



UNIVERSITY OF
ILLINOIS CHICAGO

Review – Response Analysis using Deep Learning Models

IDS 576 – DEEP LEARNING & APPLICATIONS

AIDA SANATIZADEH, ANIRUDHA BALKRISHNA, SHUBHAM KHODE & VAMSY TAMMINEEDI

Table of Contents

<i>Introduction</i>	1
<i>Objective</i>	1
<i>Existing Literature</i>	1
<i>Data</i>	2
<i>Process</i>	5
Objective 1 – Understanding emotion and building a model to determine the criticality of reviews	5
Objective 2 – Understanding Semantic Similarity of Reviews and responses	8
<i>Future Scope</i>	9
<i>Conclusion</i>	9
<i>References</i>	10

Introduction

To sustain a profitable revenue stream there are many factors that companies in the Hospitality industry must maintain. One of those is their online presence and reputation. The clearest indicator of a hotel's reputation is often found in online reviews.

Review websites such as TripAdvisor can have a significant impact on how travellers choose their accommodation. Most of the time, the consensus is what people will accept so it's vital that a hotel has a reputation for quality service and professional standards. A recent Barclays study reveals there is an extra £3.2 billion to be earned over the next decade if the hotel sector becomes more attentive to online reviews. This is because customer reviews are a great source of "Voice of the customer" and could offer tremendous insights into what customers like and dislike about a product or service.

Considering the facts above, since customer reviews are detailed texts, it is an arduous task to generate insights from texts on a large scale. In this project, we introduce and apply some techniques to help companies in analysing the reviews efficiently.

Objective

Our objective is to analyse and understand Hotel reviews posted on TripAdvisor & summarise our insights. We have shortlisted 2 objectives that will help Hotels to analyse their reviews and responses

1. Understand the emotions of the reviewers & build a model to determine the criticality of a review with the newly generated emotion as a feature
2. Compare the review and responses to understand whether the reviews were addressed correctly

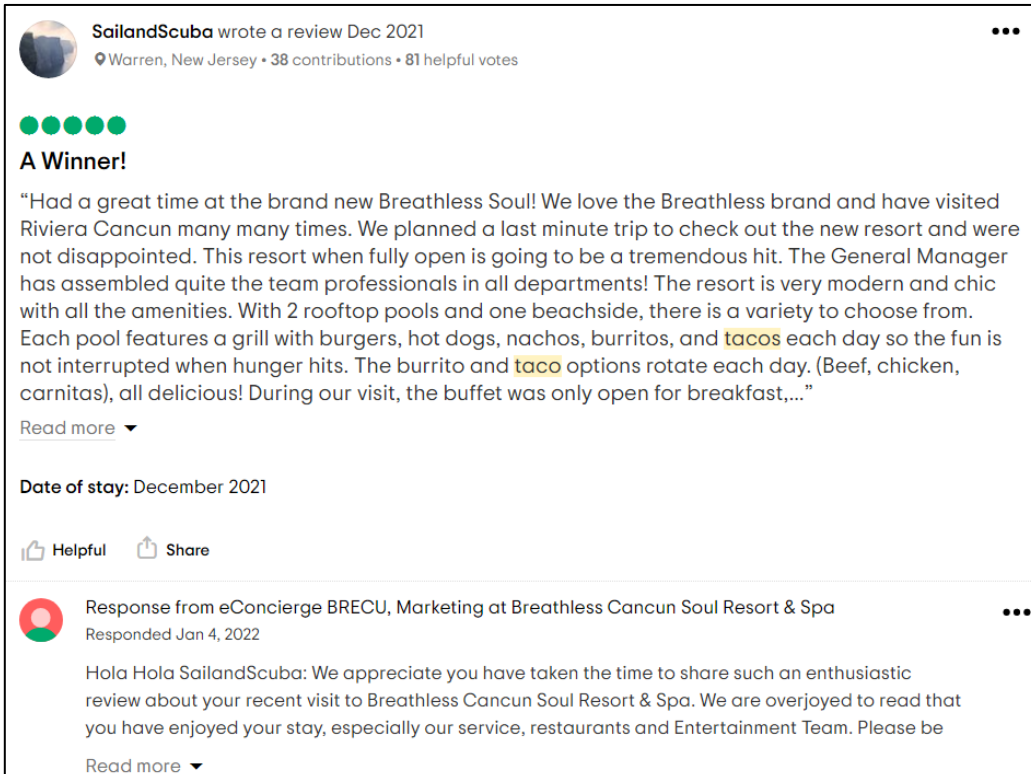
Existing Literature

A considerable amount of research has been conducted to understand and leverage insights from online reviews. Dimitrois et. al, studied the Convolutional Neural Network for sentiment analysis of TripAdvisor reviews (Amanatidis et al., 2019), Chih et. al implemented several deep-Learning techniques including CNN and RNN on hotel reviews and responses (Ku et al., 2019), and Afina et. al, also studied the LSTM-based deep learning architecture of tourist review in TripAdvisor (Ramadhani et al., 2021). We have referred to these existing research papers to understand how deep learning models can be applied in the context of review analysis.

Data

The data source is TripAdvisor, the world's largest travel website. We used the Beautiful Soup technique to scrape the data from the website. We focused our attention on 225 major global hotel chains in this data collection effort, obtaining all reviews dating back from 2018 to 2021, on every hotel location associated with each chain. We selected the chain hotels because of having a similar standardization in service delivery, plus they also tend to be franchised, with local ownership. Accordingly, when it comes to our aggregate analyses, we can exploit variation in rating distributions within a hotel brand, across geographies. The corpus consists of over 1,467,919 reviews. Out of these reviews we have taken out 123194 rows for modelling since these instances have both reviews and responses.

For each hotel, we collect the hotel name/chain, quality level (e.g., 3-star hotel), Travel type, service quality, Etc. For each associated review, we collect the review's URL, valence (an integer between 1 and 5), review text, response text, the author's profile name, the date the review was posted, and the reported date of response (see Figure 1). Using the review text, we construct a measure of emotion. we focus on all our analyses on English-language reviews.



SailandScuba wrote a review Dec 2021
 Warren, New Jersey • 38 contributions • 81 helpful votes

★★★★★

A Winner!

"Had a great time at the brand new Breathless Soul! We love the Breathless brand and have visited Riviera Cancun many many times. We planned a last minute trip to check out the new resort and were not disappointed. This resort when fully open is going to be a tremendous hit. The General Manager has assembled quite the team professionals in all departments! The resort is very modern and chic with all the amenities. With 2 rooftop pools and one beachside, there is a variety to choose from. Each pool features a grill with burgers, hot dogs, nachos, burritos, and **tacos** each day so the fun is not interrupted when hunger hits. The burrito and **taco** options rotate each day. (Beef, chicken, carnitas), all delicious! During our visit, the buffet was only open for breakfast,..."

[Read more](#) ▼

Date of stay: December 2021

Helpful Share

Response from eConcierge BRECU, Marketing at Breathless Cancun Soul Resort & Spa
 Responded Jan 4, 2022

Hola Hola SailandScuba: We appreciate you have taken the time to share such an enthusiastic review about your recent visit to Breathless Cancun Soul Resort & Spa. We are overjoyed to read that you have enjoyed your stay, especially our service, restaurants and Entertainment Team. Please be

[Read more](#) ▼

Figure : Review and Response Snapshot from TripAdvisor.com

Exploratory Analysis

Our first task is to understand the statistical relationships between different features of hotel. For this, we consider the scores of a hotel in terms of value, service, sleep quality, room, location and cleanliness and see its impact on review stars. The correlation plot reveals all hotel quality features are positively correlated with reviews, with room, value, service and cleanliness having the highest impact.

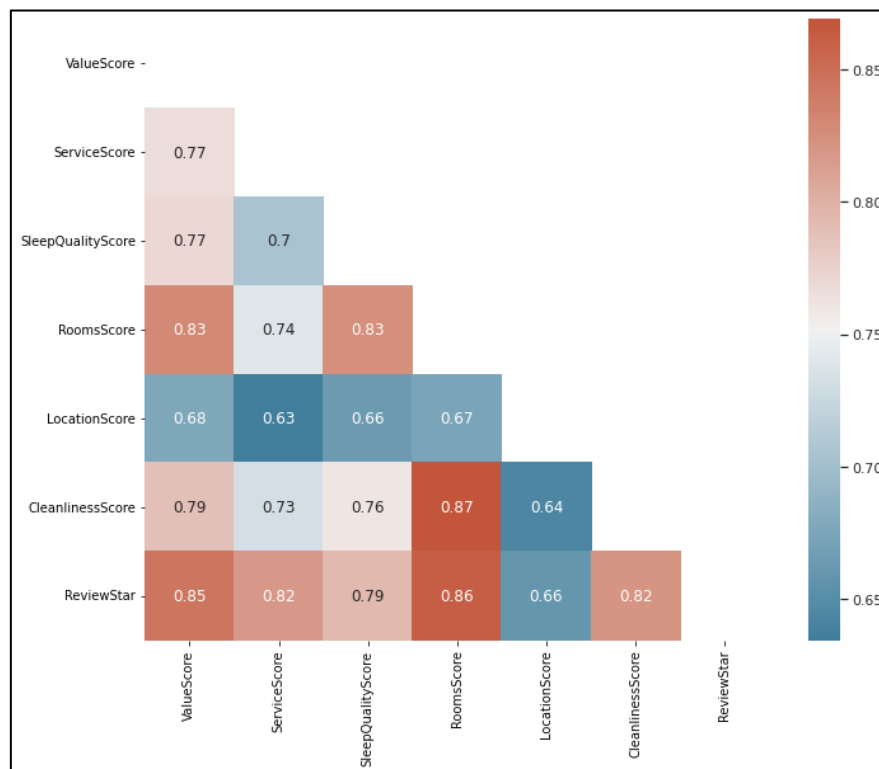


Figure: Correlation Matrix

Managerial responses to reviews can identify service failures and enhance customer satisfaction. Thus, our next task is to find out response rate of managers for available reviews. We manipulate our data to add another column 'ResponseAvailable' - Yes or No – which tells if a response was made by manager or not. Here we find an interesting trend that the response rate has been decreasing over the years. One potential reason could be the effect of COVID-19 on travel industry. Less people travelled in general, while on the hotel side there were less staff/managers due to lay-offs.

We also take a look at response rate for different travel types. As expected, managers have a relatively higher response rate for business travels type vs solo travellers.

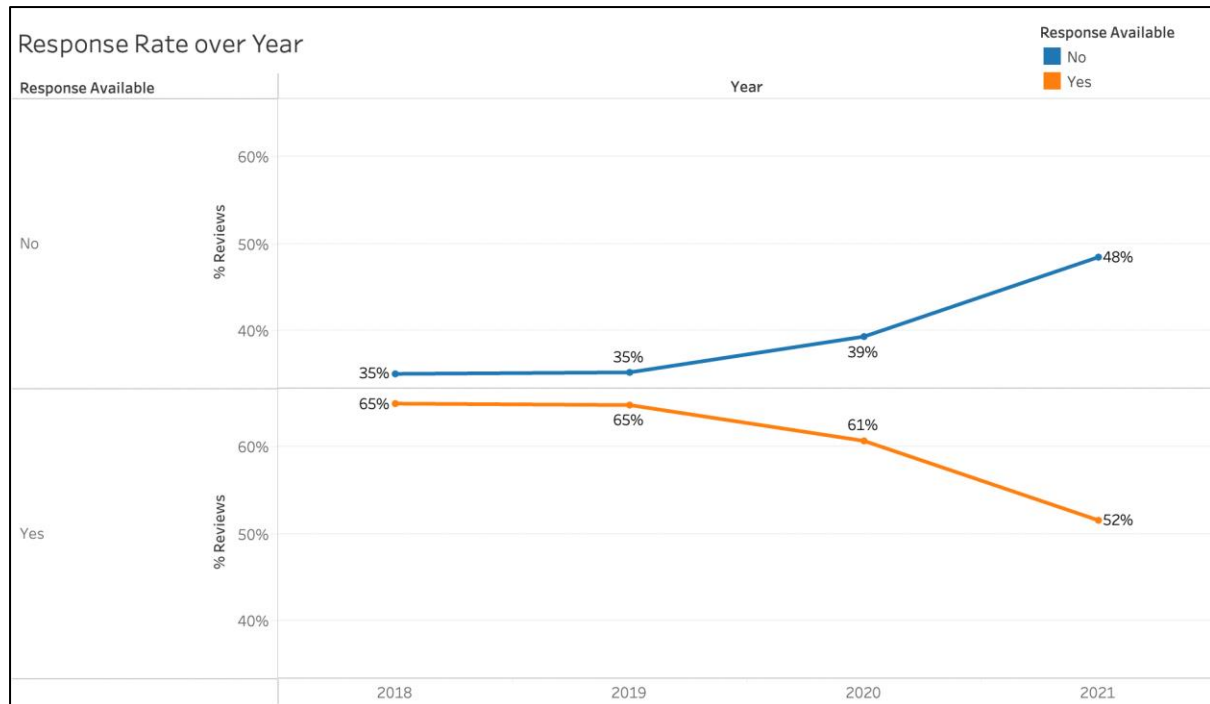


Figure: Response rate over years

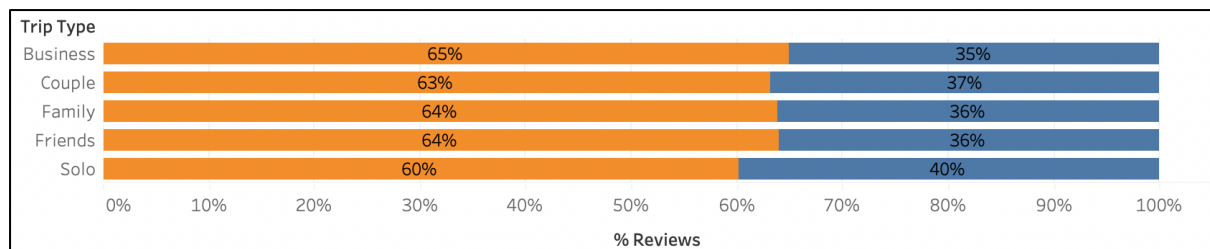


Figure: Response rate for trip type

Before we begin the modelling process, we further clean up our data by removing all NaN values, so the final corpus shrinks down to 123194 rows. We quickly take a look at the distribution of review stars in our data, and observe that more than 70% of the reviews are rated 4/5 star, which makes our data a bit imbalanced towards positive reviews.

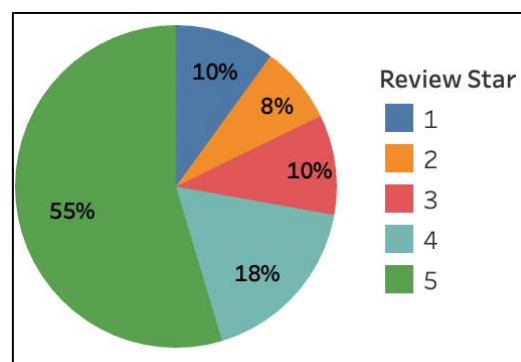


Figure: Distribution of Review stars

Process

We have used the BERT model to accomplish two of our objectives – to determine the criticality of a review and to find the similarity between a review and its response. BERT, which stands for Bidirectional Encoder Representations from Transformers is an open-source machine learning framework for NLP. BERT is designed to help computers understand the meaning of ambiguous language in the text by using surrounding text to establish context. The BERT framework was pre-trained using text from Wikipedia and can be fine-tuned with a question & answer dataset.

It is based on Transformers, a deep learning model in which every output element is connected to every input element, and the weightings between them are dynamically calculated based on their connection. In simple words, it is a pre-trained model that helps embed sentences while preserving their meanings. The figure below shows how a BERT model performs its task of Embedding a sentence.

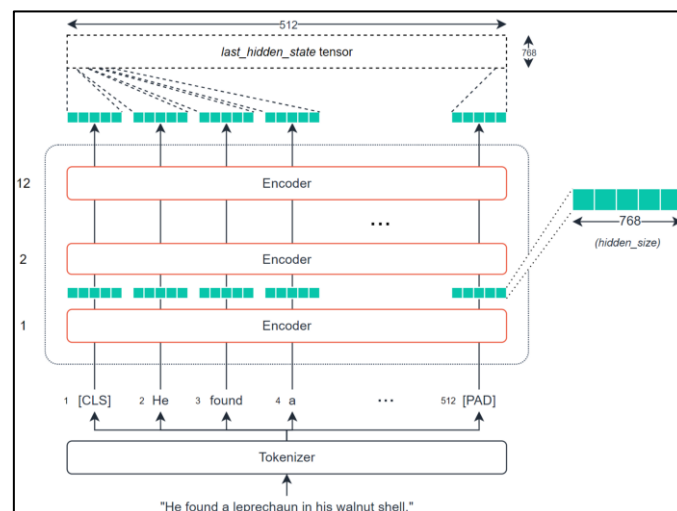


Figure – BERT Model

Objective 1 – Understanding emotion and building a model to determine the criticality of reviews

In order to get the emotion of the reviewer, the review text is used as an input to a pre-trained model[1] trained and fine-tuned using a wide range of emotions dataset[2] to classify 6 different emotions – anger, sadness, fear, surprise, joy and love.

This pre-trained model is fine-tuned on Google T5 Engine [3] which can either output a class label or a span of the input. T5's text-to-text architecture allows users to apply the same model, loss function, and hyperparameters to any NLP application, such as machine translation, document summarization, question answering, and classification (e.g., sentiment analysis). T5 can also be used for regression problems if we teach it to predict the string representation of a number rather than the actual number.

The 6 emotions detected are then grouped into 3 levels of criticalities – High, medium and low. As shown in table 1 the emotions are categorized into different criticality indices. A review with a 0 criticality index indicates high criticality and 2 is the lowest criticality.

Emotion	Criticality Index
Sadness or Anger	0
Surprise or Fear	1
Joy or Love	2

The average delay in response is calculated using the review date and response date for each review. It is observed that the hotel manager's response delay is approximately around 1 month (30 days) irrespective of the emotion of the review.

Emotion	Avg. Response Delay (Days)
Anger	31.55
Fear	29.91
Joy	31.04
Love	28.18
Sadness	29.69
Surprise	34.04

The whole dataset is then split into training and validation sets with 30% data as validation data. BERT base uncased model is used here to predict the criticality of a given review. BERT base has a total of 12 attention heads (encoders) with 768 hidden layers.

AdamW optimizer which is a variant of the optimizer Adam that has an improved implementation of weight decay. AdamW adds weight decay during the gradient calculation step[4] and uses weight decay as a form of regularization to lower the chance of overfitting.

The model is fine-tuned using the recommended parameters with batch size 16 and a learning rate of $1e-5$ but the model is overfitting and performed poorly on validation data. So, a different learning rate – $2e-5$, is used to fine-tune the model for 10 epochs. The Below graphs shows the loss obtained during fine-tuning the model on training and validation data at each epoch for different learning – $1e-5$ and $2e-5$.

It is observed that with $1e-5$ learning rate after 3 epochs the model started to overfit with a higher loss on validation data. So, a little higher learning rate $2e-5$ is used which resulted in an almost the same loss for both training and validation data. We considered $1e-5$ as our final learning rate to perform a validation run and achieved an accuracy of 0.51 and 0.59 f1-score.

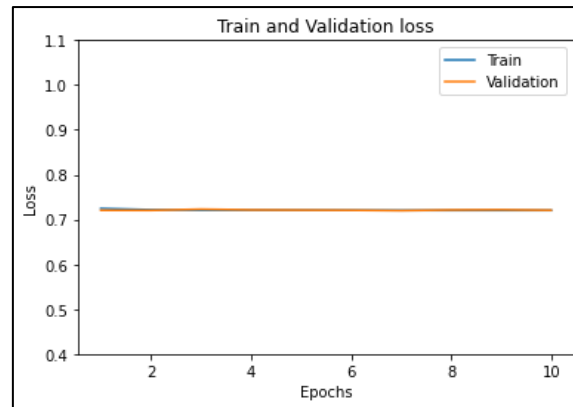


Figure – For learning rate 2e-5

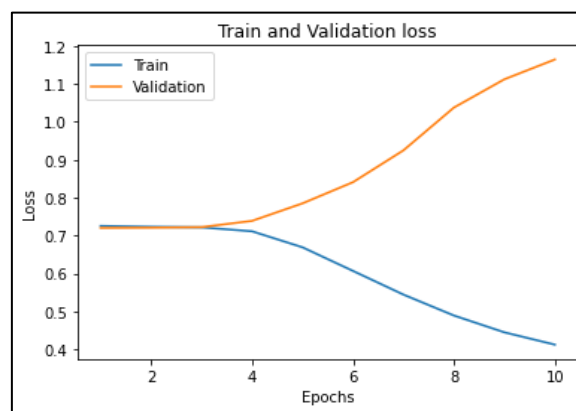


Figure – For learning rate 1e-5

Testing and Interpreting the Results:

[81]	txt = "Nice place to stay. Food was great. I would not hesitate to go back again"	
	predict_test(txt)	
		'Low Criticality'
[85]	txt = 'Not many options are available to eat. I am happy to see that atleast they provide dessert.'	# misclassified
	predict_test(txt)	
		'Low Criticality'
[92]	txt = 'Hotel is not secure!.'	
	predict_test(txt)	
		'Medium Criticality'
	txt = "I am surprised to see that they don't offer any starters"	# misclassified
	predict_test(txt)	
		'Low Criticality'
[87]	txt = "Food is very bad. Never coming back here again"	
	predict_test(txt)	
		'High Criticality'

Figure – Test Examples

We have used our model to predict the criticality of some new reviews. It is observed that highly positive and negative reviews are being predicted correctly. One of the reasons for the misclassification could be the label imbalance. 'Low Criticality' labels are the majority class

and there were also not many 'Medium Criticality' labels in our data. A stratified sampling or scraping more data from TripAdvisor could help us achieve better results.

Objective 2 – Understanding Semantic Similarity of Reviews and responses

Typically, a natural language processing (NLP) solution will take some text, process it to create a big vector/array representing said text — and then perform several transformations.

For our specific scenario, to understand if the response correctly addresses the review, we use a BERT model to vectorize the response and review. Then, we measure the cosine distance between the two embeddings or vectors. This helps us understand the semantic similarity or context behind the sentences.

Examples –

```
[ ] sentences[2]

'It's a great place to stay. Rooms are always super clean . Breakfast is great . The personnel is friendly and extremely helpful. The gym is awesome and u can reserve it by the hour . The grill works great and the coffee is good . The place is well maintained and clean'

[ ] sentences[12]

'Thank you for participating in the TripAdvisor survey. I am happy to hear you enjoyed your stay! We strive every day to provide a positive and memorable guest experience to our guests. I am glad to hear the staff was friendly and accommodating. We look forward to seeing you again! Have a great day and Safe Travels. '
```

```
[ ] from sklearn.metrics.pairwise import cosine_similarity
sim = cosine_similarity([sentence_embeddings[2]], [sentence_embeddings[12]])
print((sim*100))

[[77.010445]]
```

Figure – Example 1

In the picture above, we can see that the response addresses the review and therefore yields a similarity of 77%.

```
[ ] sentences[9]

'Bad service bad knowledge and cleanliness wasn't the best they ran us all over hotel looking for clean room because they had no log of what was what bad deal want my cash back or 3day credit would not recommend and no towels for pool bad rude service also because I reported it bout how they was not trained right or totally confused.,!'
```

```
[ ] sentences[19]

'Thank you for the feedback'
```

```
[ ] sim1 = cosine_similarity([sentence_embeddings[9]], [sentence_embeddings[19]])
print(sim1*100)

[[2.3859882]]
```

Figure – Example 2

In the second example, the response is possibly auto-generated and does not address the review. Therefore, the semantic similarity score is a mere 2%.

Analysing the semantic similarities of reviews and responses will help a manager decide if the employees that respond to reviews are posting relevant responses.

Future Scope

CRM requires learning the managerial responses of high-quality hotels. Responding to a review may increase transaction and labour expenses, while failing to reply to a review may result in lost customer retention possibilities.

Therefore, as an additional objective we can create a bot that will automatically respond to reviews based on the criticality and evaluate its response based on the similarity score. This can be achieved by using Google T5 Engine.

Conclusion

As mentioned earlier, the service industry has a lot to gain from monitoring and analysing reviews. This seems like a rather simple task but, as the scale of operations increases, we observe that a robust way to analyse reviews & responses becomes necessary.

In this project, we have attempted to create a robust method to prioritize reviews and evaluate responses. In order to prioritize reviews, we extract the emotion associated with the review using deep learning models and then create a criticality index to determine which review is detrimental to the company's image and has to be acted on immediately.

Another part of this project is to evaluate responses. We use deep learning models to embed review and response texts to understand whether a response accurately addresses the review. This will help a company understand if responders are doing a good job of answering reviews.

In summary, the criticality helps us determine the priority of responses and the similarity helps us determine the correctness of responses.

References

- [1] <https://huggingface.co/mrm8488/t5-base-finetuned-emotion>
- [2] https://github.com/dair-ai/emotion_dataset
- [3] [arXiv:1910.10683](https://arxiv.org/abs/1910.10683) [cs.LG]
- [4] <https://towardsdatascience.com/why-adamw-matters-736223f31b5d>
- [5] <https://towardsdatascience.com/bert-for-measuring-text-similarity-eec91c6bf9e1>
- [6] <https://www.sbert.net/>
- [7] <https://huggingface.co/sentence-transformers>
- [8] Amanatidis, D., Dossis, M., & Mylona, I. (2019). A convolutional neural network for sentiment analysis of tripadvisor reviews. *2019 4th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference, SEEDA-CECNSM 2019*, 1–6. <https://doi.org/10.1109/SEEDA-CECNSM.2019.8908466>
- [9] Ku, C. H., Chang, Y. C., Wang, Y., Chen, C. H., & Hsiao, S. H. (2019). Artificial intelligence and visual analytics: A deep-learning approach to analyze hotel reviews & responses. *Proceedings of the Annual Hawaii International Conference on System Sciences, 2019-January*, 5268–5277. <https://doi.org/10.24251/hicss.2019.634>
- [10] Ramadhani, A., Sutoyo, E., & Widartha, V. P. (2021). LSTM-based Deep Learning Architecture of Tourist Review in Tripadvisor. *2021 6th International Conference on Informatics and Computing, ICIC 2021*. <https://doi.org/10.1109/ICIC54025.2021.9632967>