Identificação de Temas em Comentários de Restaurantes usando BERT e Modelos de Linguagem Generativa

José A. de Almeida Neto[1], **Tiago de Melo**[1]

[1]Escola Superior de Tecnologia – Universidade do Estado do Amazonas (UEA)
Manaus – AM – Brasil

{jadan.eng20,tmelo}@uea.edu.br

Abstract. *This study investigates the application of advanced natural language processing (NLP) techniques to classify reviews of fine dining restaurants in Brazil. Using 4,000 sentences from platforms such as Google Reviews, TripAdvisor, and Yelp, the performances of Multi-label Logistic Regression, BERTimbau, and Sabia are compared. BERTimbau demonstrated the best performance, with a macro F1-Score of 0.88 and a micro F1-Score of 0.92. The analysis reveals significant thematic variations when observing individual restaurants, highlighting the effectiveness of pre-trained NLP models and suggesting directions for future research.*

Resumo. *Este estudo investiga a aplicação de técnicas avançadas de processamento de linguagem natural (PLN) para classificar comentários sobre restaurantes de alta gastronomia no Brasil. Utilizando 4.000 sentenças de plataformas como Google Reviews, TripAdvisor e Yelp, são comparados os desempenhos de Regressão Logística Multirrótulo, BERTimbau e Sabia. O BERTimbau apresentou melhor desempenho, com macro F1-Score de 0.88 e micro F1-Score de 0.92. A análise revela variações temáticas significativas quando se observam os restaurantes individualmente, destacando a eficácia dos modelos pré-treinados em PLN e sugerindo direções para pesquisas futuras.*

1. Introdução

A escolha de um restaurante na alta gastronomia é frequentemente influenciada por recomendações de guias renomados, como o Guia Michelin, que recentemente reconheceu 21 restaurantes brasileiros com uma ou duas estrelas[1]. Além disso, opiniões de outros clientes em plataformas de avaliação *online*, como TripAdvisor e Google Reviews, são consideradas determinantes na escolha de restaurantes. Além disso, estas avaliações apresentam diversas experiências gastronômicas, tornando-se uma fonte valiosa de dados para pesquisa [Yu and Zhang 2020, Gan et al. 2017].

A análise de comentários de usuários sobre restaurantes tem atraído considerável interesse acadêmico, principalmente devido ao desenvolvimento de métodos capazes de processar e resumir essas críticas para fornecer uma compreensão detalhada das preferências dos clientes [Tedjojuwono and Neonardi 2021, Fang 2022]. Entretanto, uma lacuna persiste, uma vez que muitos estudos ainda se baseiam em uma única fonte de dados, resultando em uma visão parcial das interações do usuário nas plataformas *online*.

[1]https://forbes.com.br/forbeslife/2024/05/brasil-tem-21-restaurantes-com-estrelas-michelin-em-2024-10-sao-estreantes

Este trabalho visa abordar essa lacuna através da aplicação de técnicas avançadas de processamento de linguagem natural (NLP) para a classificação multirrótulo de comentários sobre restaurantes de alta gastronomia. Além disso, explora-se o desempenho de modelos de aprendizagem supervisionada e autossupervisionada (*self-learning*), incluindo o modelo BERT [Devlin et al. 2018] e a linguagem generativa Sabiá [Pires et al. 2023], treinada exclusivamente em português do Brasil, em uma série de experimentos usando dados coletados de plataformas como Facebook, Foursquare, Google Reviews, TripAdvisor e Yelp.

O objetivo principal desta pesquisa é construir um modelo robusto para a classificação de textos multirrótulo em análises de opinião, com foco em responder às seguintes perguntas de pesquisa:

PP1 - É possível desenvolver um modelo eficiente de aprendizagem de máquina para identificar os temas prevalentes nos comentários sobre restaurantes brasileiros?

PP2 - Os modelos de linguagem generativa alcançam desempenho comparável aos modelos especificamente treinados para a tarefa de identificação de temas?

PP3 - Como os modelos identificam e interpretam os temas prevalentes nos comentários em português sobre restaurantes de alta gastronomia em plataformas online?

Para abordar as PP1 e PP2, foram avaliados três modelos: regressão logística multirrótulo [de Almeida Neto and de Melo 2023], modelo de aprendizado profundo baseado em Transformers (BERT) e o modelo de linguagem generativa Sabiá. O modelo BERT apresentou o melhor desempenho, alcançando um macro F1-Score de 0.88 e um micro F1-Score de 0.92.

Para responder à PP3, aplicou-se o modelo BERT a um extenso conjunto de comentários coletados de todos os restaurantes brasileiros premiados pelo Guia Michelin. Este estudo revelou que a distribuição de temas tanto em restaurantes quanto em plataformas diferentes, é apresentada de maneira homogênea. No entanto, ao analisar individualmente os restaurantes, é possível identificar variações nessa distribuição por plataforma online. Esses resultados indicam que, apesar de uma tendência geral, há diferenças perceptíveis na ênfase de temas dependendo da plataforma de avaliação.

O artigo está estruturado da seguinte maneira: a Seção 2 revisa a literatura; a Seção 3 descreve a metodologia; a Seção 4 apresenta e discute os resultados experimentais; e a Seção 5 conclui o estudo e sugere direções futuras.

2. Trabalhos Relacionados

Esta seção revisa os trabalhos relevantes anteriores ao tema deste estudo, organizados em duas subseções: plataformas de avaliação e métodos de classificação.

2.1. Plataformas de Avaliação

Diversos estudos têm analisado as plataformas de avaliação *online* [de Melo 2021, Fang 2022]. Em [de Melo 2021], é apresentada uma análise abrangente dos comentários em português publicados em diversas plataformas de avaliação online, incluindo Facebook, Foursquare, Google, TripAdvisor, Yelp e Zomato. O estudo reporta o comportamento dos usuários e a qualidade das avaliações.

O estudo de [Fang 2022] quantifica os efeitos das plataformas de avaliação sobre a receita dos restaurantes e o bem-estar do consumidor. Os resultados mostram que a duplicação da atividade de avaliações pode aumentar a receita de restaurantes independentes de alta qualidade e reduzir a receita de restaurantes de baixa qualidade em proporções semelhantes. Este estudo destaca a importância das avaliações *online* na dinâmica do mercado de restaurantes.

2.2. Métodos de Classificação

Diversos estudos têm explorado a classificação de texto multirrótulo com análise de opinião, utilizando uma variedade de modelos e técnicas. Em [de Almeida Neto and de Melo 2023], foi realizada uma investigação detalhada sobre a classificação de comentários de clientes de restaurantes brasileiros. O estudo examinou o uso de métodos de Processamento de Linguagem Natural (PLN) para a classificação desses comentários, explorando diversas técnicas de pré-processamento para aprimorar modelos clássicos de aprendizagem supervisionada, como Support Vector Machine (SVM), Random Forest (RF) e Logistic Regression (LR), além do AutoGluon, um método de AutoML. Adotou-se o melhor modelo deste estudo como *baseline* e utilizou-se o mesmo conjunto de dados neste trabalho.

Neste trabalho, ainda foram investigados os modelos de redes neurais pré-treinados. O modelo escolhido foi o BERT, amplamente utilizado em tarefas de PLN, como classificação de sentimentos [Hammes and Freitas 2021] e classificação de texto multirrótulo [Serras and Finger 2021]. Especificamente, utilizou-se o BERTimbau, uma versão do BERT pré-treinada para o português brasileiro [Souza et al. 2020], reconhecida por seu desempenho superior em diversas tarefas de PLN na língua portuguesa.

Além disso, utilizou-se o MariTalk, um agente de conversação construído a partir do modelo de linguagem Sabiá [Pires et al. 2023], um dos principais *Large Language Models* (LLMs) especializados em português. Como uma ferramenta de linguagem generativa, o MariTalk tem sido empregado em diversos estudos de PLN [Ioscote 2023, da Silva Oliveira et al. 2024]. A utilização do MariTalk permite aproveitar as capacidades avançadas do Sabiá para melhorar a precisão e a eficiência na classificação de texto multirrótulo, beneficiando-se de sua compreensão profunda da língua portuguesa e de sua habilidade em gerar respostas contextualmente relevantes.

3. Materiais e Métodos

3.1. Tema das Opiniões

Este estudo considerou que as sentenças dos usuários estão relacionadas ao seginte conjunto pré-definido de temas: ambiente, bebida, comida, geral, localização, preço, serviço e outros. A definição dos temas teve como base outros estudos [Yu and Zhang 2020, de Melo 2021, de Almeida Neto and de Melo 2023]. Os temas são autoexplicativos, merecendo explicação que as sentenças relacionadas ao restaurante em si, tais como "excelente restaurante", foram identificadas como *geral*. Finalmente, sentenças que não se enquadravam em nenhum dos temas anteriores foram classificadas como *outros*.

3.2. Dataset

Neste trabalho, foi utilizado um conjunto de dados manualmente anotado pelos autores do estudo [de Almeida Neto and de Melo 2023]. O *dataset* em questão trata-se de 4.000

sentenças, cada uma anotada com um ou múltiplos temas, conforme as definições de temas descritas na Subseção 3.1.

3.3. Modelos Utilizados

Este estudo utilizou o BERTimbau [Souza et al. 2020], um modelo de aprendizado profundo baseado na arquitetura BERT (*Bidirectional Encoder Representations from Transformers*) [Devlin et al. 2018], especificamente treinado para o português brasileiro. Também foi utilizado o modelo de linguagem generativa Sabiá [Pires et al. 2023], especificamente o chatbot MariTalk, também treinado e especializado em língua portuguesa.

3.4. Métricas de Avaliação

Para avaliar a identificação de temas em comentários sobre restaurantes, foram usadas as métricas de precisão (P), revocação (R) e F1-Score (F_1). Considerando A como o conjunto de temas corretamente identificados e B como o conjunto de temas identificados pelo método em avaliação, as métricas são definidas como:

$$P = \frac{|A \cap B|}{|B|} \quad R = \frac{|A \cap B|}{|A|} \quad F_1 = \frac{2 \times (P \times R)}{P + R}$$

A micro F1-Score é calculada ao se computar os valores globais de precisão e revocação para todas as classes e, em seguida, calcular a medida F1. Essa métrica atribui a mesma relevância à classificação de cada item, independentemente de sua classe. Em contraste, a macro F1-Score atribui igual importância ao desempenho do classificador em cada classe, independentemente do número de itens em cada conjunto. Assim, a análise dos classificadores utilizando essas métricas oferece avaliações complementares da efetividade de um classificador. Dado um conjunto de classes $|C|$, a macro F1-Score é a média dos F1-Scores de todas as classes, enquanto a micro F1-Score é calculada a partir da soma das precisões (P) e da soma das revocações (R) de todas as classes.

4. Experimentos

4.1. Análise dos Resultados

Para desenvolver um modelo efetivo capaz de identificar temas em comentários sobre restaurantes brasileiros, foram conduzidos experimentos utilizando-se os modelos mencionados na Subseção 3.3. O BERT foi aplicado utilizando a técnica de pré-processamento *stemming*, conforme empregado no melhor modelo do estudo anterior (LR_{Stem}) [de Almeida Neto and de Melo 2023]. A validação cruzada com 5 *folds* foi adotada, onde 3 *folds* foram usados para treinamento, um *fold* para validação e outro para teste. O *fine tuning* dos hiperparâmetros do BERT incluiu variações no *batch size* para treinamento, validação e teste, *learning rate*, número de *epochs* (1, 5 ou 10) e *threshold* (entre 0,25 e 0,35). Diante disso, após diversos experimentos, o melhor modelo BERT ($BERT_{Stem}$) foi definido com *batch size* para treinamento de 16, e para validação e teste de 4, *learning rate* de 4e-05, 5 *epochs* e *threshold* de 0,28.

Para o MariTalk, foram empregadas duas abordagens: $MariTalk_{zero\text{-}shot}$ e $MariTalk_{one\text{-}shot}$. Na abordagem *zero-shot*, não foram fornecidos exemplos no *prompt*, enquanto na abordagem *one-shot* foi incluído um exemplo de sentença para cada tema. Ambas as abordagens foram aplicadas para classificar o *dataset* completo de 4.000 sentenças.

A Tabela 1 apresenta o desempenho dos modelos em termos de precisão (P), revocação (R) e F1-Score (F_1), considerando tanto as versões macro quanto micro de cada métrica. Notam-se que os resultados do modelo Logistic Regression usando a técnica de pré-processamento *stemming* (LR_{Stem}) são provenientes do estudo prévio realizado com o mesmo *dataset* [de Almeida Neto and de Melo 2023]. Além disso, vale ressaltar que os resultados do modelo $\text{BERT}_{\text{Stem}}$ denotam uma média de validação cruzada com os mesmos 5-*folds* utilizados para obtenção das métricas do modelo LR_{Stem}. Essa consistência no método de validação permite uma comparação direta e justa entre os modelos, evidenciando a eficácia das novas abordagens propostas neste trabalho.

Tabela 1. Tabela de resultados.

Modelos	Macro			Micro		
	P	R	F_1	P	R	F_1
LR_{Stem}	0,855	0,808	0,815	0,894	0,892	0,893
$\text{BERT}_{\text{Stem}}$	0,880	0,892	0,877	0,895	0,949	0,921
$\text{MariTalk}_{\text{zero-shot}}$	0,619	0,766	0,673	0,741	0,857	0,795
$\text{MariTalk}_{\text{one-shot}}$	0,596	0,711	0,616	0,696	0,801	0,711

Os resultados apresentados na Tabela 1 indicam que o $\text{BERT}_{\text{Stem}}$ alcançou o melhor resultado em todas as métricas. Isto responde tanto a primeira quanto a segunda pergunta de pesquisa (PP1 e PP2). Este melhor modelo é eficiente na tarefa de identificação dos temas em comentários sobre restaurantes brasileiros, superando os modelos de linguagem generativa $\text{MariTalk}_{\text{zero-shot}}$ e $\text{MariTalk}_{\text{one-shot}}$.

Além disso, houve uma melhoria significativa em todas as métricas em comparação com o modelo LR_{Stem}, considerado *baseline* neste estudo, evidenciando a superioridade do BERT nesta tarefa sobre os modelos clássicos, como o de Regressão Logística. Isso reforça a tendência atual na área de Processamento de Linguagem Natural, onde modelos pré-treinados, como o BERT, são capazes de capturar melhor a complexidade e as nuances da linguagem natural, resultando em desempenhos superiores.

Por outro lado, é interessante notar que o $\text{MariTalk}_{\text{zero-shot}}$ apresentou resultados melhores que o $\text{MariTalk}_{\text{one-shot}}$. Possivelmente, essa diferença pode ser atribuída à quantidade limitada de exemplos fornecidos durante o treinamento no modo *one-shot*, o que não foi suficiente para a generalização do modelo, especialmente para os temas geral e bebida, onde houve uma queda significativa na performance. Já o modelo *zero-shot*, ao não ser ajustado com exemplos específicos, pode ter mantido melhor a generalização aprendida durante o treinamento inicial com dados mais diversificados.

4.2. Estudo de caso

Para responder à pergunta de pesquisa PP3, o melhor modelo avaliado, $\text{BERT}_{\text{Stem}}$, foi aplicado em um conjunto de dados contendo comentários sobre restaurantes brasileiros recomendados pelo Guia Michelin. Estes comentários foram separados em sentenças, conforme o conjunto de dados de treino do modelo. Os dados[2] consistem em 13.718 sentenças de comentários publicados no período de 2014 a 2024, coletados das plataformas de avaliação *online* Google, Yelp, Facebook, Foursquare e Zomato.

[2] https://github.com/zeneto11/wtag2024

Em uma primeira análise, nota-se uma predominância do tema comida nos comentários, seguido pelo tema serviço e o tema geral. Em contrapartida, o tema localização é o tema menos comentado no geral. Tanto a distribuição de temas por plataforma de avaliação online quanto por restaurante, foi observada uma homogeneidade nas proporções dos temas abordados. Entretanto, ao analisar inidividualmente os restaurantes, é possível notar variações significativas por plataforma.

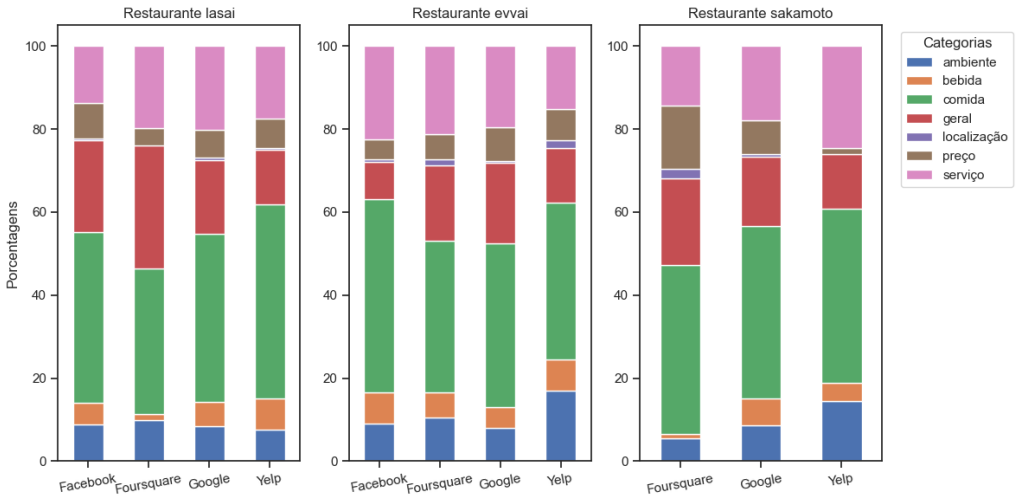

Figura 1. Análise individual de restaurantes por plataforma.

A Figura 1 mostra, ilustrativamente, a distribuição de temas para três restaurantes em diferentes plataformas. Com o intuito de melhorar a vizualização dos resultados, o tema outros foi ocultado, por conta do grande número de sentenças que não continham opinião. No restaurante Lasai, o tema geral é significativamente mais mencionado no Foursquare, além disso, observam-se variações dos temas bebida e serviço nas plataformas. Já no restaurante Evvai, na plataforma Yelp, observa-se uma maior ocorrência dos temas localização e ambiente, além de variações dos temas geral, serviço e comida nas plataformas. Já no restaurante Sakamoto, no Foursquare, verifica-se maior menção ao tema preço, superando o tema serviço, enquanto os temas bebida e ambiente são os menos mencionados. Por outro lado, no Google, há uma distribuição mais equilibrada entre os temas, enquanto no Yelp, ao contrário do Foursquare, o tema preço é significativamente menos mencionado, e o tema ambiente bem mais frequente.

5. Conclusões

Neste estudo, realizou-se uma avaliacão comparativa de modelos supervisionados e autossupervisionados na análise de comentários sobre restaurantes. Os resultados apontaram que modelo pré-treinado BERT utilizando a técnica de pré-processamento *stemming* alcançou o melhor resultado em todas as métricas consideradas. Adcionalmente, avaliou-se que a linguagem generativa MariTalk, no modo *zero-shot* e *one-shot*, ainda precisa de evolução ao lidar com a tarefa proposta. O modelo BERT, ao ser aplicado a um grande conjunto de dados, foi possível constatar variações e tendências distintas entre as diversas plataformas. Como trabalho futuro, pretende-se o desempenho de BERTimbau e Sabiá com outros modelos de NLP de última geração, como GPT-4, para avaliar melhorias e identificar os melhores modelos para diferentes tarefas.

Referências

[da Silva Oliveira et al. 2024] da Silva Oliveira, A., de Carvalho Cecote, T., Alvarenga, J. P. R., da Silva Luz, E. J., et al. (2024). Toxic speech detection in portuguese: A comparative study of large language models. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 108–116.

[de Almeida Neto and de Melo 2023] de Almeida Neto, J. A. and de Melo, T. (2023). Exploring supervised learning models for multi-label text classification in brazilian restaurant reviews. In *Anais do XX Encontro Nacional de Inteligência Artificial e Computacional*, pages 126–140. SBC.

[de Melo 2021] de Melo, T. (2021). Análise de comentários das plataformas online de restaurante michelin no brasil. In *A produção do conhecimento nas ciências da comunicação*, pages 226–238.

[Devlin et al. 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[Fang 2022] Fang, L. (2022). The effects of online review platforms on restaurant revenue, consumer learning, and welfare. *Management Science*, 68(11):8116–8143.

[Gan et al. 2017] Gan, Q., Ferns, B. H., Yu, Y., and Jin, L. (2017). A text mining and multidimensional sentiment analysis of online restaurant reviews. *Journal of Quality Assurance in Hospitality & Tourism*, 18(4):465–492.

[Hammes and Freitas 2021] Hammes, L. and Freitas, L. (2021). Utilizando bertimbau para a classificação de emoções em português. *Anais do Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 56–63.

[Ioscote 2023] Ioscote, F. C. (2023). Produção de notícia ou de texto? um estudo exploratório sobre potenciais e limitações do chatgpt, bard ai e maritalk para o jornalismo.

[Pires et al. 2023] Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023). Sabiá: Portuguese large language models. In *Brazilian Conference on Intelligent Systems*, pages 226–240. Springer.

[Serras and Finger 2021] Serras, F. R. and Finger, M. (2021). verbert: Automating brazilian case law document multi-label categorization using bert. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL 2021)*, STIL 2021. Sociedade Brasileira de Computação.

[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.

[Tedjojuwono and Neonardi 2021] Tedjojuwono, S. M. and Neonardi, C. (2021). Aspect based sentiment analysis: Restaurant online review platform in indonesia with unsupervised scraped corpus in indonesian language. In *1st International Conference on Computer Science and Artificial Intelligence*, volume 1, pages 213–218. IEEE.

[Yu and Zhang 2020] Yu, C.-E. and Zhang, X. (2020). The embedded feelings in local gastronomy: a sentiment analysis of online reviews. *Journal of Hospitality and Tourism Technology*, 11(3):461–478.