Ampliando a vitalidade dos termos do ALiB através da Extração de Informação Geolocalizada nas mídias sociais

Pedro Guimarães Mendes Santos[1], **Daniela Barreiro Claro**[1], **Laila P. Mota Santos**[1], **Rerisson Cavalcante**[2]

[1]FORMAS Research Group – Instituto de Computação – Universidade Federal da Bahia

[2]FORMAS Research Group – Instituto de Letras – Universidade Federal da Bahia
Salvador – Bahia – Brasil

{pedrogms,dclaro,laila.pereira}@ufba.br, rerissoncavalcante@gmail.com

Resumo. *O Projeto ALiB (Atlas Linguístico do Brasil) descreve a geolinguística contemporânea, priorizando a variação diatópica na análise do português brasileiro. Seus termos foram coletados entre 1996 e 2013. Com o advento das redes sociais surgiu a necessidade de analisar a vitalidade destes termos. Dentre os diversos desafios inerentes às redes sociais, tem-se a não-obrigatoriedade da marcação de geolocalização no momento da postagem e a ampla utilização da internet slang. Assim, o presente trabalho apresenta uma nova abordagem para extrair as informações de geolocalização diretamente de tweets, com o intuito de ampliar a cobertura da localização. Neste trabalho, o BERTimbau, foi treinado para realizar tarefas de Reconhecimento de Entidades Nomeadas e utilizado para extrair conteúdo de geolocalização do usuário. Os resultados dão indícios de que a extração de localização é uma possibilidade de ampliar e aprimorar a análise da vitalidade dos termos do ALiB.*

1. Introdução

As redes sociais atualmente são grandes disseminadoras e propulsoras da cultura e diversidade das localidades brasileiras. A principal forma de interação nas redes sociais é por meio da língua escrita que ressalta os diversos estilos e falares do Brasil e cuja diversidade tem sido estudada há várias décadas através da Dialetologia. De 1996 a 2013, os dados dialetais foram catalogados presencialmente, através de inquéritos e entrevistas que abrangem todo o território nacional. Esse mapeamento dos falares foi publicado em dois volumes do Atlas Linguístico do Brasil [Cardoso et al. 2014a, Cardoso et al. 2014b].

Com o advento das redes sociais, a disseminação dos falares brasileiros foi propulsionada e disseminada, ampliando a cobertura do Atlas. Uma das mais difundidas redes sociais da atualidade é o Twitter, que permite a propagação de textos curtos e coloquiais, se aproximando dos falares dialetais difundidos no Brasil. A análise da vitalidade dos termos do Atlas (ou seja, a compreensão de quais expressões permanecem em uso) e sua propagação nas redes sociais despertou interesse da comunidade, porém, dois desafios inerentes ao Twitter dificultam o processo de análise: a não-obrigatoriedade da marcação de localização e a *internet slang* (gírias e abreviações utilizadas na internet para comunicação entre usuários). Outros trabalhos já analisaram a vitalidade dos termos do ALiB [Nunes et al. 2020] porém a falta de informação referente à localidade nos tweets não permitiu uma comparação mais aprofundada, visto que a localidade

da pessoa que tweeta pode não corresponder a sua verdadeira localização, além disso, atualmente, menos de 3% dos tweets gerados por usuários possuem a geolocalização [Gupta and Nishu 2020].

Diante deste contexto, o presente trabalho tem por principal objetivo desenvolver uma abordagem para extração de informação de geolocalização diretamente do conteúdo textual produzido nos tweets, com a finalidade de maximizar as ocorrências de geocódigo (coordenadas geográficas) e, consequentemente, processar os termos do ALiB com as informações de geolocalização com maior precisão.

O desenvolvimento do método de extração da geolocalização do tweet utilizou o modelo de linguagem baseado em *Transformers*, o BERT (*Bidirectional Encoder Representations from Transformers*) em sua variação treinada para o português brasileiro, o BERTimbau [Souza et al. 2020]. O modelo passou por um processo de *fine-tuning* para ser capaz de realizar a tarefa de *Named Entity Recognition* (NER, em português Reconhecimento de Entidades Nomeadas), com o intuito de facilitar a identificação das localidades de uma maneira automatizada. O método desenvolvido foi avaliado em dois datasets totalizando mais de 720.000 tweets. Por fim, o mapa com as capitais foi gerado para a comparação e visualização dos termos nas suas respectivas localidades.

O presente artigo está organizado em seções como se segue: a seção 2 descreve o método para extração de geolocalização em tweets. A seção 3 apresenta os experimentos e seus respectivos resultados. A seção 4 traz a discussão e os resultados preliminares.

2. Geolocalidade dos termos do ALiB

O método para Extração de geolocalização em tweets com os termos do ALiB foi desenvolvido em cinco etapas: Corpora, Segregação do geocódigo, Normalização, Inferência das localizações através do modelo de linguagem e Plotagem no mapa, conforme a Figura 1. Este método foi denominado GeoALiB, cujo objetivo é extrair a informação da geolocalização de um tweet através de um modelo de linguagem para Português (BERTimbau) que foi ajustado (fine-tuning) para a tarefa de NER [Bertaglia and Nunes 2016]. Na tarefa de NER, somente o rótulo LOCAL foi utilizado.

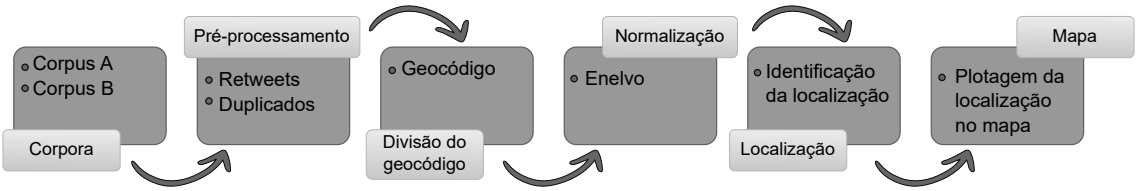


Figura 1. Etapas do método de extração de geolocalização

A primeira etapa correspondeu aos corpora utilizados para a aplicação do método: Corpus A e Corpus B. O Corpus A pré-existente, oriundo do trabalho de [Nunes et al. 2020], é composto por 500.000 tweets, redigidos em diversos idiomas, sem critério de filtragem. Devido ao baixo número de tweets com os termos do ALiB, afetando diretamente os resultados obtidos com o Corpus A, foi proposto a construção de um novo dataset: o Corpus B, composto por uma nova abordagem de filtragem de conteúdo.

Após a criação do Corpus B, iniciou-se a etapa de segregação dos corpora. Os tweets foram segregados em dois grupos: os tweets que já possuíam georreferenciamento

informado pelo usuário e os que não tinham a localização, interesse deste trabalho. A maior parte dos tweets que continham os termos do ALiB no Corpus A, não tinha nenhum dado de geolocalização (somente 4,6% dos tweets já possuíam geocódigo). Para o Corpus B, o qual todos os tweets continham termos do ALiB, cerca de 3,8% possuíam geocódigo.

Em seguida, uma etapa de normalização foi aplicada ao grupo sem georreferenciamento, visando aprimorar a capacidade de entendimento e inferência do modelo de linguagem, uma vez que os dados provenientes do Twitter tem características de escritas próprias. Expressões regulares e Enelvo [Bertaglia and Nunes 2016] fora utilizados. Durante o processo, as informações sensíveis de localização não foram removidas, como as que podem estar contidas em *hashtags*. Após este processo, os dados provenientes dos Corpus A e B estavam normalizados para serem processados pelo modelo de linguagem.

A etapa de inferência da localização ocorreu através do BERTimbau [Souza et al. 2020], processando o conjunto de tweets normalizados e anotando as entidades nomeadas que ocorrem no texto desses tweets. O BERTimbau é capaz de inferir diversas entidades em textos, como *VALOR, TEMPO, ORGANIZAÇÃO, LOCAL, PESSOA*, entre outras. Ao final do processamento, os dados foram rotulados para cada tweet individualmente e avaliados manualmente, caso a caso, as entidades que foram inferidas. Para este trabalho foram levadas em conta apenas os rótulos de *LOCAL* identificados.

Após esta etapa, catalogou-se as localizações e os termos do ALiB e os mapas foram plotados para permitir uma análise da vitalidade destes termos nas redes sociais.

3. Analisando a vitalidade dos termos

O método desenvolvido para extrair a informação da localização dos tweets foi avaliado em três experimentos distintos. Enxergou-se a necessidade de analisar o trabalho através destes experimentos pois eles são fundamentais na compreensão de: (i) quais termos do ALiB estão em uso no Twitter e com qual a frequência eles ocorrem na plataforma (**Experimentos A** e **B**); (ii) visualizar os locais dentro do território brasileiro onde os termos estão sendo usados (**Experimento C**).

No primeiro experimento, chamado neste trabalho de **Experimento A**, o método foi comparado com o método manual de [Nunes et al. 2020], utilizando o mesmo corpus, ou conjunto de dados, do trabalho anterior. O segundo experimento, chamado de **Experimento B**, objetivou analisar a precisão na extração das geolocalizações em tweets que já continham os termos do ALiB, através da utilização do Corpus B. Por fim, o último experimento, **Experimento C**, objetivou analisar os mapas gerados nos trabalhos de [Nunes et al. 2020] com o intuito de verificar se as geolocalizações novas correspondem às mesmas cartas (representações cartográficas da variação linguística) do ALiB e também quais as principais diferenças quanto ao trabalho de [Nunes et al. 2020] em relação ao método desenvolvido neste trabalho.

3.1. Experimento A

Este primeiro experimento utilizou o Corpus A. Poucos tweets válidos, excesso de retweets e ausência dos termos de interesse neste trabalho, além do grande número de ocorrências duplicadas e em outros idiomas foram alguns dos desafios deste corpus. Apesar do grande volume de dados, após análise, o número foi reduzido para 1.161 tweets

válidos, dos quais somente quinze retornaram entidades de localização após o processamento do modelo de linguagem.

3.2. Experimento B

A partir do experimento A, observou-se a necessidade de maximizar a quantidade de tweets com os termos do ALiB por meio da criação de um novo corpus, denominado Corpus-B. Esse corpus foi segmentado em CorpusB-v1 e CorpusB-v2.

Com o CorpusB-v1, foram selecionados 63 termos do ALiB de diferentes cartas, o que resultou em 49.845 novos tweets coletados. Após o pré-processamento destes tweets para segregação de georreferenciamento e normalização, o modelo de linguagem foi capaz de inferir a localização de 501 tweets, conforme descrito na Tabela 1.

Devido aos resultados obtidos com o primeiro ciclo de recuperação do Corpus B, o dataset foi expandido. Para este novo ciclo, denominado de CorpusB-v2, foram utilizadas cinco janelas de recuperação de dados para os mesmos 63 termos do ALiB. Este processo durou um pouco mais de um mês, resultando em 176.643 novos tweets. Após pré-processamento e normalização, resultou em 1.603 tweets com localização inferida, conforme demonstrado na Tabela 1.

Tabela 1. Resultados gerais com os Corpora

Resultados do Corpus A		**CorpusB-v1**	**CorpusB-v2**
Total	500 000	49 845	176 643
Tweets não-válidos	498 839		
Tweets válidos	1 161		
Sem Geocódigo	1 107	48 305	171 226
Com Geocódigo	54	1 540	5 417
BERTimbau	3 744	501	1 603

3.3. Experimento C - Plotagem no Mapa

Após a execução do modelo e uma vez que os dados de localização referentes aos dois Corpora foram contabilizados, os termos com suas respectivas localizações foram plotados no mapa. Somente duas cartas foram descritas neste estudo: Semáforo e Prostituta. A escolha destas cartas se deu principalmente pois os termos encontrados nestas cartas não são ambíguos, ou seja, semanticamente trazem em seu sentido literal o mesmo catalogado no ALiB.Esse critério otimizou a quantidade de tweets qualificados para análise.

3.3.1. A Carta Semáforo

Em análise da Carta Semáforo, após a execução do BERTimbau, o processo de inferência do modelo de linguagem conseguiu obter 163 ocorrências do termo *semáforo*, sendo 43 delas no Rio de Janeiro (RJ), 68 em São Paulo (SP), 21 em Brasília (DF), 13 em Belo Horizonte (MG), 8 em Fortaleza (CE), 3 em Duque de Caxias (RJ), 2 em Barreiras (BA) e os municípios de Contagem, Itaquera, Itatiaia, São Miguel e Mogi Guaçu tiveram 1 ocorrência cada. Já o termo *sinal* teve suas 16 ocorrências registradas na região sudeste,

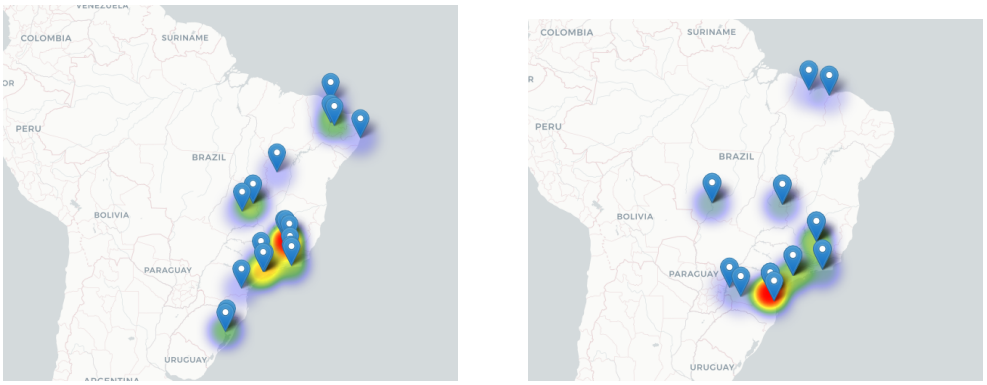
mais precisamente nas cidades de São Francisco (MG), Belo Horizonte (MG), São Paulo (SP) e Brasília (DF). Por fim, o termo *sinaleira* teve 10 registros catalogados no estado do Rio Grande do Sul, mais precisamente em Porto Alegre e Campo Bom.

O mapa de calor descrito na Figura 2a, permite analisar a frequência com que esses termos são aplicados no Twitter entre as cidades. Em muitos casos houveram múltiplos registros de um termo em uma mesma cidade. É possível identificar uma maior incidência dos termos desta carta da região sudeste do país, além de alguns casos identificados em algumas cidades da região nordeste.

3.3.2. Carta Prostituta

Em relação à Carta Prostituta, o modelo conseguiu inferir 151 novos termos. A localização de 11 ocorrências do termo *garota de programa* em Brasília (DF), São Paulo (SP), Rio de Janeiro (RJ), Cuiabá (MT), São Bento (PB), Curitiba (PR) e São José das Palmeiras (PR). O termo *prostituta* teve 27 ocorrências nas cidades de São Paulo (SP), Rio de Janeiro (RJ), Brasília (DF), Joinville(SC) e Curitiba (PR). Já os termos *puta* e *prima* tiveram 89 e 24 ocorrências respectivamente nas cidades de Belo Horizonte (MG), São Paulo (SP), Rio de Janeiro (RJ), São Bernardo e Cuiabá (MT).

O mapa de calor na Figura 2b permite analisar a frequência com que esses termos são aplicados no Twitter entre as cidades em que eles foram identificados. É possível identificar uma maior incidência dos termos desta carta da região sudeste do país, além de alguns casos identificados em algumas cidades da região nordeste.



(a) Mapa de calor da Carta Semaforo (b) Mapa de calor da Carta Prostituta

Figura 2. Cartas Semáforo e Prostituta

4. Discussão dos resultados e Conclusões Preliminares

Comparando os resultados obtidos com as duas versões do Experimento B houve um aumento expressivo na quantidade de localidades inferidas pelo BERTimbau, de 501 para 1.603, representando uma expansão de 220% no geocódigo do CorpusB-v2 em relação ao CorpusB-v1. Esse número expressivo se deve principalmente ao crescimento da base. Além disso, somando-se o CorpusB-v1 e CorpusB-v2 foi possível expandir o geocódigo para 2.104 tweets dentro do grupo de tweets originalmente sem geolocalização.

Esse número representa cerca de 1% do tamanho total da base e se torna mais expressivo quando agrega o grupo de tweets com geolocalização pré-definida originalmente (cerca de 4% de tweets com geocódigo). Esse cálculo é válido quando se deseja entender o alcance gerado pela utilização do modelo de linguagem, e validar a quantidade total de termos com georreferenciamento, sejam eles nativamente marcados pelos usuários ou inferidos pelo modelo. Também é importante ressaltar a taxa de aproveitamento do Corpus B, de 100%, uma vez que, dos 226.488 tweets, todos possuíam os termos do ALiB, enquanto no no Corpus A essa taxa foi equivalente a apenas 0,23%.

Para um entendimento da empregabilidade dos termos do ALiB dentro da região estudada, o método depende de alguns aspectos dos dados utilizados no trabalho, por exemplo: (i) menção de localização explicitamente citada no tweet do usuário para que o modelo a identifique; (ii) a localização mencionada pelo usuário precisa se referir à área que é objeto de estudo, neste caso, o território brasileiro. Outros desafios inerentes ao contexto da rede social escolhida também precisam ser elencados: o estilo de escrita que reduz o desempenho do modelo de linguagem.

Para os trabalhos futuros, observou-se a necessidade de uso de um método automatizado de desambiguação em português brasileiro e análise com outras fontes de dados.

Referências

[Bertaglia and Nunes 2016] Bertaglia, T. F. C. and Nunes, M. d. G. V. (2016). Exploring word embeddings for unsupervised textual user-generated content normalization. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 112–120.

[Cardoso et al. 2014a] Cardoso, S., Mota, J., Aguilera, V., de Aragão, M. d. S., Isquerdo, A., Razky, A., Margotti, F., and Altenhofen, C. (2014a). *Atlas linguístico do Brasil*, volume 1. Londrina: Eduel.

[Cardoso et al. 2014b] Cardoso, S., Mota, J., Aguilera, V., de Aragão, M. d. S., Isquerdo, A., Razky, A., Margotti, F., and Altenhofen, C. (2014b). *Atlas linguístico do Brasil*, volume 2. Londrina: Eduel.

[Gupta and Nishu 2020] Gupta, S. and Nishu, K. (2020). Mapping local news coverage: Precise location extraction in textual news content using fine-tuned BERT based language model. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 155–162, Online. Association for Computational Linguistics.

[Nunes et al. 2020] Nunes, A. P. M., de Jesus, L. E. N., Claro, D. B., Mota, J., Ribeiro, S., Paim, M., and Oliveira, J. (2020). Vitality analysis of the linguistic atlas of brazil on twitter. In Quaresma, P., Vieira, R., Aluísio, S., Moniz, H., Batista, F., and Gonçalves, T., editors, *Computational Processing of the Portuguese Language*, pages 184–194, Cham. Springer International Publishing.

[Souza et al. 2020] Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.