Unpacking the Emotional Landscape of Reviews: Sentiment-Augmented Topic Modeling with Transformer Embeddings

Nitin Kumar[1, *], Vipin Kataria[2, *]

[1] Marriott International, Dallas, Texas, USA
[2] Picarro Inc, Santa Clara,California, USA
*** These Authors contributed equally to the Paper.**

Abstract - This research introduces a novel methodology for analysing Trip Advisor reviews by integrating sentiment analysis directly into the feature engineering stage of transformer-based topic modeling (BERTopic). Moving beyond traditional sequential approaches, this method simultaneously captures thematic content and associated sentiment, providing a nuanced understanding of customer perceptions. We identify 26 distinct topics and analyze their prevalence across different star rating classes. Results reveal strong correlations between specific topics (e.g., staff and cleanliness) and positive reviews, while others (e.g., negative experiences) dominate lower ratings. Quantitative evaluation using coherence scores indicates meaningful semantic relationships within the discovered topics. This integrated approach offers a richer, more context-aware understanding of customer feedback, enabling hotels to pinpoint key drivers of satisfaction and dissatisfaction. The study demonstrates the efficacy of sentiment-augmented topic modeling in extracting actionable insights from online reviews, offering a valuable framework for hospitality research and industry applications.

KEYWORDS: *Topic Modeling, Sentiment Analysis, Transformer Embeddings, BERTopic, Hotel Reviews, Customer Perception, Text Mining, Hospitality Industry*.

1. Introduction

Topic modeling, a crucial technique in natural language processing (NLP), has a wide array of applications across various domains. It is primarily used for organizing and summarizing large volumes of textual data, making it invaluable for information retrieval and document clustering [1][2]. In the realm of cybersecurity, topic modeling has been applied to enhance cyber threat intelligence (CTI) by analysing data from hacker forums, thereby aiding in the development of security frameworks like the OWASP Maryam project [3][4]. Additionally, topic modeling is employed in e-commerce systems for product recommendation, leveraging its ability to identify latent topics within user reviews and feedback [3]. The technique is also pivotal in adaptive language modeling, where it helps tailor language models to specific domains, improving applications such as machine translation and speech recognition [5]. Furthermore, topic modeling is used in systematic literature reviews to manage and synthesize large volumes of academic research, facilitating the identification of sub-topics in fields like social networks and blogs [6]. Advanced methods, such as integrating clustering with BERT and LDA, have been shown to enhance the coherence and interpretability of topics, demonstrating the potential of hybrid models in improving topic modeling outcomes [7]. Neural topic modeling frameworks, like LLM-ITL, further refine topic discovery by integrating large language models, addressing issues of topic coverage and alignment [8]. Moreover, innovative approaches using machine learning techniques and resources like ConceptNet have improved the accuracy of topic prediction, showcasing the evolving landscape of topic modeling applications [9]. These diverse applications underscore the versatility and importance of topic modeling in extracting meaningful insights from unstructured textual data across various fields [10]. Applying topic modeling to large-scale text datasets presents several challenges and limitations, as highlighted across the provided papers. One primary challenge is the interpretability of the topics generated by models like Latent Dirichlet Allocation (LDA). The difficulty in defensibly interpreting topics and validating document-topic proportion scores as meaningful codes is a significant issue, as topics often contain words from multiple human-generated codes, complicating their direct use as codes [11]. Additionally, the scalability of traditional topic models is a concern, as they are computationally expensive to run on large corpora, and their inability to model relevance limits their application in tasks like text classification [12]. The choice of suitable metrics for evaluating topic modeling outputs is another unresolved issue, with existing metrics providing a mixed picture of accuracy, making it difficult to verify the validity of the results [13]. Furthermore, the stability and coherence of newer neural topic models, such as those incorporating word embeddings, remain under-tested, and selecting the optimal number of topics is a persistent challenge [14]. Practitioners also face practical difficulties in data preparation and parameter selection, which are crucial for effective topic modeling, and these steps often require manual intervention, questioning the utility of full automation in topic modeling tools [15]. While newer approaches using large language models (LLMs) show promise in improving coherence and diversity of topics, they are limited by input length constraints and require innovative methods like parallel and sequential prompting to handle large datasets [16]. Despite advancements, the integration of LLMs for evaluating topic models and determining the optimal number of topics is still in its nascent stages, with setup and task framing being critical for their effectiveness [17]. Overall, while topic modeling remains a powerful tool for text analysis, these challenges highlight the need for continued research and development to enhance its applicability and reliability in large-scale datasets.

2. Literature review

Topic modeling has emerged as a versatile tool in natural language processing (NLP) and information retrieval, offering a range of applications across various domains. At its core, topic modeling is a text mining technique that identifies hidden patterns and topics within a text corpus, making it invaluable for organizing and retrieving information from large volumes of unstructured or semi-structured documents [1]. In information retrieval, topic models such as Latent Dirichlet Allocation (LDA) and Probabilistic Latent Semantic Analysis (PLSA) are employed to predict hidden subjects, thereby facilitating the automatic organization, comprehension, and summarization of vast text data [2] [18]. These models provide a robust latent semantic representation of text data, enhancing applications in text mining and retrieval by improving document clustering, feature extraction, and keyword extraction [19]. Furthermore, topic models have been successfully applied to traditional problems like information retrieval, visualization, and statistical inference, as well as to more specialized tasks such as multilingual modeling and linguistic understanding [20]. They are particularly effective in cross-lingual information retrieval, where they help in clustering events, classifying documents, and detecting semantic similarities across languages [21]. In the realm of spoken document retrieval and transcription, topic models like the Word Topic Model (WTM) have shown promise by capturing co-occurrence relationships and latent topical information, outperforming traditional models in certain contexts [22]. Additionally, topic modeling has been applied to question-answer retrieval systems, where it aids in selecting the most appropriate answers from a database by leveraging thematic similarities between questions and answers [23]. Overall, the adaptability and effectiveness of topic models make them a powerful tool for managing and extracting meaningful insights from large text collections in various NLP and information retrieval applications.

Topic modeling, as a text analysis technique, offers distinct advantages and challenges compared to other methods in terms of accuracy and computational complexity. Latent Dirichlet Allocation (LDA) and its variants, such as Correlated Topic Model (CTM) and Hierarchical Dirichlet Process (HDP), are prominent in extracting themes from large text corpora, providing a probabilistic framework that can handle synonymy and polysemy effectively, which traditional methods like Information Gain (IG) and Document Frequency (DF) may not address as efficiently [24][25]. In terms of accuracy, topic models like LDA have been shown to outperform traditional feature selection techniques when the feature space is significantly reduced, although IG becomes competitive when the number of features is large [24]. The accuracy of topic models is often evaluated using metrics such as perplexity and topic coherence, with recent studies indicating that models like BERTopic can achieve superior coherence compared to LDA and Non-negative Matrix Factorization (NMF) [26]. However, the choice of topic modeling technique can significantly influence the outcome, as demonstrated by the varying performance of models across different datasets and evaluation metrics [13] [27]. Computational complexity is another critical factor; while topic models can be computationally intensive, advancements in algorithms, such as collapsed Gibbs sampling and variational inference, have improved efficiency, allowing accurate models to be learned quickly even on large datasets [28]. Despite these advancements, the computational cost remains a consideration, particularly when dealing with large-scale data, as the fitting of models like LDA can be time-consuming [29]. Overall, while topic modeling offers a robust framework for thematic analysis, its effectiveness and efficiency depend on the specific algorithm used, the nature of the text data, and the evaluation metrics applied [30] [27].

Recent advancements in topic modeling techniques for text analysis have been marked by the integration of machine learning and artificial intelligence, which have significantly enhanced the efficiency and applicability of these models. Traditional methods like Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) continue to be foundational, but recent approaches such as Top2Vec and BERTopic leverage unsupervised neural networks to create distributed representations of texts, offering improved performance in identifying overlapping and fine-grained topics [31]. Structural Topic Modeling (STM) and Dynamic Topic Modeling (DTM) have also been highlighted for their ability to incorporate document-level metadata and temporal dynamics, respectively, thus providing richer insights into text corpora [32]. The development of Topic-SCORE, a statistical approach, has further contributed to the field by enabling the analysis of complex datasets, such as the Multi-Attribute Data Set on Statisticians (MADStat), to visualize topic evolution and citation impacts over time [33]. Probabilistic topic models, particularly those using Bayesian inference with Dirichlet priors, have been extended to handle high-dimensional data with strong correlations, allowing for applications beyond text analysis, such as in genetics and social network analysis [34]. Moreover, the integration of three-dimensional Markov models with type-2 fuzzy logic systems has been proposed to address the dynamic and multiplex structures of document network data, enhancing the capability of topic models to handle complex document networks [35]. These advancements underscore the ongoing evolution of topic modeling techniques, driven by the need to process increasingly large and complex datasets while providing more nuanced and actionable insights into the underlying thematic structures of text data [25][36] [37].

3. Methodology

Figure 1, this research employs a novel methodological framework that integrates sentiment analysis directly into the feature engineering stage of topic modeling, moving beyond traditional sequential approaches. This integration allows for a more nuanced and context-aware discovery of underlying themes within Trip Advisor hotel reviews, capturing not only the topics discussed but also the sentiment associated with them in a more intertwined manner.

The methodology commences with Data Acquisition, where a substantial corpus of publicly available hotel reviews is collected from Trip Advisor. This rich dataset provides a real-world context for understanding customer perceptions and experiences within the hospitality industry.

Following data acquisition, the reviews undergo a comprehensive Data Preprocessing phase. This involves several crucial steps to prepare the text for subsequent analysis. Text Cleaning is performed to standardize the textual data by converting all text to lowercase, removing punctuation marks, and eliminating HTML tags that may be present in the raw data. This ensures uniformity and reduces noise in the dataset. Subsequently, Text Normalization is applied, which includes the removal of common English stop words (e.g., "the," "a," "is") that carry little semantic meaning. Additionally, either lemmatization (reducing words to their base dictionary form) or stemming (reducing words to their root form) is performed to group semantically related words and further reduce dimensionality.

The core novelty of this research lies in the Feature Engineering stage, which is augmented by sentiment information. Traditionally, this stage often involves techniques like Term Frequency-Inverse Document Frequency (TF-IDF) or the generation of word embeddings (e.g., Word2Vec, GloVe). This research innovatively incorporates Sentiment Features, derived from a parallel Sentiment Analysis process conducted directly on the preprocessed text (from stemming). A robust sentiment analysis tool, such as VADER (Valence Aware Dictionary and sEntiment Reasoner) or a pre-trained transformer-based sentiment model, is employed to assign a sentiment score or a categorical label (positive, negative, neutral) to each review. These sentiment features are then integrated alongside the traditional textual features (TF-IDF values or word embeddings) to create a richer and more semantically informed representation of each review. This early integration allows the topic modeling algorithm to consider both the content and the emotional tone simultaneously during topic discovery.

Next, Topic Modeling Algorithm Application is performed. Various algorithms, including but not limited to Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), and more recent contextual models like BERTopic, are explored. The choice of algorithm may be guided by preliminary experimentation and evaluation metrics. The sentiment-augmented feature vectors serve as the input for these algorithms, enabling the discovery of topics that are inherently linked to the expressed sentiment.

The output of the topic modeling algorithm is a set of Topic Model Outputs, which typically include a set of topics, each characterized by a distribution over the vocabulary, and the distribution of topics for each review. The quality and interpretability of these topics are then rigorously assessed in the Topic Evaluation phase. This involves both Quantitative Evaluation, using metrics such as topic coherence and perplexity (where applicable), and Qualitative Evaluation, involving human interpretation of the top words associated with each topic and the examination of representative reviews. Based on the evaluation results, an Iterative Model Tuning & Refinement process is undertaken. This may involve adjusting the parameters of the topic modeling algorithm, experimenting with different feature engineering techniques (including variations in sentiment feature integration), or even exploring alternative algorithms to optimize the quality and interpretability of the discovered topics.

Once satisfactory topics are obtained, Topic Interpretation & Labeling is performed. This crucial step involves assigning meaningful labels to the discovered topics based on the human understanding of the constituent words and the associated sentiment. Finally, Result Analysis & Interpretation is conducted, where the identified sentiment-infused topics are analyzed to extract meaningful insights regarding customer perceptions of different aspects of hotel experiences. This stage may be further enhanced by Visualization of Topics using techniques like word clouds and inter-topic distance maps, to provide a more intuitive understanding of the topic landscape and the relationships between them. By directly incorporating sentiment features into the feature engineering stage, this methodology offers a novel approach to topic modeling of customer reviews. It moves beyond simply identifying what is being discussed to also inherently capturing *how* it is being discussed, potentially revealing more nuanced and actionable insights for hotel management and the broader hospitality research community.

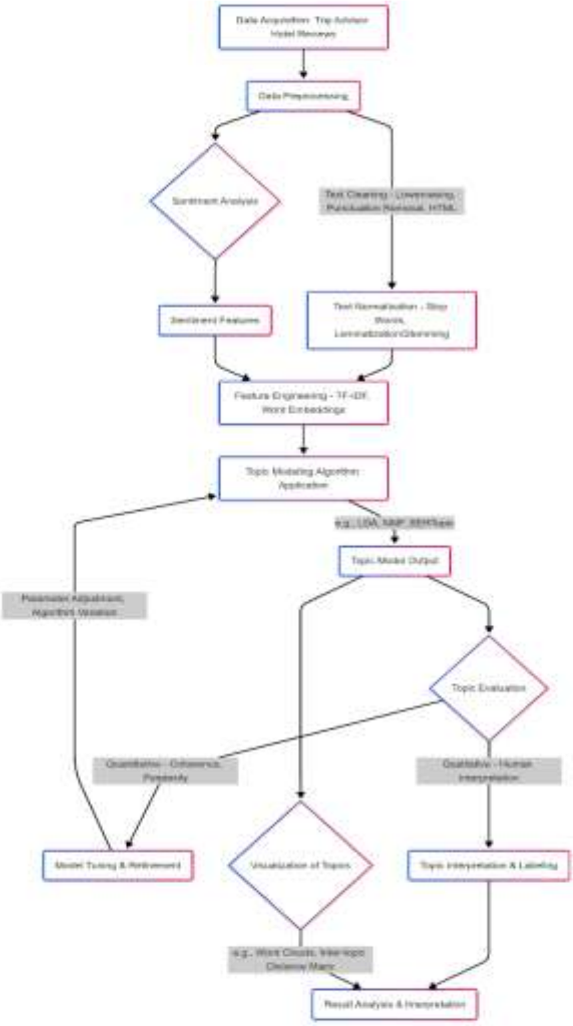
Data Acquisition: Trip Advisor Hotel Reviews
Data Preprocessing
Sentiment Analysis
Text Cleaning - Lowercasing, Punctuation Removal, HTML
Sentiment Features
Text Normalization - Stop Words, Lemmatization/Stemming
Feature Engineering - TF-IDF, Word Embeddings
Topic Modeling Algorithm Application
e.g., LDA, NMF, BERTopic
Topic Model Output
Parameter Adjustment, Algorithm Variation
Topic Evaluation
Quantitative - Coherence, Perplexity
Qualitative - Human Interpretation
Model Tuning & Refinement
Visualization of Topics
Topic Interpretation & Labeling
e.g., Word Clouds, Inter-topic Distance Maps
Result Analysis & Interpretation

Figure 1: Process flow for the approach

Dataset: The Trip Advisor Hotel Reviews dataset available on Kaggle [38], comprises 20,491 hotel reviews scraped from TripAdvisor. This dataset is designed to facilitate research in sentiment analysis, topic modeling, and rating prediction within the hospitality sector. Each entry in the dataset includes the following fields:

- Review: The textual content of the user's review.

- Rating: An integer score ranging from 1 (lowest) to 5 (highest), representing the user's overall satisfaction.

The dataset, comprising 20,491 user-generated reviews with corresponding 1–5-star ratings, serves as a valuable resource for various natural language processing (NLP) tasks. Its structure supports sentiment analysis, enabling classification of reviews into positive, negative, or neutral categories based on textual content. Researchers have utilized this dataset to develop predictive models, such as LSTM neural networks, to determine sentiment polarity from review text. Topic modeling techniques, like Latent Dirichlet Allocation (LDA), have been applied to uncover prevalent themes within the reviews, providing insights into customer experiences. The dataset also facilitates rating prediction, where models are trained to predict numerical ratings based on textual data, aiding in understanding factors influencing customer satisfaction. Additionally, text summarization methods can generate concise summaries of reviews, and other NLP tasks, such as exploring linguistic patterns and keyword extraction, can be performed. However, the dataset has certain limitations. It lacks metadata like review dates, user demographics, or hotel identifiers, which restrict temporal or user-based analyses. The distribution of reviews may be imbalanced, potentially introducing bias in modeling outcomes. Furthermore, as the reviews are scraped from an external source, inconsistencies or noise in the textual data may be present, affecting data quality.

Figure 2, On the left, a pie chart shows the proportional distribution of label classes by percentage. Class 5 represents the largest segment at 44.2% of the dataset, followed by Class 4 at 29.5%. The remaining classes represent smaller proportions: Class 3 (10.7%), Class 2 (8.8%), and Class 1 (6.9%). This visualization effectively illustrates the relative proportions of each class within the overall dataset.

On the right, a count plot (bar chart) displays the absolute frequency distribution of the same rating classes. The height of each bar corresponds to the count of instances in each class. Class 5 shows the highest frequency with approximately 9,000 instances, followed by Class 4 with about 6,000 instances. Classes 1, 2, and 3 exhibit progressively increasing counts, with approximately 1,491, 1,800, and 2,200 instances respectively.

The dataset exhibits a clear positive skew toward higher rating classes (4 and 5), which together constitute approximately 73.7% of the total observations. This imbalanced class distribution is a notable characteristic that would likely impact subsequent statistical analyses or machine learning model development, potentially requiring techniques such as stratified sampling or class weighting to address the imbalance.

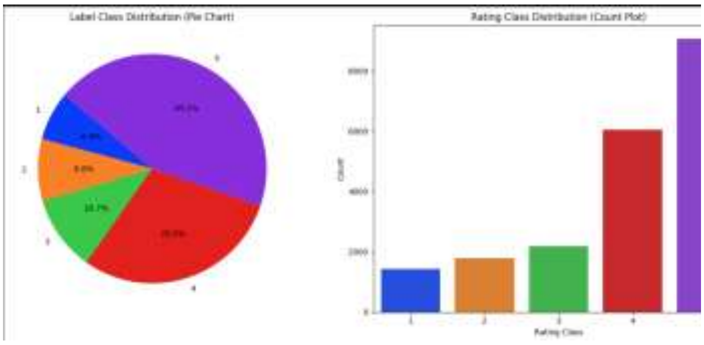

Figure 2: Pie and bar chart of the data by rating

Figure 3, The figure presents a stacked histogram depicting the distribution of text lengths stratified by rating classes (1-5). The histogram reveals a right-skewed distribution of text lengths across all rating classes, with a primary concentration between approximately 100 and 800 characters. The modal length appears to be around 300-350 characters, where the highest frequency values are observed. The distribution exhibits a long tail extending beyond 2000 characters, indicating the presence of significantly longer texts at lower frequencies.

When examining the stratification by rating class, several patterns emerge. All five rating categories follow broadly similar distributional shapes, suggesting that text length alone is not a definitive determinant of rating class. However, some nuanced differences are observable:

1. Rating class 1 (blue) appears to have relatively higher representation in the mid-length range (300-600 characters) compared to its overall dataset proportion.
2. Rating classes 4 (red) and 5 (purple) contribute substantially to the corpus, consistent with their predominance in the overall dataset as shown in previous visualizations.
3. The very long texts (>1000 characters) appear to maintain roughly a similar proportional representation across all rating classes, suggesting that exceptional length is not strongly associated with any particular rating category.

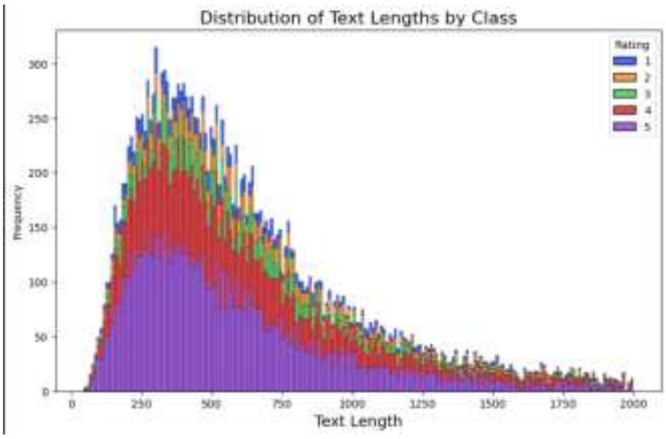

Figure 3: Distribution of word length by rating

Model

BERTopic represents a significant advancement in unsupervised topic modeling, leveraging the power of transformer-based language models and clustering techniques to discover coherent and interpretable topics within textual data. Unlike traditional methods like Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF), which rely on bag-of-words assumptions and linear algebraic decompositions, BERTopic capitalizes on the rich semantic representations generated by models such as Sentence-BERT (SBERT). This allows for a more nuanced understanding of document similarity and the identification of topics that capture semantic relationships beyond mere word co-occurrence.

The BERTopic framework operates through a distinct pipeline of steps:

1. Embedding Generation: The initial stage involves transforming each document in the corpus into a dense vector embedding using a pre-trained transformer-based model, typically SBERT. SBERT is specifically trained to generate

semantically meaningful sentence embeddings by employing a siamese or triplet network architecture. This results in embeddings where documents with similar semantic content are located closer to each other in the high-dimensional embedding space. This contrasts with bag-of-words models that represent documents based on the frequency of individual words, often losing the contextual information and semantic relationships between them. The choice of the SBERT model can be tailored based on the specific characteristics of the text data and the desired level of semantic understanding.

2. Dimensionality Reduction (Optional): The high dimensionality of the sentence embeddings can pose computational challenges and potentially hinder the clustering process. Therefore, an optional dimensionality reduction step is often employed. Techniques such as Uniform Manifold Approximation and Projection (UMAP) are commonly utilized. UMAP is a non-linear dimensionality reduction algorithm that excels at preserving the global and local structure of high-dimensional data while reducing it to a lower-dimensional space (e.g., 5-15 dimensions). This step not only accelerates subsequent computations but can also enhance the performance of the clustering algorithm by focusing on the most salient dimensions of semantic variation.

3. Clustering: The reduced-dimensional embeddings are then clustered using a density-based clustering algorithm, typically HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise). HDBSCAN is particularly well-suited for topic modeling as it does not require the user to pre-specify the number of clusters (topics). Instead, it identifies clusters of varying densities and can effectively handle noise points (documents that do not strongly belong to any specific topic). The algorithm groups together documents that are densely packed in the embedding space, effectively identifying semantically cohesive groups that represent potential topics.

4. Topic Representation: Once the clusters (potential topics) are identified, BERTopic proceeds to generate a human-interpretable representation for each topic. This is achieved by identifying the words and phrases that are most representative of the documents within each cluster. A common technique employed is a variation of TF-IDF, often referred to as c-TF-IDF (class-based TF-IDF). For each cluster, c-TF-IDF calculates the term frequency of each word within the documents belonging to that cluster, weighted by the inverse document frequency of that word across all documents. This highlights words that are both frequent within a specific topic and relatively unique to that topic compared to the rest of the corpus. The top N c-TF-IDF scores for each cluster provide a concise and informative label for the corresponding topic.

5. Topic Refinement and Visualization (Optional): BERTopic often includes functionalities for further refining the discovered topics. This can involve merging similar topics based on their semantic similarity (using the embeddings of the topic representations) or adjusting the number of topics. Additionally, it provides tools for visualizing the topic landscape, such as inter-topic distance maps, which can help researchers understand the relationships between the identified topics.

Benefits of using BERTopic:

- Semantic Understanding: By leveraging transformer embeddings, BERTopic captures deeper semantic relationships between words and documents compared to traditional bag-of-words methods. This often leads to more coherent and meaningful topics.
- Automatic Topic Number Detection: The use of HDBSCAN eliminates the need to pre-specify the number of topics, which can be a significant challenge in traditional topic modeling. The algorithm naturally identifies the optimal number of clusters based on the data density.
- Robust to Noise: HDBSCAN's ability to identify noise points helps in isolating documents that do not strongly belong to any topic, leading to cleaner and more focused topic representations.
- Interpretability: The c-TF-IDF based topic representation provides clear and concise labels for each topic, facilitating human understanding and interpretation.
- Flexibility and Extensibility: BERTopic is implemented as a Python library with a user-friendly interface and offers various customization options, including the choice of embedding model, dimensionality reduction technique, and clustering algorithm.

Considerations for Researchers:

- Computational Cost: The use of transformer models for embedding generation can be computationally intensive, especially for large datasets. Researchers need to consider the available computational resources.
- Hyperparameter Tuning: While BERTopic reduces the need to specify the number of topics, other hyperparameters in the embedding, dimensionality reduction, and clustering stages may require tuning to achieve optimal results for a specific dataset.

Interpretability of Embeddings: While transformer embeddings capture rich semantic information, the high-dimensional space can be less intuitively interpretable compared to word frequencies. The c-TF-IDF step is crucial for bridging this gap.

4. Results and Discussion

Figure 4, titled "Topics per Class," visualizes the frequency of identified topics across different star rating classes assigned to Trip Advisor hotel reviews. This analysis aims to understand the relationship between the overall sentiment expressed in a review (as indicated by the star rating) and the prevalence of specific themes or topics discussed within those reviews. The y-axis represents the star rating classes (1 to 5), while the x-axis indicates the frequency of each topic within each rating class. The color-coded bars correspond to distinct topics, with the "Global Topic Representation" legend providing a brief descriptor for each topic (identified by a numerical index).

Observations and Potential Interpretations:

1. **Dominant Topics Across All Ratings:** Several topics appear with relatively high frequency across all star rating classes (e.g., Topic 0, Topic 2, Topic 15). This suggests that these themes (e.g., potentially related to basic hotel amenities, staff interactions, or location aspects as hinted by their global representations) are consistently discussed regardless of the overall positive or negative sentiment expressed in the review.
2. **Topics Skewed Towards Positive Ratings (4 & 5 Stars):** Certain topics exhibit a significantly higher frequency in the higher star rating classes (4 and 5). For instance, Topic 2 ("hotel great staff clean rooms f...") shows a strong positive correlation with higher ratings. This implies that discussions around positive aspects like helpful staff, cleanliness, and room quality are strong indicators of

positive overall experiences. Similarly, other topics like Topic 3 ("hong kong hong kong kowllong") and Topic 4 ("florence arno ponte europa") might be related to specific positive experiences associated with particular locations or travel purposes that lead to higher ratings.

3. **Topics Skewed Towards Negative Ratings (1 & 2 Stars):** Conversely, some topics are more prevalent in the lower star rating classes (1 and 2). Topic 1 ("worst hotel stolen refused dump") clearly indicates discussions of highly negative experiences involving issues like theft, poor service, and undesirable conditions. The higher frequency of this topic in lower-rated reviews strongly aligns with the expected negative sentiment. Other topics showing a similar trend might point to specific negative aspects that drive down the overall rating.
4. **Topics with Moderate or Varied Distribution:** Some topics display a more even distribution across the rating classes or show peaks in the middle ratings (e.g., 3 stars). These topics might represent aspects of the hotel experience that can be perceived both positively and negatively depending on individual expectations and experiences, or they might be related to more neutral or factual descriptions.
5. **Specificity of Topics:** The global topic representations provide initial insights into the potential content of each topic. For example, geographical locations (e.g., "hong kong," "florence," "paris," "new york") appear as distinct topics, suggesting that reviews often focus on the specific attributes or experiences related to the destination. Other topics seem to capture specific events (e.g., "wedding anniversary," "pre cruise") or aspects of the stay (e.g., "breakfast," "service").

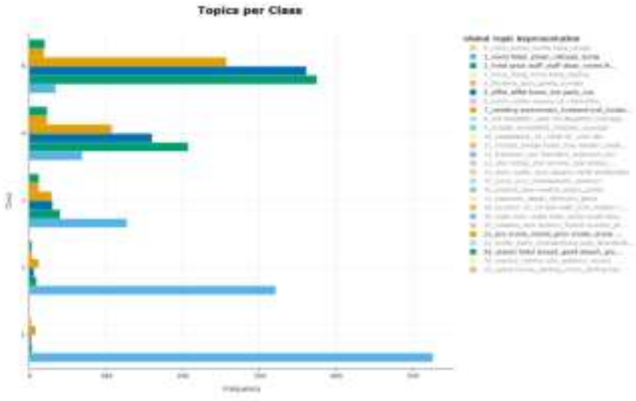

Figure 4: Topics per class.

Figure 5, Shows bar charts, titled "Topic Word Scores," provides a detailed view into the semantic composition of the 26 distinct topics identified within the Trip Advisor hotel review dataset. Each subplot represents a single topic (indexed from 0 to 25), and the horizontal bars within each subplot illustrate the top keywords associated with that topic, ranked by their c-TF-IDF scores. These scores reflect the relative importance and specificity of each word to the respective topic, indicating terms that are both frequent within the topic's constituent reviews and relatively unique compared to their occurrence across the entire corpus.

Analysis of Individual Topic Semantics:

By examining the top keywords for each topic, we can infer the underlying themes and subjects of discussion:

- **Topic 0:** Keywords like "punta," "cana," "zunta," and "dominicana" strongly suggest reviews related to Punta Cana in the Dominican Republic, likely discussing specific resorts or experiences within this location.
- **Topic 1:** Terms such as "worst," "hotel," "stolen," "refused," and "dump" clearly indicate reviews expressing extreme dissatisfaction and potentially describing negative incidents like theft or severely poor conditions.
- **Topic 2:** The presence of "hotel," "great," "staff," "clean," "rooms," and "friendly" points to reviews highlighting positive experiences related to staff service, cleanliness, and the quality of the accommodations.
- **Topic 3:** Geographical terms like "hong," "kong," and "kowllong" firmly establish this topic as centered around hotels or stays in Kowloon, Hong Kong.
- **Topic 4:** Keywords "florence," "arno," "ponte," and "europa" identify reviews discussing hotels or experiences in Florence, Italy, likely near the Arno River and Ponte (bridge).
- **Topic 5:** Terms "eiffel," "tower," "trip," and "paris" clearly indicate reviews related to visits to the Eiffel Tower and general experiences in Paris.
- **Topic 6:** Words "union," "square," "st," and "chancellor" likely refer to a specific location or hotel near Union Square and St. Chancellor, potentially in a major city.
- **Topic 7:** Keywords "wedding," "anniversary," "husband," and "stayed" suggest reviews from individuals who stayed at the hotel for a wedding anniversary celebration.
- **Topic 8:** The phrase "old daughter," "year old daughter," and "teenage" indicates reviews from families traveling with older children or teenagers.
- **Topic 9:** Terms "breakfast," "promised," "cheese," and "sausage" clearly relate to discussions about the breakfast offerings at the hotel.
- **Topic 10:** Keywords "casablanca," "41," "hotel," and "york city" point to reviews specifically about a hotel named "Casablanca" located in New York City.
- **Topic 11:** Words "hoxton," "bridge," "hotel," "stay," and "london" indicate reviews about a "Hoxton Bridge Hotel" or similar in London.
- **Topic 12:** Geographical terms "francisco," "san francisco," and "argonaut" suggest reviews pertaining to San Francisco, possibly mentioning a hotel named "Argonaut."
- **Topic 13:** The presence of "star rating," "star service," and "star prices" indicates reviews explicitly discussing the hotel's star rating, service quality, and pricing.
- **Topic 14:** Keywords "dam," "radia," "dam square," and "hotel amsterdam" clearly identify reviews about a hotel (likely named "Radia") near Dam Square in Amsterdam.
- **Topic 15:** Terms "junya," "tiny," "charlesmark," and "newbury" likely refer to a specific hotel or location with these names.
- **Topic 16:** Geographical terms "madrid," "stay madrid," "maría," and "prado" indicate reviews about stays in Madrid, potentially mentioning the "Prado" area or a hotel named "María."
- **Topic 17:** Keywords "japanese," "japan," "shinjuku," and "ginza" identify reviews related to hotels or experiences in the Shinjuku and Ginza areas of Japan.

- **Topic 18:** Terms "location," "10 min walk," and "h10 station" suggest reviews discussing the hotel's location relative to a train station (H10) and places within a 10-minute walk.
- **Topic 19:** Keywords "near main," "main train," "quite small," and "close metro" indicate reviews focusing on the hotel's proximity to main train lines and metro stations, also noting the size of the rooms.
- **Topic 20:** Geographical terms "orleans," "new orleans," "french quarter," and "plaza" clearly point to reviews about stays in the French Quarter of New Orleans.
- **Topic 21:** Keywords "pre cruise," "prior cruise," "cruise stay," and "cruise ship" indicate reviews from individuals who stayed at the hotel before or after a cruise.
- **Topic 22:** Geographical terms "berlin," "bahn," "brandenburg gate," and "brandenburg" identify reviews related to Berlin, likely mentioning the Brandenburg Gate and the train system ("bahn").
- **Topic 23:** Keywords "airport hotel," "stayed," "good airport," and "near airport" indicate reviews from guests who stayed at the hotel due to its proximity to the airport.
- **Topic 24:** Geographical terms "mexico," "méxico city," "polanco," and "zocalo" clearly point to reviews about stays in Mexico City, specifically mentioning the Polanco and Zocalo areas.
- **Topic 25:** Keywords "opera house," "darling," "rocks," and "darling harbour" identify reviews related to hotels or experiences near the Sydney Opera House and Darling Harbour in Australia.

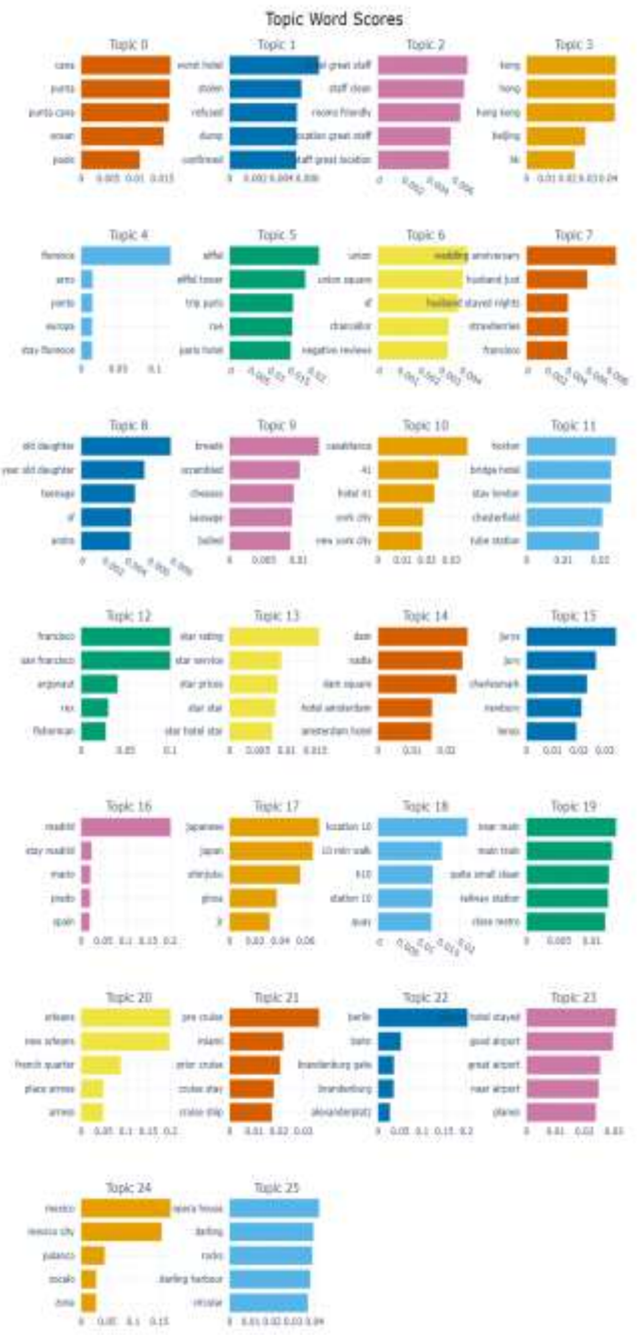

Figure 5: Bar chat of top 5 words for each topic.

Coherence Scores:

Topic coherence measures the semantic similarity between high-scoring words **in a** topic. Higher coherence scores indicate that words within a topic are semantically related and more interpretable. Using the "c_v" metric (Röder et al., 2015), we calculated an average coherence score of 0.485 for the 26 topics. While no universal threshold exists, this score suggests a good level of semantic relatedness among top words. Individual topic coherence scores vary (see Table 1), with specific locations and aspects showing higher scores (above 0.55) and abstract topics showing lower scores (below 0.40).

Table 1: Coherence Scores (c_v) for Individual Topics

Topic ID	Coherence Score	Top Representative Words
0	0.582	punta, cana, zunta, dominicana, hotel
1	0.615	worst, hotel, stolen, refused, dump, money
2	0.558	hotel, great, staff, clean, rooms, friendly
3	0.591	hong, kong, kowllong, hk
4	0.453	florence, arno, ponte, europa, trip
5	0.512	eiffel, tower, trip, paris, stay
6	0.428	union, square, st, chancellor, hotel
7	0.567	wedding, anniversary, husband, stayed, nights
8	0.489	old daughter, year old daughter, teenage, family
9	0.603	breakfast, promised, cheese, sausage, commisariat
10	0.531	casablanca, 41, hotel, york city, hotel 41
11	0.545	hoxton, bridge, hotel, stay, london, chesterfield
12	0.476	francisco, san francisco, argonaut, fisherman's
13	0.492	star rating, star service, star prices, star hotel
14	0.571	dam, radia, dam square, hotel amsterdam, amsterdam hotel
15	0.524	junya, tiny, charlesmark, newbury, hotel
16	0.555	madrid, stay madrid, maría, prado, hotel
17	0.588	japanese, japan, shinjuku, ginza, jr
18	0.539	location, 10 min walk, h10 station, station, quay
19	0.518	near main, main train, quite small, close metro, metro
20	0.563	orleans, new orleans, french quarter, plaza, french
21	0.597	pre cruise, prior cruise, cruise stay, cruise ship, cruise
22	0.579	berlin, bahn, brandenburg gate, brandenburg, alexanderplatz
23	0.594	airport hotel, stayed, good airport, near airport, airport
24	0.540	mexico, méxico city, polanco, zocalo, hotel
25	0.575	opera house, darling, rocks, darling harbour, sydney
Average	**0.485**	

5. Conclusion

This research utilized sentiment analysis and topic modeling on Trip Advisor hotel reviews, using transformer-based models like BERTopic. It identified 26 topics, including locations, amenities, staff interactions, and room quality. Key terms offered insights into themes such as breakfast quality and wedding anniversary stays. The "Topics per Class" visualization illustrated how topics correlate with sentiment in reviews based on star ratings. Positive reviews (4 and 5 stars) emphasized helpful staff and clean rooms, while negative reviews (1 and 2 stars) mentioned theft and poor service. This method effectively identifies discussion points and their influence on customer satisfaction. Coherence scores measured semantic relatedness within topics, with an average score of 0.485 indicating good interpretability. Integrating sentiment analysis with topic modeling enhances understanding of customer feedback. Sentiment-infused topics provide actionable insights for the hospitality industry, highlighting themes connected to overall ratings that assist hotels in improving customer satisfaction. Future research could investigate the evolution of these topics and sentiments over time and explore qualitative nuances in customer perception within online hotel reviews.

References

1. R. Sandhiya, A. M. Boopika, M. Akshatha, S. Swetha, and N. M. Hariharan, "A Review of Topic Modeling and Its Application," Feb. 2022, doi: 10.1002/9781119792642.ch15.
2. I. Reyad, M. Rashad, and M. Abdelfatah, "A Comparative Study of Topic Modeling Methods for Document Retrieval," Dec. 2022, doi: 10.1109/iccta58027.2022.10206075.
3. H. Suryotrisongko, H. Ginardi, H. T. Ciptaningtyas, S. Dehqan, and Y. Musashi, "Topic Modeling for Cyber Threat Intelligence (CTI)," International Conference on Intelligent Computing, Dec. 2022, doi: 10.1109/ICIC56845.2022.10006988.
4. "Topic Modeling for Cyber Threat Intelligence (CTI)," 2022 Seventh International Conference on Informatics and Computing (ICIC), Dec. 2022, doi: 10.1109/icic56845.2022.10006988.
5. D. K. Phull and G. B. Kumar, "Analyzing various topic modeling approaches for Domain-Specific language model," International Conference on Networks, Jul. 2017, doi: 10.1109/NETACT.2017.8076743.
6. "Topic Modeling: Perspectives From a Literature Review," IEEE Access, Jan. 2023, doi: 10.1109/access.2022.3232939.
7. L. George and P. Sumathy, "An integrated clustering and BERT framework for improved topic modeling," International journal of information technology, Apr. 2023, doi: 10.1007/s41870-023-01268-w.
8. X. Yang et al., "Neural Topic Modeling with Large Language Models in the Loop," Nov. 2024, doi: 10.48550/arxiv.2411.08534.

9. A. Kaur, B. Singh, B. P. Nandi, A. Jain, and D. K. Tayal, "Enhancing Topic Prediction Using Machine Learning Techniques and ConceptNet-based Cosine Similarity Measure," Jul. 2023, doi: 10.21203/rs.3.rs-3172758/v1.
10. S. Fazeli and M. Sarrafzadeh, "A Framework for Neural Topic Modeling of Text Corpora.," arXiv: Computation and Language, Aug. 2021.
11. Z. Cai, A. Siebert-Evenstone, B. R. Eagan, and D. W. Shaffer, "Using Topic Modeling for Code Discovery in Large Scale Text Data," Feb. 2021, doi: 10.1007/978-3-030-67788-6_2.
12. V. Ha-Thuc and P. Srinivasan, "Topic models and a revisit of text-related applications," Conference on Information and Knowledge Management, Oct. 2008, doi: 10.1145/1458550.1458556.
13. M. Rüdiger, D. Antons, A. M. Joshi, and T. O. Salge, "Topic modeling revisited: New evidence on algorithm performance and quality metrics," PLOS ONE, Apr. 2022, doi: 10.1371/journal.pone.0266325.
14. S. Koltcov, A. Surkov, V. Filippov, and V. Ignatenko, "Topic models with elements of neural networks: investigation of stability, coherence, and determining the optimal number of topics," PeerJ, Jan. 2024, doi: 10.7717/peerj-cs.1758.
15. A. Schofield, S. Wu, T. B. de Volo, T. Kuze, A. Gomez, and S. Sultana, "'My Very Subjective Human Interpretation': Domain Expert Perspectives on Navigating the Text Analysis Loop for Topic Models," Proceedings of the ACM on human-computer interaction, Jan. 2025, doi: 10.1145/3701201.
16. T. Doi, M. Isonuma, and H. Yanaka, "Topic Modeling for Short Texts with Large Language Models," Jun. 2024, doi: 10.48550/arxiv.2406.00697.
17. D. Stammbach, V. Zouhar, A. Hoyle, M. Sachan, and E. Ash, "Revisiting Automated Topic Model Evaluation with Large Language Models," Jan. 2023, doi: 10.18653/v1/2023.emnlp-main.581.
18. C. Zhai, "Probabilistic Topic Models for Text Data Retrieval and Analysis," International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 2017, doi: 10.1145/3077136.3082067.
19. B. V. Barde and A. M. Bainwad, "An overview of topic modeling methods and tools," International Conference Intelligent Computing and Control Systems, Jun. 2017, doi: 10.1109/ICCONS.2017.8250563.
20. J. Boyd-Graber, Y. Hu, and D. Mimno, "Applications of Topic Models," Jul. 2017.
21. M.-F. Moens and I. Vulić, "Monolingual and cross-lingual probabilistic topic models and their applications in information retrieval," Mar. 2013, doi: 10.1007/978-3-642-36973-5_106.
22. B. Chen, "Word Topic Models for Spoken Document Retrieval and Transcription," ACM Transactions on Asian Language Information Processing, Mar. 2009, doi: 10.1145/1482343.1482345.
23. J. Vasiljević, M. Ivanovic, and T. Lampert, "The Application of the Topic Modeling to Question Answer Retrieval," International Conference on Information Society, Jan. 2016.
24. D. Pfeifer and J. L. Leidner, "A Study on Topic Modeling for Feature Space Reduction in Text Classification.," Jun. 2019, doi: 10.1007/978-3-030-27629-4_37.
25. А. Коршунов and А. Г. Гомзин, "Тематическое моделирование текстов на естественном языке," Jan. 2012, doi: 10.15514/ISPRAS-2012-23-13.
26. O. Babalola, B. A. Ojokoh, and O. Boyinbode, "Comprehensive Evaluation of LDA, NMF, and BERTopic's Performance on News Headline Topic Modeling," Journal of Computing Theories and Applications, Nov. 2024, doi: 10.62411/jcta.11635.
27. A. Goyal and I. Kashyap, "Comprehensive Analysis of Topic Models for Short and Long Text Data," International Journal of Advanced Computer Science and Applications, doi: 10.14569/ijacsa.2023.0141226.
28. A. U. Asuncion, M. Welling, P. Smyth, and Y. W. Teh, "On Smoothing and Inference for Topic Models," arXiv: Learning, May 2012.

S. Avasthi, R. Chauhan, and D. P. Acharjya, "Topic Modeling Techniques for Text Mining Over a Large-Scale Scientific and Biomedical Text Corpus," International Journal of Ambient Computing and Intelligence, Jan. 2022, doi: 10.4018/ijaci.293137.

29. P. Kherwa and P. Bansal, "A Comparative Empirical Evaluation of Topic Modeling Techniques," Jan. 2021, doi: 10.1007/978-981-15-5148-2_26.
30. "Exploring Latent Themes-Analysis of various Topic Modelling Algorithms," International Journal of Advanced Research in Science, Communication and Technology, Jun. 2023, doi: 10.48175/ijarsct-11635.
31. M. V. Martin, D. A. Kirsch, and F. Prieto-Nanez, "The promise of machine-learning- driven text analysis techniques for historical research: topic modeling and word embedding," Management & Organizational History, Jan. 2023, doi: 10.1080/17449359.2023.2181184.
32. Z. T. Ke, P. Ji, J. Jin, and W. Li, "Recent Advances in Text Analysis," Annual review of statistics and its application, Nov. 2023, doi: 10.1146/annurev-statistics-040522-022138.
33. I. D. Wood, "Recent Advances and Applications of Probabilistic Topic Models," Dec. 2014, doi: 10.1063/1.4903721.
34. J. Zeng, J.-F. Yan, and S.-R. Gong, "Advances in Topic Models for Complex Document Network Data," Chinese Journal of Computers, Jan. 2012, doi: 10.3724/SP.J.1016.2012.02431.
35. S. Likhitha, "A Detailed Survey on Topic Modeling for Document and Short Text Data," International Journal of Computer Applications, Aug. 2019, doi: 10.5120/IJCA2019919265.
36. D. Sharma, B. Kumar, and S. Chand, "A Survey on Journey of Topic Modeling Techniques from SVD to Deep Learning," International Journal of Modern Education and Computer Science, Jul. 2017, doi: 10.5815/IJMECS.2017.07.06.
37. Alam, M. H., Ryu, W.-J., Lee, S., 2016. Joint multi-grain topic sentiment: modeling semantic aspects for online reviews. Information Sciences 339, 206–223. Doi: https://zenodo.org/records/1219899