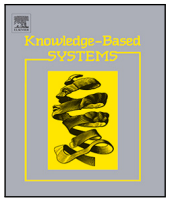
Knowledge-Based Systems 301 (2024) 112293

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Customer service chatbot enhancement with attention-based transfer learning

Jordan J. Bird *, Ahmad Lotfi

Department of Computer Science at Nottingham Trent University, Nottingham, United Kingdom

ARTICLE INFO

Keywords:
Chatbot
Customer service robotics
Social robotics
Human–robot interaction
Natural language processing

ABSTRACT

Customer service is an important and expensive aspect of business, often being the largest department in most companies. With growing societal acceptance and increasing cost efficiency due to mass production, service robots are beginning to cross from the industrial domain to the social domain. Currently, customer service robots tend to be digital and emulate social interactions through on-screen text, but state-of-the-art research points towards physical robots soon providing customer service in person. This article explores the feasibility of Transfer Learning different customer service domains to improve chatbot models. In our proposed approach, transfer learning-based chatbot models are initially assigned to learn one domain from an initial random weight distribution. Each model is then tasked with learning another domain by transferring knowledge from the previous domains. To evaluate our approach, a range of 19 companies from domains such as e-Commerce, telecommunications, and technology are selected through social interaction with X (formerly Twitter) customer support accounts. The results show that the majority of models are improved when transferring knowledge from at least one other domain, particularly those more data-scarce than others. General language transfer learning is observed, as well as higher-level transfer of similar domain knowledge. For each of the 19 domains, the Wilcoxon signed-rank test suggests that 16 have statistically significant distributions between transfer and non-transfer learning. Finally, feasibility is explored for the deployment of chatbot models to physical robot platforms including "Pepper", a semi-humanoid robot manufactured by SoftBank Robotics, and "Temi", a personal assistant robot.

1. Introduction

The growing acceptance of autonomous technology within our daily lives seems to have set us on an inevitable path towards society being aided by physical robots in a variety of different situations. The majority of robots that are used today exist within industrial work environments such as manufacturing [1] and assembly [2], as well as exploration of hazardous environments [3]. Given the societal acceptance of robots and their improvements, we can expect to find robots (or autonomous systems) of a more social nature helping in customer-facing environments such as customer service roles. This would not only help the customer but also the organisation too. In the UK alone, it has been noted that some consumers often wait up to 30 min in a queue before being able to speak to a representative [4]. Automating some of these processes with Natural Language Processing (NLP) systems which can understand issues would reduce a customer's waiting time and reduce pressure on the organisation by either solving the problem and giving advice autonomously or gathering enough useful information during the conversation which can be passed on to a human being who can then solve the issue more efficiently.

Providing customer service is an important and expensive aspect of business [5], often being the largest department in most companies. Many problems are easy to solve, for example, a forgotten password, and yet customer service representatives spend much of their important time on such issues [6]. Indeed, the conversations between representatives and customers on these issues are unique and nuanced, with exchanges dependent on prior information, vocabulary, etc. Based on this, simply regurgitating the same instructions, similarly to that of which a forgotten password button on a website's login form may produce, is subpar compared to the customer service experience. Instead, this work proposes the use of attention-based chatbots, where models learn to tune attention to prior exchanges in the conversation before producing the next. There is a considerably large amount of conversation data available on social networks where social exchanges occur between customers and customer service representatives, and thus could provide useful starting points for training chatbots that perform similar tasks during interaction with humans. Transformer-based approaches have emerged as the state-of-the-art in chatbot technologies,

* Corresponding author.
E-mail addresses: jordan.bird@ntu.ac.uk (J.J. Bird), ahmad.lotfi@ntu.ac.uk (A. Lotfi).

https://doi.org/10.1016/j.knosys.2024.112293
Received 27 February 2023; Received in revised form 16 July 2024; Accepted 28 July 2024
Available online 31 July 2024

with studies such as those performed by Roumeliotis et al. [7] showing their significant impact on the field. Along with positive impact on experimental results, it is noted that the *Jack of All Trades* approach of Large Language Models require complex computational resources and thus hold the potential to make serious environmental impact [8]. For example, the chatbots proposed in [7] are based on ChatGPT [?] and LLaMa [?], and therefore while having the ability to provide customer service, still retain much of the knowledge of the dataset originally contained, such as programming languages.

The aim of this article is to explore if it is possible to perform Transfer Learning of language between different domains to improve chatbot models; models which would form part of a framework to provide customer service to human beings through conversation and advice. To give an example, although the problems solved by an online retailer (e.g. Amazon) and a supermarket chain (e.g. Tesco) customer service representatives may differ greatly, there may exist useful knowledge at the lower level, i.e., vocabulary and language, as well as at the higher level, i.e., logical processes and problem-solving, which would aid in improving language models if it could be transferred.

The scientific contributions of this article are as follows:

(a) Experiments show that deep learning chatbots can be improved by transferring knowledge from other chatbots also trained to provide customer support.
(b) Low level transfer of knowledge occurs in terms of English language, and higher level transfer learning occurs between domains facing similar customer support requests.
(c) Feasibility studies highlight several difficulties when implementing chatbots on physical robots, and solutions are proposed to overcome them to enable a Human–Robot Interaction approach to customer support in person.

The remainder of this article is presented as follows, Section 2 reviews the scientific state-of-the-art of customer service chatbots as well as describing the theory behind attention-based modelling. Following this, Section 3 describes the method followed by the experiments in this study, before the results are presented and discussed in Section 4. The results section also provides exploration of the models, providing interesting examples of interaction with the chatbots and discussion. Feasibility observations are made and chatbots are also implemented on physical robots. Finally, Section 5 describes the future work that could be performed based on the findings of the experiments in this article before concluding the study.

2. Background

This section explores the background of the field and state of the art studies related to the experiments in this work. Chatbots are autonomous systems which aim to converse with a human user. Falling into two main categories; open domain (general conversation) or closed domain (aimed at solving a specific task) [9]. Customer service chatbots aim to mimic a human being and help solve customer queries and issues. For example, SuperAgent [10] is a question-answering chatbot that can mine data from web pages and provide information e.g. answering "does it come with the pen?" with "yes it does".

Alan Turing's question, "can machines think?", led to what we now know as the *Imitation Game* or *Turing Test*. This was the proposal that if, under specific conditions, a machine could mimic a human being, then the computer can be said to possess Artificial Intelligence (AI) [11]. Turing originally suggested the use of a teleprinter (specific conditions), but modern systems are tested online [12]. Deep learning is often used for customer service chatbots in contemporary studies [13], such as Long Short-Term Memory (LSTM) sequence-to-sequence learning in [14], which led to an improvement over information retrieval for social media-based services. Ranoliya et al. [15] proposed a more classical Extensible Markup Language-based approach for University-related queries, achieving impressive results for an automatic question-answering problem in educational support. Often, data received from customers is further analysed with sentiment analysis using either scoring and polarity [16,17] or classification [18,19]. In this study, the sequence of inputs and attention masking are considered, and so, although not explicitly scoring or classifying sentiment, valence data still exist within the dataset.

Several methods have been proposed to improve chatbots. For example, recent work on the combination of reinforcement learning with human–robot interaction showed further improvement by learning from experience [20]. Reinforcement learning has also been suggested to improve chatbots with ensemble learning strategies [21]. Ensemble chatbots have been used in studies in medicine [22], mental health care [23], and education [24]. In another study, the addition of synthetic data with transformers was shown to improve the ability of chatbots when an attention-based model was used to generate training data to create additional examples [25]. In a similar line of questioning to this work, the authors in [26] suggested the possibility of transferring knowledge between domains for the improvement of chatbots. In the aforementioned, reinforcement learning improved when transferring knowledge between restaurant booking, movie booking, and tourist information. These results are particularly exciting for tourist information since they differ from the other two booking domains, and yet improvements were still achieved. Similar goal-orientated experiments were also performed in [27] with attention mechanisms instead of transfer learning as previously described. Transfer learning has also found success in question-answering systems [28]. In [7], researchers propose to fine-tune transfer learn several transformer architectures (including LLMs) towards providing customer service and support. The findings showed that ChatGPT version *gpt-3.5-turbo-1106* achieved an accuracy of 64.24% on the dataset. Similarly, [29] propose the use of transformer-based chatbots for customer support in the Telecoms industry. The researchers found that ChatGPT 4 experienced an acceptance rate of 68.7% for general industry, and 51.3% for vertical industries, that is, those that have a more specialised customer audience.

The natural interaction has been shown to be important in human–chatbot interaction. For example, Xie et al. [30] explore how user opinions about competence, entertainment, social presence, as well as the general quality of service, are improved when the chatbot produces humour within its responses. Similarly, the findings of [31] showed that customer service chatbots were considered more trustworthy when a human-in-the-loop approach was used, which was less costly than human-only customer support.

The acceptance of physical robots in customer service is growing, with many studies suggesting that customers prefer robots with simulated emotional feelings [32]. Related research has also suggested that humanoid robots are currently the most emotionally acceptable form, with the current state of hyperrealistic robots remaining within the uncanny valley [33,34]. Wirtz et al. [35] propose that service employees and robots will eventually work in sync, with some tasks dominated by humans and vice versa for robots. Tangible actions were predicted to include autonomous receptionists and porters, and intangible actions were predicted to include information counters, claim processing, chatbots, etc. Indeed, the line between the two is blurred, i.e., a physical receptionist robot, if emulating a human being, would require the additional implementation of a closed domain chatbot. According to Tuomi et al. [36], there are many roles that robots can play when providing customer service. These roles are external, internal, and operational. This study focuses on the operational, specifically in support, defined by the aforementioned study as *dealing with routine tasks and freeing human employees to focus on more complex and dynamic situations*. To give a concrete example of how this would be made possible by a chatbot, consider the example given in the introduction; a forgotten password is a relatively easy task that representatives must solve many times per day, and thus this advice could instead be given

by an autonomous system, allowing the human operator to instead focus on more difficult tasks that may not be autonomously solved.

Attention-based modelling and transformers are a relatively new concept in deep learning [37]. Although the original study was only performed in 2017, there are many examples in which the approach has achieved state-of-the-art performance within several areas of NLP such as text synthesis [38,39] and autonomous question answering [40,41], which are of interest to this study in particular. Attention is calculated via a scaled dot product, where attention weights are calculated for each word in the input vector. The general approach is as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (1)$$

where the query Q is a token within the sequence (e.g. a letter or a word), keys K are vectors of the input sequence and values V are derived by querying keys. Unsupervised models receive Q, K and V from the same source and thus pay *self-attention*. For classification, K and V are derived from the source and Q is derived from the target. For example, Q could be the class that the text belongs to. The approach followed by this work follows multi-headed self-attention. Multi-headed attention, MH, is implemented simply by concatenation of multiple i attention heads h_i to form a larger network of interconnected attention for parameter matrix W^O:

$$MH(Q, K, V) = Concatenate(h_1, \ldots, h_h)W^O. \quad (2)$$

In [42], the authors argued that the increasing popularity of social humanoid robots would inevitably lead to daily use in service domains. It is interesting to note from this study that there has been a rapid rise in scientific publications on the subject since the early 2000's. Furthermore, a survey carried out during the aforementioned work found that the majority of members of the public (43%) would most like to see humanoid robots assisting in airports; however, there were only 548 responses. It is due to this need for travel support that American Air, British Airways, and Virgin Trains are included within the dataset toward engineering a robot that could aid within these domains. Niemelä et al. carried out a study on the use of Pepper robots in shopping centres [43], finding that the main expectations of customer service robots differed slightly between management, retailers, and customers. Management expected robots to mainly welcome, host events, and guide customers. Customers, on the other hand, expected that robots would mainly provide guidance, information, special discounts, and entertainment. As key performance indicators, management and retailers expected the number of customers to increase first and foremost. Customers, on the other hand, as expected, would measure success based on the valence of their experience and the quality of the information received. The study was inspired by work from Shi et al. [44], who studied the expectations of retail managers. These studies highlighted that social robotics was a promising addition to customer service, finding that two out of three shops experienced an increase in potential customers when aided by a humanoid robot. It was also noted that managers were enticed by the cheap labour and unique value enabled by social humanoid robots.

Merkle [45] performed a study by observing the satisfaction of human–robot customer service compared to human interaction. It was observed through the Likert scale that, on average, the Pepper robot would fail at a mean of 4.69 as opposed to human failure at 2.41. Taking this into account, the Likert scale on satisfaction produced results that could be considered unexpected when only considering the quality of service; the mean value for customer satisfaction when dealing with the more error-prone robot was 6.08, and 5.79 for a human frontline employee. In addition to providing unique value and cheap labour, the results of this study also suggest that customers are more forgiving when social robotics fail compared to human error. Social robots providing increased quality in customer-facing environments were also echoed by [46]; 15 experiences were related to the basic human needs of autonomy, competence, relatedness, stimulation and security, and a humanoid customer service robot deployed at a city service point was observed to both enrich the visitor experience and even nourish basic human needs. The background study reported here indicates the gap in the application of Transfer Learning in the NLP domain. Hence, the method detailed in the next section is reporting our proposed approach.

3. Method

Within this section, the methodology followed by all the experiments is detailed. Firstly, detailed information about data collection and processing are provided to derive datasets for learning. Following this, details on the implementation of chatbot models are given before the implementation on robotic platforms.

3.1. Data collection and preprocessing

A dataset of Tweets is initially retrieved from [47]. The dataset contains 3,003,124 instances of Tweets and responses to and from 19 different support accounts. Tweets that are not written in English are removed. All text is converted to the lower case and so are treated as the same token regardless of capitalisation, and then the human names and punctuation are removed from the strings. To keep profane words from the chatbot vocabulary, a list of words banned from Google autocomplete are retrieved from [48], and any data containing one or more of these terms are removed. Though this results in such terms being removed from the vocabulary and thus never understood, it also prevents the model from erroneously generating profane tokens and speaking them aloud to a customer. The details of each domain can be seen in Table 1. Table 2 shows an overview of how many accounts of each type are available within the dataset.

3.2. Machine and transfer learning

A general diagram of the transfer learning process is shown in Fig. 1. Data from each domain (customer service interactions) are presented as input for training the chatbot model. For example, if the text input was a customer query "I can't remember my password, what should I do?", the output text should be some form of instruction to lead the customer to the process for recovering or resetting their password. Generally, the goal of the model is to predict the next token, which in the context of these experiments are words. Thus, the model is tasked with generating tokens following a query until an answer is formed. Markup tokens "$\langle Q\rangle \cdots \langle /Q\rangle$" and "$\langle A\rangle \ldots \langle /A\rangle$" are used to denote the start and end of questions and answers, respectively. These tokens are used to stop the generation loop when the query is answered. The Domain n within the diagram denotes that each and every domain will have a model trained on said data. Following this, there are transfer learning experiments. The transfer learning experiments are explored for each domain as follows: First, a randomly initialised self-attention neural network is trained on a dataset of queries and responses from a customer service domain such as Amazon or Apple Support. Second, the neural network trained on the initial domain is then presented as an alternative starting weight distribution for a second domain n, such as American Airlines. That is, for each domain, the model is trained on the data from a starting random weight distribution, as well as from weights trained on all other models to benchmark for the possibility of knowledge transfer and its effect. $W1$ denotes the weights transferred, and the results are then compared. In this example, if the weight-transferred model performs better than the randomly initialised model, it is suggested that some useful knowledge has been transferred from the Amazon customer service domain to the American Airlines domain.

Fig. 2 shows an example of the methodology for two different learning processes. Firstly, classical learning, which begins with a randomly initialised weight matrix for the model as a starting point for machine

Table 1
Overview of each domain in the dataset after pre-processing and cleaning.

Domain	Type	X/Twitter Handle	Conversations	File size (KB)
Amazon	Online Retailer	@AmazonHelp	30 402	6780
American Air	Airline	@AmericanAir	9044	1903
Apple	Consumer Technology	@AppleSupport	29 689	6385
British Airways	Airline	@British_Airways	6094	1378
Chipotle	Catering	@ChipotleTweets	4350	651
Comcast	Telecommunications	@Comcastcares @XfinitySupport	5581	1246
Delta	Airline	@Delta	9163	1784
Hulu	Streaming	@Hulu_Support	5585	1208
PlayStation	Consumer Technology	@AskPlayStation	2783	507
Sainsburys	Supermarket	@Sainsburys	5188	1065
Spectrum	Telecommunications	@Ask_Spectrum	4749	1045
Spotify	Streaming	@SpotifyCares	8696	1753
Sprint	Telecommunications	@Sprintcare	3054	620
Tesco	Supermarket	@Tesco	9273	2180
TMobile	Telecommunications	@TMobileHelp	6835	1429
Uber	Travel	@Uber_Support	9166	1924
UPS	Delivery Service	@UPSHelp	3182	782
Virgin Trains	Travel	@VirginTrains	5912	1037
Xbox	Consumer Technology	@XboxSupport	5041	1007

Table 2
Overview of the total number of each type of support account within the dataset.

Type	Count
Airline	3
Catering	1
Consumer Technology	3
Delivery Service	1
Online Retailer	1
Streaming Platform	2
Supermarket	2
Telecommunications	4
Travel	2

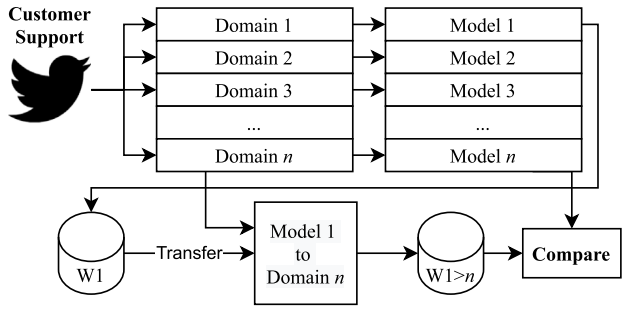


Fig. 1. General diagram of the weight transfer learning process between customer support domains. Weights, Wl, are transferred as a starting distribution before learning to provide support for a different domain.

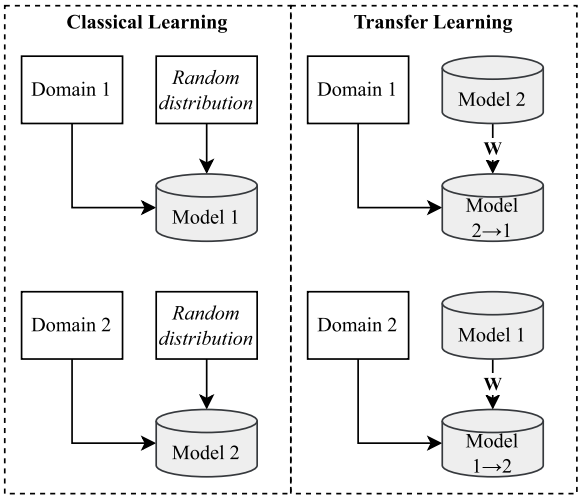


Fig. 2. Comparisons of classical learning, where a model is trained from a random distribution to a dataset, and transfer learning, where weight matrices W are repurposed for a second task.

learning. Secondly, the transfer learning process, where weight matrices, **W**, can be repurposed to another task. In this example, domains 1 and 2 share knowledge by providing initial weight distributions to each other to transfer knowledge. The diagram in Fig. 3 shows an example of tokenisation, where a customer query is transformed to input to the model, followed by token generation, which is also translated back to human-readable text.

Across all X (formerly Twitter) conversations, there are 91 967 unique tokens that constitute the universal vocabulary. The sparse categorical cross-entropy $-\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$ is used due to the large number of classes. Due to this, sparse metrics are measured in terms of loss and accuracy. In addition, the top-5 and top-10 accuracy metrics are also considered.

Fig. 4 shows the implementation process when considering robotic platforms. Since physical robotic platforms often use separate APIs (with the exception of Softbank robots), the model provides input for each specific API for text-to-speech generation. The same approach is used for platform-specific speech recognition. Effectively, input text in the form of typed commands or speech-to-text are used as input to the trained model, and a response is given, also in the form of text. This text is then communicated to the robot API for speech generation. Following the previous example input "I can't remember my password, what should I do?", the response may be "if you go to the login page on the website, click the button that says I have forgotten my password". This response is sent to the robot and is spoken aloud after text-to-speech processing.

A total of 361 chatbots are trained, 19 are trained classically from their respective datasets and 342 are trained via transfer learning (from the prior 19 to all of the others). Observations showed that training would take approximately ten days for data split validation models in terms of computational time. For this reason, data splitting (70:30) is chosen as the validation approach given that k-fold and leave-one-out strategies would take considerably more time, rendering the experiments impossible to perform. Inference speed is benchmarked by passing 1000 questions from the dataset at random to the final model topology, where an overall average tokens per second $TPS = \frac{tokens}{words}$ is calculated.

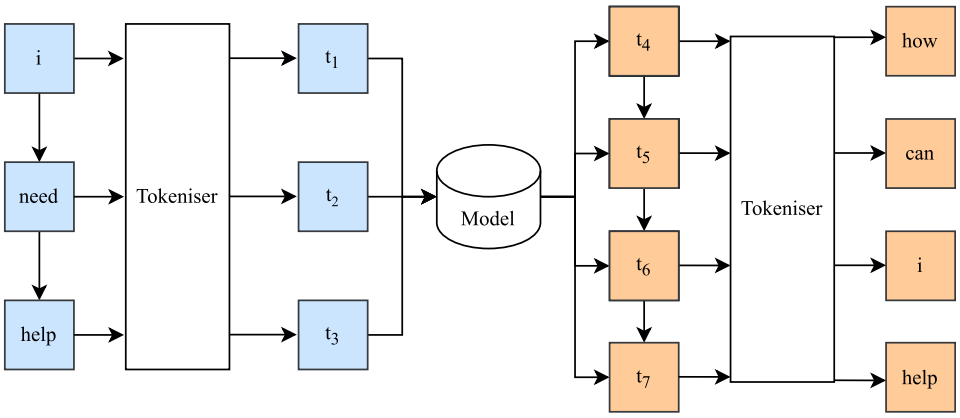


Fig. 3. Example of the process to transform the text into vocabulary tokens. Inputs (customer query) are highlighted in blue and outputs (customer service agent response) are highlighted in orange.

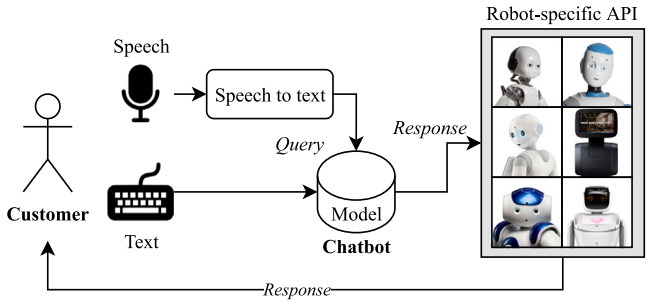


Fig. 4. Generalised diagram of the Human–Robot Interaction pipeline. Speech-to-text or text are provided as input to the model derived from the process in Fig. 1. Robot-specific APIs are used to respond to the customer.

3.3. Implementation

The chatbots are finally embedded within robot-compatible wrappers. Two robotic platforms are chosen, Temi,[1] a personal assistant robot, and Pepper,[2] a semi-humanoid robot manufactured by SoftBank Robotics which are shown in Fig. 5. These platforms were chosen for implementation due to their shared prominence in the state of the art in terms of usefulness in customer service situations [42,49–51]. A simple interface is designed for Temi, whereas interaction with the Pepper robot depends mainly on on-board head tracking and eye contact during speech recognition. Due to differences in the host server and robot operating systems, each robot required software to enable a link between the two. The Pepper robot depends entirely on Python 2.x, which in turn requires older machine learning libraries which do not support the new technology employed in attention-based token prediction. The solution to this is to bridge software on the server, wrapping robotic API calls within the more up-to-date software. Temi, on the other hand, has full support for modern libraries, but another issue was observed during speech; synchronicity was not possible for text-to-speech. In order to solve this, the robot is tasked with speaking ten Harvard sentences from the *IEEE recommended practice for speech quality measurements*[52]. Harvard sentences are chosen because phonetic sounds are balanced at the frequency with which they tend to appear in the English language. Each of the sentences is recorded and a measurement of Words Per Minute (WPM) is taken, ultimately leading to a reliable standard to command the robot to wait for further instruction while speech audio is playing.

[1] More details available from: https://www.robotemi.com/
[2] More details available from: https://www.aldebaran.com/en/pepper

Fig. 5. An image of the Temi and Pepper robots (not to scale) which provide a HRI interface for the customer service chatbot models.

All models in this article were implemented with TensorFlow and executed on a server with shared resources. The model had access to an Intel Xeon E5-2640 v4 CPU (2.4 GHz) and performed operations via CUDA on a single Nvidia Tesla M60 accelerator (with two GPUs on board); the Accelerator operated with 4096 CUDA cores and 16 GB GDDR5 VRAM. For privacy reasons, it could not be observed whether other users made use of the shared computational resources throughout. For inference speed testing, a consumer-level machine was used which had an Nvidia RTX 2080Ti GPU (4352 CUDA cores and 11 GB of GDDR6 VRAM).

4. Results

Initially, the network topology is explored. Table 3 shows the results for the tuning of the 16 different topologies, where a batch search of $\{2, 4, 8, 16\}$ attention heads and $\{64, 128, 256, 512\}$ neurons in the dense layer are compared. The validation results are relatively marginal after the 10-epoch tests on all data, and the lowest overall was that of 8 attention heads with 256 neurons (2.64). Although the differences in the results are small, this topology is chosen for the remaining chatbot experiments for simplicity.

It must be noted that the transformers explored in this work are much smaller than state-of-the-art general transformers (which are

Table 3
Validation loss during the tuning of network topologies (10 epochs on all data) for the selection of attention heads and feed-forward neurons. The selected topology is then later trained with early stopping, i.e., until no improvement is found rather than ending during learning.

		Attention heads			
		2	***4***	***8***	***16***
Dense Neurons	***64***	2.67	2.65	2.66	2.67
	128	2.68	2.65	2.65	2.67
	256	2.67	2.65	2.64	2.65
	512	2.69	2.66	2.65	2.65

Table 4
Exploration of the effects of reducing vocabulary size (10 epochs on all data).

Vocabulary size	Trainable parameters	Loss
10000	7 386 640	2.51
15000	9 951 640	2.55
20000	12 516 640	2.59
25000	15 081 640	2.63
30000	17 646 640	2.62

Table 5
Validation metrics after training on domains from an initial random weight distribution.

Dataset	Loss	Accuracy	Top-5	Top-10
Amazon	2.3	0.602	0.73	0.779
American Air	2.17	0.645	0.754	0.8
Apple Support	1.93	0.642	0.778	0.826
British Airways	2.19	0.658	0.764	0.802
Chipotle	1.17	0.808	0.877	0.901
Comcast	1.74	0.712	0.824	0.857
Delta	2	0.678	0.776	0.814
Hulu	1.84	0.712	0.81	0.844
PlayStation	0.54	0.877	0.95	0.96
Sainsburys	1.64	0.743	0.833	0.862
Spectrum	0.42	0.767	0.857	0.883
Spotify	1.57	0.74	0.831	0.862
Sprint	0.83	0.825	0.916	0.939
Tesco	2.47	0.616	0.738	0.783
TMobile	2.1	0.653	0.765	0.808
Uber	1.68	0.713	0.813	0.847
UPS	1.04	0.783	0.897	0.926
Virgin Trains	1.67	0.746	0.819	0.846
Xbox	1.57	0.761	0.842	0.87

large due to the multitude of different tasks they consider). For example, a model made up of 8 attention heads, 256 neurons, and a vocabulary of 30 000 has over 17 million trainable parameters as can be observed in Table 4 compared to GPT-2's 1.5 billion and GPT-3's 175 billion. Table 4 shows that when adding 20k extra words, and therefore over 10 million additional trainable parameters, the preliminary loss metrics do not increase drastically. Given that a difference of +0.11 is observed, a large vocabulary of 30k is chosen for these experiments; this is mainly due to the multiple domains involved to provide ample coverage to each, since the vocabulary is shared between all models to enable direct transfer learning. This size of vocabulary was also chosen given the computational limitations of the study, and larger sizes exceeded the available resources. As described previously, the global vocabulary of all data is made up of 91 967 unique tokens, with 30 000 being approximately a third of the total tokens present (32.62%).

Table 5 shows the results for training the models from an initial random weight distribution, i.e., non-transfer learning. For several of the larger datasets, the classification accuracy is relatively low. However, metrics were observed to be higher on the smaller datasets, suggesting that there may potentially be less variability. It was observed that, although this is the case, the chatbot was capable of answering queries in correct English, with few obvious mistakes (examples of interaction

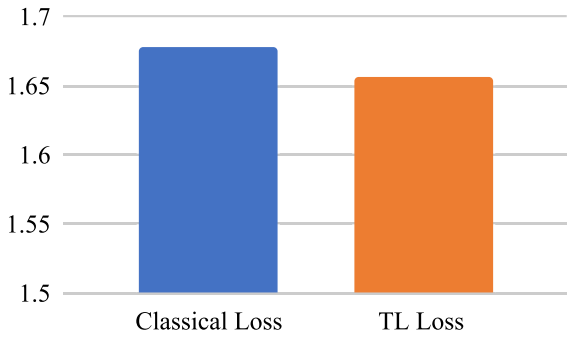


Fig. 6. Chart of the mean losses between classical and transfer learning experiments.

are presented later in this section). As expected, the top-5 and top-10 accuracies are much higher, especially in situations of low loss, showing that although accuracy may be relatively low, the correct prediction often features within the top k predictions.

The mean results can be observed in Table 6. In general, transfer learning led to lower mean losses and higher accuracies of all types; top-1, top-5, and top-10. This suggests that knowledge can be transferred between customer service domains. An overall comparison of all experiments can be found in Table 7, where it can be observed that transfer learning outperformed classical learning in most cases. In terms of loss, 13 of the 19 domains produced a lower loss when transferring knowledge. Similarly, the accuracy was higher when transfer learning for 15 of the domains, top-5 accuracy for 16 of the domains, and finally a higher top-10 accuracy for 17 of the domains. A full matrix of all individual results can be found in Appendices A–D.

Fig. 6 shows the comparison of mean losses for all classical and transfer learning experiments. A complete set of all results can be found in Appendix A. It was be observed that classification losses were reduced for 13 of the 19 domains (those that did not decrease were Amazon, Apple Support, British Airways, Uber and Xbox) when transfer learning from at least one other. It is interesting to note that the largest datasets in particular were far from benefiting from transfer learning, with the learning of the majority of smaller datasets being greatly improved by transferring from other domains. This suggests that transfer of knowledge is possibly a solution to data scarcity in the customer support chatbot problem. Of the five models that lacked improvement, British Airways and Xbox were the most interesting in particular, given that transferring from other airlines (American Air, Delta) and other console manufacturers (PlayStation) did not provide improvements. Otherwise, all domains found improvement in one or more cases when learning was transferred from other domains. Following the topology of the chatbot answering 1000 random prompts, an average tokens per second of 17.329 was observed.

Table 8 shows the observations from the Wilcoxon signed-rank test and Cohen's d effect size. Upon pair-wise interpretation via each non-transfer result compared with each transfer learning result to the same domain, it can be observed that when $\alpha = 0.05$, the null hypothesis can be rejected in the majority of cases suggesting statistically significant values. Domains such as Amazon, Apple Support, and British Airways, among others experience a Wilcoxon statistic $W = 0$, suggesting significant differences when transfer losses are higher, indicating worse performance. Chatbots within the domains of Chipotle, PlayStation, Sprint, Tesco, and UPS experience high W values, arguing that transfer learning results generally result in a lower loss value. High Cohen's D values within these domains show that there is a notable effect size, suggesting that transfer learning has a strong impact on reducing sparse categorical cross-entropy loss. In particular, this can be seen with the PlayStation chatbot, where the non-transfer loss was 0.54, but transfer learning from several domains resulted in a loss of 0.31, the most interesting of these possibly being the Xbox domain, given

Table 6
Mean observed values for each set of experiments.

Classical loss	TL loss	Classical accuracy	TL accuracy	Classical Top-5	TL Top-5	Classical Top-10	TL Top-10
1.677	**1.657**	72.01%	**72.85%**	81.97%	**82.27%**	85.31%	**85.57%**

Table 7
Comparison of the performance between classical and Transfer Learning experiments. The counts show the number of superior results out of the 19 total domains.

Lower loss		Higher accuracy		Top-5 Acc.		Top-10 Acc.	
Classical	TL	Classical	TL	Classical	TL	Classical	TL
6	13	4	15	3	16	2	17

Table 8
Results for the Wilcoxon signed-rank test and Cohen's d effect size for model losses.

Domain	Non-transfer result	Transfer mean	Wilcoxon statistic	p-value	Significant?	Cohen's d
Amazon	2.3	2.36	0	7.63E−06	Yes	−6.53
American Air	2.17	2.15	32.5	2.08E−02	Yes	0.49
Apple Support	1.93	1.99	0	7.63E−06	Yes	−3.16
British Airways	2.19	2.25	0	7.63E−06	Yes	−2.22
Chipotle	1.17	1.10	0	7.63E−06	Yes	2.41
Comcast	1.74	1.73	68	4.68E−01	No	0.19
Delta	2	2.05	4.5	5.34E−05	Yes	−1.46
Hulu	1.84	1.82	29.5	2.55E−02	Yes	0.53
PlayStation	0.54	0.34	1	1.53E−05	Yes	2.89
Sainsburys	1.64	1.64	44	2.12E−01	No	0.09
Spectrum	1.42	1.48	1	1.53E−05	Yes	−2.16
Spotify	1.57	1.55	13.5	2.58E−03	Yes	1.21
Sprint	0.83	0.65	3	3.81E−05	Yes	1.80
Tesco	2.47	2.47	30	1.39E−02	Yes	0.01
TMobile	2.1	2.07	22	4.01E−03	Yes	0.75
Uber	1.68	1.71	0	7.63E−06	Yes	−1.26
UPS	1.04	0.85	4	5.34E−05	Yes	1.70
Virgin Trains	1.67	1.64	34.5	7.45E−02	No	0.22
Xbox	1.57	1.62	0	7.63E−06	Yes	−2.19

the similarities between the videogame consoles that both companies offer support for; in this case, $W = 1.0$ showing that transfer learning tended to produce lower losses overall, and $D = -1.26$ arguing that the statistically significant values were accompanied by a large effect size.

Appendices B–D present all individual results for each domain in terms of accuracy, top-5 accuracy, and top-10 accuracy, respectively. Of the 19 domains, 15 experienced higher classification accuracy when transferring knowledge from at least one other domain. Several instances were slight increases, there were experiments that showed a much larger increase in ability when transfer learning. Three main examples of this can be seen, transfer of knowledge from Tesco to Chipotle leads accuracy to rise from 0.808 to 0.844, secondly, transferring from Tesco to Sprint leads the accuracy to rise from 0.825 to 0.883. The most interesting example, though, is when a transfer of knowledge is performed between two similar domains, Xbox to PlayStation, which causes the accuracy to increase from 0.877 to 0.93. This is particularly interesting since the problems experienced by users of these two services tend to be similar, albeit on different platforms. The UPS chatbot was improved by transferring from every other domain except for Amazon; Tesco, TMobile, and Xbox transfer learning caused the accuracy to rise from 0.783 to 0.841. A similar observation can be made from the top-k results, with transfer learning aiding in correctly predicting the next object in the sequence to be contained within the top 5 and 10 predictions.

Although losses were not reduced and errors were considered more severe, accuracy was improved for models such as the Xbox Support chatbot when transfer learning. For example, rising from 0.761 to 0.772 when transferring weights from the Sainsburys and Tesco models. It is worth noting that these two biggest improvements came from the transfer of knowledge from two large British retailers. The best model for the PlayStation chatbot in terms of accuracy is observed to be that which is transferred from Xbox (0.877 to 0.93) which is particularly interesting due to their similar domains. This was also the case for the top-5 and top-10 accuracy observations, although other models also achieved this score (0.96 to 0.981). Results such as this suggest that in addition to transferring knowledge of the English language and conversations in general, that there is also the possibility that specific domain knowledge can be carried over as well, in some cases. To expand on the accuracy metrics, an exploration of the top k predictions is performed. A similar pattern is observed to the prior experiments, the majority of domains experience an increase in metrics when the transfer of knowledge is performed from at least one other domain. This is the case for 16/19 domains with regard to $k = 5$ and 17/19 for $k = 10$, the two domains that did not experience increases stay the same between the two, while transfer learning to Tesco when $k = 10$ is improved by transfer learning from American Air (0.783 to 0.8). A particularly large increase that stands out from many of the experiments is when learning is transferred from the Apple Support chatbot to the Spectrum chatbot, with $k = 5$ accuracy rising from 0.857 to 0.884.

Examples of communication with the chatbot can be seen in Fig. 7 (Amazon) and Fig. 8 (Tesco). Punctuation has been inferred manually for readability. Note that, although the same query has been asked, the response is tailored to the input. This can be seen especially in the Amazon examples, where a friendly tone in the first example soon changes to being apologetic and understanding once the input is edited to express negative emotions towards the situation. The Tesco example shows a different problem, since losing a membership card is a concrete problem whereas an order not arriving yet may be a non-issue. An interesting nuance can be found in the responses if they are compared, in the first instance, "I lost my clubcard". is relatively emotionless and states the issue, with the chatbot responding that support understands how inconvenient the issue must be. In the second, more polite input, "Hi,

Input: "My package hasn't arrived."
Output: "Oh no! Have we missed the delivery date? Can you share the order details?"

Input: "I am so angry that my package hasn't arrived!"
Output: "I'm sorry for the frustration, we'd love to help. Kindly share the details and we'll get in touch with you."

Fig. 7. Example of two interactions with the non-transfer chatbot trained on Amazon data. Note that although the problem is the same, the response changes due to emotion. The chatbot recognises the negative valence and high arousal in the second message and apologises for the frustrating experience.

Input: "I lost my clubcard."
Output: "I completely understand how inconvenient this must be. Please can I get your full name and email address so we can look into this?"

Input: "Hi, I can't find my tesco clubcard. I need a new one please."
Output: "Hi, can you confirm your full name and email address? Kind regards."

Fig. 8. Example of two similar interactions with the non-transfer Tesco chatbot. The response changes due to the customer starting their message with 'Hi,', and the chatbot reiterates the greeting.

Table 9
Time taken for the Temi robot to speak the first ten Harvard sentences.

Harvard sentence	Number of words	Time taken (s)	WPM
1	8	2.8	171.43
2	8	2.93	163.82
3	9	2.91	185.57
4	9	3.16	170.89
5	7	3.31	126.89
6	7	2.76	152.17
7	8	2.94	163.27
8	8	3.65	131.51
9	7	3.32	126.51
10	8	3.51	136.75
Average	7.9	3.129	152.88

I can't find my clubcard. I need a new one please". is responded to in kind, with the chatbot choosing to wrap the solution between "Hi", and "Kind regards". These examples of nuance show that a learning-based approach from real conversations can introduce a more natural feel to the interaction, as opposed to static responses with solutions. This achievement was one of the two major goals of this study, effectively to explore how approaches can move away from the expert system-like solution-responses and more towards a natural conversation with a machine as would be performed with another human being.

Consumer robot feasibility and implementation

This section details implementation of the chatbots to the Temi and Pepper robots, for purposes of both feasibility and proof-of-concept. During implementation, observations of drawbacks were made (e.g. with available consumer APIs), and solutions are presented. An image of Temi and Pepper can be seen in Fig. 5, with an image of the chatbot user interface in Fig. 9.

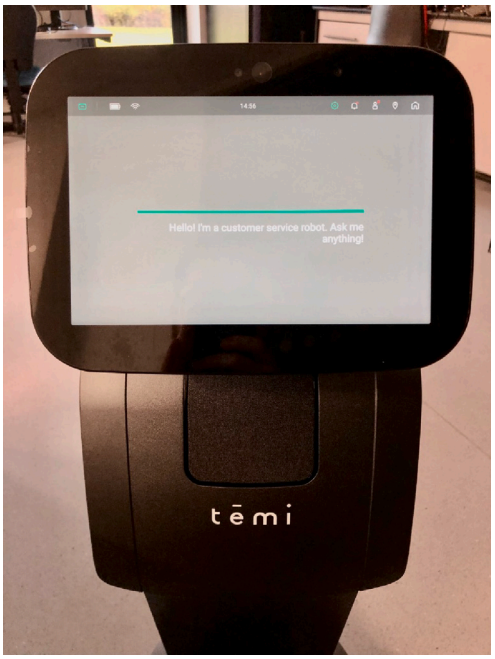


Fig. 9. The user interface on Temi which allows for the user to speak. Once text has been extracted from audio, the user's message is inferred and responded to by the transformer-based chatbot.

During the Temi API implementation, it was observed that even when an await command is used, text-to-speech is not awaited until completion. This causes only the final speech command to be executed if there are multiple. Exploration discovered that this was due to a finished signal being sent upon communication of the command from a computer to the robotic device. To remedy this problem, the speed at which Temi speaks was measured in Table 9, where an average speaking speed of 152.88 words per minute was found. This provided a pointer to a sleep command parameter, so the text-to-speech transition was expected correctly, simply calculated as $wait_{sec} = \frac{words}{WPM} \times 60$. Although the average is used, future work should focus on further exploration of the speed of synthetic speech.

This problem did not occur during implementation on the Pepper robot, although other issues were observed. The issues were twofold; (i) native speech recognition on the Softbank robots is limited to the recognition of keywords rather than speech-to-text, and (ii) NAOqi is only available as a Python 2.7 implementation, and direct conversion to Python 3.x is not possible. Two solutions are proposed in the strategy detailed in Fig. 10; speech recognition must instead be performed on the server side with text being processed and responses passed to the robot, and a bridge is instantiated to communicate between Python versions. TensorFlow is used to produce a prediction on the input data in Python 3.x (3.7 for this study), then operating in Python 2.7 to connect to the robot over a network and send the command across.

5. Future work and conclusion

The marginal differences observed when tuning topology, as well as the results found in [53], suggest that future models could be made less computationally complex while remaining accurate by pruning towards the most useful attention heads. Future work could explore pruning as a method to improve compatibility with weaker robotic hardware, as well as the response time of the chatbot. Within the current version of TensorFlow (2.6.0), the metrics tested for sparse categorical learning are cross-entropy loss, accuracy, and top-k accuracy; this study focused on the analysis of these metrics with $k = 5$ and $k = 10$. If other metrics are tested and implemented in the future, then precision, recall, and

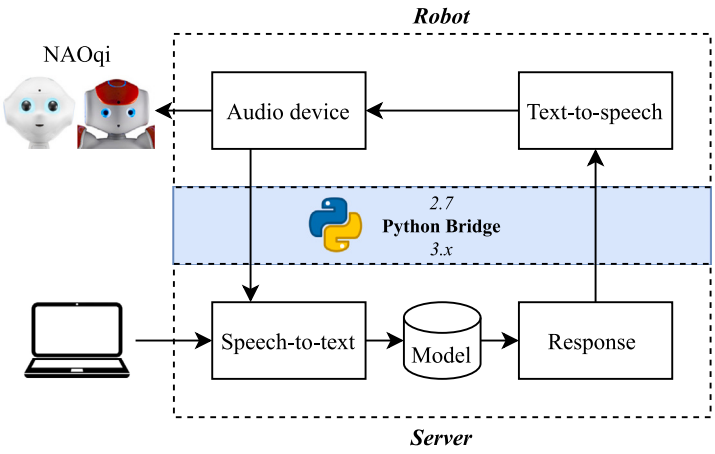


Fig. 10. An implementation strategy of server-sided speech-to-text and a Python bridge to enable interfacing with the NAOqi SDK.

F1 score, etc. should also be measured and compared. Alternatively, with a large amount of RAM required, future experiments could one-hot encode the model output to sparse matrices (rather than integer labels) and utilise categorical cross-entropy, enabling the aforementioned metrics. Given more computational resources, future work could also concern the transfer of knowledge from all-to-one domains rather than the one-to-one transfer learning experiments that were performed in this study. Beyond tuning of the transformer approach, future work could also concern the exploration of other types of machine learning model in comparison, such as temporal (e.g. recurrent and long short term memory neural networks) approaches. Finally, towards consumer use, optimisation methods such as 4-bit quantisation could be explored to increase the speed at which tokens are generated beyond the average 17.329 TPS that was observed during this study.

Several issues facing the feasibility of implementation were encountered during the latter half of these studies, and proposed solutions to overcome them, ultimately providing strategies to the implementation of the chatbots on physical robots. The largest issues faced were due to the possibilities of the NAOqi SDK, with the main drawback in particular being that the SDK is unable to perform speech-to-text, rather, certain keywords are recognised instead. The solution to perform speech-to-text on another device allowed for a working implementation, and future work must therefore concern the possibility of embedded devices on the robots for more distant HRI to take place. Another issue faced was the lack of Python 3.x support for the NAOqi SDK, which is incompatible with recent implementations of TensorFlow. A Python bridge overcame this issue by communicating direct commands from a server that performed inference on the model.

To finally conclude, this study has shown that knowledge transfer is possible between chatbots of several domains in order to improve the ability of autonomous customer support. In the majority of cases, several machine learning metrics were improved when knowledge was transferred from at least one other domain. Statistical testing also revealed that, regardless of whether the result improved over classical learning or did not, that all but three of the domains led to results which, according to the Wilcoxon signed-rank test, were statistically significant. During exploration of the models, it was observed that, in contrast to the more static nature of nonhuman support, relatively natural communication took place. Examples of this included the chatbots seemingly empathising with angrier customers and those faced with particularly difficult issues, as well as a change in tone of voice given the user's input, and also nuanced behaviours such as responding 'hi' to customers who had begun their message with the same. The results found in the experiments within this article are promising, providing a method to enable better customer support chatbots in the future by sharing knowledge from other domains.

CRediT authorship contribution statement

Jordan J. Bird: Conceptualization, Data curation, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. **Ahmad Lotfi:** Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data is available publicly for download.

Appendix A. Loss matrix

See Table A.1.

Appendix B. Accuracy matrix

See Table B.1.

Appendix C. Top-5 accuracy matrix

See Table C.1.

Appendix D. Top-10 accuracy matrix

See Table D.1.

Table A.1
Validation loss when transfer learning (grey cells denote non-transfer).

	Source																		
Target	***Amazon***	***American Air***	***Apple Support***	***British Airways***	***Chipotle***	***Comcast***	***Delta***	***Hulu***	***PlayStation***	***Sainsburys***	***Spectrum***	***Spotify***	***Sprint***	***Tesco***	***TMobile***	***Uber***	***UPS***	***Virgin Trains***	***Xbox***
Amazon	2.3	2.38	2.37	2.37	2.36	2.36	2.38	2.36	2.35	2.37	2.36	2.36	2.35	2.38	2.37	2.36	2.35	2.36	2.36
American Air	2.28	2.17	2.19	2.13	2.14	2.13	2.13	2.13	2.21	2.15	2.13	2.14	2.14	2.15	2.13	2.14	2.12	2.14	2.13
Apple Support	2.04	1.99	1.93	1.99	1.98	1.98	2	1.98	1.97	1.99	1.97	1.99	1.96	2	1.98	1.99	1.97	1.99	1.97
British Airways	2.34	2.25	2.309	2.19	2.25	2.24	2.24	2.24	2.23	2.26	2.24	2.26	2.24	2.23	2.25	2.26	2.22	2.25	2.26
Chipotle	1.15	1.1	1.16	1.09	1.17	1.09	1.09	1.1	1.1	1.09	1.08	1.12	1.09	1.08	1.09	1.09	1.01	1.11	1.08
Comcast	1.8	1.71	1.78	1.73	1.71	1.74	1.72	1.7	1.71	1.73	1.76	1.71	1.72	1.71	1.72	1.72	1.69	1.78	1.8
Delta	2.14	1.96	2.05	2.07	2.07	2.05	2	2.06	2.05	2.07	2.05	2.02	2.05	2.06	2.06	2.02	2.03	2.04	2.05
Hulu	1.88	1.84	1.91	1.81	1.83	1.82	1.8	1.84	1.78	1.85	1.78	1.82	1.8	1.8	1.81	1.82	1.83	1.83	1.8
PlayStation	0.58	0.31	0.45	0.31	0.33	0.32	0.32	0.31	0.54	0.31	0.33	0.33	0.32	0.3	0.32	0.33	0.33	0.37	0.31
Sainsburys	1.76	1.62	1.72	1.62	1.63	1.61	1.67	1.61	1.62	1.64	1.64	1.64	1.62	1.6	1.61	1.63	1.61	1.63	1.61
Spectrum	1.55	1.5	1.52	1.5	1.49	1.47	1.49	1.48	1.49	1.5	1.42	1.49	1.47	1.46	1.45	1.49	1.46	1.41	1.49
Spotify	1.61	1.55	1.55	1.54	1.55	1.55	1.55	1.54	1.52	1.54	1.52	1.57	1.53	1.57	1.54	1.55	1.52	1.54	1.54
Sprint	0.98	0.6	0.85	0.6	0.63	0.6	0.61	0.6	0.63	0.6	0.61	0.62	0.83	0.59	0.6	0.64	0.62	0.66	0.61
Tesco	2.87	2.45	2.51	2.44	2.46	2.45	2.45	2.44	2.45	2.45	2.42	2.46	2.42	2.47	2.43	2.45	2.41	2.45	2.43
TMobile	2.18	2.07	2.11	2.07	2.07	2.05	2.07	2.05	2.05	2.08	2.04	2.07	2.06	2.06	2.1	2.07	2.04	2.13	2.06
Uber	1.81	1.72	1.74	1.71	1.73	1.7	1.72	1.7	1.69	1.71	1.7	1.71	1.7	1.72	1.7	1.68	1.69	1.7	1.71
UPS	1.21	0.79	1.07	0.79	0.84	0.79	0.81	0.79	0.83	0.84	0.81	0.85	0.8	0.79	0.8	0.85	1.04	0.88	0.81
Virgin Trains	1.74	1.17	1.75	1.66	1.66	1.66	1.67	1.66	1.65	1.66	1.66	1.68	1.65	1.66	1.66	1.67	1.65	1.67	1.66
Xbox	1.66	1.63	1.65	1.59	1.59	1.64	1.64	1.61	1.6	1.62	1.62	1.63	1.62	1.6	1.6	1.63	1.58	1.62	1.57

Table B.1
Validation accuracy when transfer learning (grey cells denote non-transfer).

	Source																		
Target	***Amazon***	***American Air***	***Apple Support***	***British Airways***	***Chipotle***	***Comcast***	***Delta***	***Hulu***	***PlayStation***	***Sainsburys***	***Spectrum***	***Spotify***	***Sprint***	***Tesco***	***TMobile***	***Uber***	***UPS***	***Virgin Trains***	***Xbox***
Amazon	0.602	0.6	0.595	0.6	0.6	0.6	0.6	0.6	0.601	0.6	0.6	0.6	0.601	0.6	0.599	0.598	0.601	0.598	0.6
American Air	0.644	0.645	0.63	0.647	0.643	0.646	0.647	0.645	0.652	0.646	0.644	0.644	0.648	0.647	0.646	0.644	0.648	0.643	0.652
Apple Support	0.632	0.639	0.642	0.64	0.638	0.639	0.64	0.64	0.641	0.639	0.64	0.638	0.641	0.639	0.64	0.638	0.64	0.637	0.639
British Airways	0.648	0.663	0.634	0.658	0.658	0.663	0.663	0.665	0.66	0.664	0.66	0.659	0.662	0.665	0.663	0.657	0.662	0.657	0.663
Chipotle	0.81	0.842	0.815	0.842	0.808	0.842	0.842	0.843	0.839	0.841	0.839	0.839	0.838	0.844	0.842	0.838	0.839	0.835	0.839
Comcast	0.7	0.714	0.693	0.713	0.711	0.712	0.712	0.714	0.708	0.712	0.727	0.71	0.727	0.716	0.714	0.708	0.708	0.719	0.72
Delta	0.68	0.68	0.668	0.687	0.683	0.687	0.678	0.686	0.684	0.687	0.686	0.681	0.685	0.687	0.687	0.68	0.686	0.682	0.686
Hulu	0.69	0.7	0.685	0.713	0.71	0.714	0.712	0.712	0.714	0.714	0.713	0.71	0.712	0.714	0.713	0.701	0.71	0.704	0.714
PlayStation	0.865	0.928	0.892	0.927	0.923	0.925	0.924	0.926	0.877	0.926	0.923	0.924	0.924	0.929	0.927	0.921	0.923	0.916	0.93
Sainsburys	0.731	0.759	0.733	0.757	0.764	0.757	0.763	0.769	0.766	0.743	0.762	0.751	0.767	0.761	0.759	0.752	0.756	0.762	0.768
Spectrum	0.743	0.766	0.748	0.765	0.761	0.768	0.766	0.768	0.762	0.767	0.767	0.762	0.765	0.768	0.765	0.762	0.762	0.774	0.766
Spotify	0.737	0.742	0.732	0.742	0.739	0.741	0.74	0.742	0.741	0.741	0.742	0.74	0.742	0.74	0.742	0.74	0.741	0.739	0.742
Sprint	0.791	0.882	0.827	0.883	0.877	0.882	0.881	0.882	0.876	0.882	0.879	0.877	0.825	0.883	0.881	0.871	0.88	0.871	0.882
Tesco	0.613	0.615	0.596	0.616	0.612	0.615	0.615	0.615	0.616	0.616	0.614	0.614	0.615	0.616	0.616	0.613	0.615	0.61	0.616
TMobile	0.65	0.664	0.637	0.656	0.653	0.657	0.655	0.657	0.653	0.656	0.657	0.657	0.662	0.658	0.653	0.653	0.655	0.66	0.656
Uber	0.712	0.72	0.702	0.716	0.713	0.714	0.713	0.715	0.712	0.714	0.714	0.714	0.713	0.715	0.716	0.713	0.714	0.711	0.714
UPS	0.741	0.838	0.769	0.843	0.832	0.84	0.839	0.84	0.835	0.839	0.835	0.829	0.84	0.841	0.841	0.826	0.783	0.825	0.841
Virgin Trains	0.751	0.744	0.722	0.744	0.739	0.739	0.742	0.743	0.738	0.741	0.74	0.739	0.74	0.744	0.742	0.738	0.738	0.746	0.742
Xbox	0.735	0.737	0.732	0.751	0.758	0.737	0.737	0.771	0.767	0.772	0.769	0.734	0.771	0.772	0.737	0.766	0.77	0.763	0.761

Table C.1
Validation top-5 accuracy when transfer learning (grey cells denote non-transfer).

	Source																		
Target	*Amazon*	*American Air*	*Apple Support*	*British Airways*	*Chipotle*	*Comcast*	*Delta*	*Hulu*	*PlayStation*	*Sainsburys*	*Spectrum*	*Spotify*	*Sprint*	*Tesco*	*TMobile*	*Uber*	*UPS*	*Virgin Trains*	*Xbox*
Amazon	0.73	0.725	0.724	0.724	0.724	0.725	0.725	0.725	0.727	0.725	0.726	0.726	0.726	0.724	0.725	0.724	0.727	0.725	0.726
American Air	0.755	0.754	0.749	0.756	0.754	0.755	0.757	0.755	0.756	0.754	0.755	0.756	0.758	0.755	0.756	0.755	0.757	0.753	0.755
Apple Support	0.77	0.773	0.778	0.773	0.773	0.773	0.772	0.773	0.775	0.773	0.774	0.774	0.775	0.772	0.774	0.772	0.774	0.773	0.774
British Airways	0.759	0.766	0.75	0.764	0.759	0.763	0.763	0.764	0.76	0.762	0.762	0.761	0.762	0.764	0.762	0.761	0.762	0.759	0.77
Chipotle	0.89	0.897	0.884	0.9	0.877	0.896	0.896	0.895	0.893	0.895	0.893	0.895	0.895	0.897	0.896	0.895	0.894	0.892	0.9
Comcast	0.824	0.826	0.82	0.824	0.822	0.824	0.825	0.825	0.823	0.824	0.829	0.822	0.832	0.826	0.826	0.823	0.825	0.82	0.828
Delta	0.776	0.78	0.769	0.778	0.776	0.778	0.776	0.778	0.777	0.778	0.778	0.776	0.776	0.778	0.779	0.775	0.779	0.776	0.778
Hulu	0.81	0.81	0.799	0.812	0.809	0.811	0.811	0.81	0.811	0.81	0.811	0.81	0.812	0.812	0.812	0.804	0.81	0.806	0.81
PlayStation	0.94	0.97	0.956	0.97	0.969	0.97	0.97	0.97	0.95	0.97	0.968	0.969	0.97	0.972	0.97	0.968	0.969	0.965	0.97
Sainsburys	0.828	0.84	0.829	0.838	0.842	0.837	0.838	0.842	0.841	0.833	0.84	0.836	0.843	0.84	0.838	0.836	0.838	0.839	0.843
Spectrum	0.848	0.856	0.884	0.855	0.852	0.855	0.855	0.856	0.854	0.854	0.857	0.854	0.854	0.856	0.855	0.854	0.853	0.859	0.855
Spotify	0.832	0.833	0.83	0.832	0.83	0.831	0.831	0.831	0.831	0.831	0.833	0.831	0.832	0.831	0.831	0.831	0.832	0.83	0.833
Sprint	0.893	0.942	0.913	0.941	0.939	0.94	0.941	0.942	0.937	0.942	0.939	0.938	0.916	0.943	0.942	0.936	0.94	0.934	0.941
Tesco	0.735	0.735	0.725	0.734	0.731	0.734	0.733	0.734	0.735	0.733	0.734	0.734	0.734	0.738	0.733	0.732	0.735	0.731	0.735
TMobile	0.77	0.772	0.759	0.767	0.765	0.768	0.768	0.768	0.767	0.767	0.768	0.767	0.772	0.768	0.765	0.767	0.768	0.766	0.767
Uber	0.812	0.813	0.81	0.814	0.812	0.832	0.813	0.813	0.813	0.813	0.813	0.813	0.813	0.813	0.814	0.813	0.813	0.811	0.814
UPS	0.871	0.923	0.888	0.926	0.92	0.925	0.924	0.925	0.912	0.923	0.922	0.918	0.924	0.925	0.925	0.918	0.897	0.915	0.923
Virgin Trains	0.82	0.817	0.805	0.817	0.814	0.814	0.815	0.816	0.813	0.815	0.815	0.814	0.815	0.817	0.816	0.814	0.814	0.819	0.816
Xbox	0.83	0.827	0.83	0.835	0.838	0.826	0.826	0.844	0.845	0.845	0.843	0.825	0.843	0.845	0.828	0.843	0.843	0.842	0.842

Table D.1
Validation top-10 accuracy when transfer learning (grey cells denote non-transfer).

	Source																		
Target	*Amazon*	*American Air*	*Apple Support*	*British Airways*	*Chipotle*	*Comcast*	*Delta*	*Hulu*	*PlayStation*	*Sainsburys*	*Spectrum*	*Spotify*	*Sprint*	*Tesco*	*TMobile*	*Uber*	*UPS*	*Virgin Trains*	*Xbox*
Amazon	0.779	0.774	0.774	0.774	0.773	0.775	0.775	0.775	0.776	0.774	0.775	0.775	0.775	0.773	0.774	0.774	0.776	0.774	0.775
American Air	0.8	0.8	0.796	0.8	0.798	0.799	0.8	0.799	0.798	0.796	0.799	0.799	0.801	0.797	0.799	0.798	0.802	0.797	0.797
Apple Support	0.82	0.819	0.826	0.82	0.82	0.819	0.819	0.82	0.821	0.82	0.82	0.821	0.822	0.819	0.821	0.819	0.821	0.82	0.82
British Airways	0.81	0.804	0.795	0.802	0.797	0.801	0.802	0.802	0.799	0.8	0.801	0.801	0.8	0.802	0.801	0.8	0.8	0.799	0.82
Chipotle	0.91	0.915	0.909	0.915	0.901	0.915	0.914	0.913	0.912	0.914	0.912	0.914	0.913	0.915	0.914	0.914	0.912	0.911	0.914
Comcast	0.86	0.859	0.858	0.858	0.857	0.857	0.859	0.859	0.857	0.858	0.858	0.857	0.862	0.858	0.86	0.858	0.859	0.857	0.858
Delta	0.814	0.817	0.81	0.814	0.813	0.815	0.814	0.814	0.814	0.814	0.813	0.814	0.813	0.815	0.815	0.813	0.816	0.813	0.814
Hulu	0.842	0.84	0.841	0.847	0.846	0.846	0.847	0.844	0.846	0.846	0.847	0.846	0.848	0.85	0.848	0.844	0.846	0.843	0.845
PlayStation	0.961	0.981	0.973	0.98	0.98	0.981	0.98	0.981	0.96	0.98	0.979	0.98	0.98	0.982	0.981	0.98	0.98	0.977	0.981
Sainsburys	0.861	0.866	0.861	0.865	0.868	0.864	0.866	0.868	0.866	0.862	0.866	0.864	0.868	0.867	0.865	0.863	0.864	0.866	0.868
Spectrum	0.881	0.884	0.883	0.882	0.879	0.882	0.882	0.883	0.881	0.88	0.883	0.881	0.881	0.883	0.883	0.882	0.881	0.885	0.882
Spotify	0.865	0.864	0.864	0.853	0.862	0.862	0.862	0.863	0.863	0.862	0.864	0.862	0.863	0.861	0.863	0.863	0.863	0.862	0.863
Sprint	0.924	0.958	0.94	0.957	0.955	0.956	0.957	0.957	0.954	0.957	0.955	0.955	0.939	0.958	0.958	0.954	0.955	0.952	0.957
Tesco	0.782	0.8	0.776	0.778	0.776	0.778	0.778	0.778	0.781	0.778	0.779	0.78	0.779	0.783	0.778	0.778	0.78	0.777	0.78
TMobile	0.811	0.813	0.806	0.811	0.808	0.811	0.811	0.811	0.81	0.809	0.811	0.811	0.814	0.811	0.808	0.81	0.811	0.808	0.809
Uber	0.85	0.847	0.845	0.847	0.846	0.847	0.846	0.846	0.847	0.846	0.847	0.846	0.846	0.847	0.847	0.847	0.847	0.846	0.847
UPS	0.91	0.945	0.923	0.947	0.943	0.946	0.945	0.946	0.943	0.945	0.944	0.941	0.946	0.946	0.946	0.942	0.926	0.939	0.945
Virgin Trains	0.85	0.847	0.84	0.847	0.845	0.845	0.846	0.846	0.844	0.845	0.845	0.844	0.845	0.847	0.847	0.845	0.845	0.846	0.846
Xbox	0.87	0.861	0.865	0.867	0.866	0.858	0.86	0.87	0.871	0.871	0.87	0.859	0.87	0.87	0.86	0.87	0.869	0.869	0.87

References

[1] B. Esmaeilian, S. Behdad, B. Wang, The evolution and future of manufacturing: A review, J. Manuf. Syst. 39 (2016) 79–100.

[2] G. Michalos, S. Makris, N. Papakostas, D. Mourtzis, G. Chryssolouris, Automotive assembly technologies review: challenges and outlook for a flexible and adaptive approach, CIRP J. Manuf. Sci. Technol. 2 (2) (2010) 81–91.

[3] G. Lunghi, R. Marin, M. Di Castro, A. Masi, P.J. Sanz, Multimodal human-robot interface for accessible remote robotic interventions in hazardous environments, IEEE Access 7 (2019) 127290–127319.

[4] H. Wallop, On hold: How long it takes to speak to a human at major organisations, 2021, This Is Money, [Online]. Available: https://www.thisismoney.co.uk/money/news/article-9737653/On-hold-long-takes-speak-human-major-organisations.html.

[5] L.E. Bygballe, E. Bø, S.E. Grønland, Managing international supply: The balance between total costs and customer service, Ind. Mark. Manag. 41 (3) (2012) 394–401.

[6] J. Potter-Brotman, The new role of service in customer retention, Manag. Serv. Qual. Int. J. (1994).

[7] K.I. Roumeliotis, N.D. Tselikas, D.K. Nasiopoulos, LLMs in e-commerce: a comparative analysis of GPT and LLaMA models in product review evaluation, Nat. Lang. Process. J. 6 (2024) 100056.

[8] M.C. Rillig, M. Ågerstrand, M. Bi, K.A. Gould, U. Sauerland, Risks and benefits of large language models for the environment, Environ. Sci. Technol. 57 (9) (2023) 3464–3466.

[9] A.S. Lokman, M.A. Ameedeen, Modern chatbot systems: A technical review, in: Proceedings of the Future Technologies Conference, Springer, 2018, pp. 1012–1023.

[10] L. Cui, S. Huang, F. Wei, C. Tan, C. Duan, M. Zhou, Superagent: A customer service chatbot for e-commerce websites, in: Proceedings of ACL 2017, System Demonstrations, 2017, pp. 97–102.

[11] A.M. Turing, J. Haugeland, Computing Machinery and Intelligence, MIT Press Cambridge, MA, 1950.

[12] L. Floridi, M. Taddeo, M. Turilli, Turing's imitation game: still an impossible challenge for all machines and some judges—an evaluation of the 2008 Loebner contest, Minds Mach. 19 (1) (2009) 145–150.

[13] M. Nuruzzaman, O.K. Hussain, A survey on chatbot implementation in customer service industry through deep neural networks, in: 2018 IEEE 15th International Conference on E-Business Engineering, ICEBE, IEEE, 2018, pp. 54–61.

[14] A. Xu, Z. Liu, Y. Guo, V. Sinha, R. Akkiraju, A new chatbot for customer service on social media, in: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 2017, pp. 3506–3510.
[15] B.R. Ranoliya, N. Raghuwanshi, S. Singh, Chatbot for university related FAQs, in: 2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI, IEEE, 2017, pp. 1525–1530.
[16] J. Feine, S. Morana, U. Gnewuch, Measuring service encounter satisfaction with customer service chatbots using sentiment analysis, 2019.
[17] A.D. Tran, J.I. Pallant, L.W. Johnson, Exploring the impact of chatbots on consumer sentiment and expectations in retail, J. Retail. Consum. Serv. 63 (2021) 102718.
[18] J.J. Bird, A. Ekárt, C.D. Buckingham, D.R. Faria, High resolution sentiment analysis by ensemble classification, in: Intelligent Computing-Proceedings of the Computing Conference, Springer, 2019, pp. 593–606.
[19] K. Jia, Chinese sentiment classification based on word2vec and vector arithmetic in human–robot conversation, Comput. Electr. Eng. 95 (2021) 107423.
[20] E. Ricciardelli, D. Biswas, Self-improving chatbots based on reinforcement learning, in: 4th Multidisciplinary Conference on Reinforcement Learning and Decision Making, 2019.
[21] H. Cuayáhuitl, D. Lee, S. Ryu, Y. Cho, S. Choi, S. Indurthi, S. Yu, H. Choi, I. Hwang, J. Kim, Ensemble-based deep reinforcement learning for chatbots, Neurocomputing 366 (2019) 118–130.
[22] M. Bali, S. Mohanty, S. Chatterjee, M. Sarma, R. Puravankara, Diabot: a predictive medical chatbot using ensemble learning, Int. J. Recent Technol. Eng. (2019) 2277–3878.
[23] N. Harilal, R. Shah, S. Sharma, V. Bhutani, CARO: an empathetic health conversational chatbot for people with major depression, in: Proceedings of the 7th ACM IKDD CoDS and 25th COMAD, 2020, pp. 349–350.
[24] A. Mondal, M. Dey, D. Das, S. Nagpal, K. Garda, Chatbot: An automated conversation system for the educational domain, in: 2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing, ISAI-NLP, IEEE, 2018, pp. 1–5.
[25] J.J. Bird, A. Ekárt, D.R. Faria, Chatbot Interaction with Artificial Intelligence: human data augmentation with T5 and language transformer ensemble for text classification, J. Ambient Intell. Humaniz. Comput. (2021) 1–16.
[26] V. Ilievski, C. Musat, A. Hossmann, M. Baeriswyl, Goal-oriented chatbot dialog management bootstrapping with transfer learning, in: IJCAI, 2018.
[27] A. Hatua, T.T. Nguyen, A.H. Sung, Goal-oriented conversational system using transfer learning and attention mechanism, in: 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference, UEMCON, IEEE, 2019, pp. 0099–0104.
[28] J. Yu, M. Qiu, J. Jiang, J. Huang, S. Song, W. Chu, H. Chen, Modelling domain relationships for transfer learning on retrieval-based question answering systems in e-commerce, in: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, 2018, pp. 682–690.
[29] M. Xiaoliang, Z. RuQiang, L. Ying, D. Congjian, D. Dequan, Design of a large language model for improving customer service in telecom operators, Electron. Lett. 60 (10) (2024) e13218.
[30] Y. Xie, C. Liang, P. Zhou, L. Jiang, Exploring the influence mechanism of chatbot-expressed humor on service satisfaction in online customer service, J. Retail. Consum. Serv. 76 (2024) 103599.
[31] D. Huang, D.G. Markovitch, R.A. Stough, Can chatbot customer service match human service agents on customer satisfaction? An investigation in the role of trust, J. Retail. Consum. Serv. 76 (2024) 103600.
[32] K.C. Yam, Y.E. Bigman, P.M. Tang, R. Ilies, D. De Cremer, H. Soh, K. Gray, Robots at work: People prefer—and forgive—service robots with perceived feelings, J. Appl. Psychol. (2020).
[33] M. Mori, Bukimi no tani [the uncanny valley], Energy 7 (1970) 33–35.
[34] J. Murphy, U. Gretzel, C. Hofacker, Service robots in hospitality and tourism: investigating anthropomorphism, in: 15th APacCHRIE Conference, Vol. 31, 2017.
[35] J. Wirtz, P.G. Patterson, W.H. Kunz, T. Gruber, V.N. Lu, S. Paluch, A. Martins, Brave new world: service robots in the frontline, J. Serv. Manag. (2018).
[36] A. Tuomi, I.P. Tussyadiah, J. Stienmetz, Applications and implications of service robots in hospitality, Cornell Hosp. Q. 62 (2) (2021) 232–247.
[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
[38] J. Devlin, M.-W. Chang, Open sourcing BERT: State-of-the-art pre-training for natural language processing, 2018, Google AI Blog. Weblog. [Online] Available from: https://ai.googleblog.com/2018/11/open-sourcing-bertstate-of-art-pre.html. (Accessed 4 December 2019).
[39] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, 2019.
[40] T. Shao, Y. Guo, H. Chen, Z. Hao, Transformer-based neural network for answer selection in question answering, IEEE Access 7 (2019) 26146–26156.
[41] D. Lukovnikov, A. Fischer, J. Lehmann, Pretrained transformers for simple question answering over knowledge graphs, in: International Semantic Web Conference, Springer, 2019, pp. 470–486.
[42] A. Karar, S. Said, T. Beyrouthy, et al., Pepper humanoid robot as a service robot: a customer approach, in: 2019 3rd International Conference on Bio-Engineering for Smart Technologies, BioSMART, IEEE, 2019, pp. 1–4.
[43] M. Niemelä, P. Heikkilä, H. Lammi, A social service robot in a shopping mall: expectations of the management, retailers and consumers, in: Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, 2017, pp. 227–228.
[44] C. Shi, S. Satake, T. Kanda, H. Ishiguro, How would store managers employ social robots? in: 2016 11th ACM/IEEE International Conference on Human-Robot Interaction, HRI, IEEE, 2016, pp. 519–520.
[45] M. Merkle, Customer responses to service robots–comparing human-robot interaction with human-human interaction, in: Proceedings of the 52nd Hawaii International Conference on System Sciences, 2019.
[46] K. Kaipainen, A. Ahtinen, A. Hiltunen, Nice surprise, more present than a machine: Experiences evoked by a social robot for guidance and edutainment at a city service point, in: Proceedings of the 22nd International Academic Mindtrek Conference, 2018, pp. 163–171.
[47] Thought Vector, Dataset: Customer support on Twitter, 2017, Kaggle, [Online]. Available: https://www.kaggle.com/thoughtvector/customer-support-on-twitter.
[48] R.J. Gabriel, GitHub repository: Robertjgabriel/Google-profanity-words, 2016, GitHub, [Online]. Available: https://github.com/RobertJGabriel/Google-profanity-words.
[49] A. Batool, S.W. Loke, N. Fernando, J. Kua, Towards a system for aged care centres based on multiuser–multidevice interactions in iot collectives, in: MobiQuitous 2020-17th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, 2020, pp. 470–475.
[50] C.-F. Hung, Y. Lin, H.-J. Ciou, W.-Y. Wang, H.-H. Chiang, FoodTemi: The AI-oriented catering service robot, in: 2021 IEEE International Conference on Consumer Electronics-Taiwan, ICCE-TW, IEEE, 2021, pp. 1–2.
[51] A. Tuomi, I.P. Tussyadiah, P. Hanna, Spicing up hospitality service encounters: the case of pepper™, Int. J. Contemp. Hosp. Manag. (2021).
[52] E. Rothauser, IEEE recommended practice for speech quality measurements, IEEE Trans. Audio Electroacoust. 17 (1969) 225–246.
[53] P. Michel, O. Levy, G. Neubig, Are sixteen heads really better than one? Adv. Neural Inf. Process. Syst. 32 (2019) 14014–14024.