

Caracterização Subjetiva de Zonas Urbanas Através de Análise de Reviews e NLP

Miguel Miragaia¹[108307] e Tomé de Almeida¹[127896]

¹Universidade de Aveiro, Portugal

Resumo. O presente trabalho propõe uma abordagem orientada por dados para a avaliação subjetiva de zonas urbanas através da análise de reviews públicas associadas a pontos de interesse (POIs) na cidade de Aveiro, Portugal. Através de técnicas de Processamento de Linguagem Natural (NLP), como análise de sentimento multilíngue e modelação de tópicos, o projeto extrai padrões semânticos e emocionais que complementam métricas tradicionais baseadas em classificações numéricas. Os dados foram recolhidos automaticamente via API Google Places, enriquecidos com detecção automática de idioma e análise de sentimento específica para inglês (modelo VADER) e português (BERTweet-PT). A modelação de tópicos por Latent Dirichlet Allocation (LDA) permitiu identificar os temas principais discutidos pelos utilizadores. Para facilitar a exploração dos resultados, foi desenvolvido um dashboard interativo com múltiplas visualizações geoespaciais, incluindo mapas de calor, agrupamentos e visualização de tópicos, acessível a stakeholders não técnicos. Os resultados indicam diferenças no sentimento e temas entre as avaliações em inglês e português, além de evidenciar padrões espaciais que destacam zonas com avaliações consistentemente positivas ou críticas. O projeto contribui com uma pipeline reproduzível e extensível, que apoia planeadores urbanos, entidades de turismo e outros negócios na compreensão detalhada da experiência subjetiva dos visitantes. Apesar das limitações inerentes, como restrições da API e amostras reduzidas, a metodologia demonstra o valor da análise sistemática e da visualização interativa para informar decisões urbanas e promover um entendimento mais profundo do espaço público.

Palavras-chave: Análise de sentimento multilíngue; Avaliação urbana subjetiva; Latent Dirichlet Allocation (LDA); Visualização geoespacial interativa

1 Introdução

A avaliação da qualidade e das características das zonas urbanas tem sido tradicionalmente realizada com base em indicadores quantitativos, como classificações numéricas, categorias administrativas ou métricas oficiais [1]. Contudo, estas abordagens tendem a desvalorizar a dimensão subjetiva da experiência urbana, isto é, a forma como os utilizadores percecionam, vivenciam e avaliam os espaços da cidade no seu quotidiano [2]. Nesse contexto, as reviews públicas disponibilizadas em plataformas digitais constituem uma fonte relevante de informação qualitativa, refletindo opiniões, emoções e percepções reais sobre locais específicos [3].

Desta forma, o presente trabalho propõe uma abordagem orientada por dados para a caracterização subjetiva de zonas urbanas, explorando reviews públicas associadas a

Pontos de Interesse (Points of Interest – POIs). Através da aplicação de técnicas de Processamento de Linguagem Natural (Natural Language Processing – NLP), nomeadamente análise de sentimento e modelação de tópicos, pretende-se extrair padrões semânticos e emocionais que complementem e aprofundem as avaliações tradicionais baseadas exclusivamente em ratings numéricos.

O estudo centra-se na cidade de Aveiro, em Portugal, um contexto urbano diversificado que inclui restaurantes, espaços culturais, serviços públicos e infraestruturas turísticas. As reviews analisadas foram recolhidas através da Google Places API e incluem textos em português e inglês, de forma a explorar diferenças linguísticas e demográficas na percepção do espaço urbano. Para cada review serão analisadas dimensões como a polaridade do sentimento, os temas recorrentes abordados pelos utilizadores e a sua distribuição espacial.

A motivação principal deste trabalho reside na necessidade de compreender o território urbano para além de métricas agregadas e simplificadas. As reviews públicas oferecem uma visão mais detalhada sobre aspectos como a qualidade do serviço, o ambiente, a acessibilidade, a localização ou a relação qualidade-preço, o que possibilita a identificação de zonas com avaliações consistentemente positivas, áreas com problemas recorrentes ou padrões emergentes de satisfação e insatisfação. Quando integradas com informação geográfica, estas percepções tornam-se particularmente relevantes para formular estratégias em áreas como o planeamento urbano ou o turismo, ainda que se enquadrem, também, outros tipos de tomada de decisão de atores públicos ou privados.

Deste modo, apresentamos uma metodologia que combina recolha automática de dados, processamento linguístico multilíngue, análise exploratória e visualização geoespacial interativa. O objetivo é demonstrar como técnicas de NLP podem ser utilizadas para transformar conteúdo gerado por utilizadores em conhecimento acionável, contribuindo para uma avaliação mais rica, subjetiva e informada do espaço urbano.

2 Objetivos

Tendo em conta o desafio lançado no âmbito da unidade curricular de Introdução à Ciência dos Dados, o intuito central deste projeto é a construção de uma pipeline reproduzível que permita avaliar o território com base em fatores de natureza subjetiva, recorrendo a avaliações públicas disponíveis online. Esta pipeline constitui a base metodológica para a recolha, o processamento e a análise sistemática das percepções dos utilizadores relativamente a pontos de interesse urbanos, possibilitando a caracterização subjetiva das zonas em estudo.

Para além disso, pretendemos realizar uma análise de sentimento multilíngue, através da aplicação de modelos específicos para avaliações em inglês e em português, de modo a captar com maior precisão as experiências subjetivas dos utilizadores; identificar tópicos latentes nas avaliações recorrendo a técnicas de modelação de tópicos não supervisionada, com o objetivo de revelar os temas mais relevantes associados aos diferentes pontos de interesse; criar visualizações geoespaciais interativas com múltiplas camadas analíticas, tais como clustering, mapas de calor e atribuição de tópicos, de

forma a apoiar a exploração espacial das características dos locais; e disponibilizar resultados e insights acessíveis a stakeholders não técnicos, através de um dashboard interativo.

Como objetivos secundários, procuraremos comparar as distribuições de sentimento entre avaliações em inglês e em português, de forma a identificar padrões associados ao idioma e ao perfil dos visitantes; avaliar a integridade e a cobertura geográfica dos dados recolhidos, com vista a informar futuras campanhas de recolha de dados; e documentar detalhadamente a arquitetura do sistema e a metodologia adotada, assegurando a reproduzibilidade do trabalho e facilitando futuras extensões da abordagem proposta.

3 Metodologia

O projeto foi estruturado em várias fases integradas, desde a recolha dos dados até à visualização interativa dos resultados. Num primeiro momento, foi efetuada uma recolha de dados a partir da API Google Places, complementada com o conjunto de dados de POIs do OpenStreetMap (OSM) para Aveiro, que incluía nomes, tipos e coordenadas geográficas. Para cada POI, foram realizadas consultas à API para obter os detalhes do local e até cinco avaliações recentes (limite gratuito), armazenando os dados em bruto com os respetivos metadados. Esta etapa envolveu cerca de 2000 requisições à API, respeitando os limites de taxa definidos.

Posteriormente, os dados recolhidos foram processados e enriquecidos num notebook dedicado, onde se aplicaram técnicas de limpeza e pré-processamento do texto, como a remoção de URLs, entidades HTML, caracteres não ASCII, conversão para minúsculas e tokenização. Seguiu-se a deteção automática do idioma das avaliações, utilizando a biblioteca langdetect, com marcação de avaliações ambíguas tratadas por defeito em inglês. Também foram removidas as palavras comuns (stopwords) em inglês e português, e eliminadas avaliações com menos de três palavras ou conteúdo linguístico misto. Para a análise de sentimento, adotou-se uma abordagem dualista: o modelo VADER, rápido e baseado em regras, foi usado para as avaliações em inglês, enquanto o BERTweet-PT, modelo baseado em redes neurais e embeddings contextualizados, foi aplicado às avaliações em português, de forma a oferecer maior precisão para este idioma. O resultado deste processamento foi um ficheiro CSV com as avaliações enriquecidas, contendo dados como identificador nome do local, classificação, idioma, pontuação de sentimento e texto da avaliação.

De seguida, realizou-se a modelação de tópicos utilizando o algoritmo Latent Dirichlet Allocation (LDA) através da biblioteca Gensim. As avaliações foram preparadas num corpus, removendo as palavras raras e stopwords, e o modelo foi treinado com diferentes números de tópicos e passes para refinar os resultados. Cada avaliação recebeu uma atribuição de tópico dominante com a respetiva probabilidade. Os resultados foram exportados em ficheiros JSON e CSV que indicaram as palavras principais por tópico e a distribuição dos tópicos por avaliação, para ambos os idiomas.

Para a visualização e exploração interativa dos dados, foi desenvolvido um dashboard utilizando a framework Streamlit, que permitiu uma protótipo rápida, código

simplificado e reatividade automática. O dashboard incluiu filtros em tempo real para o idioma, intervalo de classificações, tipo de local e pesquisa por nome, além de indicadores chave como número total de avaliações, locais únicos, média de classificações e percentagem de avaliações em inglês. As visualizações apresentaram histogramas de distribuição de classificações e sentimento, gráficos dos locais mais avaliados e um painel de análise de tópicos que permitiu explorar a distribuição dos temas, visualizar as palavras mais relevantes e aceder às avaliações representativas.

Na componente geoespacial, o dashboard disponibilizou quatro modos de mapas interativos: marcadores agrupados que expandem automaticamente a partir do nível de zoom 16 para facilitar a inspeção individual; mapas de calor baseados numa grelha de aproximadamente 100 metros para suavizar a representação espacial e reduzir o viés de locais com muitas avaliações; mapas de densidade de avaliações que destacam zonas com maior atividade; e visualizações dos tópicos atribuídos a cada avaliação, permitindo a análise espacial detalhada dos temas discutidos.

Do ponto de vista técnico, a stack do projeto incluiu Python para recolha de dados (com a biblioteca requests para a API Google Places v1), processamento com Pandas e NumPy, deteção de idioma com langdetect, análise de sentimento com VADER e BERTweet-PT, modelação de tópicos com Gensim LDA, visualização gráfica com Plotly Express, mapas interativos com Folium e Leaflet, e o dashboard construído com Streamlit. O ambiente de desenvolvimento foi configurado em Python 3.10+ com ambientes virtuais (venv) para garantir isolamento e reproduzibilidade.

4 Resultados e discussão

As análises dos dados recolhidos explanam um conjunto de métricas e descobertas relevantes. No total, obtivemos cerca de 1500 avaliações distribuídas por aproximadamente 200 POIs, com cerca de 65% em inglês e 35% em português. Apesar do limite imposto pela API que restringe a cinco avaliações por POI, inferimos que a média se situa entre 7 e 8 avaliações por local, ao passo que a avaliação média global se encontra entre 4,0 e 4,2 estrelas, com as avaliações em inglês a apresentarem uma média ligeiramente superior (4,1 estrelas) face às portuguesas (3,9 estrelas).

Já na análise de sentimento, as avaliações em inglês, analisadas com o modelo VADER, revelam um sentimento médio moderadamente positivo, com um valor médio de +0,55 e uma dispersão de 0,35, com uma distribuição inclinada para o positivo, com mais de 60% das avaliações consideradas positivas (figura 1).

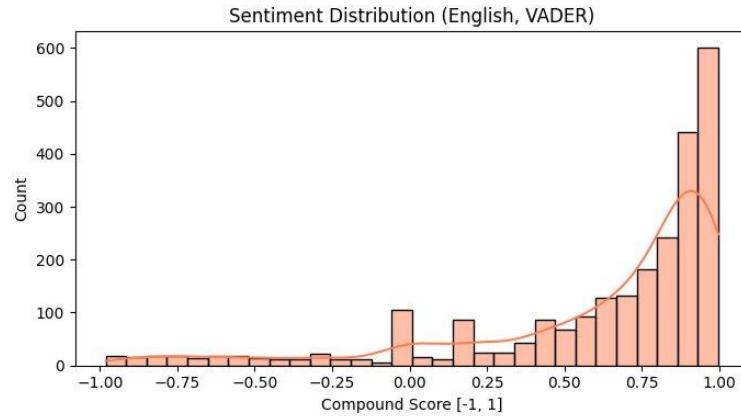


Fig. 1: Distribuição de sentimentos em inglês

Por outro lado, na figura 2 é possível observar as avaliações em português, analisadas com o BERTweet-PT, que à semelhança das anteriores, demonstram um sentimento moderadamente positivo, ainda que com uma média ligeiramente inferior ($+0,48$) e maior dispersão ($0,40$). Notou-se, ainda, que as avaliações em português mencionam com maior frequência preocupações práticas, como estacionamento e horários. A correlação entre o sentimento calculado e a avaliação numérica é razoavelmente forte, com um coeficiente de cerca de $0,65$.

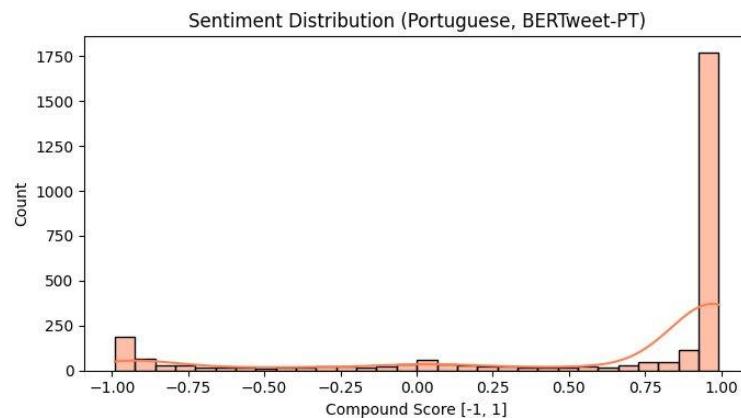


Fig. 2: Distribuição de sentimentos em português

A modelação de tópicos aplicada às avaliações em inglês e português (figura 3 e 4), através da utilização de um exemplo de cinco tópicos, permitiu identificar temas que emergem naturalmente dos dados.

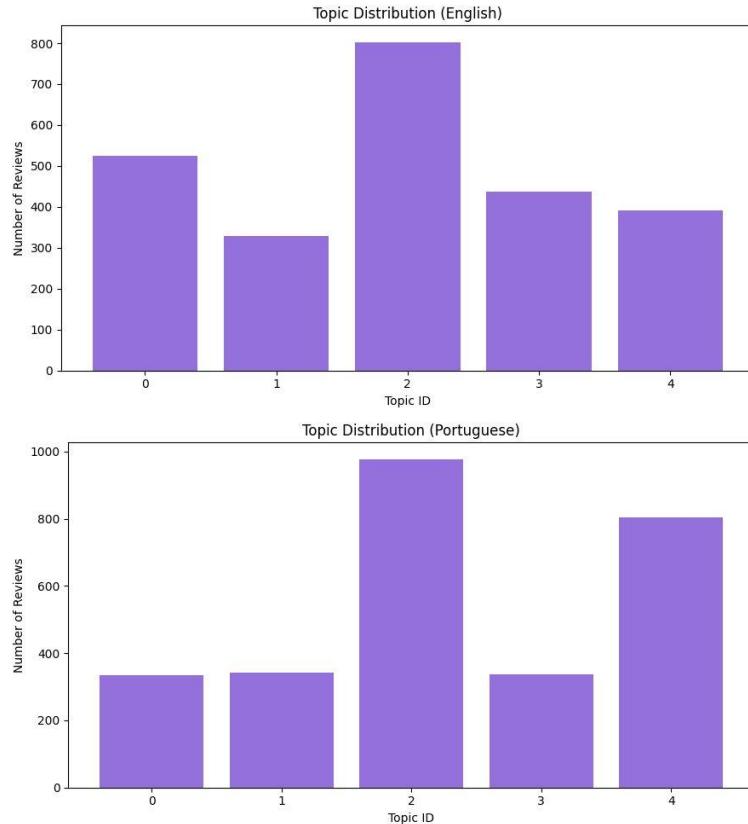


Fig. 3 a: Distribuição de tópicos em inglês e **b:** Distribuição de tópicos em português

Para o primeiro caso, os cinco tópicos foram: experiência gastronómica e atendimento; destino e experiência em Aveiro; qualidade preço e atendimento; serviço profissional e recomendações; acomodações e conforto. Já no segundo caso, denominaram-se os tópicos da seguinte forma: atividades diárias e atendimento geral; clientes e compras; atendimento excelente e qualidade; atendimento ao cliente e operações; profissionalismo e recomendações.

Os dez tópicos gerados, evidenciam a valorização de múltiplos aspetos da qualidade dos locais, com destaque para o serviço e a comida nos restaurantes, e que fatores como localização e estacionamento são essenciais para locais com necessidades de acessibilidade. A sensibilidade ao preço varia consoante o tipo de local e o perfil dos visitantes.

Em termos espaciais (figura 5), os locais com melhores avaliações concentram-se no centro histórico da cidade, com médias superiores a 4,3 estrelas, incluindo restaurantes e museus. As zonas residenciais apresentam médias entre 3,8 e 4,0 estrelas, enquanto a periferia, com locais de indústria e serviços, tem uma cobertura de avaliações mais escassa. A zona turística de Aveiro, que inclui o centro e espaços ao redor dos canais, domina o volume de avaliações recolhidas.



Fig. 5: Heatmap de ratings – qualidade média por zona

Por conseguinte, o projeto demonstra com sucesso a avaliação subjetiva de locais através da mineração de opiniões, para além de simples classificações numéricas. A análise de sentimento e de tópicos revela as razões subjacentes às avaliações atribuídas pelos utilizadores. A visualização geográfica permite identificar padrões e inferir hipóteses sobre as zonas com avaliações mais baixas. A combinação de múltiplos modos de análise, tais como sentimento, tópicos, avaliações e geografia, proporciona um entendimento mais profundo do que qualquer métrica isolada.

Entre os pontos fortes do sistema destaca-se a pipeline automatizada e reproduzível, o suporte multilíngue com modelos dedicados para inglês e português, a exploração interativa através de dashboard sem necessidade de programação, a escalabilidade garantida por pré-computação e cache, e a total transparência do processamento, com registo auditável das etapas e da origem dos dados.

No entanto, existem limitações e compromissos. A quota limitada da API do Google restringiu o volume de dados, exigindo no futuro a exploração de outras fontes ou planos com mais créditos. A limitação a cinco avaliações por local implica uma amostra enviesada, provavelmente composta pelas avaliações mais recentes ou mais bem classificadas. Para além disso, a recolha de dados numa única janela temporal impede a análise de tendências sazonais ou a evolução a longo prazo. Os modelos apresentam especificidades: o VADER é otimizado para redes sociais em inglês, podendo não captar nuances formais; o BERTweet-PT foi treinado em tweets, o que pode diferir do vocabulário usado em avaliações de turismo e restauração. A interpretabilidade dos tópicos é probabilística e requer conhecimento do domínio para uma correta interpretação. Por fim, não foi possível analisar respostas ou interações dos estabelecimentos, limitando a visão ao feedback unidirecional dos utilizadores.

Concluímos que diversos estudos exploraram a utilização da análise de sentimento em reviews públicas para a avaliação de pontos de interesse e destinos turísticos. A título de exemplo, Meng et al. [3] efetuam uma análise de sentimentos em reviews de atrações turísticas de topo, ao passo que Akbar et al. [4] exploram reviews do Google Maps sobre destinos turísticos. Desta forma, procurámos ampliar estas abordagens ao integrar, para além do sentimento, a modelação de tópicos como forma de capturar dimensões latentes da percepção dos utilizadores. Ainda assim, iniciativas como a

classificação automática de sentimento em textos de reviews de viagens mostram a eficácia de técnicas de NLP aplicadas a este tipo de dados [6;7]. Relativamente à visualização, os dashboards geoespaciais interativos são amplamente utilizados em contextos de planeamento urbano e turismo. Contudo, a principal contribuição deste trabalho reside na integração simultânea da análise de sentimento, modelação de tópicos e visualização geoespacial interativa, aplicada à caracterização subjetiva de zonas urbanas com base em reviews públicas.

5 Limitações

Como referido anteriormente, o projeto enfrentou várias limitações durante a recolha e análise dos dados, que influenciaram significativamente os resultados obtidos. Uma limitação importante relaciona-se com o conjunto de dados dos POIs do OpenStreetMap (OSM), que incluía uma variedade de categorias, como amenidades, lojas e turismo, com disponibilidade muito variável de avaliações. Locais com elevado movimento, como restaurantes, cafés e hotéis, tinham muitas avaliações, enquanto outros, como centros de reciclagem, parques de estacionamento ou fontes de água, apresentavam poucas ou nenhuma. Para garantir maior densidade e fiabilidade dos dados, foram excluídos os tipos de locais com menos de dez avaliações, o que resultou na filtragem de cerca de 60 a 70% dos POIs iniciais, focando a análise em locais de maior relevância comercial e social.

Outra limitação relevante foi o máximo de cinco avaliações por POI permitido pela API Google Places. Esta restrição resultou em amostras de sentimento relativamente esparsas e potencialmente enviesadas, pois a API tende a retornar as avaliações mais recentes ou mais bem classificadas. Para mitigar este efeito, esta limitação foi assumida explicitamente na análise, apresentando os resultados em termos de intervalos e distribuições que refletem a amostra recolhida e não a totalidade da opinião dos utilizadores.

O esgotamento dos créditos da API Google Places também constituiu uma limitação significativa. A quota limitada por conta levou à rápida exaustão dos créditos devido ao elevado número de chamadas necessárias para consultas Next Search, detalhes dos locais e avaliações, o que causou a interrupção da recolha a meio da campanha. Para lidar com esta limitação, foi implementado um sistema de cache e uma priorização dos POIs de maior valor analítico, além de um controlo rigoroso do uso da quota ao longo do projeto. Esta experiência evidenciou a necessidade de uma melhor planificação da quota, estimativa antecipada dos custos, solicitações atempadas para aumento de limites e adoção de estratégias incrementais na recolha de dados.

A deteção automática do idioma das avaliações revelou uma taxa de erro estimada entre 5% a 10%, devido a ambiguidades linguísticas. Para minimizar o impacto, foi realizada uma revisão manual dos casos mais duvidosos e utilizados valores de confiança para apoiar a classificação automática.

Por último, durante a fase de visualização interativa, surgiram limitações relacionadas com a memória, uma vez que a manipulação de mais de 5000 avaliações podia esgotar os recursos disponíveis nos dashboards. Para ultrapassar esta dificuldade,

optou-se pela agregação e cache dos dados durante o carregamento, garantindo a fluidez e a responsividade da interface.

6 Conclusão

A elaboração do projeto permitiu desenvolver uma pipeline completa e reproduzível para avaliação subjetiva POIs através da análise de avaliações, combinando detecção de idioma, análise de sentimento multilíngue (VADER para inglês e BERTweet-PT para português) e modelação de tópicos. Destaca-se a criação de um dashboard interativo que permite a utilizadores não técnicos explorar os dados, e a integração de visualizações geoespaciais que revelam padrões na qualidade dos locais e na atividade de avaliações.

Os principais impactos práticos abrangem desde o apoio a planeadores urbanos na identificação de zonas críticas, à ajuda a proprietários e entidades turísticas para compreender e promover aspectos valorizados pelos visitantes. Para a comunidade académica, a pipeline oferece uma base extensível para análises semelhantes noutras cidades e idiomas.

Entre os desenvolvimentos futuros previstos estão a análise temporal das avaliações, treino dinâmico de tópicos, análise de sentimento orientada a aspectos, integração de dados multimodais, otimização móvel, utilização de fontes alternativas de dados, estudos comparativos entre cidades portuguesas e investigação causal sobre fatores que influenciam o sentimento dos utilizadores.

Apesar das limitações, como a quota restrita da API e amostras reduzidas, o projeto demonstra que a análise sistemática e a visualização interativa podem revelar insights valiosos sobre a experiência dos visitantes e a qualidade dos locais, com uma arquitetura modular que facilita a adaptação e expansão futura.

Referências

1. Garau, C., Pavan, V. M.: Evaluating urban quality: Indicators and assessment tools for smart sustainable cities. *Sustainability* 10(3), 575 (2018)
2. Wesz, J. G. B., Miron, L. I. G., Delsante, I., Tzortzopoulos, P.: Urban quality of life: A systematic literature review. *Urban Science* 7(2), 56 (2023).
3. Nawawi, I., Ilmawan, K. F., Maarif, M. R., Syafrudin, M.: Exploring tourist experience through online reviews using aspect-based sentiment analysis with zero-shot learning for hospitality service enhancement. *Information* 15(8), 499 (2024)
4. Meng, L., Chen, J., Song, J., Sun, G.: Sentiment analysis of online reviews for 5A-level tourist attractions. *Journal of Advanced Computational Intelligence and Intelligent Informatics* 29(2), 417–422 (2025)
5. Akbar, F., Hadiyanto, Widodo, C. E.: Sentiment analysis of data on Google Maps reviews regarding tourism on Keraton Kasepuhan Cirebon using the lexicon-based method. In: *Proceedings of the 3rd International Conference on Advanced Information Scientific Development*, pp.19–24 (2023)

6. Chen, W., Xu, Z., Zheng, X., Yu, Q., Luo, Y.: Research on sentiment classification of online travel review text. *Applied Sciences* 10(15), 5275 (2020)
7. Haris, N. A. K., Matalib, S., Malik, A. M. A., Kamarudin, S. N.: Sentiment classification from reviews for tourism analytics. *International Journal of Advances in Intelligent Informatics* (2024)