


Article

When BERT Started Traveling: TourBERT—A Natural Language Processing Model for the Travel Industry

Veronika Arefeva¹ and Roman Egger^{2,3,*} ¹ Institute of Business Informatics, Johannes Kepler University of Linz, 4040 Linz, Austria² Department of Innovation and Management in Tourism, Salzburg University of Applied Sciences, 5412 Salzburg, Austria³ Department of Tourism and Service Management, Modul University Vienna, 1190 Wien, Austria

* Correspondence: roman.egger@fh-salzburg.ac.at

Abstract: In recent years, Natural Language Processing (NLP) has become increasingly important for extracting new insights from unstructured text data, and pre-trained language models now have the ability to perform state-of-the-art tasks like topic modeling, text classification, or sentiment analysis. Currently, BERT is the most widespread and widely used model, but it has been shown that a potential to optimize BERT can be applied to domain-specific contexts. While a number of BERT models that improve downstream tasks' performance for other domains already exist, an optimized BERT model for tourism has yet to be revealed. This study thus aimed to develop and evaluate TourBERT, a pre-trained BERT model for the tourism industry. It was trained from scratch and outperforms BERT-Base in all tourism-specific evaluations. Therefore, this study makes an essential contribution to the growing importance of NLP in tourism by providing an open-source BERT model adapted to tourism requirements and particularities.



Citation: Arefeva, V.; Egger, R. When BERT Started Traveling: TourBERT—A Natural Language Processing Model for the Travel Industry. *Digital* **2022**, *2*, 546–559. <https://doi.org/10.3390/digital2040030>

Academic Editors: Costas Vassilakis, George Lepouras and Manolis Wallace

Received: 9 September 2022

Accepted: 28 October 2022

Published: 11 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: TourBERT; BERT; tourism; natural language model

1. Introduction

Tourism products and services tend to be highly descriptive [1] as they cannot be tested in advance. In addition, tourism services are co-created with the customer and are relatively expensive compared to everyday products. As a result, the descriptions of products and services tend to be very excessive and text heavy. Alongside detailed descriptions from the supply side, user-generated content (UGC) continues to gain more relevance [2]. Whether on review platforms, such as TripAdvisor, or social media channels, such as Twitter, Facebook, or Instagram, individuals are constantly sharing their travel experiences and, in turn, influencing other users [3]. This content is of particular importance for tourism providers as they seem to be losing their power to UGC [4]. Therefore, to better understand consumer behavior and adapt to marketing initiatives, the automated analysis of texts using NLP methods is becoming increasingly important for both academia and the tourism industry [5]. At the same time, more powerful language models are emerging, enabling more advanced text analyses to be conducted.

BERT, developed by Google, is considered one of the most powerful and widely used language models. On the one hand, this pre-trained language model has been trained on a huge generic corpus and can be used universally. On the other hand, however, it has its weaknesses when it comes to domain-specific applications. Therefore, this paper aims to develop and evaluate a domain-specific BERT model for tourism. The proposed TourBERT model was pre-trained from scratch using 3.6 million tourist reviews and 46,000 descriptions of tourist services, attractions, and sights from more than 20 different countries around the world. This study makes a unique contribution to the extant body of natural language models and tourism research as the evaluation of TourBERT has proven its superiority to

BERT in all tasks concerning tourism-relevant content. TourBERT can be rendered the state-of-the-art language model for the tourism industry and for academic text analytics alike owing to the fact that the pre-trained model can be fine-tuned to perform numerous tasks such as text representation, text classification and clustering, topic modeling, sentiment analysis, or question answering.

2. Literature Review

With an increase in computational power and more effective and efficient algorithms, abundant research has been conducted in recent years, both within academia and the tourism industry, on how to best process textual data. According to Wennker [6], 80% of all data that is produced is text-based, which underscores Poon's [7] statement that "information is the lifeblood of tourism." Especially since the rise of UGC, a vast amount of unstructured text has become available at one's disposal, the analysis of which can provide important insights into tourists and their wants, needs, and experiences that are highly relevant for tourism marketing [5].

Regardless, the analysis of text data is challenging and requires the conversion of text into numerical values, which are necessary to use as input data for powerful machine learning algorithms. Over the past years, a wide variety of language models have been developed, ranging from the pure analysis of word frequencies to complex transformer models that are able to process multilingual data and take content as well as context into account. Especially through the concept of transfer learning, which is based on the use of pre-trained models, huge progress in NLP has been archived. However, since such language models are trained on huge corpora, the training process is extremely time-consuming and computationally intense. The applied training corpus is therefore responsible for the field of application and the domain the model will work well in [8].

Since its launch in 2018, Google's Bidirectional Encoder Representations from Transformers (BERT) is currently one of the most significant natural language models [9]. BERT-Large, which is based on a transformers architecture, is considered one of the most powerful language models, with 24 layers, 16 attention heads, and 340 million parameters in total [10]. It is a model pre-trained from scratch and can be fine-tuned to perform numerous downstream tasks such as text classification, question answering, sentiment analysis, extractive summarization, named entity recognition, or sentence similarity [8]. BERT-Base was pre-trained in a self-supervised way on a large English corpus consisting of raw texts from the BookCorpus dataset. This includes over 11,000 books in addition to the entire English Wikipedia. The nature of this training corpora implies that BERT was trained on a generic and unspecified domain corpus [11]. Yet, for domain-specific applications and downstream tasks, it has been proven that pre-training BERT on a large domain-specific corpus can be useful as it allows for better apprehension of linguistic peculiarities [12]. For example, several BERT variants have been pre-trained for the financial (FinBERT) [13], medical (Clinical BERT) [14], biological (BioBERT) [15], and computer science sectors (SciBERT) [16]. For tourism-related content, however, a domain-specific adaptation of BERT is not available on the market yet, hence why this paper introduces TourBERT. TourBERT will now be presented and evaluated in more detail in the next paragraphs.

3. Methodology and Results

The following sections describe the methodological procedure for the development of the TourBERT language model. The pre-training of TourBERT will be presented first, followed by its model evaluations. For the sake of clarity, the results of the five different evaluations are reported immediately after the description of each evaluation process.

3.1. Pre-Training TourBERT

TourBERT embodies BERT-Base-Uncased as its underlying architecture and was trained from scratch—unlike BioBERT or FinBERT, which were both pre-trained further from the BERT-Base initial checkpoint. The training corpus was pre-processed by convert-

ing the data into lowercase and splitting it into sentences, ultimately resulting in 22,601,333 sentences in total. Thereafter, two TourBERT models with SentencePiece and WordPiece tokenizers were trained, respectively. The motivation to use SentencePiece rather than conventional WordPiece tokenizers in conjunction with BERT was to establish an opportunity to extend TourBERT to a multi-language model in the future since SentencePiece is able to account for grammatical peculiarities of different complex languages like Chinese. To obtain a custom vocabulary, SentencePiece (32,000) and WordPiece (30,522) tokenizers were trained, with the latter being equal to the size of the BERT-Base tokenizer. Pre-training of both models was done for 1M steps on a single Google Colab Pro TPU instance, which lasted about three days in total.

3.2. TourBERT Model Evaluation

The evaluation of TourBERT was performed using both quantitative and qualitative measures. Two sentiment classification tasks were used for the supervised evaluation, while topic modeling, synonyms search, and a within-vocabulary words similarity distribution analysis were applied as part of the unsupervised evaluation. It is important to note that the evaluation of supervised tasks used SentencePiece tokenizers only since both models had comparable performance, as will be shown below.

3.2.1. Supervised Evaluation: Sentiment Classification

For classification purposes, BERT's architecture must be extended with a classifier layer in order to enable predictions. This can be achieved in numerous ways; for example, one of the most widely used approaches is attaching a softmax layer on top of the BERT model. A more advanced way of designing a classifier, however, involves an Long short-term memory (LSTM) layer, which is useful for the representation of long sequences exceeding BERT's maximum input length. In the case of TourBERT, outputs were passed through a single feed-forward layer, a simple classifier known for benchmarking different transformer models against each other. Keeping in mind that an architecture as such would not yield state-of-the-art results, the aim was simply to demonstrate that TourBERT can surpass BERT-Base without tending to achieve superior results on a particular dataset.

The sentiment classification task was performed on two publicly available datasets involving hotel reviews. The first dataset contains 69,308 hotel reviews from TripAdvisor [17] and includes three sentiment classes: {-1: "negative", 0: "neutral", 1: "positive"}. The second dataset contains 515,000 reviews from Europe hotels [18]. Here, only reviews with either negative or positive labels were used, which, in turn, transformed this problem into a binary classification with the following two classes: {-1: "negative", 1: "positive"}. The dataset contains attributes such as hotel name, number of reviews, and geographical position as well as negative and positive reviews from each reviewer. If a user had left only positive reviews, then the value for the negative reviews was left blank, and vice-versa. The following pre-processing approach was thus used to extract only positive and negative examples in order to prepare this dataset for a binary classification problem: Only reviews from users who left *either* only negative or only positive reviews were included. Using this approach, 35,000 positive and 35,000 negative reviews were sampled resulting in 70,000 samples in total.

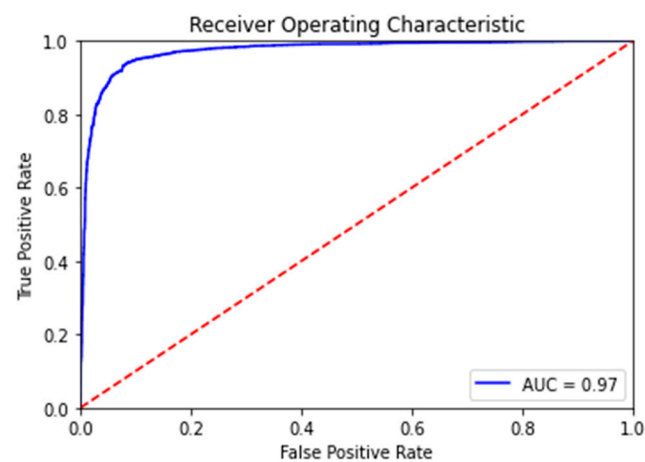
Both datasets were first pre-processed and then split into training, validation, and testing sets according to a 80%/10%/10% proportion. The pre-processing procedures included lowercasing and the removal of punctuation and non-ASCII characters from the text. Evaluation results for both tasks are shown in Tables 1 and 2 below, while Figure 1 presents the ROC curve and AUC score for TourBERT and BERT-Base models in the second task.

Table 1. Evaluation results for TourBERT and BERT-Base models for datasets from Tripadvisor.

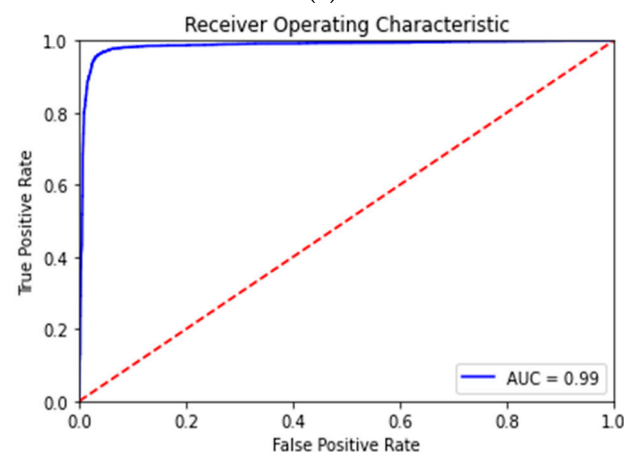
	Validation Set		Test Set			
	Loss	Accuracy	Accuracy	Precision	Recall	F1
BERT-Base	0.4250	0.8190	0.81	0.66	0.4	0.42
TourBERT (WordPiece)	0.3146	0.8708	0.86	0.7	0.65	0.68
TourBERT (SentencePiece)	0.3166	0.8712	0.87	0.7	0.65	0.68

Table 2. Evaluation results for TourBERT and BERT-Base models for datasets from Europe hotels.

	Validation Set		Test Set	
	Loss	Accuracy	Accuracy	AUC
BERT-Base	0.2296	0.9218	0.9279	0.97
TourBERT (WordPiece)	0.1371	0.9569	0.9633	0.99
TourBERT (SentencePiece)	0.1329	0.9586	0.9626	0.99



(a)



(b)

Figure 1. Cont.

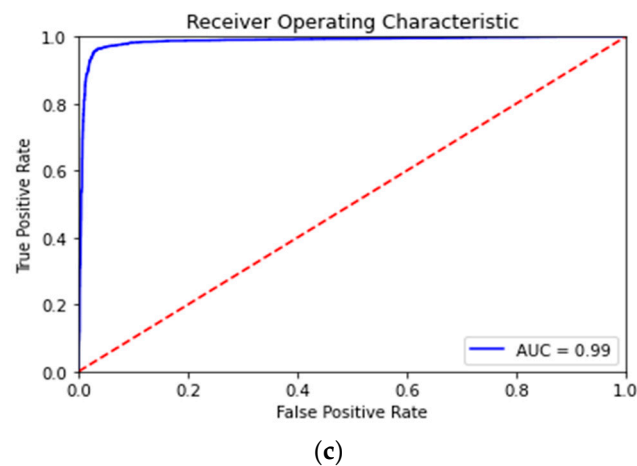


Figure 1. Area under ROC-Curve (AUC) scores for BERT-Base (a), TourBERT SentencePiece (b), and TourBERT WordPiece (c).

3.2.2. Unsupervised Evaluation: Visualization of Photo Annotations

The first unsupervised evaluation task was the visualization of photo annotations via TensorBoard Projector. For this task, a dataset of 48 photos depicting different tourism activities, such as sports activities, sightseeing, and shopping, amongst others, was applied. Next, 622 people were asked to manually label these photos by assigning two bi-gram tags to each individual photo. These annotations were then visualized using the TensorBoard Projector API, which allows for the visualization of original photos on a 2D or 3D plot located within their respective cluster centers. Finally, after performing UMAP, i.e., inspecting and comparing the groups' separation quality on the plot, the evaluation was complete. The visualization results for BERT-Base and TourBERT are presented in Figures 2 and 3, respectively.



Figure 2. TensorBoard Projector for BERT-Base (contains two views as a result of symmetric axes rotation).

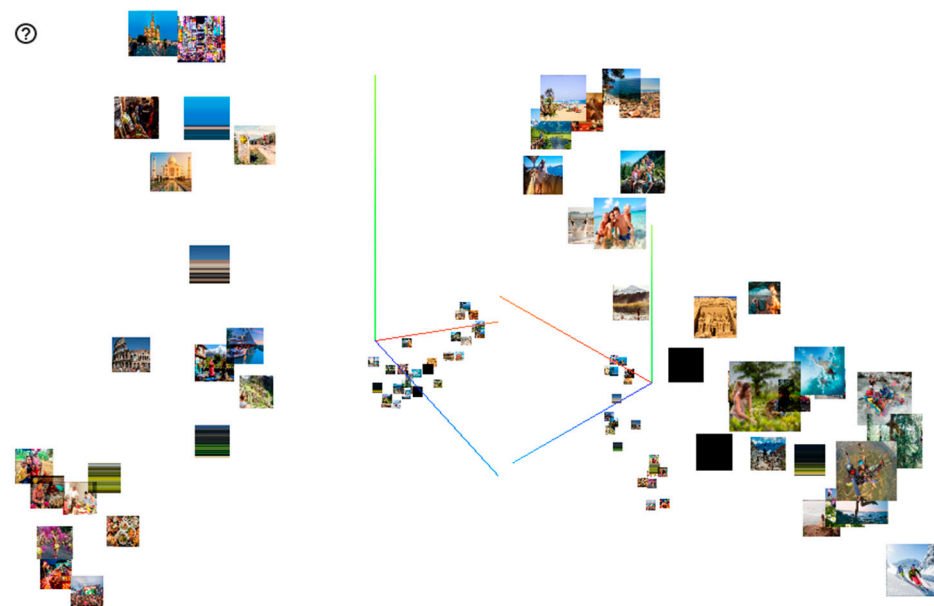


Figure 3. TensorBoard Projector for TourBERT (contains two views as a result of symmetric axes rotation).

The purpose of such a visualization is to evaluate the separation of clusters that naturally form from the down-projection method. Overall, one can observe that the TourBERT vectors lead to better group separation and that the pictures within each group contain similar content. Contrarily, when observing the results produced with BERT-Base vectors, the content of the pictures appear to be heavily mixed, without any visible cluster separation.

3.2.3. Unsupervised Evaluation: Topic Modeling

A subsequent unsupervised evaluation was undertaken by applying a topic modeling approach. For this, 5000 Instagram posts with the hashtag *#wanderlust* were extracted from public accounts and crawled using the Python Scrapy library. Instagram, as a social platform, principally utilizes photos to reflect its primary source of information, while the textual description of Instagram posts is often either limited to hashtags and emojis, unrelated to the photo, or missing entirely. Therefore, images were annotated using Google Cloud Vision API, and a TourBERT vector was generated for each photo annotation. Photo annotations were analyzed based on their similarity using a K-means clustering approach. The number of clusters was chosen using the silhouette score, which resulted in 25 clusters. In order to enable cluster center visualization on a 2D plot, a PCA down-projection method was selected to transform a 768-dimensional BERT embedding into a two-dimensional map.

Figure 4 below shows the cluster centers on a 2D plot, where the size of a cluster center is proportional to the cluster's population size. A visualization as such allows the quality of the topic separation to be evaluated.

From Figure 4, one can notice that the cluster centers produced with the down-projected TourBERT vectors reveal better separation than those produced with BERT-Base ones.

Another aspect of the topic modeling analysis was the estimation of word similarity within the same cluster. Topic words for both BERT-Base and TourBERT can be seen in Tables 3 and 4.

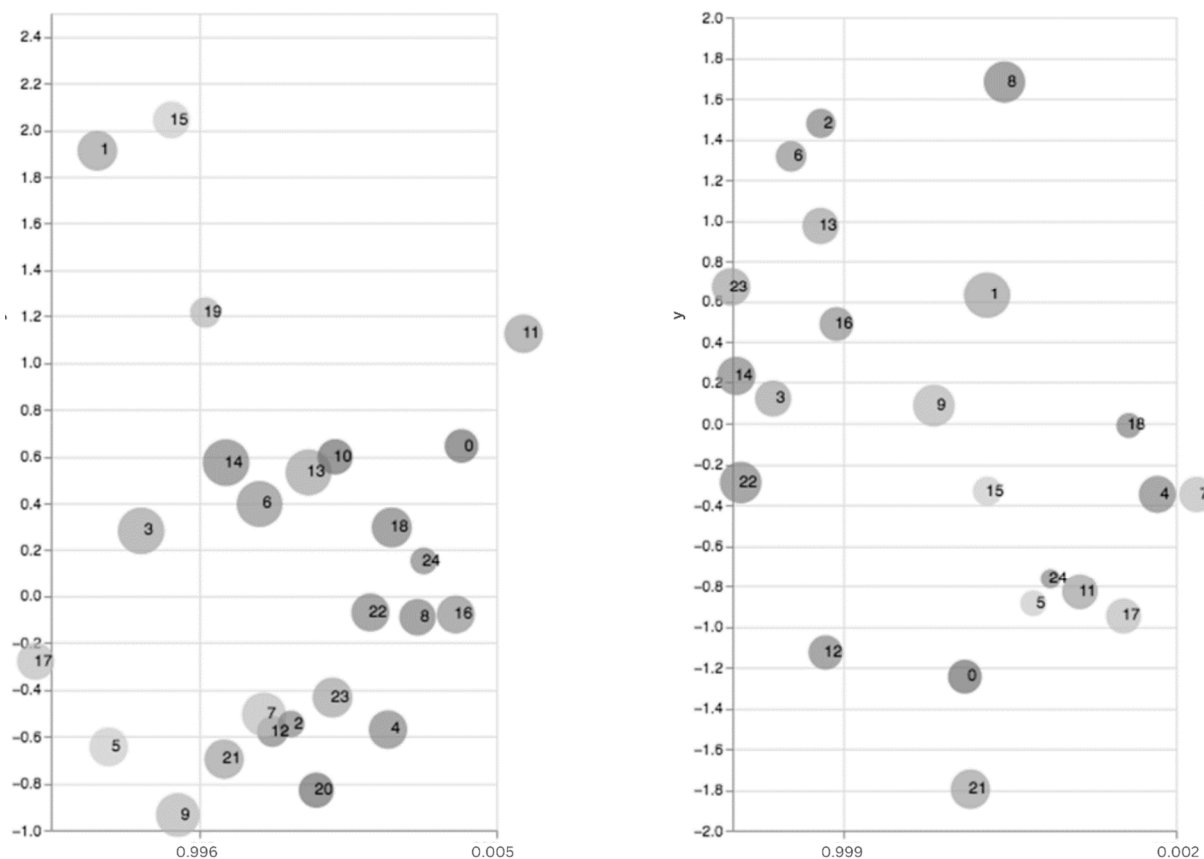


Figure 4. Topic modeling results for BERT-Base (left) and TourBERT (right).

Table 3. Topic words for 25 topics produced with BERT-Base vectors.

Topic	Words
0	fashion, sleeve, shoulder, flash, flash photography, photography, street, street fashion, smile, hair, neck, eyewear, eyebrow, happy, sky
1	shades, tints, tints shades, plant, black, sky, shirt, bicycle, photography, white, font, sleeve, wood, building, automotive
2	sky, nature, water, landscape, plant, natural, cloud, people, tree, people nature, natural landscape, water sky, happy, cloud sky, azure
3	automotive, vehicle, sky, tire, font, plant, landscape, design, wood, art, building, rectangle, cloud, water, lighting
4	plant, natural, water, landscape, natural landscape, sky, ecoregion, cloud, tree, mountain, nature, cloud sky, highland, community, plant community
5	water, landforms, sky, coastal, coastal oceanic, oceanic, oceanic landforms, landscape, cloud, natural, beach, water sky, natural landscape, azure, plant
6	people, nature, sky, smile, people nature, sunglasses, flash, flash photography, photography, water, sleeve, care, vision, vision care, eyewear
7	landscape, sky, plant, cloud, natural, natural landscape, water, tree, building, nature, cloud sky, mountain, vehicle, people, blue
8	water, sky, cloud, landscape, plant, natural, natural landscape, resources, water resources, building, tree, mountain, cloud sky, water sky, nature
9	landscape, plant, natural, sky, water, natural landscape, nature, cloud, tree, grass, people, people nature, cloud sky, sky plant, wood
10	fashion, happy, sky, people, nature, photography, flash, flash photography, eyewear, smile, people nature, care, vision, vision care, plant

Table 3. *Cont.*

Topic	Words
11	plant, sky, water, natural, landscape, ecoregion, tree, natural landscape, cloud, photography, fashion, flash, flash photography, smile, happy
12	plant, natural, landscape, water, natural landscape, sky, tree, dog, nature, grass, cloud, terrestrial, wood, people, landforms
13	building, sky, plant, window, vehicle, facade, tree, wood, design, house, automotive, tire, cloud, road, city
14	vehicle, automotive, sky, building, plant, tire, font, design, art, window, cloud, tree, wood, rectangle, lighting
15	plant, shades, tints, tints shades, sky, wood, black, fashion, bicycle, photography, rectangle, people, white, building, font
16	plant, water, natural, sky, landscape, natural landscape, cloud, ecoregion, mountain, tree, cloud sky, community, plant community, resources, water resources
17	landscape, plant, water, sky, natural, natural landscape, shades, tints, tints shades, tree, cloud, landforms, wood, coastal, coastal oceanic
18	fashion, sleeve, flash, flash photography, photography, street, street fashion, lip, shoulder, eyelash, eyebrow, smile, hairstyle, sky, neck
19	water, sky, equipment, cloud, equipment supplies, supplies, boating, boating equipment, boats, boats boating, landforms, boat, watercraft, coastal, coastal oceanic
20	water, landscape, natural, plant, sky, cloud, natural landscape, mountain, tree, nature, cloud sky, azure, highland, resources, water resources
21	plant, water, sky, nature, landscape, natural, cloud, tree, people, natural landscape, people nature, grass, cloud sky, mountain, building
22	sky, plant, cloud, water, landscape, building, natural, tree, natural landscape, mountain, cloud sky, window, nature, travel, road
23	plant, natural, sky, landscape, water, natural landscape, tree, cloud, nature, terrestrial, terrestrial plant, flower, grass, petal, wood
24	food, sky, cuisine, ingredient, recipe, tableware, dish, food tableware, ingredient recipe, water, tableware ingredient, staple, staple food, plate, produce

Table 4. Topic words for 25 topics produced with TourBERT vectors.

Topic	Words
0	plant, sky, tree, building, road, landscape, wood, cloud, road surface, surface, grass, window, sky plant, leisure, water
1	diving, underwater, water, fluid, marine, equipment, biology, marine biology, organism, fish, water underwater, liquid, diving equipment, underwater diving, blue
2	beach, people, water, sky, people beach, cloud, nature, people nature, water sky, azure, happy, travel, beach people, coastal, coastal oceanic
3	landscape, mountain, natural, sky, cloud, natural landscape, plant, slope, tree, cloud sky, highland, snow, sky mountain, terrain, sky plant
4	font, art, arts, event, rectangle, brand, design, pattern, graphics, photography, happy, painting, magenta, logo, visual
5	building, sky, window, facade, tower, design, urban, city, cloud, urban design, plant, sky building, road, house, building window
6	water, sky, afterglow, cloud, dusk, atmosphere, landscape, natural, natural landscape, sky atmosphere, cloud sky, sunlight, sunset, water sky, tree
7	tableware, drinkware, table, bottle, cup, dishware, food, glass, wood, plant, furniture, device, stemware, kitchen, wine
8	people, nature, sky, people nature, flash, flash photography, photography, happy, water, smile, plant, cloud, leg, gesture, tree
9	water, sky, equipment, boat, watercraft, cloud, vehicle, lake, supplies, boating, boating equipment, boats, boats boating, equipment supplies, water sky
10	care, vision, vision care, sunglasses, sleeve, eyewear, goggles, glasses, sky, dress, fashion, smile, shirt, flash, flash photography

Table 4. Cont.

Topic	Words
11	automotive, vehicle, tire, bicycle, wheel, motor, motor vehicle, automotive tire, vehicle automotive, sky, lighting, automotive lighting, car, plant, tire wheel
12	plant, landscape, natural, natural landscape, sky, tree, nature, grass, community, plant community, cloud, people, people nature, water, sky plant
13	sky, water, cloud, landscape, natural, atmosphere, cloud sky, blue, natural landscape, azure, plant, nature, tree, horizon, sunlight
14	water, natural, landscape, sky, natural landscape, cloud, plant, nature, mountain, resources, water resources, ecoregion, tree, cloud sky, water sky
15	temple, sky, building, architecture, plant, facade, city, cloud, art, travel, tree, leisure, sculpture, world, monument
16	nature, plant, people nature, people, sky, happy, tree, landscape, cloud, natural, water, grass, natural landscape, travel, leisure
17	wood, design, building, rectangle, interior, interior design, window, shades, tints, tints shades, property, font, furniture, flooring, plant
18	food, cuisine, ingredient, tableware, recipe, dish, food tableware, ingredient recipe, produce, staple, staple food, cuisine dish, tableware ingredient, plate, cake
19	fashion, street, street fashion, sleeve, eyewear, flash, flash photography, photography, shirt, happy, waist, smile, dress, design, shoe
20	lip, eyebrow, eyelash, smile, hair, chin, shoulder, skin, nose, forehead, hairstyle, neck, eye, lip chin, facial
21	plant, flower, tree, terrestrial, twig, landscape, terrestrial plant, natural, petal, natural landscape, branch, grass, wood, sky, flowering
22	water, natural, plant, landscape, landforms, natural landscape, fluvial, fluvial landforms, landforms streams, streams, resources, water resources, sky, watercourse, water water
23	water, landscape, landforms, natural, sky, coastal, coastal oceanic, oceanic, oceanic landforms, cloud, natural landscape, water sky, azure, resources, water resources
24	dog, plant, animal, carnivore, breed, dog breed, fawn, sky, terrestrial, working, working animal, companion, companion dog, collar, grass

Although the hashtag *#wanderlust* may lead one to think of photos that, to some extent or another, contain natural landscapes, the topic model produced with TourBERT vectors was able to identify distinct topics like “underwater world” (topic 1), “beach activities” (topic 2), “food and drink” (topic 7), “vehicle” (topic 11), or “animals” (topic 24). An attempt to find similarly grouped clusters for the BERT-Base model did not result in such success since nearly every topic includes landscape descriptions. While several distinct topics were indeed found by the model, the majority of them contain mixed concepts, each one including terms describing nature or landscapes.

For better visibility and to gain a better understanding of the quality and distinction of the topics, another visualization for each of the two topic models was produced, as can be seen in Figures 5 and 6. Each figure contains a table, with the first column presenting words for a given topic (see Tables 3 and 4) and all subsequent columns depicting the top 10 most similar samples, i.e., photos for that topic.

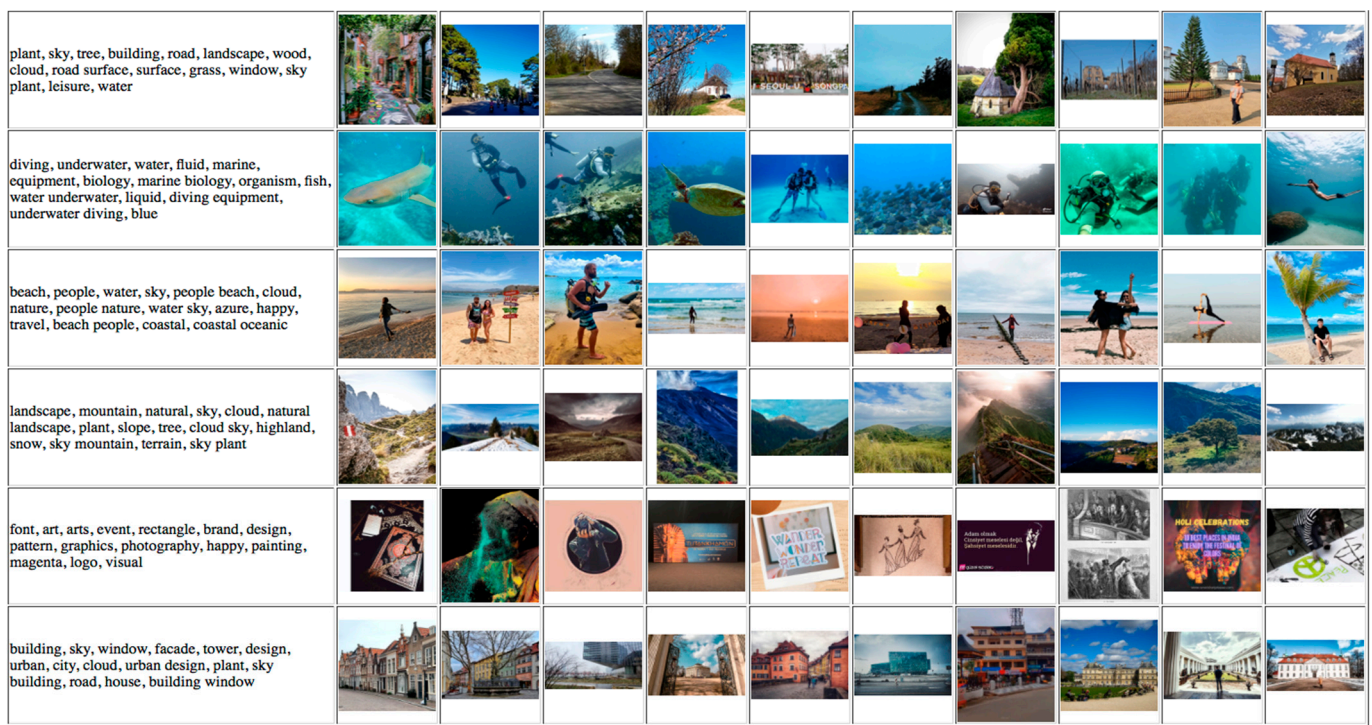


Figure 5. The first six topics, with their respective cluster words and top 10 most similar images, produced by the K-means model with TourBERT vectors.

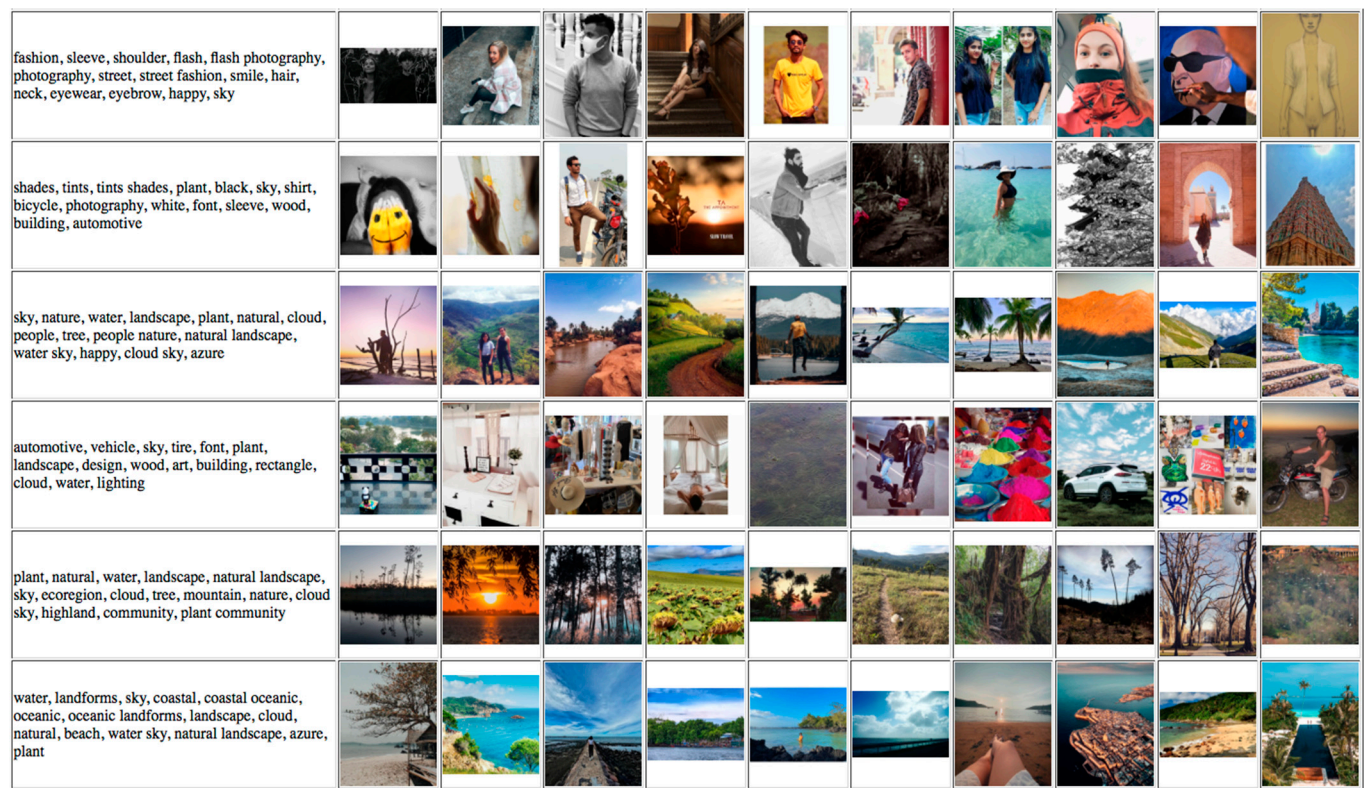


Figure 6. The first six topics, with their respective cluster words and top 10 most similar images, produced by the K-means model with BERT-Base vectors.

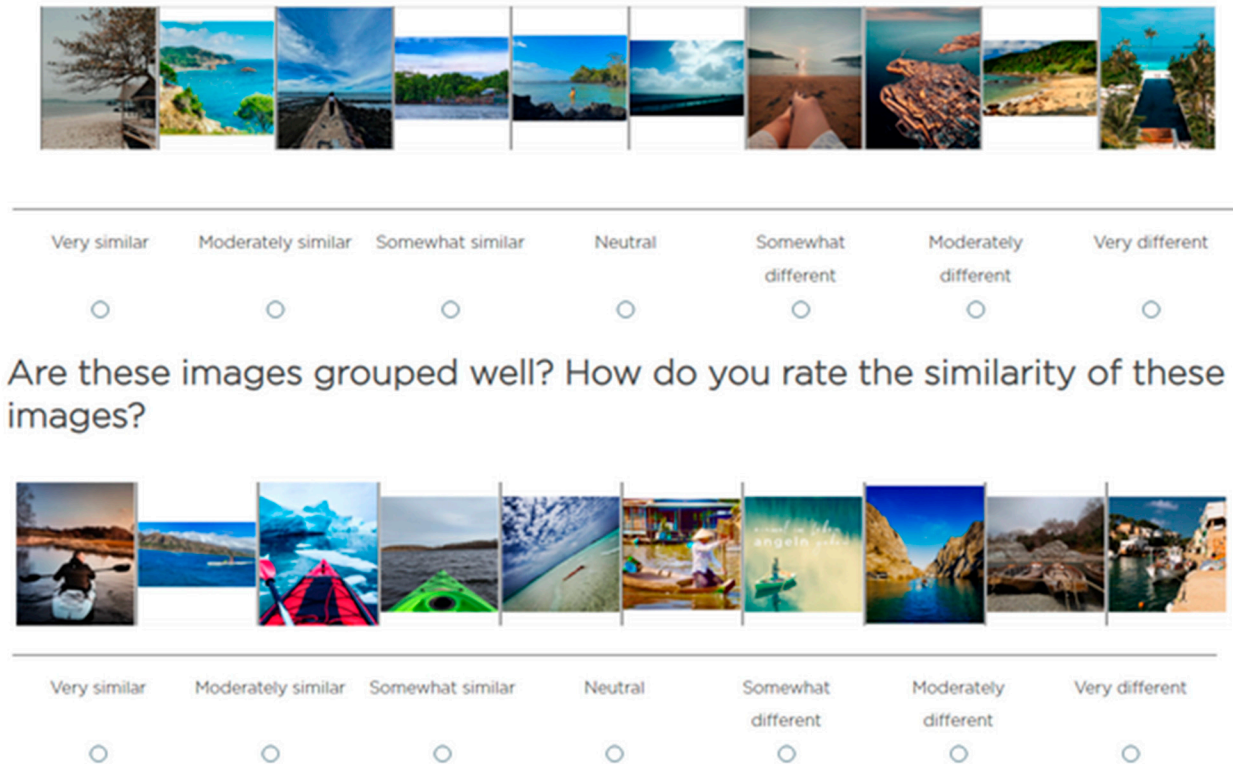
When inspecting the results from both models, it becomes apparent that the clusters created through TourBERT are much more homogenous within the clusters themselves

and quite heterogeneous across clusters. On the other hand, those generated by BERT-Base occasionally include photos that are relatively dissimilar to each other despite belonging to the same topic, such as in topic 3.

3.2.4. Unsupervised Evaluation: User Study

To further investigate the quality of each topic produced by the abovementioned models and prove the assumptions made thus far, a user study was conducted on the same set of images and annotations to statistically evaluate the results. First, a set of the 10 most similar photos for each of the 25 clusters produced by BERT-Base and TourBERT was created. Thereafter, users were asked to evaluate the similarity of the photos within each of the 50 clusters using a seven-point Likert scale, with possible answers ranging from “very similar” to “very different” (see Figure 7). Similar to measuring the intercoder reliability in qualitative studies, this evaluation approach allowed for an intersubjective perception of the quality of the clusters. Throughout this process, the image clusters were shown to the participants in a rotating manner, i.e., alternating randomly.

Are these images grouped well? How do you rate the similarity of these images?



Very similar Moderately similar Somewhat similar Neutral Somewhat different Moderately different Very different

Are these images grouped well? How do you rate the similarity of these images?

Very similar Moderately similar Somewhat similar Neutral Somewhat different Moderately different Very different

Figure 7. Two examples of image clusters shown to the participants.

To investigate this study’s results, a pairwise t-test was performed with SPSS, the results of which are presented in Table 5 below. The coding ranged from 1—very similar to 7—very different, with the mean values being 3.75 and 2.5 for BERT-Base and TourBERT, respectively, at a highly significant level (Sig. two-sided = 0.000). Effect size was measured with Cohen’s d, yielding a medium-level effect of 0.517.

Table 5. Results of the paired *t*-test for samples mean comparison for TourBERT and BERT-Base models.

Paired Samples Statistics							
Pair 1		Mean	N	Std. Deviation		Std. Error Mean	
	BERT	3.7759	82	0.71655		0.07913	
	TourBERT	2.5239	82	0.61724		0.06816	
Paired Sample Test							
		Mean	SD	Std. EM	t	df	Sig. (2-tailed)
Pair 1	BERT— TourBERT	1.252	0.51773	0.0571	21.898	81	0.000
Paired Samples Effect Sizes							
				Standardizer	Point Estimate	95% Confidence Interval	
						Lower	Upper
Pair 1	BERT— TourBERT	Cohen’s d		0.51773	2.418	1.986	2.846
		Hedges’ correction		0.52015	2.407	1.977	2.833

From the results above, it can be concluded that the similarity between the annotated images was perceived significantly better with TourBERT than with BERT-Base.

3.2.5. Unsupervised Evaluation: Synonyms Search

Assuming that BERT-Base, due to the fact that it had been trained on a generic corpus, would achieve more generic results than the TourBERT model, which had been trained on a tourism-specific corpus, it was hypothesized that a similarity search of tourism-related terms would lead to better results with TourBERT than with BERT-Base. Therefore, with the help of a tourism-domain expert, words containing multiple semantic meanings in general as well as tourism-specific contexts were selected. For example, the word “transfer” has multiple meanings and is usually associated with “transformation”, “transplantation”, and so on; however, from a tourist’s perspective, associations such as “taxi”, “pick up”, or “hotel transfer” might come to mind. The output of the top eight most similar words for each term can be seen in Tables 6 and 7 for both BERT-Base and TourBERT alike.

Table 6. Synonyms search with BERT-Base.

Authenticity	Experience	Entrance	Attraction	Ticket	Destination	Guide	Transfer	Sightseeing	Service
legitimacy	teach	shelter	attractions	tickets	dying	companion	recovery	trees	vessel
sincerity	heal	entrances	restaurant	fare	choice	entry	exchange	fireworks	authority
competence	communicate	archway	hotel	fares	lame	visit	imaging	shops	headquarters
authorship	consume	gate	exhibit	card	address	database	restoring	pacing	facility
flexibility	learn	roof	pavilion	trains	exit	forum	sale	comedy	workshop
integrity	eat	causeway	nightclub	bus	partner	workshop	comparison	prostitutes	circulation
conscience	consider	tenants	mall	metro	correction	access	recovering	sidewalk	companion
characterization	experiences	exit	ballroom	freight	priorities	google	screening	nights	operation

Table 7. Synonyms search with TourBERT.

Authenticity	Experience	Entrance	Attraction	Ticket	Destination	Guide	Transfer	Sightseeing	Service
uniqueness	experinece	entry	destination	tickets	spot	##guide	transfers	exploring	sevice
ambience	expereince	enterance	feature	entry	attraction	guides	transport	sights	services
originality	experiance	admittance	landmark	entrance	place	tourguide	pickup	attractions	staff
intimacy	adventure	admission	place	wristband	point	guid	transportation	exploration	personnel
charm	experiences	ticket	institution	admission	itinerary	driver	journey	nightlife	hospitality
accuracy	enjoyment	fee	museum	fee	hotspot	interpreter	limousine	hiking	personel
flare	opportunity	carpark	spot	pass	venture	guiding	shuttle	outings	frontdesk
warmth	expere	payment	site	tix	hangout	narrator	pickups	excursions	housekeeping

From a technical perspective, the native implementation of BERT does not allow for the querying of most similar words since, unlike Word2Vec or FastText models, BERT does not contain static vectors but, rather, produces them dynamically. As a result, it can output two completely different vectors for the same word based on the context it was mentioned in. As the intention is still to compare words as standalone context-independent units, an algorithm that enables any BERT-like model to query its vocabulary in order to find the most similar words was constructed. The algorithm works as follows: For the first step, pairwise similarities between all the words in BERT's vocabulary were computed resulting in a $30,522 \times 30,522$ matrix. Then, using the KDTree algorithm from Python's Sklearn library, a search index was built on that matrix, which allows for fast querying.

When comparing synonyms produced by BERT-Base and TourBERT, one can see that TourBERT captures the tourism-specific meaning of a given word almost perfectly. On the contrary, BERT-Base captures a more generic meaning of the same word. For example, TourBERT associates the word "ticket" with "entrance" and "wristband", whereas BERT-Base considers the same word in the scope of public transport, presenting words like "trains", "bus", and "metro". To provide another example, the word "destination" is associated via the BERT-Base model with words such as "dying", "choice", "lame", and "address", whereas TourBERT outputs "spot", "attraction", "place", and other words that are closely related to "destination" in a tourism context.

4. Conclusions

In tourism research as well as in the tourism industry, the automatic analysis of texts is becoming increasingly important. Language models are needed to perform a variety of downstream tasks such as topic modeling, text classification, entity recognition, sentiment analysis, or information extraction. However, it has been shown that the quality of the domain-specific use of pre-trained models depends significantly on the training corpus itself. While optimized language models have already been developed for business- and scientific domains, such as the financial [13], medical [14], or biological [15] sectors, this has yet to be the case for tourism. Therefore, the aim of this study was to optimize the most important and widely used language model to date, BERT, for tourism-specific applications. By means of five different evaluation tasks, the successful completion of all tasks could be demonstrated, proving the applicability and performance of TourBERT for tourism contexts. TourBERT outperformed BERT-Base in all domain-specific tasks and thus represents a suitable language model for academia and the tourism industry. This study further contributes to the discussion of the importance of domain-specific language models from a theoretical perspective, while, from a methodological point of view, it provides detailed insights into the development and training of TourBERT. As a result, this study can also be seen as a guide on how to train and evaluate BERT models for other domains. The practical contribution lies in making TourBERT available to the open-source community: The model is hosted on the Hugging Face Model Hub and accessible via <https://huggingface.co/veroman/TourBERT> (accessed on 23 May 2022). TourBERT is thus freely accessible and ready to use for tourism-specific NLP tasks. Although an attempt was made to ensure that the training corpus was as multi-layered as possible and that the intercultural dimension, a very important aspect for tourism, was taken into account, an even larger training corpus would most likely lead to increased performance rates. In particular, the inclusion of scientific texts would be useful at this point in order to better analyze texts, such as scientific books and papers, in the context of tourism.

Author Contributions: Conceptualization, V.A. and R.E.; methodology, V.A. and R.E.; evaluation, V.A. and R.E.; writing V.A. and R.E. All authors have contributed equally. All authors have read and agreed to the published version of the manuscript.

Funding: This project was carried out without funding.

Data Availability Statement: We publicly release the TourBERT model which is available on Hugging Face Model Hub and is accessible through <https://huggingface.co/veroman/TourBERT> (accessed on 23 May 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Doolin, B.; Burgess, L.; Cooper, J. Evaluating the use of the Web for tourism marketing: A case study from New Zealand. *Tour. Manag.* **2002**, *23*, 557–561. [CrossRef]
2. Yu, J.; Egger, R. Tourist Experiences at Overcrowded Attractions: A Text Analytics Approach. In *Information and Communication Technologies in Tourism 2021*; Springer: Cham, Switzerland, 2021; pp. 231–243.
3. Daxböck, J.; Dulbecco, M.L.; Kursite, S.; Nilsen, T.K.; Rus, A.D.; Yu, J.; Egger, R. The Implicit and Explicit Motivations of Tourist Behaviour in Sharing Travel Photographs on Instagram: A Path and Cluster Analysis. In *Information and Communication Technologies in Tourism 2021*; Springer: Cham, Switzerland, 2021; pp. 244–255.
4. Saraiva, J.P.D.P.M. Web 2.0 in restaurants: Insights regarding TripAdvisor’s use in Lisbon. Doctoral Dissertation, Universidade Catolica Portuguesa, Lisboa, Portugal, 2013.
5. Egger, R.; Gokce, E. Natural Language Processing: An Introduction. In *Applied Data Science in Tourism. Interdisciplinary Approaches, Methodologies and Applications*; Egger, R., Ed.; Springer: Berlin/Heidelberg, Germany, 2022; pp. 307–334.
6. Wennker, P. Künstliche Intelligenz in der Praxis. In *Anwendung in Unternehmen und Branchen: KI wettbewerbs- und zukunftsorientiert Einsetzen*; Springer Gabler: Wiesbaden, Germany, 2020; Available online: <https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=6326361> (accessed on 23 May 2022).
7. Poon, A. *Tourism, Technology and Competitive Strategies*; CAB International: Wallingford, UK, 1993.
8. Egger, R. Text Representations and Word Embeddings. Vectorizing Textual Data. In *Applied Data Science in Tourism. Interdisciplinary Approaches, Methodologies and Applications*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 335–361.
9. Tenney, I.; Dipanjan, D.; Pavlick, E. BERT rediscovers the classical NLP pipeline. *arXiv* **2019**, arXiv:1905.05950.
10. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv* **2019**, arXiv:1909.11942.
11. Edwards, A.; Camacho-Collados, J.; De Ribaupierre, H.; Preece, A. Go simple and pre-train on domain-specific corpora: On the role of training data for text classification. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; pp. 5522–5529.
12. Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; Smith, N.A. Don’t stop pretraining: Adapt language models to domains and tasks. *arXiv* **2020**, arXiv:2004.10964.
13. Araci, D. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv* **2019**, arXiv:1908.10063.
14. Alsentzer, E.; Murphy, J.R.; Boag, W.; Weng, W.H.; Jin, D.; Naumann, T.; McDermott, M. Publicly available clinical BERT embeddings. *arXiv* **2019**, arXiv:1904.03323.
15. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [CrossRef] [PubMed]
16. Beltagy, I.; Lo, K.; Cohan, A. Scibert: A pretrained language model for scientific text. *arXiv* **2019**, arXiv:1903.10676.
17. Avishek Garain. Hotel Reviews from around the world with Sentiment Values and Review Ratings in different Categories for Natural Language Processing. IEEE Dataport. Available online: <https://ieee-dataport.org/documents/hotel-reviews-around-world-sentiment-values-and-review-ratings-different-categories> (accessed on 22 April 2020).
18. Liu, J. 515K Hotel Reviews Data in Europe. 2019. Available online: <https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe> (accessed on 2 June 2021).