outra possibilidade de pesquisa refere-se à utilização de modelos mais recentes na área de PLN. Esses modelos foram desenvolvidos para processar textos

considerando todas as palavras e suas variações. Pesquisas futuras podem estudar outras formas de converter o contexto de um POI em um texto mais adequado ao modelo, seja utilizando relações topológicas, ou utilizando outras informações para relacionar os POIs no documento;

4. **Investigar maneiras de combinar as informações das feições com outras abordagens de forma única**: o foco deste trabalho foi investigar *embeddings* produzidos com feições geográficas e seus resultados isolados, bem como quando concatenados com outros *baselines*. No entanto, em trabalhos futuros, pode-se combinar as informações do contexto dos POIs de uma forma única. Além disso, será relevante explorar se as diversas relações entre os elementos do contexto dos POIs podem gerar representações mais precisas.

Bibliografia

[1] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[2] Tiba Zaki Abdulhameed, Imed Zitouni, and Ikhlas Abdel-Qader. Enhancement of the word2vec class-based language modeling by optimizing the features vector using pca. In *2018 IEEE International Conference on Electro/Information Technology (EIT)*, pages 0866–0870. IEEE, 2018.

[3] Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. Docbert: Bert for document classification. *arXiv preprint arXiv:1904.08398*, 2019.

[4] Pasquale Balsebre, Weiming Huang, and Gao Cong. Lamp: A language model on the map. *arXiv preprint arXiv:2403.09059*, 2024.

[5] Syed Raza Bashir and Vojislav B. Misic. Bert4loc: BERT for location - POI recommender system. *CoRR*, abs/2208.01375, 2022.

[6] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[7] Junxiang Bing, Meng Chen, Min Yang, Weiming Huang, Yongshun Gong, and Liqiang Nie. Pre-trained semantic embeddings for poi categories based on multiple contexts. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[8] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[9] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[10] Miguel Á. Carreira-Perpiñán. A review of mean-shift algorithms for clustering. *CoRR*, abs/1503.00687, 2015.

[11] Meng Chen, Lei Zhu, Ronghui Xu, Yang Liu, Xiaohui Yu, and Yilong Yin. Embedding hierarchical structures for venue category representation. *ACM Transactions on Information Systems (TOIS)*, 40(3):1–29, 2021.

[12] Anne Cocos and Chris Callison-Burch. 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*, pages 99–104, 2017.

[13] Alessandro Crivellari and Euro Beinat. From motion activity to geo-embeddings: Generating and exploring vector representations of locations, traces and visitors through large-scale mobility data. *ISPRS International Journal of Geo-Information*, 8(3):134, 2019.

[14] Alessandro Crivellari and Bernd Resch. Investigating functional consistency of mobility-related urban zones via motion-driven embedding vectors and local poi-type distributions. *Computational Urban Science*, 2(1):1–15, 2022.

[15] Emanuele Damiano, Aniello Minutolo, Stefano Silvestri, and Massimo Esposito. Query expansion based on wordnet and word2vec for italian question answering systems. In *Advances on P2P, Parallel, Grid, Cloud and Internet Computing: Proceedings of the 12th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC-2017)*, pages 301–313. Springer, 2018.

[16] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7:1–30, 2006.

[17] Yue Deng, Jiping Liu, Yang Liu, and An Luo. Detecting urban polycentric structure from poi data. *ISPRS International Journal of Geo-Information*, 8(6):283, 2019.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference*

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 4171–4186. Association for Computational Linguistics, 2019.

[19] Jingtao Ding, Guanghui Yu, Yong Li, Depeng Jin, and Hui Gao. Learning from hometown and current city: Cross-city poi recommendation via interest drift and transfer learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(4):1–28, 2019.

[20] Ruixue Ding, Boli Chen, Pengjun Xie, Fei Huang, Xin Li, Qiang Zhang, and Yao Xu. A multi-modal geographic pre-training method. *arXiv preprint arXiv:2301.04283*, 2023.

[21] Song Gao, Krzysztof Janowicz, and Helen Couclelis. Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS*, 21(3):446–467, 2017.

[22] Yoav Goldberg. Neural network methods for natural language processing. *Synthesis lectures on human language technologies*, 10(1):1–309, 2017.

[23] Shanshan Han, Cuiming Liu, Keyun Chen, Dawei Gui, and Qingyun Du. A tourist attraction recommendation model fusing spatial, temporal, and visual embeddings for flickr-geotagged photos. *ISPRS International Journal of Geo-Information*, 10(1):20, 2021.

[24] Sheng Hu, Zhanjun He, Liang Wu, Li Yin, Yongyang Xu, and Haifu Cui. A framework for extracting urban functional regions based on multiprototype word embeddings using points-of-interest data. *Computers, Environment and Urban Systems*, 80:101442, 2020.

[25] Yunfeng Hu and Yueqi Han. Identification of urban functional areas based on poi data: A case study of the guangzhou economic and technological development zone. *Sustainability*, 11(5):1385, 2019.

[26] Jizhou Huang, Haifeng Wang, Yibo Sun, Yunsheng Shi, Zhengjie Huang, An Zhuo, and Shikun Feng. Ernie-geol: A geography-and-language pre-trained model and its applications in baidu maps. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3029–3039, 2022.

[27] Jizhou Huang, Haifeng Wang, Yibo Sun, Yunsheng Shi, Zhengjie Huang, An Zhuo, and Shikun Feng. Ernie-geol: A geography-and-language pre-trained model and its applications in baidu maps. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3029–3039, New York, NY, USA, 2022. Association for Computing Machinery.

[28] Tianyuan Huang, Zhecheng Wang, Hao Sheng, Andrew Y Ng, and Ram Rajagopal. Learning neighborhood representation from multi-modal multi-graph: Image, text, mobility graph and beyond. *arXiv preprint arXiv:2105.02489*, 2021.

[29] Weiming Huang, Lizhen Cui, Meng Chen, Daokun Zhang, and Yao Yao. Estimating urban functional distributions with semantics preserved poi embedding. *International Journal of Geographical Information Science*, 36(10):1905–1930, 2022.

[30] Md Ashraful Islam, Mir Mahathir Mohammad, Sarkar Snigdha Sarathi Das, and Mohammed Eunus Ali. A survey on deep learning based point-of-interest (poi) recommendations. *Neurocomputing*, 472:306–325, 2022.

[31] Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3967–3974, 2019.

[32] Renhe Jiang, Xuan Song, Zipei Fan, Tianqi Xia, Zhaonan Wang, Quanjun Chen, Zekun Cai, and Ryosuke Shibasaki. Transfer urban human mobility via poi embedding over multiple cities. *ACM Transactions on Data Science*, 2(1):1–26, 2021.

[33] Jiaqi Jin, Zhuojian Xiao, Qiang Qiu, and Jinyun Fang. A geohash based place2vec model. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 3344–3347. IEEE, 2019.

[34] Phillip A Laplante. *Encyclopedia of Information Systems and Technology-Two Volume Set*. CRC Press, 2015.

[35] Claudia Leacock and Martin Chodorow. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283, 1998.

[36] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[37] Jiwon Lee, Kiyun Yu, and Jiyoung Kim. Public bike trip purpose inference using point-of-interest data. *ISPRS International Journal of Geo-Information*, 10(5):352, 2021.

[38] Tong Li, Yanxin Xi, Huandong Wang, Yong Li, Sasu Tarkoma, and Pan Hui. Learning representations of satellite imagery by leveraging point-of-interests. *ACM Transactions on Intelligent Systems and Technology*, 14(4):1–32, 2023.

[39] Zekun Li, Wenxuan Zhou, Yao-Yi Chiang, and Muhao Chen. Geolm: Empowering language models for geospatially grounded language understanding. *arXiv preprint arXiv:2310.14478*, 2023.

[40] Zhouhan Lin, Cheng Deng, Le Zhou, Tianhang Zhang, Yi Xu, Yutong Xu, Zhongmou He, Yuanyuan Shi, Beiya Dai, Yunchong Song, et al. Geogalactica: A scientific large language model in geoscience. *arXiv preprint arXiv:2401.00434*, 2023.

[41] Kang Liu, Ling Yin, Feng Lu, and Naixia Mou. Visualizing and exploring poi configurations of urban regions on poi-type semantic space. *Cities*, 99:102610, 2020.

[42] Xi Liu, Clio Andris, and Sohrab Rahimi. Place niche and its regional variability: Measuring spatial context patterns for points of interest with representation learning. *Computers, Environment and Urban Systems*, 75:146–160, 2019.

[43] Xiao Liu, Juan Hu, Qi Shen, and Huan Chen. Geo-bert pre-training model for query rewriting in poi search. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2209–2214, 2021.

[44] Xin Liu, Yong Liu, and Xiaoli Li. Exploring the context of locations for personalized location recommendations. In *IJCAI*, pages 1188–1194, 2016.

[45] Yongheng Liu, Zhen Yang, Tong Li, and Di Wu. A novel poi recommendation model based on joint spatiotemporal effects and four-way interaction. *Applied Intelligence*, 52(5):5310–5324, 2022.

[46] Zhewei Liu, Xiaolin Zhou, Wenzhong Shi, and Anshu Zhang. Recommending attractive thematic regions by semantic community detection with multi-sourced vgi data. *International Journal of Geographical Information Science*, 33(8):1520–1544, 2019.

[47] Yan Luo, Chak-Tou Leong, Shuhai Jiao, Fu-Lai Chung, Wenjie Li, and Guoping Liu. Geo-tile2vec: A multi-modal and multi-stage embedding framework for urban analytics. *ACM Transactions on Spatial Algorithms and Systems*, 2022.

[48] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[49] Xiaohui Mou, Fei Cai, Xin Zhang, Jie Chen, and Rongrong Zhu. Urban function identification based on poi and taxi trajectory data. In *Proceedings of the 2019 3rd International Conference on Big Data Research*, pages 152–156, 2019.

[50] Paul Mousset, Yoann Pitarch, and Lynda Tamine. End-to-end neural matching for semantic location prediction of tweets. *ACM Transactions on Information Systems (TOIS)*, 39(1):1–35, 2020.

[51] "Gustavo Niemeyer". "geohash intro", "2011".

[52] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[53] Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and R Ward. Semantic modelling with long-short-term memory for information retrieval. *arXiv preprint arXiv:1412.6629*, 2014.

[54] Debjyoti Paul, Feifei Li, and Jeff M Phillips. Semantic embedding for regions of interest. *The VLDB Journal*, 30:311–331, 2021.

[55] Hao Peng, Jianxin Li, Yangqiu Song, and Yaopeng Liu. Incrementally learning the hierarchical softmax function for neural language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[56] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[57] Dulce G Pereira, Anabela Afonso, and Fátima Melo Medeiros. Overview of friedman's test and post-hoc analysis. *Communications in Statistics-Simulation and Computation*, 44(10):2636–2653, 2015.

[58] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. *CoRR*, abs/1403.6652, 2014.

[59] Achilleas Psyllidis, Song Gao, Yingjie Hu, Eun-Kyeong Kim, Grant McKenzie, Ross Purves, May Yuan, and Clio Andris. Points of interest (poi): a commentary on the state of the art, challenges, and prospects for the future. *Computational Urban Science*, 2(1):1–13, 2022.

[60] Feng Qi, Mian Dai, Zixian Zheng, and Chao Wang. Geodecoder: Empowering multimodal map understanding. *arXiv preprint arXiv:2401.15118*, 2024.

[61] Quan Qin, Shishuo Xu, Mingyi Du, and Songnian Li. Identifying urban functional zones by capturing multi-spatial distribution patterns of points of interest. *International Journal of Digital Earth*, 15(1):2468–2494, 2022.

[62] Chen Qu, Liu Yang, Minghui Qiu, W Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. Bert with history answer embedding for conversational question answering. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 1133–1136, 2019.

[63] Dragomir R Radev, Hong Qi, Harris Wu, and Weiguo Fan. Evaluating web-based question answering systems. In *LREC*. Citeseer, 2002.

[64] Landy Rajaonarivo, Tsunenori Mine, and Yutaka Arakawa. Little known poi category estimation via syntactical knowledge graph generated via tweets. In *International Conference on Intelligent Systems and Knowledge Engineering*, 2023.

[65] Yeasir Rayhan and Tanzima Hashem. Aist: An interpretable attention-based deep learning model for crime prediction. *ACM Trans. Spatial Algorithms Syst.*, 9(2), apr 2023.

[66] Nornadiah Mohd Razali, Yap Bee Wah, et al. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, 2(1):21–33, 2011.

[67] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

[68] Wenzhong Shi, Zhewei Liu, Zhenlin An, and Pengfei Chen. Regnet: a neural network model for predicting regional desirability with vgi data. *International Journal of Geographical Information Science*, 35(1):175–192, 2021.

[69] Yoshiyuki Shoji, Katsurou Takahashi, Martin J Dürst, Yusuke Yamamoto, and Hiroaki Ohshima. Location2vec: Generating distributed representation of location by using geo-tagged microblog posts. In *International Conference on Social Informatics*, pages 261–270. Springer, 2018.

[70] Salatiel Dantas Silva, Cláudio E. C. Campelo, and Maxwell Guimarães de Oliveira. Generating POI type embeddings based on variableword2vec sentences. In Leonardo Bacelar Lima Santos and Marconi de Arruda Pereira, editors, *XXIII Brazilian Symposium on Geoinformatics - GEOINFO 2022, São José dos Campos, SP, Brazil, November 28 30, 2022*, pages 15–26. MCTIC/INPE, 2022.

[71] Salatiel Dantas Silva, Claudio Elízio Calazans Campelo, and Maxwell Guimarães De Oliveira. Poi types characterization based on geographic feature embeddings. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, pages 507–514, 2023.

[72] Salatiel Dantas Silva, Cláudio E. C. Campelo, and Maxwell G. Oliveira. Austin pois and geographic features dataset, April 2024.

[73] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 100(3/4):441–471, 1987.

[74] Vincent Spruyt. Loc2vec: Learning location embeddings with triplet-loss networks. *Sentiance web article: https://www. sentiance. com/2018/05/03/venue-mapping*, 2018.

[75] Chi Sun, Luyao Huang, and Xipeng Qiu. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 380–385, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[76] Fengze Sun, Jianzhong Qi, Yanchuan Chang, Xiaoliang Fan, Shanika Karunasekera, and Egemen Tanin. Urban region representation learning with attentive fusion. *arXiv preprint arXiv:2312.04606*, 2023.

[77] Huanliang SUN, Cheng PENG, Junling LIU, and Jingke XU. Community structure representation learning for. *Journal of Computer Applications*, 42(6):1782, 2022.

[78] Wannita Takerngsaksiri, Shoko Wakamiya, and Eiji Aramaki. City link: Finding similar areas in two cities using twitter data. In *International Symposium on Web and Wireless Geographical Information Systems*, pages 13–27. Springer, 2019.

[79] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science, 2022.

[80] Kota Tsubouchi, Hayato Kobayashi, and Toru Shimizu. Poi atmosphere categorization using web search session behavior. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, pages 630–639, 2020.

[81] Yi-Fu Tuan. *Space and place: The perspective of experience*. U of Minnesota Press, 1977.

[82] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[83] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[84] Lin Wan, Han Wang, Yuming Hong, Ran Li, Wei Chen, and Zhou Huang. itours-pot: a context-aware framework for next poi recommendation in location-based social networks. *International Journal of Digital Earth*, 15(1):1614–1636, 2022.

[85] Mu-Fan Wang, Yi-Shu Lu, and Jiun-Long Huang. Spent: A successive poi recommendation method using similarity-based poi embedding and recurrent neural network with temporal influence. In *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 1–8. IEEE, 2019.

[86] Pengyang Wang, Yanjie Fu, Jiawei Zhang, Xiaolin Li, and Dan Lin. Learning urban community structures: A collective embedding perspective with periodic spatial-temporal mobility graphs. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(6):1–28, 2018.

[87] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*, pages 165–174, 2019.

[88] Xuan Wang, Cui Zhu, Bingxin Xue, and Wenjun Zhu. A recommendation algorithm that combines rating data and review text. In *EMIE 2022; The 2nd International Conference on Electronic Materials and Information Engineering*, pages 1–7. VDE, 2022.

[89] Zhangyu Wang and Vahid Moosavi. From piace2vec to multi-scale built-environment representation: A general-purpose distributional embedding for urban data analysis. In *Proceedings of the 4th ACM SIGSPATIAL Workshop on Location-Based Recommendations, Geosocial Networks, and Geoadvertising*, pages 1–12, 2020.

[90] Zhecheng Wang, Haoyuan Li, and Ram Rajagopal. Urban2vec: Incorporating street view imagery and pois for multi-modal urban neighborhood embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1013–1020, 2020.

[91] Hong Wei, Janit Anjaria, and Hanan Samet. Learning embeddings of spatial, textual and temporal entities in geotagged tweets. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 484–487, 2019.

[92] Matt Womer. Points of interest core. W3C working draft, W3C, May 2011. https://www.w3.org/TR/2011/WD-poi-core-20110512/.

[93] Junhang Wu, Ruimin Hu, Dengshi Li, Yilin Xiao, Lingfei Ren, and Wenyi Hu. Where have you gone: Category-aware multigraph embedding for missing point-of-interest identification. *Neural Processing Letters*, 55(3):3025–3044, 2023.

[94] Rong Wu, Jieyu Wang, Dachuan Zhang, and Shaojian Wang. Identifying different types of urban land use dynamics using point-of-interest (poi) and random forest algorithm: The case of huizhou, china. *Cities*, 114:103202, 2021.

[95] Sexi Wu et al. *Design and Implementation of LBW–A Mental Health Application for Children*. PhD thesis, WORLD HEALTH & POPULATION, 2021.

[96] Z Wu and M Palmer. Verbs semantics and lexical selection. *Proceedings of the 32nd annual meeting on Association for Computational Linguistics–Association for Computational Linguistics*, 1994.

[97] Mingjun Xiang. Region2vec: An approach for urban land use detection by fusing multiple features. In *Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence*, pages 13–18, 2020.

[98] Jian Xie, Yidan Liang, Jingping Liu, Yanghua Xiao, Baohua Wu, and Shenghua Ni. Quert: Continual pre-training of language model for query understanding in travel domain search. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5282–5291, 2023.

[99] Chao Xu, Walter Forkel, Stefan Borgwardt, Franz Baader, and Beihai Zhou. Automatic translation of clinical trial eligibility criteria into formal queries. In *JOWO*, 2019.

[100] Hengpeng Xu, Jinmao Wei, Zhenglu Yang, and Jun Wang. Graph attentive network for region recommendation with poi-and roi-level attention. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*, pages 509–516. Springer, 2020.

[101] Ronghui Xu, Weiming Huang, Jun Zhao, Meng Chen, and Liqiang Nie. A spatial and adversarial representation learning approach for land use classification with pois. *ACM Transactions on Intelligent Systems and Technology*, 14(6):1–25, 2023.

[102] Mu Xue. A text retrieval algorithm based on the hybrid lda and word2vec model. In *2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*, pages 373–376. IEEE, 2019.

[103] Bo Yan, Krzysztof Janowicz, Gengchen Mai, and Song Gao. From itdl to place2vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. In *Proceedings of the 25th ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 1–10, 2017.

[104] Min Yang, Bo Kong, Ruirong Dang, and Xiongfeng Yan. Classifying urban functional regions by integrating buildings and points-of-interest using a stacking ensemble method. *International Journal of Applied Earth Observation and Geoinformation*, 108:102753, 2022.

[105] Xin Yang, Shuaishuai Bo, and Zhaojie Zhang. Classifying urban functional zones based on modeling pois by deepwalk. *Sustainability*, 15(10):7995, 2023.

[106] Yao Yao, Xia Li, Xiaoping Liu, Penghua Liu, Zhaotang Liang, Jinbao Zhang, and Ke Mai. Sensing spatial distribution of urban land use by integrating points-of-interest and google word2vec model. *International Journal of Geographical Information Science*, 31(4):825–848, 2017.

[107] Yao Yao, Qia Zhu, Zijin Guo, Weiming Huang, Yatao Zhang, Xiaoqin Yan, Anning Dong, Zhangwei Jiang, Hong Liu, and Qingfeng Guan. Unsupervised land-use change detection using multi-temporal poi embedding. *International Journal of Geographical Information Science*, 37(11):2392–2415, 2023.

[108] Dongjin Yu, Wenbo Wanyan, and Dongjing Wang. Leveraging contextual influence and user preferences for point-of-interest recommendation. *Multimedia Tools and Applications*, 80(1):1487–1501, 2021.

[109] Yang Yue, Yan Zhuang, Anthony GO Yeh, Jin-Yun Xie, Cheng-Lin Ma, and Qing-Quan Li. Measurements of poi-based mixed use and their relationships with neighbourhood vibrancy. *International Journal of Geographical Information Science*, 31(4):658–675, 2017.

[110] Yao-Yi Chiang Zekun Li, Jina Kim and Muhao Chen. Spabert: A pretrained language model from geographic data for geo-entity representation. *EMNLP*, 2022.

[111] Wei Zhai, Xueyin Bai, Yu Shi, Yu Han, Zhong-Ren Peng, and Chaolin Gu. Beyond word2vec: An approach for urban functional region extraction and identification by combining place2vec and pois. *Computers, Environment and Urban Systems*, 74:1–12, 2019.

[112] Chengkun Zhang, Liuchang Xu, Zhen Yan, and Sensen Wu. A glove-based poi type embedding model for extracting and identifying urban functional regions. *ISPRS International Journal of Geo-Information*, 10(6):372, 2021.

[113] Lu Zhang, Zhu Sun, Jie Zhang, Yiwen Wu, and Yunwen Xia. Conversation-based adaptive relational translation method for next poi recommendation with uncertain check-ins. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[114] Pengpeng Zhao, Anjing Luo, Yanchi Liu, Fuzhen Zhuang, Jiajie Xu, Zhixu Li, Victor S Sheng, and Xiaofang Zhou. Where to go next: A spatio-temporal gated network for next poi recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[115] Silin Zhou, Dan He, Lisi Chen, Shuo Shang, and Peng Han. Heterogeneous region embedding with prompt learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4981–4989, 2023.

[116] Yang Zhou and Yan Huang. Deepmove: Learning place representations through large scale movement data. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 2403–2412. IEEE, 2018.

[117] Xiangdian Zhu, Ye Wu, Luo Chen, and Ning Jing. Spatial keyword query of region-of-interest based on the distributed representation of point-of-interest. *ISPRS international journal of geo-information*, 8(6):287, 2019.

[118] Donald W Zimmerman and Bruno D Zumbo. Relative power of the wilcoxon test, the friedman test, and repeated-measures anova on ranks. *The Journal of Experimental Education*, 62(1):75–86, 1993.

Apêndice A

Análise dos conjuntos de teste BHE e RHE

Os conjuntos de testes das tarefas BHE e RHE foram construídos para capturar a opinião das pessoas sobre a similaridade entre os tipos de POI. O primeiro conjunto foi utilizado em uma tarefa de votação sobre o tipo de POI mais diferente entre três tipos. Este conjunto consiste em 77 listas de três tipos cada. Para entender melhor a configuração desse conjunto, investigou-se como os tipos de cada lista se relacionam. Para isso, foi examinada a existência de qualquer interseção considerando a hierarquia dos tipos verificando se havia relacionamentos de irmandade ou ancestralidade entre os tipos de cada lista.

Utilizando a hierarquia do Yelp, constatou-se que 74 das 77 listas apresentavam relacionamentos de irmandade ou ancestralidade, representando aproximadamente 96% dos dados. Ao analisar essas 74 listas, observou-se que em 67 delas o relacionamento hierárquico ocorria entre dois tipos, enquanto o terceiro era considerado externo. Por exemplo, uma das listas continha os tipos `Restaurants`, `Airports` e `Italian`. Neste caso, o tipo `Restaurants` é ancestral de `Italian`. Notavelmente, a maioria dos participantes escolheu o tipo que não apresentava uma relação hierárquica direta como o mais diferente, ou seja, 100% dos tipos diferentes eram aqueles sem relação hierárquica direta. Isso evidencia que a opinião das pessoas foi baseada na suposição de que um tipo é "igual"ao outro porque em algum nível da hierarquia eles são a mesma coisa.

Os outros 7 exemplos consistiam em listas formadas por tipos que eram parentes entre si na hierarquia. Como exemplo, existe uma lista que possui os tipos `Pilates`, `Yoga` e

Hiking, todos descendentes do tipo Active Life (relacionado a atividades físicas). A Tabela A.1 demonstra os tipos presentes nesse caso.

Tipo1	**Votos1**	**Tipo2**	**Votos2**	**Tipo3**	**Votos3**
DUI Law	17	Legal Services	1	Lawyers	7
Pilates	3	Trainers	10	Hiking	12
Pilates	0	Yoga	1	Hiking	24
Pilates	0	Gyms	1	Hiking	25
Hostels	0	Limos	25	Hotels	0
Golf	0	Lakes	19	Trainers	6
Golf	25	Pilates	0	Yoga	0

Tabela A.1: Listas do conjunto BHE que possuem interseção total agrupados por tipo e votação.

Observando os valores de votação, nota-se que em alguns casos, é provável que os votos foram dados considerando a atividade ou serviço relacionado ao tipo. Por exemplo, na linha que apresenta os tipos Pilates, Yoga e Hiking, a votação maior foi para o tipo Hiking (indicando atividade de escalada). Apesar de os três tipos indicarem esportes, os esportes yoga e pilates são mais similares entre si do que a escalada, pois geralmente são realizados em locais fixos, como ginásios ou parques e possuem movimentos similares. Além disso, a prática de escalada requer equipamentos diferentes e, evidentemente, um ambiente distinto. Uma análise semelhante pode ser feita para a linha que apresenta os tipos Pilates, Gyms e Hiking, assim como para a linha que apresenta os tipos Golf, Yoga e Pilates. Novamente, observa-se que as três atividades são distintas mas apresentam relação de localidade ou de características.

Considerando os exemplos nos quais todos os tipos não possuem relações hierárquicas (a Tabela A.2 sumariza esses exemplos), nota-se que o tipo mais votado como diferente é aquele que descreve uma função mais distinta. Na primeira linha, o tipo mais votado é *Shopping*, que abrange todos os tipos relacionados à venda de produtos e serviços, enquanto os outros dois tipos estão relacionados a veículos (Transportation e Parking). Na segunda linha, observa-se que o tipo mais votado como mais diferente foi Churches (igrejas), enquanto os tipos mais similares são Car Rental (aluguel de

carros) e `Hotels & Travel` (hotéis e viagens). Nesse caso, nota-se uma relação entre viagens e carros, pois é comum alugar carros durante uma viagem. Por fim, na última linha, o tipo `Test Preparation` foi votado como mais diferente. Esse tipo está dentro da categoria `Education`, relacionada a atividades de ensino. Os dois tipos mais similares são `Arcades` e `Arts & Entertainment`, relacionados a videogames e locais de arte e entretenimento, respectivamente. Nesse caso, observa-se que locais de videogames também proporcionam entretenimento, assim como locais de arte.

Tipo1	**Votação1**	**Tipo2**	**Votação2**	**Tipo3**	**Votação3**
Transportation	1	Shopping	24	Parking	0
Churches	26	Car Rental	0	Hotels & Travel	0
Arcades	2	Arts & Entertainment	0	Test Preparation	24

Tabela A.2: Listas do conjunto BHE que não possuem interseção em nenhum tipo agrupados por tipo e votação.

O conjunto RHE é composto por 70 listas contendo dois tipos. Para cada lista, os participantes deveriam indicar numericamente o quão similares os tipos eram, utilizando valores de 1 a 7, onde 1 indicava pouca similaridade e 7 indicava muita similaridade. Assim como no conjunto BHE, verificou-se se existia interseção entre os tipos selecionados na tarefa. Os resultados demonstraram que 36 exemplos (51%) compartilham similaridade hierárquica. Ou seja, o conjunto é praticamente dividido, com metade das listas apresentando tipos hierarquicamente similares e a outra metade não. Para avaliar as médias de similaridade nessas duas situações, calculou-se a média dos valores de similaridade. Como resultado, para as listas onde não há interseção de tipos, a média de similaridade é $2,43$, enquanto que para as listas onde há interseção, a média é $5,09$. Isso indica que a opinião dos participantes considerou se os tipos são "iguais"na hierarquia para atribuir um valor mais alto, enquanto que, quando isso não ocorre, o valor atribuído é mais baixo.

Apêndice B

Análise dos Dados de Zona de Austin

Os dados das zonas urbanas de Austin foram utilizados nesta tese para produção do modelo *Shortest Path* e para realização da tarefa de classificação de zonas. Os dados foram obtidos de dados oficiais do governo[1]. Os dados de zonas original é composto por aproximadamente 22.200 zonas. Segundo o guia de zoneamento de Austin[2], cada zona pode assumir quatro categorias do nível mais alto, sendo elas `Residential`, `Commercial`, `Industrial` e `Spetial Purpose`.

A categoria `Residential` refere-se a áreas designadas para uso residencial, onde residências, casas, apartamentos e outras estruturas habitacionais são permitidas. A categoria `Commercial` denota áreas destinadas a atividades comerciais e empresariais, incluindo espaços onde empresas, lojas, escritórios, restaurantes e outros estabelecimentos comerciais podem operar. Por sua vez, a categoria `Industrial` engloba áreas destinadas a atividades industriais, como fabricação, armazenamento e distribuição de produtos, geralmente reservadas para instalações industriais como fábricas, armazéns, depósitos e áreas de logística. Por fim, a categoria `Special Purpose` refere-se a áreas destinadas a fins especiais ou específicos que não se enquadram nas categorias residencial, comercial ou industrial, podendo incluir instalações governamentais, instituições educacionais, parques, áreas de preservação ambiental, instalações de saúde, entre outros usos especiais. Cada categoria possui regula-

[1]Disponível em data.austintexas.gov/dataset/Zoning-Small-Map-Scale-/tv5s-wvtc/. Acesso em 20 de maio de 2024.

[2]Disponível em https://www.austintexas.gov/sites/default/files/files/Planning/zoning_guide.pdf. Acesso em 20 de maio de 2024.

mentações específicas sobre o tamanho e tipo de habitação ou negócio permitido, densidade populacional, entre outras restrições ou diretrizes urbanísticas.

Cada categoria possui subcategorias que especificam detalhes e regulamentações sobre como o espaço deve ser utilizado. A Figura B.1 ilustra as subcategorias conforme a regulamentação de Austin. Na figura, observa-se que muitas subcategorias são diferenciadas por um código, o que aumenta ainda mais a especificidade de cada uma.

Table 1. Base Zoning Districts

Residential		Commercial	
LA	Lake Austin Residence	NO	Neighborhood Office
RR	Rural Residence	LO	Limited Office
SF-1	Single Family—Large Lot	GO	General Office
SF-2	Single Family—Standard Lot	CR	Commercial Recreation
SF-3	Family Residence	LR	Neighborhood Commercial
SF-4A	Single Family—Small Lot	GR	Community Commercial
SF-4B	Single Family—Condominium	L	Lake Commercial
SF-5	Urban Family Residence	CBD	Central Business District
SF-6	Townhouse & Condominium	DMU	Downtown Mixed Use
MF-1	Multifamily —Limited Density	W/LO	Warehouse/limited Office
MF-2	Multifamily—Low Density	CS	General Commercial Services
MF-3	Multifamily—Medium Density	CS-1	Commercial-Liquor Sales
MF-4	Multifamily—Moderate Density	CH	Commercial Highway Serv
MF-5	Multifamily—High Density	**Special Purpose**	
MF-6	Multifamily—Highest Density	DR	Development Reserve
MH	Mobile Home Residence	AV	Aviation Services
Industrial		AG	Agricultural
IP	Industrial Park	P	Public
LI	Limited Industrial Services	PUD	Planned Unit Development
MI	Major Industry	TN	Traditional Neighborhood
R&D	Research & Development		

Figura B.1: Tabela demonstrando as categorias e subcategorias das zonas de Austin.

Fonte: Governo de Austin[3]

Os dados das zonas incluem um conjunto de informações, como o ID da zona, sua geometria, data de criação e modificação, categoria base da zona (formada pelo segundo nível da categoria, sem a presença da combinação indicada pelo hífen), categoria completa (formada pela combinação das categorias do segundo nível, podendo ser representada por vários hifens, como LO-SF-MH), entre outros.

Conforme mencionado anteriormente, existem aproximadamente 22.000 zonas na ci-

dade. A Figura B.2 ilustra um trecho das zonas. Na figura, é possível perceber que as zonas existentes são muito pequenas, em que muitas delas apresentam uma geometria semelhante à dos prédios. Relacioná-las diretamente com os POIs pode não fornecer informações significativas, pois é provável que muitas zonas tenham poucos ou até mesmo apenas um POI em sua área.

Figura B.2: Zonas em uma região da cidade.

Fonte: Autoria própria

Para contornar essa limitação, foram utilizadas as informações sobre os bairros da cidade para agrupar zonas da mesma categoria em um único conjunto por bairro. Essa prática é comum em trabalhos relacionados, nos quais frequentemente zonas com áreas pequenas são descartadas ou combinadas com zonas de categorias semelhantes para formar uma zona maior. Após esse processo de agrupamento, foi obtido um total de 406 zonas. Em seguida, procedeu-se com a associação dos POIs às suas respectivas zonas, como demonstrado na Tabela B.1.

Alguns experimentos iniciais foram realizados com as zonas que apresentaram mais de 20 ocorrências, totalizando assim 9 zonas. Entretanto, percebeu-se um desempenho muito baixo, em torno de 40% de F1-Score. Verificando-se a matriz de confusão resultante (Figura B.3).

Os resultados revelaram que o algoritmo *Random Forest* apresentou dificuldades em distinguir entre as zonas das categorias `Comercial` e as da categoria SF (*Single-Family*) de `Residencial`, resultando em uma alta taxa de confusão entre elas. Além disso, todas as

Zona	Quantidade
GR	69
CS	65
SF	55
MF	42
LR	33
LO	31
GO	24
PUD	24
LI	23
TOD	9
RR	8
P	6
UNZ	3
ERC	3
CH	3
DR	2
NBG	2
R&D	1
NO	1
CBD	1
MI	1
AV	1
DMU	1
IP	1

Tabela B.1: Quantidade de zonas urbanas com POIs.

classificações para a zona *Planed Unit Development* (PUD) foram incorretas. Diante desse cenário, realizou-se uma análise dos tipos predominantes em cada uma dessas zonas. Especificamente, examinaram-se os cinco tipos mais frequentes em cada zona, conforme ilustrado na B.4.

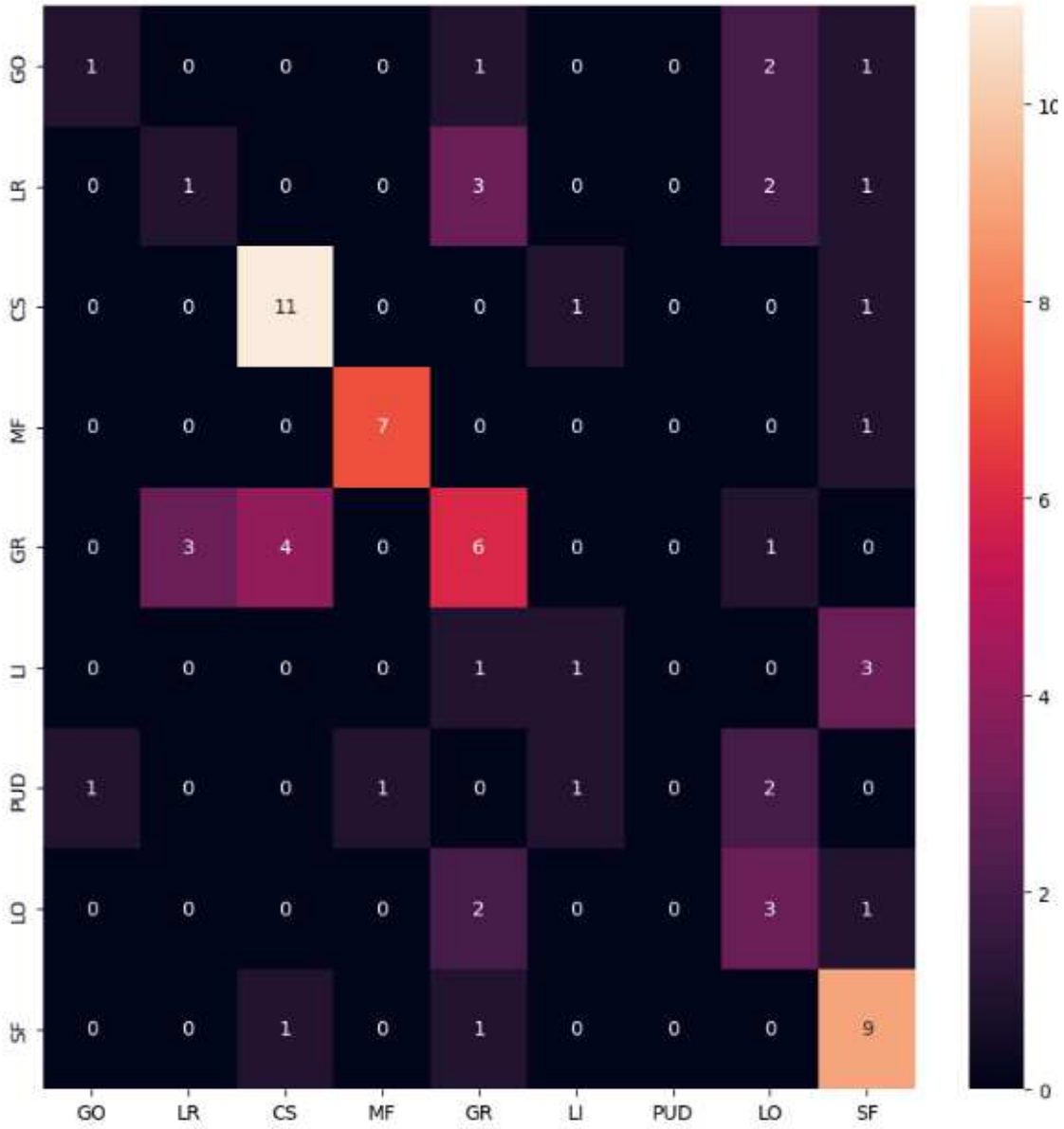


Figura B.3: Matriz de confusão de testes iniciais da tarefa de classificação de zonas.

Fonte: Autoria própria

Observando a Figura B.4, é notável que todos os tipos de zonas compartilham um tipo comum entre os cinco mais frequentes. Esse padrão sugere que as zonas naturalmente exibem similaridades entre si, fazendo com que classificadores tendam a errar. Além disso, as categorias GR, LR e GO, subcategorias de `Commercial`, mostram o maior grau de sobreposição. Da mesma forma, a categoria PUD, uma subcategoria de `Special Purpose`, também exibe uma considerável sobreposição com as demais categorias. Por outro lado, a categoria GO, subcategoria de `Commercial`, apresenta a menor interseção, embora com-

Figura B.4: Top 5 tipos de POI em cada zona e suas interseções.

Fonte: Autoria própria

partilhe tipos relacionados hierarquicamente, o que implica uma certa similaridade.

A partir desse resultado, optou-se por excluir as subcategorias de `Special Purpose` e `Industrial`, pois elas têm uma representação menor nos dados e uma significativa interseção com outras categorias. Além disso, considerando que as próprias subcategorias do zoneamento indicam um nível de especificidade muito elevado, foi realizado um mapeamento entre subcategorias irmãs para evitar redundâncias. Por exemplo, as subcategorias GO e LO estão relacionadas a escritórios, portanto, foram consolidadas em uma nova subcategoria `Escritórios`. Da mesma forma, as subcategorias GR, LR e CS estão associadas ao comércio e foram agrupadas na mesma nova subcategoria `Comércio`. Por fim, as categorias SF e MF estão ligadas a áreas residenciais e também foram combinadas em uma nova subcategoria `Familiar`. Essa abordagem resultou em melhorias nos resultados, conforme evidenciado na Seção 6.5.