NIVEDITA
JOURNAL INFORMATICS
e-ISSN : xxx-xxx
p-ISSN : xxx-xxx

Vol. 01, No. 1, Desember 2024

Irrelevancy Detection in Multilingual Tourism Review

Putu Kussa Laksana Utama[1], Luh Gede Surya Kartika[2], I Putu Adi Pratama[3], I Gede Wahyu Sanjaya[4], Putu Adi Saskara[5]

[1,2,3,4,5]Program Studi Informatika, Universitas Hindu Negeri I Gusti Bagus Sugriwa Denpasar, Denpasar, Indonesia

email: [1]kussa_laksana@uhnsugriwa.ac.id

Commented [MSOffice1]: This study categorizes reviews as relevant or ambiguous, but does not provide a very clear definition of ambiguity. Ambiguity can come from a variety of factors, such as unclear language, sarcasm, or lack of context. Does this study deal with all forms of ambiguity explicitly?

Commented [KLU2R1]: We have already put additional context in the beginning of introduction section

Abstract

This study investigates irrelevancy within multilingual tourism reviews, focusing on how off-topic or ambiguous user-generated content can undermine reliable insight for travelers. A consolidated dataset is constructed by combining a publicly available resource from Kaggle with additional posts acquired from X (formerly Twitter). Each review is manually labeled as relevant or ambiguous to capture instances where the content fails to clearly address travel or hotel-related topics. We employ a multilingual BERT embedding model to encode the diverse language inputs, enriched with a sentiment vector derived via knowledge distillation from twitter-xlm-roberta-base to DistilBERT. A gating mechanism then fuses the semantic and emotional signals, highlighting parts of each review most influenced by user attitudes. The final classification stage involves fine-tuning a BERT-based network to distinguish between unambiguous and ambiguous content. Experimental comparisons with a Monolingual BERT approach and a baseline (multilingual embedding without sentiment) reveal that incorporating sentiment features yields consistent improvements in accuracy, precision, recall, and F1-score. This outcome underscores the importance of capturing emotional cues to mitigate errors arising from partial dissatisfaction, unclear references, or cultural nuances. From a practical standpoint, the results point to potential applications in automated moderation, improved recommendation systems, and policy guidelines for tourism platforms. Overall, this work demonstrates that sentiment-aware, multilingual models can enhance detection of irrelevancy and ambiguity, fostering more trustworthy and context-rich online review ecosystems in the travel domain.

Keywords: *Tourism Reviews, Multilingual Embeddings, Irrelevancy Detection, Sentiment Analysis, Knowledge Distillation.*

Abstrak

Penelitian ini mengkaji relevan atau tidak relevannya ulasan pariwisata dalam media online dalam berbagai bentuk bahasa, dengan menitikberatkan pada bagaimana konten yang tidak sesuai topik yang dihasilkan pengguna dapat mengaburkan informasi bagi wisatawan. Salah satu metode dalam penelitian ini mengkonstruksi dataset gabungan dimana data publik yang bersumber dari Kaggle ditambahkan dengan data post X (Twitter). Setiap ulasan diberi label secara manual (relevan atau tidak relevan) untuk mengidentifikasi kasus di mana konten tidak secara jelas membahas topik terkait hotel atau akomodasi pariwisata lainnya. Pada penelitian ini, model embedding BERT multibahasa digunakan untuk menyandikan input dalam berbagai ragam bahasa yang disertai dengan sentimen atau emosi yang dimiliki oleh input tersebut. Aspek sentimen pada setiap ulasan ini diperoleh melalui metode knowledge distillation dari *twitter-xlm-roberta-base* ke DistilBERT. Mekanisme *gating* kemudian digunakan untuk menggabungkan sinyal semantik dan emosional, dengan penekanan pada bagian-bagian ulasan yang paling berpengaruh. Tahap klasifikasi akhir melibatkan fine-tuning jaringan BERT guna membedakan antara konten yang relevan dan tidak relevan. Hasil pengujian menunjukan performa pendekatan Monolingual BERT sedikit lebih buruk dibandingkan dengan pendekatan Multilingual BERT dan juga Multilingual BERT ditambah aspek sentimen ditunjukan oleh nilai akurasi, presisi, recall, dan F1-score. Hasil ini menekankan pentingnya menangkap nuansa emosional guna mengurangi kesalahan akibat referensi yang tidak relevan dan juga perbedaan budaya. Dari sudut pandang praktis, hasil yang diperoleh pada penelitian ini menunjukan adanya potensi implementasi untuk autamatisasi moderasi konten, sistem rekomendasi yang lebih baik, dan panduan kebijakan untuk platform-paltform pariwisata. Secara keseluruhan, penelitian ini menunjukkan bahwa model multibahasa yang menyertakan sinyal sentimen dapat meningkatkan deteksi ketidakrelevanan dan ambiguitas, sehingga mendorong terciptanya ekosistem ulasan daring yang lebih andal dan kaya konteks di bidang pariwisata.

NIVEDITA
JOURNAL INFORMATICS
e-ISSN : xxx-xxx
p-ISSN : xxx-xxx

Vol. 01, No. 1, Desember 2024

Kata kunci: Ulasan Pariwisata, Embedding Multibahasa, Deteksi Ketidakrelevanan, Analisis Sentimen, Knowledge Distillation.

Diajukan: 4 Maret 2025; Diterima: 10 Maret 2025

INTRODUCTION

Ambiguity in online review document can be translated as content that does not clearly convey a hotel or tourism-related context, resulting in insufficient information to understand the user's true experience. This lack of context can manifest in multiple ways, such as unclear language, tangential references, or even sarcasm that obscures the actual sentiment. Ambiguity is a serious challenge in the study of multilingual tourism reviews, as words and phrases often carry multiple meanings across different languages and cultural contexts [1]. This issue can lead to misunderstandings when travelers read online comments about accommodation, attractions, or travel services. In recent years, researchers have started to study ambiguity detection methods [2], [3], [4], [5], aiming to improve the clarity and reliability of automated text analysis. Although many approaches attempt to improve clarity by identifying ambiguous terms or sentiments, they often overlook the role of irrelevancy as a trigger for misinterpretation. Even if a review is grammatically coherent, it may still fail to contribute meaningful information about the guest's actual experience if it is off-topic or contains non-tourism-related content. This scenario becomes more pronounced in multilingual contexts, where mismatched keywords or translation nuances can further obscure the relationship between the posted content and the intended topic.

Irrelevancy becomes a notable concern when users post remarks that do not clearly pertain to the hotel or travel context. This lack of relevance can mask or distort the overall sentiment and meaning, leading readers to question whether a review actually provides useful insight. Consequently, irrelevancy can amplify the broader challenge of ambiguity in multilingual tourism reviews, where linguistic and cultural differences already complicate textual interpretation. Multilingual text document often include a wide range of language structures, vocabulary choices, and writing styles [6]. These differences can amplify the problem of irrelevancy, making it more difficult for natural language processing (NLP) agent to identify intended meanings.

In response, the present research aims to highlight irrelevancy as a key factor influencing ambiguity in tourism reviews. By focusing on the detection of irrelevant content within multilingual user-generated data, we seek to ensure that only texts genuinely reflecting hotel or travel aspects are subject to deeper sentiment analysis. This approach not only enhances the accuracy of ambiguity detection methods but also yields a cleaner corpus of user feedback that better supports travel planning and service improvement. Ultimately, by acknowledging and addressing irrelevancy alongside other sources of ambiguity, we can develop more comprehensive tools for analyzing multilingual tourism reviews in a way that benefits both travelers and industry stakeholders.

A key research question in this field is how to construct a comprehensive multilingual tourism reviews dataset. Such a dataset should include a large variety of languages, review platforms, and cultural contexts in order to capture the different ways people describe hotels, attractions, and travel experiences. Careful collection and labeling of these reviews is crucial, with attention not only to textual content but also to linguistic features that can influence interpretation.

Another important research question concerns the development of algorithms that can accurately capture sentiment tones in reviews. Tourism feedback can contain complex emotional expressions, ranging from enthusiastic praise to subtle dissatisfaction. Designing models that can distinguish these different levels of sentiment requires advanced techniques in sentiment analysis and natural language understanding. Commonly used approaches include deep learning techniques [7], [8] or transformer-based architectures [9], [10], combined with domain-specific knowledge to identify tourism-specific terms and phrases. By addressing these challenges, future work can offer more robust tools for both travelers and tourism providers, ensuring that reviews are interpreted accurately across diverse linguistic and cultural settings.

Irrelevancy in multilingual tourism reviews poses a significant barrier to reliable automated analysis. We seek to address this challenge by systematically gathering data, extracting sentiment information, and merging both semantic and emotional cues into a unified feature set for machine learning models. Specifically, our goal is to compile a comprehensive multilingual dataset that captures diverse linguistic and cultural elements, implement methods to accurately identify sentiment tones from user reviews, integrate the extracted sentiment data into a semantic vector representation of the user reviews, and compare the performance of multiple machine learning models using this enriched feature set against baseline models without sentiment-based features.

Based on these objectives, our research will address the following questions:

i. How can a robust multilingual dataset of tourism reviews be constructed to capture linguistic diversity and potential irrelevancy?
ii. Which methods are most effective for extracting and representing sentiment tones from multilingual user reviews, and how should these be integrated into the semantic representation?
iii. How does the performance of machine learning models that utilize combined semantic and sentiment features compare to that of baseline models that do not incorporate sentiment information?

In summary, the objectives and questions presented here lay the groundwork for exploring irrelevancy in multilingual tourism reviews through a strategy that involves dataset construction, sentiment extraction, and feature integration. In the next section, we will discuss related work in this domain, focusing on existing techniques for managing linguistic and cultural diversity in tourism reviews, as well as advances in sentiment analysis and machine learning methods.

RELATED WORKS

Previous studies on multilingual tourism reviews have explored various aspects of dataset construction, sentiment analysis, and machine learning techniques. Early work often limited datasets to a single language or relied on manually translated reviews, hindering the capture of subtle linguistic and cultural nuances [11], [12]. More recent efforts have shifted toward larger and more diverse corpora that encompass multiple languages and dialects, thereby better reflecting the complexity of tourism-related expressions and idiomatic phrases [13] [14], [15].

In sentiment analysis, traditional lexicon-based methods depend on predefined lists of positive and negative terms. Although straightforward to implement, these methods can struggle with cross-lingual morphological and syntactic variations [11]. Machine learning models, particularly deep learning, often show improved adaptability because they learn from labeled examples rather than relying solely on fixed dictionaries [16]. Transformer-based architectures such as BERT have demonstrated strong performance in classifying sentiment for short texts [17], [18] but typically require domain-specific fine-tuning to handle tourism-specific terminology.

Several works have also investigated the integration of semantic and sentiment features for tasks like opinion mining and review summarization [19]. This approach captures both the literal meaning of words and the underlying emotional or attitudinal tones, yielding richer insights for downstream applications. Studies indicate that incorporating sentiment-based signals can substantially improve performance [20], [21], although the extent of these gains can vary based on factors such as dataset size, annotation quality, and the linguistic complexity of the target languages.

In addition, work focusing on irrelevant review detection in multilingual online texts has often been framed as a text classification challenge, where the goal is to accurately predict whether certain words, phrases, or sentences exhibit irrelevancy. Our current research, which focuses on irrelevancy detection using sentiment tones as a central feature, falls under this broader classification paradigm. By treating the identification of irrelevant reviews as a downstream text classification task—enriched by sentiment extraction and integrated semantic information—this approach has the potential to boost accuracy and interpretability, particularly for users who seek clearer insights into diverse tourism experiences.

METHODOLOGY

We adopt a multi-step process that involves crafting a robust multilingual dataset, extracting sentiment information, and combining semantic and emotional features for downstream ambiguity detection. Each step is designed to ensure that our model captures both the linguistic variety and the sentiment nuances found in tourism-related reviews.

Data Crafting

We begin by acquiring an established tourism review dataset from Kaggle [1], which provides a foundational collection of user feedback on various hotels. Next, we expand the dataset by gathering real-time content from X (formerly Twitter) through its official API. For each hotel name present in the Kaggle dataset, we query the API for posts in English, Spanish, German, and Portuguese that mention the same hotel. Rather than discarding posts deemed irrelevant, we manually label them according to their relevance and clarity.

[1] Source: https://www.kaggle.com/datasets/datafiniti/hotel-reviews

Commented [MSOffice3]: The dataset sources are from Kaggle and X (Twitter), but there is no in-depth discussion about whether these two sources have similarities in language style, sentence structure, and relevance to the purpose of the study. Do datasets from Kaggle have the same characteristics as data from X? If not, how does this study ensure that the model is not biased against a single source?

Commented [KLU4R3]: We've also already revised this section, please have a look

Although Kaggle and X (formerly Twitter) may differ in language style, sentence structure, and user motivations for posting, we have taken steps to align these sources for our study. First, we extracted hotel and restaurant names from the Kaggle dataset and then searched on X using both the establishment name and relevant hashtags (for instance, #HotelName, #RestaurantName, or #TravelDestination). By focusing on posts containing these hashtags or direct references, we ensured that the retrieved tweets were thematically similar to the reviews found on Kaggle.

During this data acquisition process, we applied manual curation to filter out tweets that were clearly off-topic or irrelevant to the travel context. For example, a post mentioning the hotel name but referring to an unrelated event or personal anecdote was marked as ambiguous rather than discarding it outright. We also examined language style and structure, ensuring that tweets providing genuine travel experiences or feedback were grouped with the Kaggle entries in a consistent manner. In many cases, tweets may be shorter and more informal than Kaggle reviews, but our curation steps aimed to preserve comparable content about accommodations, dining, or general travel impressions.

By aligning the thematic focus and labeling off-topic content, we reduce the risk that the model becomes biased toward one source's style or structure. The final curated dataset thus integrates two originally distinct data sources—Kaggle and X—into a more unified corpus in which relevant aspects of hotel or restaurant reviews are retained. This manual tailoring approach allows the study to maintain consistency in language style, sentence structure, and subject matter across both data sets.

Multilingual Embedding Vector Representation

After completing the data acquisition and labeling process, we transform each review into a vector representation using a multilingual BERT-based embedding model (mBERT) [22], [23], [24] Passing each review through the embedding model yields a high-dimensional vector $\mathbf{e} \in \mathbb{R}^d$ that encodes contextual and semantic information across different languages. This representation enables the system to handle linguistic variation more effectively and compare terms that differ in surface forms.

Sentiment Classification

Initially, we consider of employing a popular pre-trained model (DistilBERT based model) [25], [26], [27] to categorize each review as positive, neutral, or negative. This step produces a one-hot or probability distribution over sentiment classes, which we represent as a sentiment vector $\mathbf{s} \in \mathbb{R}^3$. However, using pre-trained model by default is overlooking the opportunity to maximize the capability of pre-trained model trained on a downstream task such as sentiment classification. Hence, we leverage a zero-shot or weakly supervised knowledge distillation technique that relies on the twitter-xlm-roberta-base model as a teacher. First, we apply the teacher model to our unlabeled or partially labeled review set, generating "soft" sentiment labels (probabilities indicating positive, neutral, or negative). These probabilities reflect the teacher model's understanding of sentiment across multiple languages.

Next, we take advantage of the teacher model's predictions to supervise a smaller, more efficient DistilBERT model, used here as the student model. During this knowledge distillation process, DistilBERT is trained to match or approximate the teacher's soft label distributions rather than single "hard" class labels. This approach benefits from the rich contextual cues encoded in the teacher model while allowing the student model to retain a lighter architecture. As a result, we obtain a compact sentiment analysis model that still captures nuanced emotional signals in user reviews. By incorporating both zero-shot or weakly supervised labeling and knowledge distillation, our sentiment classification pipeline is able to accommodate a variety of multilingual review styles while remaining efficient in terms of computational resources.

We chose a knowledge distillation pipeline instead of a full XLM-RoBERTa or a specialized tourism-domain transformer for two main reasons. First, no comprehensive, pre-trained language model currently exists for the tourism review domain, which requires large-scale, domain-specific corpora spanning multiple languages. Second, while XLM-RoBERTa is a robust multilingual model, it is more general and not specifically tailored to the informal, emotive style of user-generated travel reviews. In contrast, *twitter-xlm-roberta-base* focuses on social media text, making it better at handling short, colloquial expressions.

Feature Combination

To integrate the multilingual embedding vector **e** and the sentiment vector **s**, we employ a gating mechanism. First, we transform **s** into a gating vector $\mathbf{g} \in \mathbb{R}^d$ via a small feed-forward network:

$$\mathbf{g} = \sigma(W_s \mathbf{s} + b_s) \quad (1)$$

Commented [MSOffice5]: The use of multilingual BERT combined with sentiment from twitter-xlm-roberta-base to DistilBERT through knowledge distillation is an interesting approach, but it is not explained why this architecture was chosen over other alternatives such as full XLM-Roberta or tourism domain-based transformer models.

Commented [KLU6R5]: We added an argument that's probably related this comment

where $W_s \in \mathbb{R}^{dx3}$, $b_s \in \mathbb{R}^d$, and σ is a non-linear activation function. Next, we apply element-wise multiplication between **e** and **g** to produce a fused representation:

$$\mathbf{f} = \mathbf{e} \circ \mathbf{g} \quad (2)$$

This operation modulates each dimension of **e** based on sentiment-related signals, highlighting or suppressing particular semantic features. By doing so, we enable the model to focus on critical aspects of the embedding that correlate with user sentiment.

Classification Model

Finally, we feed **f** into a BERT classification model fine-tuned for ambiguity detection in multilingual tourism reviews. During training, we minimize the classification error to learn patterns that capture both

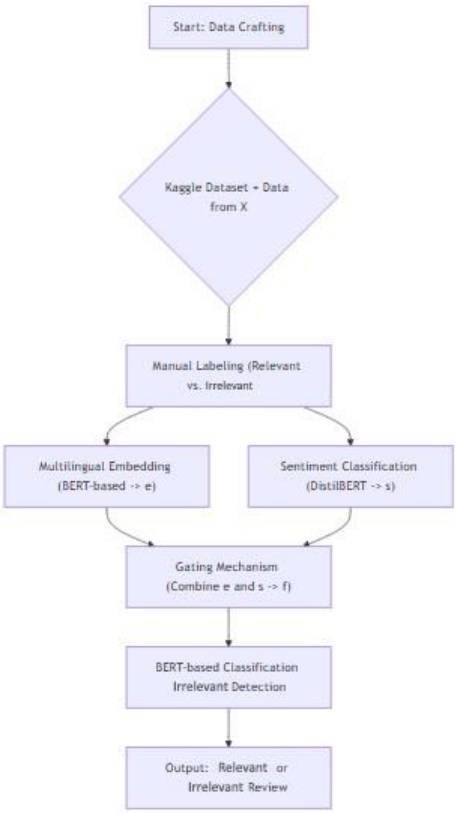

Figure. 1. Multilingual Ambiguity Detection Framework

Linguistic and sentiment-driven nuances. By consolidating semantic embeddings and emotional signals through the gating mechanism, our methodology aims to provide a comprehensive approach for identifying and characterizing ambiguous content across multiple languages.

The figure 1 shows how we integrate multilingual embeddings and sentiment features through a gating mechanism to detect ambiguity in multilingual tourism reviews.

NIVEDITA JOURNAL INFORMATICS
e-ISSN : xxx-xxx
p-ISSN : xxx-xxx
Vol. 01, No. 1, Desember 2024

Training and Hyperparameter Setup

After constructing the final classification model, we split the labeled dataset into two subsets: a training set and a test set. The training set is used to optimize the model's parameters via backpropagation, while the test set serves as the final evaluation benchmark. Throughout training, batches of samples are fed into the model, and network parameters are updated to minimize the cross-entropy loss function.

We utilize the Adam optimizer with a learning rate of 1×10^{-5}. This rate is chosen to ensure gradual and stable updates to the parameters, particularly important for pretrained language models. The total number of epochs is set to 100, which provides ample time for convergence without prematurely halting improvements in performance. We further incorporate dropout or weight decay as regularization techniques if necessary, striking a balance between underfitting and overfitting. For the gating mechanism, we initialize the feed-forward layer's weights and biases using Xavier initialization and apply the ReLU activation function to introduce nonlinearity. Each training batch has a size of 32, adhering to hardware constraints and empirical performance considerations. In evaluating the model, we deploy a confusion matrix to analyze how well the classifier distinguishes between ambiguous and unambiguous reviews.

RESULT

The table 1 presents accuracy, precision, recall, and F1-score for each method, with slight performance gains observed when sentiment information is included.

Table 1. Performance comparison of with-sentiment method, without-sentiment method, and BERT monolingual (Eng),

Method	Accuracy	Precision	Recall	F1-Score
Baseline (without-Sentiment)	0.84	0.82	0.81	0.81
With Sentiment	0.87	0.85	0.84	0.85
BERT-base (English)	0.80	0.79	0.78	0.78

This figure 2 compares the baseline approach (multilingual embeddings + simple feed-forward) with the enhanced approach (multilingual embeddings + sentiment analysis) and monolingual approach.

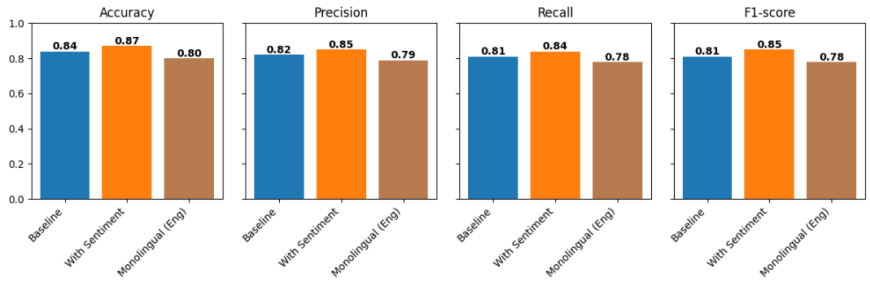

Figure. 2. Performance comparison in graph visualization

Commented [MSOffice7]: Data visualization is unneeded if there is data presented in the table, contrary to the table provided hardly to understand

Commented [KLU8R7]: We provide visual representation, as it's more attractive for reader instead of single table.

These results compare three strategies for classifying tourism-related user reviews in terms of accuracy, precision, recall, and F1-score. The Monolingual BERT (English Only) model demonstrates the lowest performance, with an accuracy of 80% and corresponding precision, recall, and F1-scores of 79%, 78%, and 78%, respectively. This outcome suggests that while a language-specific model can capture English nuances, its applicability is limited when dealing with the diverse linguistic characteristics of a broader dataset.

Meanwhile, the Baseline approach, which uses multilingual embeddings without sentiment features, attains slightly higher performance, achieving 84% accuracy and an F1-score of 81%. This increase may stem from the model's ability to handle multiple languages more effectively, thereby reducing errors caused by code-switching or non-English text. However, despite offering broader coverage of multilingual content, the absence of explicit sentiment information still leaves room for misinterpretation, particularly when a user's emotional tone is crucial for deciphering ambiguous or context-dependent feedback.

Lastly, the With Sentiment (Multilingual) model shows the strongest overall results, with an accuracy of 87% and an F1-score of 85%. By incorporating emotional cues obtained from a knowledge-distilled DistilBERT sentiment vector, this approach adds another layer of interpretability and fine-grained nuance to multilingual embeddings. Subtle distinctions in positive, negative, or neutral attitudes thus become clearer, reducing misclassifications and enhancing the model's ability to handle the inherent complexity of user-generated tourism reviews.

DISCUSSION

The comparative findings from our experiments suggest that incorporating sentiment features within a multilingual framework offers significant advantages in handling the complexity of tourism-related user reviews. By analyzing three distinct approaches—Monolingual BERT (English Only), Baseline (No Sentiment, Multilingual), and With Sentiment (Multilingual)—we observe that leveraging emotional cues in tandem with multilingual embeddings has a tangible impact on classification quality. Several key insights and connections to past research help clarify why this outcome arises and what it implies for future work in this domain.

First, the inferior performance of the Monolingual BERT model aligns with other studies that highlight the limitations of single-language approaches in tasks involving cross-lingual data [28]. Although monolingual models may excel when applied to high-resource languages such as English, they fail to capture essential features of non-English texts, including idiomatic expressions, cultural references, and syntactic structures. In contrast, our Baseline (No Sentiment, Multilingual) model, which employs multilingual embeddings, is better equipped to manage code-switching and text in multiple languages, thus achieving modestly higher accuracy and F1-scores. This underscores observations made by [29], who emphasize the diversity of language usage in tourism reviews and the need for methods that address multilingual contexts head-on.

Nevertheless, the most striking improvement emerges when sentiment is explicitly modeled—using a knowledge-distilled DistilBERT vector derived from twitter-xlm-roberta-base. The sentiment-enriched approach matches findings from [30], who advocates for deeper integration of opinion signals to bolster text classification outcomes. By treating positive, neutral, or negative emotion as an additional dimension of the data, we capture subtle user attitudes that may otherwise remain hidden in purely semantic representations. This helps resolve ambiguities arising from partial dissatisfaction or lukewarm endorsements, aligning with the notion that emotional cues can disambiguate nuanced user intent [31].

Another noteworthy factor is how irrelevancy detection interacts with sentiment. Our revised methodology underscores that reviews deemed irrelevant, or containing off-topic content, can still exhibit emotional polarities. Past research [11] indicates that even seemingly extraneous remarks can provide indirect clues—if not about the hotel itself, then about the user's broader travel mindset. By consolidating these signals into the gating-based feature combination process, we ensure that relevant parts of the embedding are amplified, while sentiment cues direct attention toward emotionally significant segments. Consequently, the model better navigates ambiguous expressions that might stem from unclear or context-agnostic user statements.

Furthermore, the knowledge distillation paradigm offers practical benefits. DistilBERT, being lighter than its teacher model, can deploy faster and scale more easily across large datasets or real-time applications, confirming prior observations about the efficiency advantages of distilled models [32]. When dealing with large volumes of multilingual tourist data, computational constraints can be critical; therefore, balancing model complexity with inference speed becomes an important design consideration.

Taken together, these insights reinforce the proposition that sentiment integration—supported by an effective multilingual embedding strategy—can reduce classification errors, particularly for reviews that are partially relevant or exhibit ambiguous linguistic signals. Our results are broadly consistent with previous studies highlighting the role of emotional context in clarifying otherwise uncertain text [19]. While there remain open questions about the generalizability of these findings to less commonly spoken languages or specialized dialects, the evidence here makes a strong case for sentiment-aware, multilingual approaches in tourism analytics. By refining techniques that fuse semantic and emotional dimensions, future research can continue improving the interpretation of user-generated content and address evolving challenges such as deeper cultural nuance or newly emergent travel-related platforms.

An important takeaway from this work is the value of integrating multilingual sentiment analysis and irrelevancy detection into the operational practices of tourism platforms and related industries. By automating the identification of off-topic or uncertain reviews, platform administrators can streamline the moderation process, ensuring that prospective travelers access higher-quality and more relevant feedback.

This filtering approach can also curb the spread of misleading or spam-like content, improving trust in user-generated information. In practice, an application could flag potentially irrelevant reviews for human validation, allowing moderators to assess whether the comments genuinely pertain to a hotel stay, a local attraction, or a broader travel experience. The result is a curated corpus of user feedback that not only better supports traveler decision-making but also sheds light on key areas of improvement for service providers.

From a policy perspective, stakeholders such as tourism boards, government agencies, or professional associations may wish to issue guidelines that encourage the adoption of robust, language-inclusive sentiment analysis tools. These recommendations could be incorporated into quality assurance frameworks for online travel agencies (OTAs) and hospitality websites, where user-generated reviews significantly influence consumer perceptions. Additionally, platform policies might specify minimum standards for transparency regarding which linguistic or emotional cues trigger review filtering or moderation. Such guidelines can promote fairness and inclusivity by avoiding the systematic removal of certain dialects, minority languages, or critical viewpoints. More broadly, these insights can help shape responsible data governance strategies, motivating the tourism industry to respect user privacy while simultaneously enabling advanced analytics that enrich the traveler experience. By combining sound technical approaches with thoughtful policy measures, the tourism sector can continue evolving toward more authentic and reliable user interactions.

CONCLUSION

This study set out to address the issue of irrelevancy within multilingual tourism reviews, examining how off-topic or ambiguous user-generated content can reduce the overall reliability of online platforms. By crafting a refined dataset of both relevant and irrelevant reviews, and by leveraging a novel combination of multilingual embeddings, gating mechanisms, and sentiment vectors from a knowledge-distilled DistilBERT model, we have demonstrated that explicit attention to emotional cues offers tangible benefits for classification tasks. Not only does sentiment information help to disambiguate user attitudes in otherwise unclear statements, but it also provides additional context for identifying when reviews deviate from the core topic.

In our comparisons, the Baseline model that relied solely on multilingual embeddings performed moderately well, and a Monolingual BERT model focusing on English showed even lower performance when faced with a diverse set of languages. Incorporating sentiment signals, however, improved results on standard metrics, including accuracy, precision, recall, and F1-score. These findings highlight the importance of acknowledging emotional tone as a direct influencer of how travelers convey their experiences. Subtle positivity, negativity, or neutrality in a review can play a key role in clarifying user intent-even when the textual content is limited, code-switched, or contains cultural references unfamiliar to generic natural language processing models.

By demonstrating that sentiment-enriched approaches can filter out or at least flag off-topic postings more effectively, our methodology has practical implications for tourism-related businesses, policy-makers, and platform administrators. Through the lens of a refined, domain-specific dataset, our results underscore the need for ongoing refinement of multilingual text analytics—particularly in fields where user expression is both regionally and linguistically diverse. Future research might explore deeper integration of cultural nuances, additional low-resource languages, or new forms of social media data to further enhance the accuracy of irrelevancy and ambiguity detection. In doing so, the broader tourism ecosystem will benefit from more trustworthy and comprehensible online feedback.

REFERENCES

[1] B. Thompson, S. G. Roberts, and G. Lupyan, ‘Cultural influences on word meanings revealed through large-scale semantic alignment’, *Nature Human Behaviour*, vol. 4, no. 10, pp. 1029–1038, 2020.

[2] Y. Chen *et al.*, ‘Cross-modal Ambiguity Learning for Multimodal Fake News Detection’, in *Proceedings of the ACM Web Conference 2022*, Virtual Event, Lyon France: ACM, Apr. 2022, pp. 2897–2905. doi: 10.1145/3485447.3511968.

[3] A. Ferrari and A. Esuli, ‘An NLP approach for cross-domain ambiguity detection in requirements engineering’, *Autom Softw Eng*, vol. 26, no. 3, pp. 559–598, Sep. 2019, doi: 10.1007/s10515-019-00261-7.

[4] F. Peng, X. Wu, Y. Zhao, and Y. Li, ‘Anaphora Ambiguity Detection Method Based on Cross-domain Pronoun Substitution (S).’, in *SEKE*, 2023, pp. 646–649. Accessed: Jan. 17, 2025. [Online]. Available: https://ksiresearch.org/seke/seke23paper/paper173.pdf

[5] F. Pittke, H. Leopold, and J. Mendling, 'Automatic detection and resolution of lexical ambiguity in process models', *IEEE Transactions on Software Engineering*, vol. 41, no. 6, pp. 526–544, 2015.
[6] M. Figlerowicz and M. Figlerowicz, 'Multilingual style', *Textual Practice*, vol. 35, no. 6, pp. 1015–1036, Jun. 2021, doi: 10.1080/0950236X.2021.1936760.
[7] S. Seo, C. Kim, H. Kim, K. Mo, and P. Kang, 'Comparative study of deep learning-based sentiment classification', *IEEE Access*, vol. 8, pp. 6861–6875, 2020.
[8] O. Araque, I. Corcuera-Platas, J. F. Sánchez-Rada, and C. A. Iglesias, 'Enhancing deep learning sentiment analysis with ensemble techniques in social applications', *Expert Systems with Applications*, vol. 77, pp. 236–246, 2017.
[9] Z. Gao, A. Feng, X. Song, and X. Wu, 'Target-dependent sentiment classification with BERT', *Ieee Access*, vol. 7, pp. 154290–154299, 2019.
[10] J. Yu and J. Jiang, 'Adapting BERT for target-oriented multimodal sentiment classification', IJCAI, 2019. Accessed: Feb. 01, 2025. [Online]. Available: https://ink.library.smu.edu.sg/sis_research/4441/
[11] M. Hu and B. Liu, 'Mining and summarizing customer reviews', in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, Seattle WA USA: ACM, Aug. 2004, pp. 168–177. doi: 10.1145/1014052.1014073.
[12] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, 'Finding Deceptive Opinion Spam by Any Stretch of the Imagination', Jul. 22, 2011, *arXiv*: arXiv:1107.4557. doi: 10.48550/arXiv.1107.4557.
[13] Z. Xiang, Z. Schwartz, J. H. Gerdes Jr, and M. Uysal, 'What can big data and text analytics tell us about hotel guest experience and satisfaction?', *International journal of hospitality management*, vol. 44, pp. 120–130, 2015.
[14] S. M. Kumar, N. Reddy, A. Malapati, and L. Kumar, 'An Ensemble Model for Sentiment Classification on Code-Mixed Data in Dravidian Languages.', in *FIRE (Working Notes)*, 2021, pp. 1085–1093. Accessed: Feb. 01, 2025. [Online]. Available: https://easychair.org/publications/preprint_download/sKB5
[15] X. Chen, Y. Sun, B. Athiwaratkun, C. Cardie, and K. Weinberger, 'Adversarial deep averaging networks for cross-lingual sentiment classification', *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 557–570, 2018.
[16] X.-Y. Zhang, S. Wang, and X. Yun, 'Bidirectional active learning: A two-way exploration into unlabeled and labeled data set', *IEEE transactions on neural networks and learning systems*, vol. 26, no. 12, pp. 3034–3044, 2015.
[17] H. Xiao and L. Luo, 'An Automatic Sentiment Analysis Method for Short Texts Based on Transformer-BERT Hybrid Model', *IEEE Access*, 2024, Accessed: Feb. 01, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10580959/
[18] M. S. Viñán-Ludeña and L. M. De Campos, 'Discovering a tourism destination with social media data: BERT-based sentiment analysis', *JHTT*, vol. 13, no. 5, pp. 907–921, Nov. 2022, doi: 10.1108/JHTT-09-2021-0259.
[19] W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth, 'Cursing in English on twitter', in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, Baltimore Maryland USA: ACM, Feb. 2014, pp. 415–425. doi: 10.1145/2531602.2531734.
[20] Y. Zhu, W. Zheng, and H. Tang, 'Interactive Dual Attention Network for Text Sentiment Classification', *Computational Intelligence and Neuroscience*, vol. 2020, pp. 1–11, Nov. 2020, doi: 10.1155/2020/8858717.
[21] S. Dong and C. Liu, 'Sentiment Classification for Financial Texts Based on Deep Learning', *Computational Intelligence and Neuroscience*, vol. 2021, no. 1, p. 9524705, Jan. 2021, doi: 10.1155/2021/9524705.
[22] L. Khan, A. Amjad, N. Ashraf, and H.-T. Chang, 'Multi-class sentiment analysis of urdu text using multilingual BERT', *Scientific Reports*, vol. 12, no. 1, p. 5436, 2022.
[23] K. R. Mabokela, T. Celik, and M. Raborife, 'Multilingual sentiment analysis for under-resourced languages: a systematic review of the landscape', *IEEE Access*, vol. 11, pp. 15996–16020, 2022.
[24] M. Pota, M. Ventura, H. Fujita, and M. Esposito, 'Multilingual evaluation of pre-processing for BERT-based sentiment analysis of tweets', *Expert Systems with Applications*, vol. 181, p. 115119, 2021.
[25] S. K. Akpatsa *et al.*, 'Online News Sentiment Classification Using DistilBERT.', *Journal of Quantum Computing*, vol. 4, no. 1, 2022, Accessed: Feb. 05, 2025. [Online]. Available: https://cdn.techscience.cn/ueditor/files/jqc/TSP_JQC-4-1/TSP_JQC_26658/TSP_JQC_26658.pdf

[26] V. Dogra, A. Singh, S. Verma, Kavita, N. Z. Jhanjhi, and M. N. Talib, 'Analyzing DistilBERT for Sentiment Classification of Banking Financial News', in *Intelligent Computing and Innovation on Data Science*, vol. 248, S.-L. Peng, S.-Y. Hsieh, S. Gopalakrishnan, and B. Duraisamy, Eds., in Lecture Notes in Networks and Systems, vol. 248. , Singapore: Springer Nature Singapore, 2021, pp. 501–510. doi: 10.1007/978-981-16-3153-5_53.

[27] M. Jojoa, P. Eftekhar, B. Nowrouzi-Kia, and B. Garcia-Zapirain, 'Natural language processing analysis applied to COVID-19 open-text opinions using a distilBERT model for sentiment categorization', *AI & Soc*, vol. 39, no. 3, pp. 883–890, Jun. 2024, doi: 10.1007/s00146-022-01594-w.

[28] S. Ruder, I. Vulić, and A. Søgaard, 'A survey of cross-lingual word embedding models', *Journal of Artificial Intelligence Research*, vol. 65, pp. 569–631, 2019.

[29] Z. Xiang and U. Gretzel, 'Role of social media in online travel information search', *Tourism management*, vol. 31, no. 2, pp. 179–188, 2010.

[30] A. Abdi, S. M. Shamsuddin, S. Hasan, and J. Piran, 'Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion', *Information Processing & Management*, vol. 56, no. 4, pp. 1245–1259, 2019.

[31] N. Aldunate, M. Villena-González, F. Rojas-Thomas, V. López, and C. A. Bosman, 'Mood detection in ambiguous messages: the interaction between text and emoticons', *Frontiers in psychology*, vol. 9, p. 423, 2018.

[32] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, 'DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter', Mar. 01, 2020, *arXiv*: arXiv:1910.01108. doi: 10.48550/arXiv.1910.01108.