ISSN 2007-9737

Improving Sentiment Classification for Hotel Recommender System through Deep Learning and Data Balancing

Reza Nouralizadeh Ganji, Chitra Dadkhah, Nasim Tohidi

Toosi University of Technology,
Computer Engineering Faculty, Artificial Intelligence Department,
Iran

{reza.nooralizadehganji, n.tohidi}@email.kntu.ac.ir, dadkhah@kntu.ac.ir

Abstract. A recommender system is a type of information filtering system that predicts and recommends items or products to users based on their preferences and past behaviors. It is commonly used in e-commerce and social media to suggest items that a user may be interested in purchasing, reading, watching, or listening to. Sentiment analysis is an area of natural language processing that has emerged as a popular way for organizations to detect and categorize opinions about a product, idea, or service. In recent years, many attempts have been made to apply sentiment analysis in designing recommender systems, in order to recommend various items, such as hotels. It is thought that providing a quality hotel suggestion based on the requirements and preferences of users is a challenge and, naturally, an alluring effort for tourism applications. In this paper, the quality of decision making for hotel recommender systems based on sentiment analysis, deep learning, and data balancing techniques has been improving. Multiple approaches are used with our proposed system to provide high-quality hotel recommendations. To achieve this goal, first, the existing dataset is balanced, using the translating and text paraphrasing policy by a transformer-based model called T5. Afterwards, an integrated method, including the transformer-based XLM-RoBERTa model is used along with the attention mechanism for sentiment analysis. The result of the comparison of our proposed model with the four best non-transformer-based models; RNN, GRU, LSTM, Bi-LST, and the most recent transformer-based model, En-RFBERT, on the TripAdvisor dataset, showed the superiority of our proposed method. Our proposed system beats En-RFBERT by 3%, 7%, and 5% in Macro Precision, Recall, and F1-score, respectively, and performs better than En-RFBERT when it comes to responsiveness time.

Keywords. Recommender system, sentiment analysis, data balancing, natural language processing, deep learning, transformer, attention.

1 Introduction

With the emergence of the internet in our life, the world can be seen from a wide-shot perspective, and when it grew and became complicated, diverse knowledge was shining brighter around us, underlying this enchanting expansion, an important challenge appears swiftly: how can relevant information be retrieved from the massive amount of unstructured data?

One of the practical solutions for this challenging question can be found in the Recommender Systems (RSs). By providing personalized recommendations, RSs have become an essential system for business to enhance user experience, increase customer loyalty, and drive sales.

Various techniques are used to assemble recommendation systems including Collaborative Filtering (CF), Content-Based filtering (CB), Knowledge-Based filtering (KB), and hybrid approaches. Collaborative filtering relies on the behavior and preferences of similar users to make recommendations, while content-based filtering uses features of the items themselves to suggest related items. In a knowledge-based recommender system, the knowledge base is a database of information about the items being recommended, such as their features, attributes, and characteristics.

The system uses this information to determine which items would best meet the user's needs and preferences. Hybrid approaches combine the techniques to provide more accurate and diverse recommendations [1].

When users with different needs are looking for items like products [2], hotels [3], and movies [4], e-commerce platforms and their customers stand to gain a great deal from the development of a method to present products in a manner that is determined by the preferences of consumers. The latest studies have found that one of the most advanced techniques for developing such an RS is to extract user opinions about a particular product [5].

Since the hotel RSs are intended from a travel marketing perspective, a huge number of shared textual reviews concerning various hotel attributes (e.g., cleanliness, food, service, etc.) can be viewed as a key and informative source for extracting users' opinions [6].

The sentiments derived from these shared reviews can assist recommendation algorithms in gaining a deeper comprehension of hotel attributes, thereby producing hotel recommendations that are more appropriate to the user preferences [7].

The concept of opinion mining or sentiment analysis refers to studying what people feel about items such as hotels, products, restaurants, and their attributes [8]. In search of resources for analyzing sentiments, we can find platforms like Airbnb[1], TripAdvisor[2], and Expedia[3], which provide opinionated textual reviews in digital form.

By processing these reviews, insights about hotels and how they treat their guests can be gathered and incorporated into a recommendation system to assist travelers (users) in finding the most suitable hotels.

For instance, users may place great importance on food quality, and when traveling with their family, they may prefer to stay in hotels that serve delicious food. Therefore, it is essential to take into account both user preferences and hotel attributes [9-11].

In this paper, the quality of decision making for hotel recommender systems based on sentiment analysis, deep learning, and data balancing techniques has been improved. Multiple approaches are used with our proposed system to provide high-quality hotel recommendations.

The remainder of this work is arranged in the following manner. In Section 2, the literature related to recommender systems and sentiment analysis is reviewed. Section 3 gives the details of our proposed method.

The evaluation and experimental results are presented in section 4. Finally, the study is concluded in section 5, and some possible future works are also mentioned there.

2 Related Works

There is an excellent opportunity for hotel RSs that leverage critical advantages inherent in the recommendation process; indeed, travelers want to specify their requirements explicitly, such as destination, priorities, and duration of trips; on the other hand, RSs can properly prepare options for the user to specify his or her prerequisites and when these important pieces of knowledge are gathered, suitable hotels can be retrieved in the recommendation list [12, 13].

De Pessemier et al. concentrated on a list of users' ratings, personal preferences, and destination-specific requirements and formed a hybrid approach consisting of the CB and KB techniques for recommending travel destinations to groups.

They gave each dimension, such as location, tourist profile, type of attraction, and transportation costs, and then performed a weighted average rating prediction [14].

Gulzar et al. provided a framework for course recommendation that assists learners in selecting courses that meet their given requirements, implying that the system can read the learner's requirements explicitly and recommend courses to them.

Their algorithm searches the space of potential courses using published content about them and attempts to improve the quality of user queries by locating synonyms and creating N-grams in order to return a larger number of relevant courses. Their proposed hybrid approach performed well and has acquired a recommendation accuracy of 95.25% [15].

[1] www.airbnb.com

[2] www.tripadvisor.com

[3] www.expedia.com

For sentiment analysis of user comments on YouTube about smartphone devices, Mai and Le, suggested a deep learning technique based on multilingual Bidirectional Encoder Representations from Transformers (BERT) [16].

In order to create a more generalizable model, they combined sentiment analysis tasks trained at the sentence and aspect levels to minimize the requirement for feature engineering and other language sources.

Their model has achieved an F1-score of 81.78% and outperformed recent baselines by a large margin of 3.06% accuracy. When sentiment analysis is performed at the aspect level, information about both the whole text and the different aspect categories becomes essentially significant [17].

Liao et al. proposed that feature extraction on full text and aspect tokens be performed using a pre-trained model of the Robustly Optimized BERT Pre-training Approach (RoBERTa) [18].

By combining the pre-trained RoBERTa model, which has been trained on a large amount of textual data, and the attention mechanism, the most relevant features can be detected, and stable performance can be achieved.

In light of the growing popularity of pre-trained and transformer-based models like BERT, several researchers are exploring new ways to improve on these established models [19]. In another study, Song et al. examined the potential of BERT intermediate layers to enhance BERT fine-tuning and attempted to extract knowledge from the intermediate layers [20].

As a result, aspect-based sentiment analysis and natural language inference tasks were improved. The modified model was evaluated on two tasks, aspect-based sentiment analysis, and natural language inference, and both yielded better results than the previous models by achieving a 76.69% accuracy metric.

One of the problems with mining opinions from tourism review data is that there are a variety of phrases or sentences that are in more than one language. From a wider perspective, analyzing user-generated content from across countries, like the United States, Germany, France, Italy, etc., necessitates the processing of various languages. Barriere et al. proposed fine-tuning the multilingual transformer model termed XLM-RoBETRa [21] and utilizing data-augmentation using an automatic translation approach to address the issues associated with non-English tweets sentiment.

Their suggested architecture showed better performance when compared to monolingual pre-trained models in this context and obtained a Macro F1-score of 71.4% [22].

In a separate work, Ghosh et al. investigated the presence of multi-language code-mixed texts and proposed a multi-task system that includes polarity recognition and sentiment classification [23].

This system made use of the XLM-RoBERTa cross-lingual embedding-based transformer model. The two tasks included in this work were polarity detection and sentiment classification. When compared to state-of-the-art XLMR-based models employed for the same job, the results achieved by their model using transfer learning were superior.

In order to provide more precise recommendations, RSs can benefit from processing and identifying item information, such as shared reviews, and understanding the underlying sentiments expressed in those feedbacks [24]. In this regard, a number of studies have focused on the application of sentiment analysis in RSs and have developed models to support their conclusions.

Asani et al. suggested a sentiment-based restaurant recommendation system. The method begins by extracting individual food preferences from gathered textual comments via a lexicon-based Sentiment Analysis technique based on SentiwordNet.

According to these extracted preferences, their suggested RS can recommend restaurants and assist users in selecting the best options and making an informed choice [25].

An approach to tourism recommendation based on semantic clustering and sentiment analysis was devised by Abbasi-Moud et al. in [26]. Textual data from TripAdvisor was investigated for user preferences using semantic clustering and text sentiment analysis.

The authors adopted an adjustment mechanism that modified tourist recommended destinations based on the current situation and a variety of contextual parameters, such as time and weather data. Liu et al. offer a multilingual review-

aware deep recommender architecture that accurately recommends based on investigated feelings in reviews.

They employed pre-trained multilingual word embedding to deal with reviews from different languages, fed the embedding to Bidirectional GRU to obtain enriched word representations, then extracted aspects obtained to perform aspect-level sentiment analysis using the attention mechanism.

Following that, the target user rating of a particular item can be predicted using the explored sentiments and attention mechanisms [10]. Recently, Ray et al. proposed a hotel RS that generates tailored recommendations using trinary sentiment analysis and aspect categorization of hotel reviews.

To start with, they tried to group the textual data into categories based on hotel characteristics using fuzzy logic and the cosine similarity methods to do this. Second, they constructed an ensemble of five sentiment classifiers, three of which were trained to recognize binary polarities with BERT, and two of which were learned to detect trinary polarities with BERT and Random Forest, to handle their imbalanced dataset.

The final step of their proposed approach is to generate a list of recommended hotels based on user preferences and sentiment associated with hotel reviews. Their proposed ensemble model performed well in sentiment classification and obtained a Macro F1-score of 84% and improved accuracy [27].

In another endeavor, Roy and Dutta were involved in the design and development of a sentiment analysis-based movie recommendation system [28]. To perform sentiment analysis on user-submitted reviews of various films, this system used an evolutionary algorithm known as Water Cycle Earthworm Optimization (WCEWO).

They present a system in which Hierarchical Attention Networks (HAN) are used to aid in the process of sentimental classification. The WCEWO algorithm is used to support the HAN training process, and as a consequence, an acceptable movie recommendation is achieved by presenting users with relevant recommendations for positively reviewed movies.

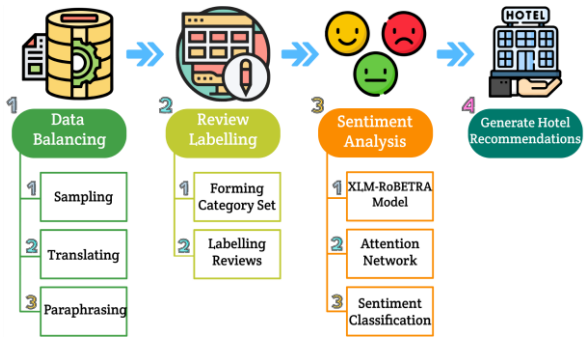


Fig. 1. Structure of the proposed system

3 Proposed System

Despite the fact that using shared textual reviews about hotels can strengthen our recommendations, our strategy may face four fundamental flaws. First, an imbalanced dataset containing many more instances of one sentiment class than another can pose serious difficulties for most learning algorithms that assume a relatively balanced distribution [29].

Second, although the textual reviews are written in English, due to the diversity of nationalities of the travelers (users) who write these reviews, we see the presence of words from different languages in the written sentences, which can disrupt the performance of the sentiment classifier model.

For example, the phrase «*hinnalle ainakin off season*» which contains Finnish and English words, appears in one textual review feedback.

Third, because each hotel has diverse attributes like cleanliness, food, and room, any textual review published about it might also have varied thoughts about its different attributes.

For instance, a former guest of a hotel might think about the quality of the room and submit feedback in the form of a text in order to assist other travelers or users in evaluating the hotel.

Therefore, for the purpose of exploring the opinions towards a particular aspect, it is necessary to identify what aspect corresponds to which text. Fourth, one of the crucial difficulties for RSs that benefit from text mining approaches is predicting the sentiment polarity of new data in a timely manner. For example, using an ensemble of deep learning models for sentiment analysis may

achieve acceptable performance on imbalanced data, but classifying new data need much more time than unified and integrated models.

Thus, in the case of exponential growth in the number of users and textual reviews, recommender frameworks should be agile and have low inference latency.

The proposed system incorporates the benefits of sentiment data into the recommendation phase, which implies that the users specify their needs explicitly and receive tailored suggestions for Top@K items.

Our propose system classify the users opinion based on their review into 3 polarity classes as positive, negative and neutral using the deep learning methods. The polarity of the user review can reflect whether the reviewer had a favorable or unfavorable experience with the hotel, which can help determine whether the hotel is a good fit for other people with similar tastes who want to travel in the future.

Minority classes are treated with translation and paraphrase, and the majority class is under sampled to balance the massively unbalanced dataset. Additionally, an effective sentiment analysis is used that can predict the polarity class for multilingual data in an adequate response time.

Our proposed recommender system consists of several steps, as illustrated in Figure 1. In the first step, we employ data balancing as a pre-processing strategy to enhance the quality and quantity of the training data and improve the performance of sentiment classifiers.

In the second step, we classify textual reviews into categories based on their similarity to a set of predefined categories using supervised machine learning, which allows us to determine which hotel attribute is discussed in each shared review. In the third step, we use an attention network and transformer-based XLM-RoBERTa model to classify the reviews based on their sentiment.

Finally, user preferences, which include destination locations and desired hotel attributes, are incorporated as input queries, and with the help of the previous two steps, personalized recommendations consisting of hotels with a higher sentiment score in a particular hotel attribute are generated and recommended to the user. This multi-step approach enables us to provide more accurate and relevant recommendations while taking into account the user's individual preferences.

For instance, if a user wishes to travel to Paris and prefers to stay at a hotel that serves the most delicious foods, the proposed system receives destination locations and preferred hotel attributes as input queries and recommends Top@k hotels that are located in Paris and have the highest positive sentiment polarities in their related shared reviews, which were written about the food attribute of those hotels.

3.1 Data Balancing

The accuracy metric is generally one of the most important characteristics to be considered in order to evaluate the quality of RSs. The accuracy of RSs recommendations is closely related to the quality of the dataset used in the analysis.

Therefore, a dataset with a large proportion of one sentiment class over the other may have major consequences for item recommendations that use reviews of items, as a primary source of information to construct recommendations.

There is a significant data imbalance in the TripAdvisor dataset, with far fewer negative and neutral reviews than positive ones. This leads to algorithms that are more prone to picking up on false patterns and overfitting high-frequency data. It is also possible that models trained on unbalanced data will underperform when confronted with new and unknown real-world examples.

Sampling, translating, and paraphrasing are the three strategies that were used for balancing the dataset in order to generate high-quality sentiment classifiers. Given the prevalence of shared reviews in the majority class, sampling is one of the most straightforward methods to implement first. In this phase, therefore, all shared ratings belonging to the majority class are under sampled, and as a result, the size of the majority class is reduced, which may result in information loss in subsequent analyses.

To address this issue, the sentiment classification step employs a pre-trained transformer that, with the aid of transfer learning, can work with a smaller amount of data. In addition, a combined under sampling strategy for the majority class and supplementing techniques for

the minority class are used to prevent potential information loss and enhance with additional informative data.

The rapid advancement of text translation algorithms and the open-source availability of powerful translation models like Google Translate which has been made it possible to employ these algorithms to generate new data.

With the translation method, we can generate completely distinct textual data written in a different language without the possibility of similar phrases appearing in the source and generated data, despite back translation approaches.

We need to make sure that our model can handle text from many different languages when the translation method is used. To supplement the dataset with translated data, each user text review was prepared following some preprocessing and put into the Google Translate API to generate fresh data in the French and German languages.

As a result, our proposed system can generate two distinct sets of data for each user text review. Finally, to supplement the dataset with paraphrased data, the suggested approach makes use of the Text-To-Text Transfer Transformer (T5) mechanism to generate more diversified paraphrases that retain the same meaning but have a more diverse vocabulary.

The T5-large paraphrasing model [30], which was trained on the ParaNMT dataset, is used to produce four new textual data for a preprocessed source text review.

3.2 Review Labeling

In this step, sets of representative nouns are extracted based on categories that indicate the attributes of hotels that have already been established. After that, shared reviews are categorized by looking for similarities between the reviews and each element in the index set of different categories.

Significant characteristics of the review set should be discovered using an approach based on frequent word sequences and equivalence categories to locate categories and their index terms.

This method for identifying topics, which was presented by Zhan et al. [31], generates a set of representative nouns for each predefined category by recognizing frequent sequences of words of varying lengths.

Table 1. Hotel attributes and their extracted representative nouns

Hotel Attribute	Representative Nouns
Cleanliness	satisfactory, ample, hygienic, proper, spotless, odor, dirty, clean, smell
Service	desk, check in, check out, reliable, fast, convenient, service
Location	railway, view, station, airport, distance, far, close, train, metro, transport, market, mall, surrounding, areas, highway, traffic, out
Value	price, amount, rate, cheap, worth, low, money, economical, reasonable, fee, expensive, charge
Room	bed, bunkbeds, toilet, bathroom, shower, dryer, fridge, space, spacious, outdated, noisy
Food	drink, breakfast, spicy, food, tasty, tea, coffee, buffet, bar, restaurant, dinner, lunch, brunch, delicious
Facility	front, pool, gym, wifi, spa, internet, wireless, broken, parking, ventilation
Staff	friendly, helpful, reliable, quick, good, polite, staff

All sets of candidate nouns that appear in the same set of reviews will be combined into a single equivalence class. Table 1 has a list of these categories and the most prevalent words inside them.

The reviews are categorized into groups by applying two distinct methods—fuzzy string matching and cosine similarity—to determine the degree of similarity between each shared review and each element in the index sets various categories.

In order to determine the fuzzy similarity value for a review to belong to a specific category, the fuzzy string-matching approach makes use of the Levenstein distance, which can determine the distance between two sequences of words.

For each category, the similarity value for a review to belong to a given category is determined by calculating the cosine similarity between each

vector representation of words in the review and the index terms in that category.

Fuzzy matching and cosine similarity both produce scores on a scale from 0 to 1, and this final step of the review labeling procedure involves averaging the two calculated similarity scores and considering the most valuable hotel attribute. Consequently, the shared reviews are categorized into the label class that has the highest similarity.

3.3 Sentiment Analysis

An enhanced transformer-based model with an attention mechanism was used to accomplish sentiment analysis in cross-lingual reviews, taking into account the benefits of transfer learning.

We employed XLM-RoBERTa in its most fundamental form from the HuggingFace model hub. Transformers serve as the fundamental building blocks for all three BERT, RoBERTa, and XLM-RoBERTa architectures; a transformer encoder layer of our proposed method is shown in Figure 2.

The XLM-RoBERTa transformer model takes tokens from a multilingual textual review as its inputs, and it produces a contextualized embedding vector as its output. The output that is created, which is the output of the last hidden state, is a vector with 768 dimensions, which is the same size as the base transformer version in its initial form.

Initially, consecutive sentences in a textual review will be separated by [CLS] and [SEP] tokens. Then, all tokens are supplied to the embedding layer, which is comprised of the sum of token embeddings, position embeddings, and segment embeddings, in order to conduct main numerical inputs as real value vectors.

On each layer of the transformer, feed-forward neural networks, multi-head attention, and layer normalization are applied to the embedding patches. The base version, was chosen because of its reasonable complexity and inference time.

The transformed representations are acquired at the top of each hidden state of each transformer layer, and some recent research chooses the outputs of the final layer as the representations with the most information for further analysis.

By utilizing an improved attention mechanism, not only the hidden states of the final layer but also

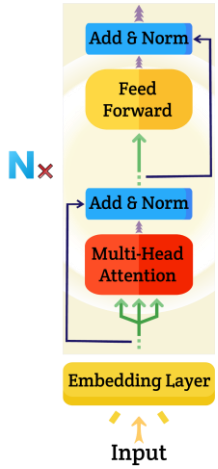


Fig. 2. Encoder part of Transformer of our proposed system

those of all intermediate transformer layers can be exploited in XLM-RoBERTa, allowing for the extraction of enriched features.

In order to accomplish this, the representation of the classification token (i.e., the [CLS] token), which includes an aggregated representation of the entire input sequence, will be gathered and stacked for each transformer layer.

After that, a technique called dot product attention will be used to dynamically combine all the intermediates and figure out how the representations of each layer contribute to the whole.

In the end, the outputs of the attention mechanism that have been completed are sent to a fully-connected neural network that has two hidden layers, then a SoftMax layer for polarity prediction comes after it; using Eq. 1 and Eq. 2, we are able to track the actions of hidden and SoftMax layers.

Because the user comments in the dataset are labeled with one of three emotion classes—positive, neutral, or negative—the number of neurons that are deployed in the final hidden layer ought to be three.

In addition to this, a discrete probability distribution function known as SoftMax (σ) might be utilized to determine the classification probability of the subclass that corresponds with it. Furthermore, the most probable class might be identified using Eq. 3:

$$O_{FC} = W_{l_2} f\left(W_{l_1}(H_{Att}) + b_{l_1}\right) + b_{l_2}, \quad (1)$$

$$\sigma(O_{FC})_i = \frac{e^{O_{FC_i}}}{\sum_{j=1}^{K} e^{O_{FC_j}}}, \quad for\ j = 1,2,\dots,K, \quad (2)$$

$$\hat{y} = \underset{i \in 1,2,\dots,K}{\operatorname{argmax}}\{\sigma(O_{FC})_i\}, \quad (3)$$

where $H_{Att} \in \mathbb{R}^{d_{Att}}$ denotes outputs of dot product attention module, and $W_{l_i} \in \mathbb{R}^{d_{FC_i} \times d_{FC_{i-1}}}$ and $b_{l_i} \in \mathbb{R}^{d_{FC_i} \times 1}$ denote weight and bias vectors of MLP's i[th] layer respectively and $f(\cdot)$ denotes activation which is the ReLU function.

Additionally, O_{FC} is the output of the feed forward layer used for classification. The proposed sentiment classifier model is shown in Figure 3.

The entirety of the model is trained via supervised learning by attempting to achieve the lowest possible error in terms of cross-entropy classification and using backpropagation to determine the gradients of all the parameters, and then employing stochastic gradient descent to bring those gradients up to date.

During training, the objective is to ensure that each phrase has the smallest possible amount of cross-entropy error between y and $\hat{y}$, where, as mentioned in Eq. 4, y represents the ground truth, and $\hat{y}$ represents the output of the sentiment classifier model:

$$loss = -\sum_i \sum_j y_i^j \log \hat{y}_i^j + \lambda \|\theta\|^2, \quad (4)$$

where i represents the sentence index, j the class index, λ the L2-regularization factor, and θ the collection of all parameters.

3.4 Generate Hotel Recommendations

Our proposed system performs the process of making recommendations in two independent phases: offline and online as shown in Figure 4. In the offline phase, the labeling of hotel attributes is accomplished by computing the similarity between the vector representation of each word in the shared review and the index terms in predefined categories/classes.

Then, using the proposed sentiment classifier, which has previously been trained on a balanced dataset, the sentiment class of each textual review could be calculated profoundly.

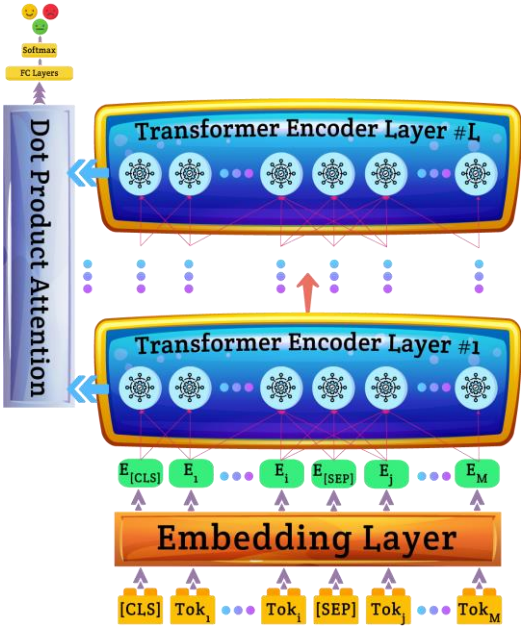


Fig. 3. Proposed sentiment classifier

Thus, the label of each shared review, which is a specific attribute of a hotel, and the polarity of its sentiment orientation are identified during the offline phase.

Users might choose their preferences, like preferred location and hotel attributes, as input to our proposed hotel recommendation system.

Therefore, in the online phase, hotels that correspond to the users' selected location are retrieved as the initial recommendation list. Then the list is arranged according to the hotel attributes such as food, service, etc.

The resulting lists of hotel comments are then sorted in descending order based on their sentiment polarities, which are analyzed in the offline phase. At the end, the user receives Top@k hotels that have been identified based on the highest sentiment polarity ratings in a certain category.

4 Experimental Results

We applied an updated, extensive, and diversified collection of hotel reviews which was gathered

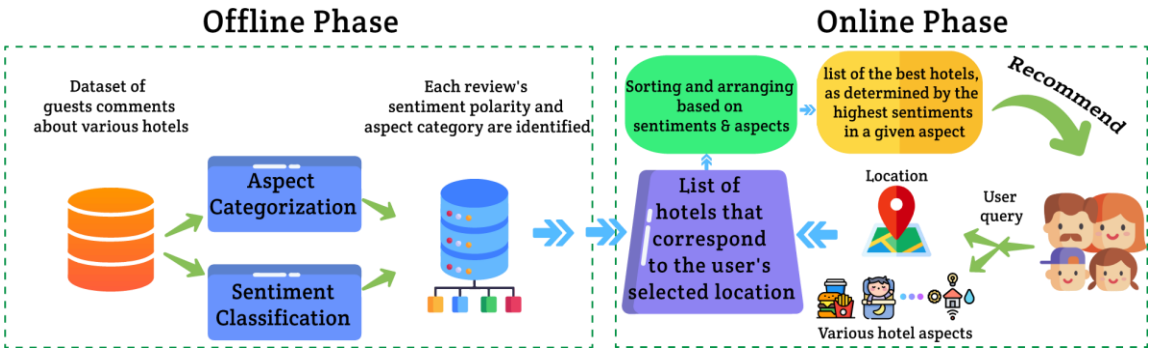
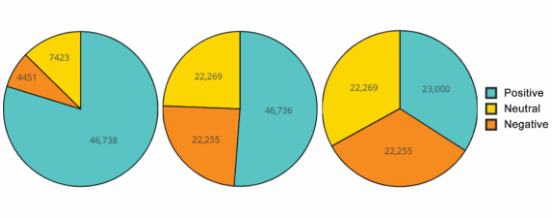


Fig. 4. The process of recommendations in our proposed system



a) b) c)

Fig. 5. Polarity-wise hotel review distribution; (a) original, (b) augmented by paraphrasing and translating, (c) augmented and sampled data

from the TripAdvisor website by Garain et al. [32], for training and evaluating our proposed system.

This dataset is comprised of critical factors, all of which relate in some way to the hotels and the services they offer, as seen through the view of past guests. It has 58,620 sample data instances in total. One useful feature of this dataset is the inclusion of textual comments that have been annotated with sentiment scores.

In order to produce high-quality sentiment classifiers, three methods are proposed to deal with inequality in the original TripAdvisor dataset.

First, the Google Translate API has been used to make two different sets of data for each source text review in the minority classes. Second, a large version of the T5 transformer model has been applied to generate four new data for a preprocessed source text review in the minority classes.

Third, the under-sampling strategy is applied to all text reviews belonging to the majority class; Figure 5 depicts the distribution of text reviews by sentiment/polarity class for the original and augmented datasets.

Python 3, a widely used programming language, is used to implement the proposed system and other baselines. Fuzzy string matching in text review labeling has been implemented using the Fuzzywuzzy Python library.

Moreover, the architecture described in the sentiment analysis section and all other baselines relating to the sentiment classification task has been implemented using the Pytorch library, which provides a high-level neural network API. All transformer-based models were trained on NVIDIA TESLA P100 GPUs, which were provided by Kaggle[4].

For training, the batch size is set to 16, the maximum length of a sequence is 350 tokens, and the dropout rate is set to 0.1 for some regularization.

Using the back-propagation algorithm and the Adam stochastic optimizer with a learning rate and

[4] www.kaggle.com

Table 2. Comparison of the proposed system with baseline models on the TripAdvisor dataset

Model	Precision	Recall	F1-score	Support
RNN	0.73	0.63	0.60	5862
GRU	0.79	0.74	0.76	5862
LSTM	0.78	0.70	0.70	5862
Bi-LSTM	0.81	0.70	0.74	5862
En-RFBERT	0.86	0.82	0.84	5862
Proposed model	**0.89**	**0.89**	**0.89**	6753

Table 3. Comparison of the models for all polarity classes

Criteria	Polarity	Models					
		RNN [33]	GRU [34]	LSTM [35]	Bi-LSTM [36]	En-RFBERT [28]	Proposed system
Precision	Positive	0.90	**0.95**	0.92	0.93	**0.95**	0.90
	Neutral	0.72	0.63	0.74	0.63	0.81	**0.87**
	Negative	0.58	0.80	0.67	0.87	0.81	**0.91**
Recall	Positive	**0.99**	0.97	**0.99**	0.98	0.95	0.90
	Neutral	0.12	0.63	0.26	0.55	0.66	**0.83**
	Negative	0.77	0.63	0.85	0.56	0.82	**0.95**
F1-score	Positive	0.94	**0.96**	0.95	**0.96**	**0.96**	0.90
	Neutral	0.20	0.63	0.39	0.59	0.73	**0.85**
	Negative	0.66	0.71	0.75	0.68	0.82	**0.93**

Table 4. Test time of the proposed system and En-RFBERT

Model	Testing Time (Second)	Max Length (Token)	Support
En-RFBERT	206.2	90	5862
Proposed model	**79.69**	350	6753

L2-regularization weight of 2×10^{-5} and 10^{-3}, the network is trained for 14 epochs.

As the loss function, cross entropy is operated, and an accuracy metric is used to find when the model has converged.

The performance of our proposed system is evaluated based on four metrics: Precision, Recall, F1-score, and Accuracy. In addition, the macro average of each criterion can be calculated using Eq. 5:

$$loss = -\sum_{i}\sum_{j} y_i^j \log \hat{y}_i^j + \lambda \|\theta\|^2. \quad (5)$$

We compared two types of the newest trained baseline models on the TripAdvisor dataset with our proposed system. Deep state-of-the-art

Networks with recurrent units can produce state-of-the-art outcomes for any text classification task, and the most current transformer-based approach showed impressive results on the TripAdvisor dataset.

The result of the comparison of our proposed system with the four best non-transformer-based models; RNN, GRU, LSTM, Bi-LSTM, and the most recent transformer-based model, En-RFBERT, are presented in Table 2. Our suggested technique beats En-RFBERT by 3%, 7%, and 5% in Macro Precision, Recall, and F1-score, respectively.

So, the system that was proposed has gotten noticeably improved results on the sentiment/polarity classes that suffer from unbalanced data.

In reality, assembling a diversified and well-balanced training and testing set of data (also known as support) and analyzing the data with an advanced sentiment classifier model play critical roles in this development.

Table 3 shows the significant improvements in minority sentiment/polarity classes, such as negative and positive, in the proposed system based on the macro average of each criterion.

The most important reason for making these modifications is to apply data balancing to the original dataset and to develop new and diversified phrases that have the same meaning as source text reviews with the assistance of methods for translating and paraphrasing.

Many more instances of positive polarity were seen as opposed to neutral and negative in all three train, validation, and test datasets, on which other baselines were trained and tested.

This caused the models to learn much more in one class and less in other sentiment classes, which resulted in overestimated results. In other words, they saw many more instances of positive polarity as opposed to neutral and negative.

In contrast to previous techniques, the under-sampling method has been applied to all text reviews in the majority class in the strategy that has been presented.

This results in balanced data being produced in all three sets of data—the train set, the validation set, and the test set.

The amount of time that passed during the inference phase was analyzed for our proposed system and the newer transformer-based model (En-RFBERT), as shown in Table 4.

According to the result of table 4, the proposed system performs better than En-RFBERT when it comes to responsiveness time. In comparison to En-RFBERT, the proposed system has been shown to take significantly less time, even when applied to a larger number of cases in the test set, and it selects a maximum sequence length threshold that is substantially higher.

The proposed system utilizes one integrated transformer-based model with an attention network, whereas the En-RFBERT model deploys an ensemble of five sentiment classifier models, four of which are transformer-based BERT models.

As a result, the proposed system has a higher computational efficiency than the En-RFBERT model; Figures 6 and 7 illustrate the distribution of distinct sentiment/polarity across all the models evaluated for this study on a balanced and original dataset.

5 Conclusion

From a travel marketing standpoint, several shared textual feedbacks on hotel's attributes (e.g., cleanliness, food, service, etc.) may be considered a vital and valuable source for extracting consumers' preferences.

This study proposes a system that employs sentiment analysis techniques and data balancing to improve the quality of hotel recommendations based on specific locations and various interested aspects of the hotel. In this regard, the initial step is to balance the current dataset.

Then the attention network is employed in conjunction with the transformer-based XLM-RoBERTa model to analyze the sentiment polarities. The competitive findings of the paper are the result of employing data balancing strategy and a modern multilingual transformer, XLM-RoBERTa, with an attention mechanism for sentiment analysis.

The performance of our proposed systems superiors the most recent transformer-based model, En-RFBERT by 3%, 7%, and 5% in Macro Precision, Recall, and F1-score, respectively. So, the model that was suggested has gotten noticeably improved results on the

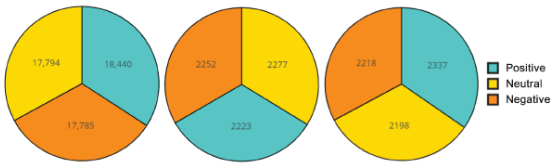


Fig. 6. Polarity-wise hotel review distribution: (a) train set, (b) validation set, (c) test set of the balanced dataset

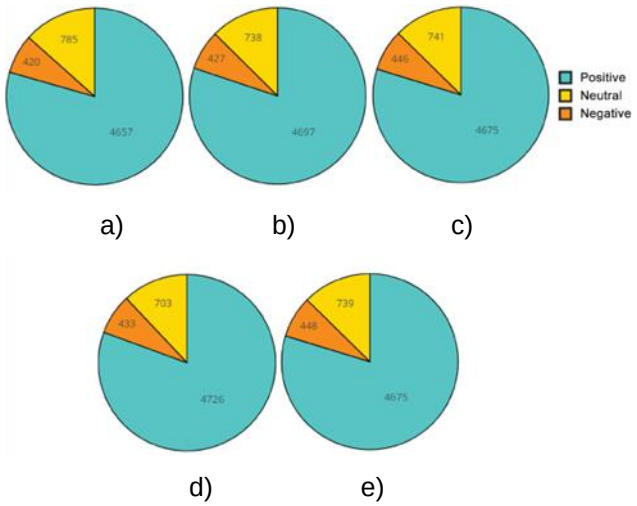


Fig. 7. Polarity-wise hotel review distribution on a validation set of the original dataset for: (a) En-RFBERT, (b) RNN, (c) GRU, (d) LSTM, (e) Bi-LSTM

sentiment/polarity classes that suffer from unbalanced data.

The experimental results on the TripAdvisor dataset demonstrated that the proposed system is able to achieve 0.89 in the F1-score, which is superior performance compared to the related systems and advocates substantial agility in predicting the sentiment/polarity of text reviews, which has a huge impact on hotel recommendations.

In comparison to En-RFBERT, the proposed system has been shown to take significantly less time, even when applied to a larger number of cases in the test set, and it selects a maximum sequence length threshold that is substantially higher.

Since this study focused on such complex topics as treating unbalanced data, dealing with multilingual reviews, and making accurate sentiment predictions, there are still certain restrictions on doing trials.

First, we used the transformer-based XLM-RoBERTa for sentiment classification without considering many trainable parameters. Although this deep learning model has the potential to produce impressive performance, it requires careful attention to model parameters that affect flexibility.

As a result, the focus of the next stage will be on delivering a miniature model with high-performance capabilities. Second, the existence of noisy and incomplete texts presents a formidable issue that may be addressed in future research.

References

1. **Aggarwal, C. C. (2016).** Recommender systems. The Textbook, Springer International Publishing. DOI: 10.1007/978-3-319-29659-3.
2. **Bobadilla, J., Ortega, F., Hernando, A., Gutiérrez, A. (2013).** Recommender systems

survey. Knowledge-Based Systems, Vol. 46, pp. 109–132. DOI: 10.1016/J.KNOSYS.2013.03.012.

3. **Veloso, B. M., Leal, F., Malheiro, B., Burguillo, J. C. (2019).** On-line guest profiling and hotel recommendation. Electronic Commerce Research and Applications, Vol. 34, p. 100832. DOI: 10.1016/J.ELERAP.2019.100832.
4. **Tohidi, N., Dadkhah, C. (2020).** Improving the performance of video collaborative filtering recommender systems using optimization algorithm. International Journal of Nonlinear Analysis and Applications, Vol. 11, No. 1, pp. 483–495. DOI: 10.22075/ijnaa.2020.19127.2058.
5. **Rosa, R. L., Schwartz, G. M., Ruggiero, W. V., Rodriguez, D. Z. (2019).** A knowledge-based recommendation system that includes sentiment analysis and deep learning. IEEE Transactions on Industrial Informatics, Vol. 15, No. 4, pp. 2124–2135. DOI: 10.1109/TII.2018.2867174.
6. **Zheng, L., Noroozi, V., Yu, P. S. (2017).** Joint deep modeling of users and items using reviews for recommendation. Proceedings WSDM '17: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, pp. 425–434. DOI: 10.1145/3018661.3018665.
7. **Liu, P., Zhang, L., Gulla, J. A. (2021).** Multilingual review-aware deep recommender system via aspect-based sentiment analysis. ACM Transactions on Information Systems, Vol. 39, No. 2, pp. 1–3. DOI: 10.1145/3432049.
8. **Liu, B. (2016).** Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge university press, Computational Linguistics, Vol. 42, No. 3, pp. 595–598. DOI: 10.1162/COLI_r_00259.
9. **Aliannejadi, M., Crestani, F. (2018).** Personalized context-aware point of interest recommendation. ACM Transactions on Information Systems (TOIS), Vol. 36, No. 4, pp. 1–28. DOI: 10.1145/3231933.
10. **Xue, W., Li, T., Rishe, N. (2016).** Aspect identification and ratings inference for hotel reviews. World Wide Web, Vol. 20, No. 1, pp. 23–37. DOI: 10.1007/S11280-016-0398-9.
11. **Zhang, Y., Miao, D., Wang, J., Zhang, Z. (2019).** A cost-sensitive three-way combination technique for ensemble learning in sentiment classification. International Journal of Approximate Reasoning, Vol. 105, pp. 85–97. DOI: 10.1016/J.IJAR.2018.10.019.
12. **Sohrabi, B., Vanani, I. R., Tahmasebipur, K., Fazli, S. (2012).** An exploratory analysis of hotel selection factors: A comprehensive survey of Tehran hotels. International Journal of Hospitality Management, Vol. 31, No. 1, pp. 96–106. DOI: 10.1016/J.IJHM.2011.06.002.
13. **Zhang, K., Wang, K., Wang, X., Jin, C., Zhou, A. (2015).** Hotel recommendation based on user preference analysis. 2015 31st IEEE International Conference on Data Engineering Workshops, IEEE Computer Society, Vol. 2015, pp. 134–138. DOI: 10.1109/ICDEW.2015.7129564.
14. **Pessemier, T. D., Dhondt, J., Martens, L. (2016).** Hybrid group recommendations for a travel service. Multimedia Tools and Applications, Vol. 76, No. 2, pp. 2787–2811. DOI: 10.1007/S11042-016-3265-X.
15. **Gulzar, Z., Leema, A. A., Deepak, G. (2018).** PCRS: Personalized course recommender system based on hybrid approach. Procedia Computer Science, Vol. 125, pp. 518–524. DOI: 10.1016/J.PROCS.2017.12.067.
16. **Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2019).** BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.
17. **Mai, L., Le, B. (2020).** Joint sentence and aspect-level sentiment analysis of product comments. Annals of Operations Research 2020, Vol. 300, pp. 493–513. DOI: 10.1007/S10479-020-03534-7.
18. **Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D. Stoyanov, V. (2019).** RoBERTa: A robustly optimized BERT pretraining

approach. Proceeding of ICLR´20 Conference Blind Submission, DOI: 10.48550/arxiv. 1907.11692.

19. **Liao, W., Zeng, B., Yin, X., Wei, P. (2020).** An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa. Applied Intelligence, Vol. 51, No. 6, pp. 3522–3533. DOI: 1007/S10489-020-01964-1.
20. **Song, Y., Wang, J., Liang, Z., Liu, Z., Jiang, T. (2020).** Utilizing BERT intermediate layers for aspect based sentiment analysis and natural language inference. Computation and Language. DOI: 10.48550/arxiv.2002.04815.
21. **Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Stoyanov, V. (2020).** Unsupervised cross-lingual representation learning at scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747.
22. **Barriere, V., Balahur, A. (2020).** Improving sentiment analysis over non-English tweets using multilingual transformers and automatic translation for data-augmentation. pp. 266–27. DOI: 10.18653/V1/2020.COLING-MAIN.23.
23. **Ghosh, S., Priyankar, A., Ekbal, A., Bhattacharyya, P. (2023).** Multitasking of sentiment detection and emotion recognition in code-mixed Hinglish data. Knowledge-Based Systems, Vol. 260, p. 110182, DOI: 10.1016/J.KNOSYS.2022.110182.
24. **Liu, J., Yu, Y., Mehraliyev, F., Hu, S., Chen, J. (2022).** What affects the online ratings of restaurant consumers: a research perspective on text-mining big data analysis. International Journal of Contemporary Hospitality Management, Vol. 34, No. 10, pp. 3607–3633. DOI: 10.1108/IJCHM-06-2021-0749/FULL/XML.
25. **Asani, E., Vahdat-Nejad, H., Sadri, J. (2021).** Restaurant recommender system based on sentiment analysis. Machine Learning with Applications, Vol. 6, p. 100114. DOI: 10.1016/J.MLWA.2021.100114.
26. **Abbasi-Moud, Z., Vahdat-Nejad, H., Sadri, J. (2021).** Tourism recommendation system based on semantic clustering and sentiment analysis. Expert Systems with Applications, Vol. 167, p. 114324. DOI: 10.1016/J.ESWA.2020.114324.
27. **Ray, B., Garain, A., Sarkar, R. (2021).** An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews. Applied Soft Computing, Vol. 98, p. 106935. DOI: 10.1016/J.ASOC.2020.106935.
28. **Roy, D., Dutta, M. (2022).** Optimal hierarchical attention network-based sentiment analysis for movie recommendation. Social Network Analysis and Mining, Vol. 12, p.138. DOI: 10.1007/s13278-022-00954-0.
29. **Kaur, H., Pannu, H. S., Malhi, A. K. (2019).** A systematic review on imbalanced data challenges in machine learning. ACM Computing Surveys, Vol. 52, No. 4. DOI: 10.1145/3343440.
30. **Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Liu, P. J. (2020).** Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, Vol. 21, No. 1, 5485–5551.
31. **Zhan, J., Loh, H. T., Liu, Y. (2009).** Gather customer concerns from online product reviews – A text summarization approach. Expert Systems with Applications, Vol. 36, No. 2, pp. 2107–2115. DOI: 10.1016/J.ESWA.2007.12.039.
32. **Garain, A. (2020).** hotel reviews from around the world with sentiment values and review ratings in different categories for natural language processing. IEEE Dataport. DOI: 10.21227/8ggw-hm23.
33. **Rumelhart, D. E., Hinton, G. E., Williams, R. J. (1986).** Learning representations by back-propagating errors. Nature, Vol. 323, pp. 533–536. DOI: 10.1038/323533a0.
34. **Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y. (2014).** Learning phrase representations using RNN encoder–decoder for statistical machine translation. EMNLP 2014, Conference on Empirical Methods in Natural Language Processing, Proceedings of

the Conference, pp. 1724–1734. DOI: 10.3115/V1/D14-1179.

35. **Hochreiter, S., Schmidhuber, J. (1997).** Long short-term memory. Neural Computation, Vol. 9, No. 8, pp. 1735–1780. DOI: 10.1162/NECO.1997.9.8.1735.
36. **Schuster, M., Paliwal, K. K. (1997).** Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, Vol. 45, No. 11, pp. 2673–2681. DOI: 10.1109/78.650093.

Article received on 27/06/2023; accepted on 01/08/2023. Corresponding author is Chitra Dadkhah.