Contents lists available at ScienceDirect

International Journal of Information Management Data Insights

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

Airline reviews processing: Abstractive summarization and rating-based sentiment classification using deep transfer learning

Ayesha Ayub Syed[a,*], Ford Lumban Gaol[a], Alfred Boediman[b], Widodo Budiharto[c]

[a] *Department of Doctor of Computer Science – BINUS Graduate Program, Bina Nusantara University, Jakarta, Indonesia*
[b] *Department of Econometrics and Statistics - The University of Chicago, Booth School of Business*
[c] *Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia*

ARTICLE INFO

Keywords:
Airline reviews
Domain adaptation
Opinion summarization
Review rating
Sentiment classification
Two-stage finetuning

ABSTRACT

Opinion summarization and sentiment classification are key processes for understanding, analyzing, and leveraging information from customer opinions. The rapid and ceaseless increase in big data of reviews on e-commerce platforms, social media, or review portals becomes a stimulus for the automation of these processes. In recent years, deep transfer learning has opted to solve many challenging tasks in Natural Language Processing (NLP) relieving the hassles of exhaustive training and the requirement of extensive labelled datasets. In this work, we propose frameworks for Abstractive Summarization (ABS) and Sentiment Analysis (SA) of airline reviews using Pretrained Language Models (PLM). The abstractive summarization model goes through two finetuning stages, the first one, for domain adaptation and the second one, for final task learning. Several studies in the literature empirically demonstrate that review rating has a positive correlation with sentiment valence. For the sentiment classification framework, we used the rating value as a signal to determine the review sentiment, and the model is built on top of BERT (Bidirectional Encoder Representations from Transformers) architecture. We evaluated our models comprehensively with multiple metrics. Our results indicate competitive performance of the models in terms of most of the evaluation metrics.

1. Introduction

Air travellers use popular media like Skytrax, TripAdvisor, Google Reviews, or Twitter to share their experiences about airlines, airports, flights, etc. Reviews reflect customers' dynamic attitudes toward the product or service quality (Lu et al., 2022) and compose a comprehensive characterization of customer perceptions resulting from the interaction between emotion and cognition during a product or service value estimation (Xu et al., 2024). These reviews are information-rich resources packed with consumer experiences, opinions, emotions, recommendations, and information on distinctive product features. The reviews in the form of eWOM (electronic Word of Mouth) are valuable for potential customers as well as for business organizations. Customers read reviews before decision-making to compare alternatives and know any product/service-related issues or ratings while businesses utilize reviews as metrics for service quality (Lu et al., 2022), to know the product reputation among the consumers, and to maintain a standard in the competitive marketplace (Kumar et al., 2021). Based on dual process theories, customers process information through a central route where they analyze all the relevant pieces of information and a peripheral route where they evaluate the information and go through the decision-making process (Bigne et al., 2021).

Relating to the discipline of economics of information, there is a certain search cost associated with exploring and finding pertinent content from the big data of customer reviews. This cost occurs in terms of 'time' and 'effort' and it needs to be reduced by extracting the most important information (Al-Natour & Turetken, 2020). Big data analytics techniques are constantly evolving and gaining significant attention across multiple disciplines (Chintalapudi et al., 2021; Hassani et al., 2020). However, many existing studies related to big data analytics lack theoretical context, generalizability, and causal inferences, resulting in weak theoretical contributions (Kar et al., 2023) to the field of research.

Text mining techniques like Automatic Text Summarization (ATS) and Sentiment Analysis (SA) use Natural Language Processing (NLP) and machine/deep learning methods to automatically summarize text and analyse sentiment from the text. The automatic summarization of the

[*] Corresponding author at: Department of Doctor of Computer Science – BINUS Graduate Program, Bina Nusantara University, Anggrek Campus, Jl. Kebon Jeruk Raya No. 27, Kebon Jeruk, Jakarta, Indonesia.
E-mail address: ayesha.syed@binus.ac.id (A.A. Syed).

https://doi.org/10.1016/j.jjimei.2024.100238

Available online 20 April 2024

reviews also known as opinion summarization retains only useful information from the review, and sentiment analysis assists in capturing the polarity of the sentiment expressed in the review. Sentiment analysis classifies the review sentiment as positive, negative, or neutral based on the opinion and emotion expressed within the review. A review rating or star rating is a quantitative value assigned by the reviewer based on the product experience. Customer ratings are considered as one of the prime measures of customer satisfaction (Chatterjee, 2019). There are several research works in the literature supporting the positive relationship between rating and sentiment valence. (Bigne et al., 2023) analyzed uniformity between the emotional tone of online reviews collected from TripAdvisor and their ratings. Their findings suggest an alignment between review sentiment and its rating. The findings from Zhu et al. (2020) on Airbnb reviews reveal that higher ratings are linked to positive sentiment and lower ratings express negative sentiment. The results from Baniya et al. (2021) also indicate that reviews with positive ratings are dominated by positive sentiments.

In the case of reviews, large, annotated datasets for abstractive summarization and sentiment classification are scarce, and manually annotating a huge dataset is quite an expensive, exhaustive, and time-consuming activity. From a Natural Language Processing (NLP) perspective, unlike other existing review datasets, airline reviews are generally longer, express opinions on multiple legs of the journey, communicate experiences regarding various aspects, such as seat comfort, inflight and ground crew services, entertainment, food, beverages, etc., and contain mixed sentiments, all in one review. These peculiarities make airline reviews challenging for certain NLP tasks including summarization and sentiment analysis. Currently, there are no benchmark datasets for abstractive summarization and rating-based sentiment classification of airline reviews. We bridge this gap by introducing datasets for abstractive summarization and sentiment classification of airline reviews accompanied by a dataset for domain adaptive training/ review title generation.

Opinion summarization can be modelled using either an extractive or abstractive approach. The extractive approach generates a review summary by extracting and concatenating chunks of information (words/phrases) from the original text while the abstractive approach being more sophisticated also focuses on linguistic features and uses Natural Language Generation (NLG) techniques to produce a human-like summary consisting of novel words that are not present in the original text (Syed et al., 2021). Currently, the idea of deep transfer learning has been utilized to achieve improved performance across various NLP tasks, where large pre-trained language models are finetuned on the target annotated datasets. Deep transfer learning lowers training time and costs by transferring the knowledge gained from a source task to learn a related target task. The weights of the model are initialized from the pretrained model weights and finetuned on the target task. The transfer learning approach has also been known to be effective for low-resource datasets (Zhang et al., 2020). Abstractive summarization using transfer learning usually suffers from the domain shift problem when the source and target domains exhibit data from varying distributions resulting in performance degradation at the target end (Ramponi & Plank, 2021).

In this research, we take a model-based approach for transferring knowledge from pretrained language models for the tasks of abstractive summarization and rating-based sentiment classification of airline reviews. For the abstractive summarization task, the novelty of our research lies in proposing a finetuning-based approach (two-stage finetuning) for the adaptation of PLM to the airline review domain (addressing the domain shift) while most of the existing works in abstractive summarization utilize pretraining-based approaches for Domain Adaptation (DA) purposes. The pretraining-based approaches are computationally expensive and require larger training times as compared to finetuning-based methods. For sentiment classification, most of the existing techniques for sentiment scoring are based on the word count, length of sentence, and the ratio of the positive and negative word counts, etc. In our research, these techniques are not deemed suitable due to the multi-aspect nature of airline reviews, due to the possibility that a customer might express his negative sentiment on a single product aspect and discuss it at length using multiple sentences while his opinion on other product aspects might be positive but expressed in just one or two sentences. In such a situation, there is a possibility that the dominant sentiment of the review becomes the negative one while the overall sentiment expressed by the customer is positive. To avoid such kinds of unfavorable results and improve the consistency of sentiment prediction, our research utilizes ‘customer rating’ and ‘recommendation value’ signals. Both these signals have a positive association with customer satisfaction leading to positive customer sentiment (Chatterjee, 2019). The higher ratings are associated with positive recommendations and vice versa. In our work, the sentiments are predicted based on customer ratings. However, there might be possible discrepancies between the textual content of the review and the customer-assigned rating. These discrepancies have been examined across the dataset using the ‘recommendation value’ signal. The data instances with higher ratings along with negative recommendations were verified as having discrepancies between textual content and rating and were removed from the data.

Thus, our contributions from this work are:

1. We have introduced datasets (Syed et al., 2023) for abstractive summarization, domain adaptive training, and rating-based sentiment classification for airline reviews collected from the Skytrax reviews portal (https://www.airlinequality.com/).
2. We propose a novel two-step abstractive summarization framework for airline reviews using a pretrained PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive SUmmarization Sequence-to-sequence models) (Zhang et al., 2020) model that also diminishes the domain shift problem via supplementary training on an intermediate task.
3. We propose a rating-based sentiment classification model for airline reviews using the pretrained BERT language model.

The remaining paper is organized as follows. Section 2 provides a background and literature review on abstractive summarization, sentiment classification, deep transfer learning, and pretrained language models. Section 3 highlights related works from existing literature in the areas of abstractive summarization and sentiment classification. Section 4 explains the research methodology. Section 5 describes the datasets and the data analysis using figures, graphs, network diagrams, bubble charts, etc. Sections 6 and 7 present frameworks for abstractive summarization and sentiment classification of airline reviews in due detail. Section 8 is the discussion section and Section 9 concludes this article.

2. Literature review

2.1. Abstractive summarization (ABS)

Abstractive summarization methods as opposed to their extractive counterparts produce more concise, fluent, and coherent summaries by assimilating the semantic, syntactic, and contextual information from the input words. Katwe et al. (2023) taxonomize abstractive summarization into document-based, structure-based, semantic-based, and deep learning-based approaches.

Document-based abstractive summarization refers to whether the content being summarized is based on a single document or encapsulates topics from multiple documents. Structure-based methods are guided by existing knowledge and schemas and include various forms such as trees, templates, ontologies, lead and body phrases, and graph or rule-based methods (Syed et al., 2021). Semantic approaches construct a semantic representation of the input text by locating the noun and verb phrases from the input. This semantic representation then becomes the input for the natural language generation system. Among the semantic categories are the models based on information items, predicate arguments, semantic graphs, and multimodal semantics (Sciforce, 2019).

Deep learning techniques for abstractive summarization are based on seq2seq models using the RNN (Recurrent Neural Network), LSTM (Long Short-Term Memory), GRU (Gated Recurrent Unit), the bidirectional variants (biRNN, biLSTM, biGRU), transformer architecture (Vaswani et al., 2017), etc. The recent works are developed utilizing Pretrained Language Models (PLM) and are focused on making abstractive summarization systems more efficient and effective. Fig. 1 illustrates the main components and stages of a deep learning-based abstractive summarization framework that includes the dataset acquisition and pre-processing stage, encoder-decoder architecture, mechanisms to enhance model functionality and performance, training and optimization process, model testing and evaluation using automatic metrics or human evaluation. For a detailed explanation of neural abstractive text summarization and abstractive summarization using pretrained language models, we refer readers to (Syed et al., 2021) and (Syed et al., 2022).

2.2. Sentiment classification

Sentiment classification is the process of automatically determining the orientation of opinion in a piece of text by classifying text into three primary sentiment classes; positive, negative, and neutral (Tan et al., 2023). It is a multidisciplinary area of research integrating sociology, psychology, data mining, computational linguistics, natural language processing, and machine learning/deep learning fields (Ligthart et al., 2021).

The sentiment analysis process can be performed at multiple levels of granularity as document level, sentence level, word level, and aspect level (Bordoloi & Biswas, 2023). The document-level sentiment classification determines the overall polarity in a text/document by combining the polarities of all words and sentences in the document. At the sentence level, sentiment analysis determines the polarity in a sentence by combining the polarities of all the words or phrases that make up that sentence. Word-level classification aims to determine the sentiment of individual words and their influence on the overall text sentiment. It is usually implemented using a dictionary-based or corpus-based approach (Bordoloi & Biswas, 2023). Aspect-based sentiment classification is a fine-grained analysis task in which first aspects and targets are identified and then the orientation of these aspects toward the targets are determined. The general workflow for sentiment classification is illustrated in Fig. 2.

Chamekh et al. (2022) categorizes sentiment analysis approaches into lexicon-based, machine learning-based, and hybrid approaches. Tan et al. (2023) groups the classifiers for sentiment classification into three classes, machine learning (decision tree, Naïve Bayes, k-nearest neighbor, support vector machine, etc.), deep learning (RNN, LSTM, BERT, etc.), and ensemble learning (combination of multiple classifiers to achieve better performance). Cambria et al. (2017) proposed a triple-layer structure for solving the problem of sentiment analysis. The first layer is the syntactic layer in which the text is pre-processed, the second layer is the semantics layer that is focussed on interpreting meanings, concepts, anaphora, and entities in the text and removing the

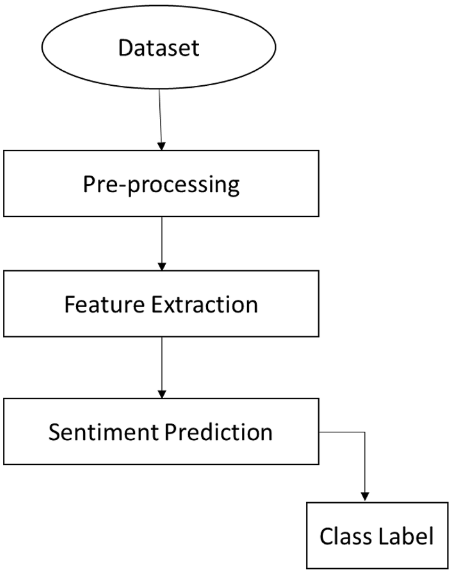


Fig. 2. General Workflow for Sentiment Classification.

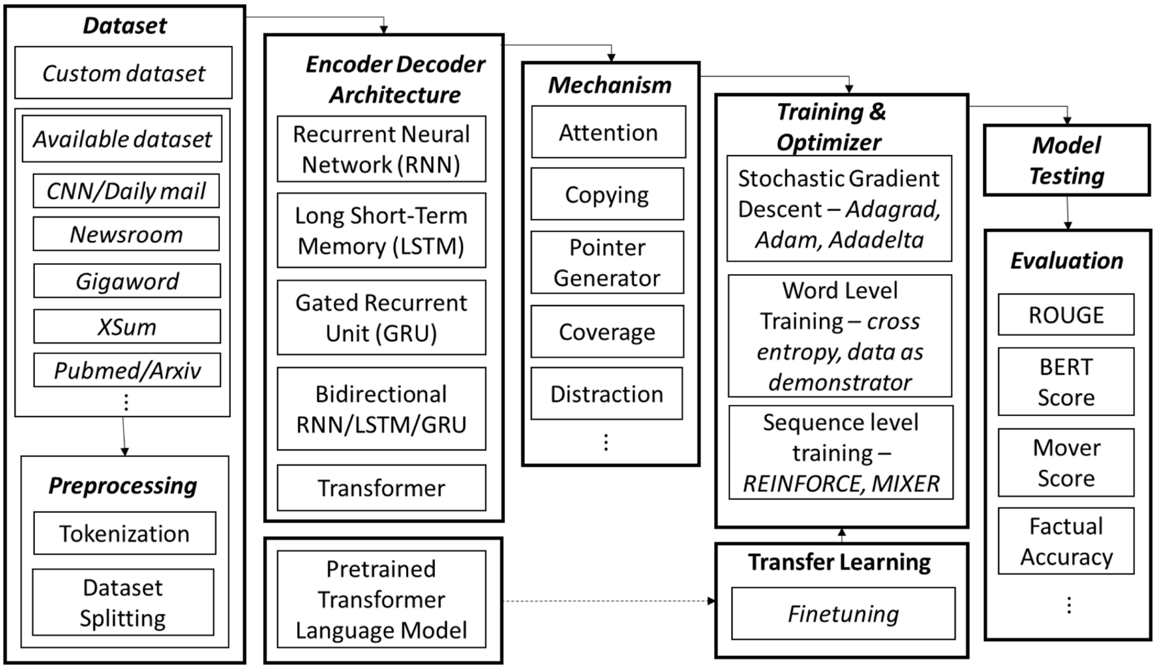

Dataset
Custom dataset
Available dataset
CNN/Daily mail
Newsroom
Gigaword
XSum
Pubmed/Arxiv
⋮
Preprocessing
Tokenization
Dataset Splitting
Encoder Decoder Architecture
Recurrent Neural Network (RNN)
Long Short-Term Memory (LSTM)
Gated Recurrent Unit (GRU)
Bidirectional RNN/LSTM/GRU
Transformer
Pretrained Transformer Language Model
Mechanism
Attention
Copying
Pointer Generator
Coverage
Distraction
⋮
Training & Optimizer
Stochastic Gradient Descent – Adagrad, Adam, Adadelta
Word Level Training – cross entropy, data as demonstrator
Sequence level training – REINFORCE, MIXER
Transfer Learning
Finetuning
Model Testing
Evaluation
ROUGE
BERT Score
Mover Score
Factual Accuracy
⋮

Fig. 1. Components of Neural Abstractive Text Summarization Framework (Syed et al., 2022).

unnecessary content from the normalized text, the third layer is the pragmatics layer that performs various tasks like personality recognition, sarcasm detection, semantics understanding, aspect extraction, and polarity determination. Recently, researchers have been investigating pretrained language models for the task of sentiment classification and realized that pretraining and finetuning language models on specific datasets produce better performance even on datasets with class imbalance and out-of-domain data problems (Kant et al., 2018).

2.3. *Deep transfer learning*

Deep transfer learning in general refers to the application of Transfer Learning to Deep Neural Networks. In NLP, it involves transferring knowledge from neural networks based on large pretrained language models to the target task with limited data for an efficient learning process and better model performance.

2.3.1. *Transfer learning*

Transfer Learning is a learning paradigm that enables the transfer of knowledge from a source domain to a related target domain. It assists in minimizing large-scale data requirements for the target domain as well as improves the final task performance (Zhuang et al., 2021). The process of transfer learning occurs in two phases. The first phase comprises gaining knowledge by pretraining a language model on a single or multiple source task/domain. The second phase constitutes the fine-tuning process where the knowledge acquired during the first phase is transferred/passed on to the target task/domain (Han et al., 2021). Formally, transfer learning is explained using the concepts of ‘domain’ and ‘task’.

2.3.1.1. Domain & task. A domain is a set of values that can be used by a function. For transfer learning, we define domain d as, $d = (\chi, P(X))$, where χ is the feature space and $P(X)$ is the marginal probability distribution. Further, we have a source domain and a target domain. $d_S = \{(x_{S_1}, y_{S_1}), \ldots, (x_{S_n}, y_{S_n})\}$, where d_S is the source domain data, $x_{S_i} \in \chi_S$ and $y_{S_i} \in Y_S$ and $d_T = \{(x_{T_1}, y_{T_1}), \ldots, (x_{T_n}, y_{T_n})\}$, where d_T is the target domain data, $x_{T_i} \in \chi_T$ and $y_{T_i} \in Y_T$. A task is the purpose for which a mathematical function is sorted out. We define task t as: $t = (Y, f(.))$, where Y is the label space, and $f(.)$ is the predictive function. Likewise, we have a source task, symbolized as t_S, a source predictive function represented as $f_S(.)$, a target task notated as t_T and a target predictive function denoted as $f_T(.)$.

Considering a source domain d_S with its source task t_S and a target domain d_T with its task t_T, where, either $d_S \neq d_T$ or $t_S \neq t_T$, transfer learning utilizes the knowledge from a source domain d_S and source task t_S to learn the target domain predictive function $f_T(.)$ (Weiss et al., 2016).

2.3.2. *Domain adaptation (DA)*

In practice, while implementing transfer learning, there usually occurs a situation where the source and target datasets belong to varying domains and data distributions do not match. The problem is known as the domain shift problem. However, in an ideal scenario, machine learning models are expected to generalize effectively on out-of-distribution data. Domain Adaptation (DA) is a practice that can be used to address the domain shift issue (Ramponi & Plank, 2021). DA is an example of transductive transfer learning where the focus is to diminish the dissimilarity between the source and target domains (Zhuang et al., 2021). Formally, given that $P_s(X) \neq P_t(X)$, domain adaptation strives to learn a function f from a source domain D_s to a target domain D_t that generalizes and performs well on the target domain D_t.

DA implementation can be supervised or unsupervised. For supervised DA, a small amount of labeled target data is usually available while in unsupervised DA, the target data is not labeled. Ramponi & Plank (2021) categorizes domain adaptation implementation approaches into three classes, model-centric, data-centric, and hybrid approaches. The model-centric methods work by manipulating the loss function or changing model architecture/parameters. The data-centric methods work by labeling/pretraining the data. The hybrid approaches use a combination of model and data-centric methods.

2.3.3. *Fine-Tuning*

Fine-tuning is the standard approach used for transferring knowledge from a pre-trained model to the target task. Generally, the idea of finetuning is to set the parameters of the neural network initially with the parameters of the pre-trained model and then optimize those parameters using the target task and dataset (Guo et al., 2019). Practically, the fine-tuning process is a four-step process. The first step is to obtain the pre-trained model on the source dataset. Second, to design a target model by initializing the parameters from the pre-trained model. An additional task-specific layer is added to the model in the third step. The final part of the process is to train the new model on the target training dataset. In the whole process, the newly added output layer would be trained fully from the beginning. The parameters of the rest of the model would be finetuned with the source model as the base model.

2.4. *Pretrained language models (PLM)*

With pretrained language models, NLP has entered the real progressive and advanced era. The paradigm of feature engineering in NLP has been shifted from hand-crafted features to distributed feature representation obtained via deep neural network models (Min et al., 2021). The availability of fast computational resources, development of the Transformer architecture, and continuous advancements in training and optimization techniques contribute to the success of large pretrained language models (Qiu et al., 2020). Pretrained language models are trained in a semi-supervised, supervised, or unsupervised manner on huge text corpora to learn the universal language representations. These representations apply to a wide variety of downstream tasks. The first generation of pretrained models includes pretrained embeddings like word2vec (Mikolov et al., 2013) and Global Vectors (GloVe) (Pennington et al., 2014). Although these embeddings were effective in capturing the semantics of words, they failed to capture the context and understand higher-level relationships between words in the text. The second generation of pretrained language models like BERT, GPT (Generative Pretrained Transformers), etc., have resolved the problem of the polysemous nature of words used in sentences and can produce a contextualized representation of words.

2.4.1. *Pretraining objectives*

The pretraining objectives are the tasks that PLMs are trained on. These objectives can be learned through supervised, unsupervised, or semi-supervised learning. In NLP, huge datasets are scarce for the supervised training of language models, so most of the PLMs are trained with self-supervised learning objectives. Masked Language Modelling (MLM) as in BERT (Devlin et al., 2019) is a well-known example of self-supervised learning where some words in the input text are masked, and the model learns to predict the masked words using the remaining words. Other common pretraining tasks include permuted language modeling, denoising auto-encoder, next-sentence prediction, etc.

Table 1
Loss functions for various pretraining objectives.

Task	Loss Function
Masked Language Modelling	$L_{MLM} = -\sum_{\hat{x} \in m(x)} \log p(\widehat{x} \mid x_{\setminus m(x)})$
Permuted Language Modelling	$L_{PLM} = -\sum_{t=1}^{T} \log p(z_t \mid z_{<t})$
Denoising Autoencoder	$L_{DAE} = -\sum_{t=1}^{T} \log p(x_t \mid \widehat{x}, x_{<t})$
Next Sentence Prediction	$L_{NSP} = -\log p(t \mid x, y)$

Table 1 mentions the loss functions for well-known pretraining tasks.

3. Related work

3.1. Abstractive summarization

Online reviews are important for managing customer policies, maintaining service quality, and helping customers in decision-making. (Yu et al., 2021) analyzed domain adaptation for low resource abstractive summarization over diverse domains, one of which is movie reviews. In this work, domain adaptation is implemented as continued pretraining or a second phase of pretraining on the BART language model under three different methods such as Source Domain Pre-Training (SDPT), Domain Adaptive Pre-Training (DAPT), and Task Adaptive Pre-Training (TAPT). In SDPT, pretraining is continued on a large, labeled corpus from the source domain e.g., the news domain. In DAPT, pretraining is continued on an unannotated data corpus from a related domain. TAPT leverages a small task-related unannotated corpus. On some of the target domain datasets, SDPT performed better while on others DAPT performed well. The domain adaptation techniques generally show an improvement in ROUGE (Recall Oriented Understudy for Gisting Evaluation) R-1 scores as compared to the standard finetuning of the BART model on the target datasets. AdaSum (Brazinskas et al., 2022) is a few-shot opinion summarization model. It proposes self-supervised pretraining of the BART (Bidirectional and Auto-Regressive Transformer) (Lewis et al., 2020) language model on customer reviews using adapters to introduce in-domain knowledge.

Jain et al. (2021) proposed MRCBert (Machine Reading Comprehension BERT) using the paradigm of transfer learning for unsupervised summarization of user opinions. In this work, the researchers utilized the machine reading comprehension technique to extract opinions from reviews and generate aspect-wise and rating-wise summaries for the Amazon reviews dataset. The target is achieved through question answering task performed by MRCBert which answers specific questions about the product to extract user opinions. The extracted opinions are fed to the abstractive summarization model to produce a review summary. The uniqueness of this work is that it does not require a labeled dataset or ground truth summaries and is unsupervised. Also, it is applicable in low-resource scenarios.

The work of Brazinskas et al. (2022) is based on the use of pre-trained adapters for the improvement of opinion summarization in a few-shot setting. Adapters are small neural network modules that are added to the pre-trained language model such as BART (Bidirectional and Auto-Regressive Transformers) to efficiently tune the model on a small target dataset. The researchers added adapters to the BART model and pretrained only the adapter modules in a self-supervised way and task-specific as well as query-based manner on an unannotated corpus of customer reviews. This gives domain-specific learning to the model. The adapters were then tuned on a small human-annotated dataset. The resulting summaries from pretrained adapter-based finetuning were better in quality and improved in terms of ROUGE scores as compared with summaries from the standard finetuning method. The summaries from query-based adapter pretraining were better in terms of coherence and exhibited low redundancy.

Bražinskas et al. (2020) presented FEWSUM, a framework for opinion summarization in a few-shot learning setting. The transformer-based conditional language model is trained using a leave-one-out objective. The model is conditioned on the review properties during training. These properties are automatically obtained from a large unannotated reviews corpus. The model is then tuned jointly with a tiny plugin network on a small target/annotated dataset consisting of 60 to 100 reviews. The model was evaluated on Amazon and Yelp datasets. The performance in terms of ROUGE scores as well as the human evaluation was better as compared to the other competing approaches.

A model for abstractive review summarization was proposed by Shobana and Murali (2021). The model's architecture consists of bidirectional LSTM as the encoder and standard LSTM as the decoder. The model implements a better attention mechanism to increase semantic understanding, a pointer generator network for handling rare words, and a coverage mechanism to avoid repetition and produce a better-quality summary. The experiment and evaluation were carried out on the Amazon mobile reviews dataset using the ROUGE metric.

3.2. Sentiment classification

The work in Bigne et al. (2023) approached sentiment classification using deep learning-based open-source tools and examined the relation between sentiment polarity and star ratings of 20,954 customer reviews related to the tourism industry collected from TripAdvisor. Their findings reveal conformity and alignment between customer sentiment orientation and the review star rating. Setiyawan et al. (2021) conducted lexicon-based sentiment analysis on 2937 reviews collected from an Indonesian e-commerce platform. The researchers also investigated the relationship between customer sentiment and review rating. Their findings reveal that most of the time, customers give a high rating with a straightforward opinion and a positive sentiment. The work of (Iddrisu et al., 2023) emphasized on identifying the sarcastic language in aviation reviews and proposed a three operator framework (*Assemble+Deft, Edify+Authenticate, Forecast*) for the analysis and classification of sarcastic and non-sarcastic sentiments within the aviation sector using RNN with GRU (Gated Recurrent Unit) and Support Vector Machine (SVM) algorithms.

Munikar et al. (2019) used the BERT model to tackle the fine-grained sentiment classification task on the Stanford Sentiment Treebank (SST) dataset. Their findings suggest that BERT outer performed other baselines based on RNN, LSTM, BiLSTM, and CNN (Convolutional Neural Network). Ullah et al. (2020) built a sentiment classifier based on text and emoticons for airline data from Twitter. The textual and emoticon data were analyzed individually and in combination to determine sentiments using machine learning and deep learning techniques. Various features like TF-IDF (Term Frequency – Inverse Document Frequency), Bag of Words (BoW), emoticon lexicon, etc. were used during the process. The research findings suggest better results for sentiment prediction with deep learning-based models as compared to machine learning models. Further, text and emoticon combined data present better results as compared to text-only data.

A hybrid RoBERTa-LSTM model for sentiment classification was proposed in the work of Tan et al. (2022). RoBERTa stands for Robustly Optimized BERT Pre-training Approach. The work utilized data augmentation techniques to alleviate the class imbalance issue by up sampling the minor classes. Roberta can produce better embeddings while LSTM effectively captures long-distance semantics. The combination of architectures leverages the strengths of both architectures to achieve the target performance. Alduailej & Alothaim (2022) proposed AraXLNet by pretraining the XLNet (eXtreme Language understanding Network) language model on a large Arabic dataset and finetuning it on an annotated Arabic tweets dataset from Twitter for Arabic language sentiment classification. The preprocessing was conducted using the Farasa segmenter. The findings indicate the model's superior performance on several Arabic language benchmark datasets.

4. Research methodology

The methodology for this research work as illustrated in Fig. 3, comprises three major phases: Data, Analysis, and Modelling & Evaluation.

4.1. Data

During the data phase, the datasets of airline reviews for abstractive summarization, domain adaptive training, and sentiment classification

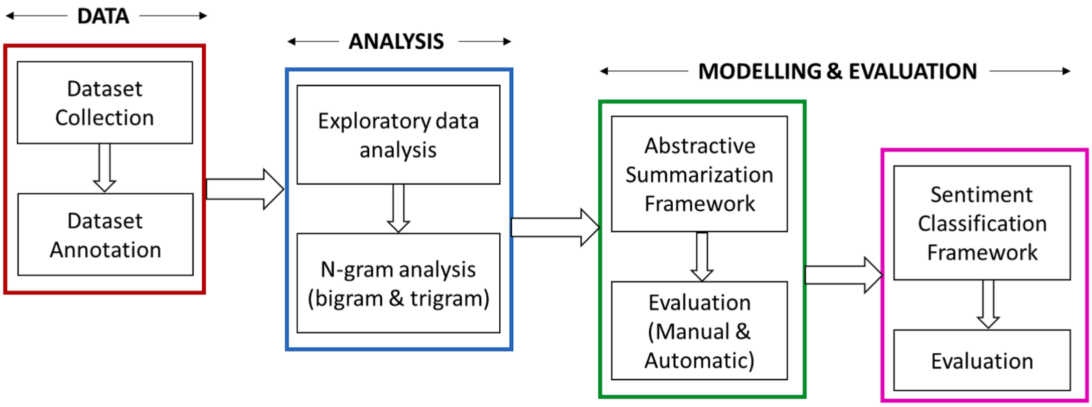

DATA
Dataset Collection
Dataset Annotation
ANALYSIS
Exploratory data analysis
N-gram analysis (bigram & trigram)
MODELLING & EVALUATION
Abstractive Summarization Framework
Evaluation (Manual & Automatic)
Sentiment Classification Framework
Evaluation

Fig. 3. Methodology of Current Research.

were collected and annotated.

4.1.1. Data collection

The datasets were collected from the Skytrax (https://www.airlinequality.com/) review portal via web scraping using Python Requests and Beautiful Soup libraries. An HTTP request was sent to the website by specifying the website's URL via Requests library. The data of interest was extracted from the HTML content using the Beautiful Soup library. For abstractive summarization and title generation datasets, the 'reviews' and 'titles' were retrieved. For the sentiment classification dataset, additional tags like 'rating' and 'recommendation value' were also retrieved. The data was stored in CSV files.

4.1.2. Data annotation

The datasets for abstractive summarization and sentiment classification were annotated based on the required tasks. For the abstractive summarization dataset, abstractive summaries were manually written for each review following a certain criterion. The summary writing criteria requires the summary to include an expression of customer opinion towards various flight aspects including seat, comfort, food, entertainment, etc. as well as an overall opinion towards the product (airline). The summaries were written with clear and consistent language rules (structure and grammar). The summary size was maintained around 30–40 % as of the original review. For the sentiment classification dataset, sentiment class was assigned based on customer rating value. The class assignment rule is presented in Table 2.

4.2. Analysis

After the collection and annotation processes, the prepared datasets enter the analysis phase. Two approaches were taken for data analysis. The first one is exploratory data analysis and the second is language-based or n-gram data analysis.

4.2.1. Exploratory data analysis (EDA)

During exploratory data analysis, we analyzed the datasets visually for various review text statistics including word frequency analysis, review length analysis in terms of the number of words and characters, average word length, token counts, rating distribution, recommendation value distribution, kernel density plot, correlation between variables like customer rating, recommendation, and sentiment, etc. Visualizations were created using Python's matplotlib. This data analysis and visualization offered several insights into how variables are related and how this correlation between variables can be used to design better models.

Table 2
Sentiment class assignment based on customer rating.

Rating	Sentiment Class
1–4	Negative
5–6	Neutral
7–10	Positive

4.2.2. N-gram analysis

N-grams are single words or combinations of words and are more commonly known as unigrams, bigrams, and trigrams. These are useful for slicing data into meaningful word combinations that can identify the trends in the data and offer several insights from the data based on the requirements. For n-gram analysis, the datasets were first pre-processed. Preprocessing is cleaning and transforming raw data to make it useful for further analysis. The cleaning process was based on removing punctuations, URLs, stop words, and lowercasing the text, followed by tokenization, stemming, and lemmatization processes. The cleaned data was then analyzed for bigram and trigram networks, the n-gram analysis across sentiments and recommendation value. The data was visualized in the form of network diagrams, word clouds, and bubble charts.

4.3. Modelling & evaluation

The final phase is the modeling and evaluation phase where the models for abstractive summarization and sentiment classification were designed and assessed. The datasets for abstractive summarization and domain adaptation were utilized in training the abstractive summarization model while the sentiment classification dataset was used to train the rating-based sentiment prediction model. The models were evaluated for performance using specified metrics. In the case of abstractive summarization, the model's performance was evaluated via automatic as well as human evaluation.

4.3.1. Abstractive summarization modelling and evaluation

Fig. 4 illustrates the methodology undertaken for abstractive summarization modeling. During the data preparation phase, the datasets were split into train, validation, and test sets. Next, the data is tokenized using the Pegasus tokenizer. A tokenizer is used for preparing the inputs for a model. During the tokenization process, the strings are split into sub-word tokens. These tokens can then be readily converted to numerical data.

The next step is the implementation of transfer learning by selecting a pretrained language model, in our case the PEGASUS language model. First, the PEGASUS model was finetuned in a standard way on the abstractive summarization dataset for airline reviews. Second, PEGASUS was finetuned on an intermediate task of review title generation (first stage finetuning). The finetuning process was then continued to the second stage to make the model learn the final task of review summary generation.

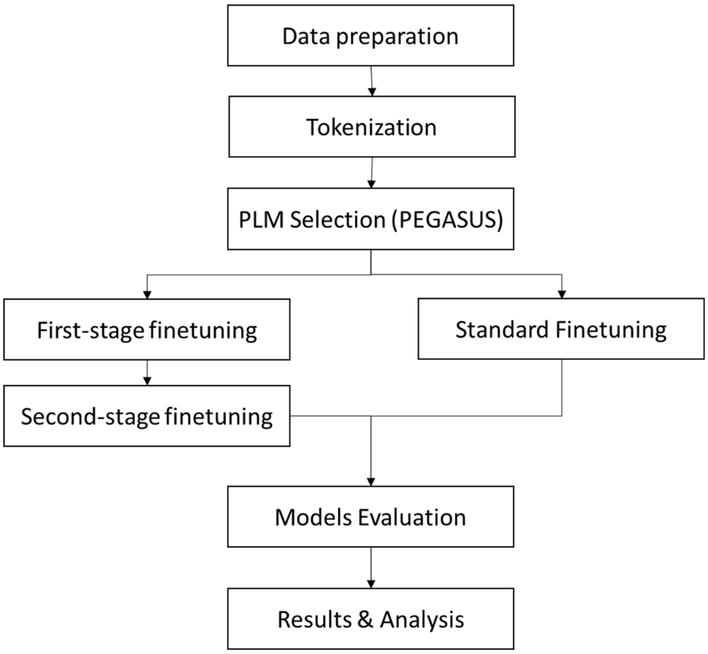


Fig. 4. Methodology for abstractive summarization modelling and evaluation.

The models were evaluated using automatic and human evaluation methods. For automatic evaluation ROUGE and BERT score metrics were utilized. The manual evaluation was conducted through the Best Worst Scaling (BWS) method. The results of models with standard finetuning and two-stage finetuning were compared and analyzed for performance. Finally, the results were also compared to other existing studies implementing domain adaptation for abstractive summarization to examine the impact of two-stage finetuning as a domain adaptation strategy.

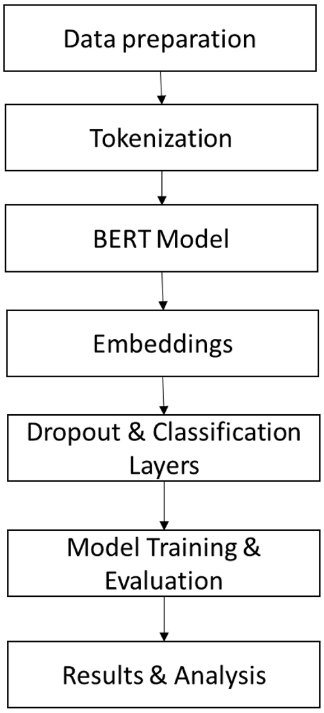


Fig. 5. Methodology for sentiment classification modeling and evaluation.

4.3.2. Rating-based sentiment classification modelling and evaluation

Fig. 5 portrays the methodology for modelling the rating-based sentiment classification problem. The initial phases are the data preparation and tokenization followed by BERT model, input embeddings, and the addition of dropout layer and fully connected classification layer on top of BERT. The next step is the model training followed by the evaluation phase.

The evaluation metrics utilized were accuracy, precision, recall, F1-score, and confusion matrix. The evaluation was conducted using the conventional random train-val-test split as well as via 5-fold cross-validation. The results were analyzed and compared with the existing studies to demonstrate the effectiveness of the rating-based sentiment prediction model. As part of the model evaluation, the finetuned BERT-based model was also compared to unsupervised approaches like VADER (Valence Aware Dictionary for Sentiment Reasoning) and BERT pipeline, standard machine learning methods like Support Vector Machines (SVM) and Decision Trees, the deep learning architectures like the Recurrent Neural Networks and Long-Short Term Memory Networks. The results from all the models were analyzed and discussed.

5. Datasets

This section introduces the datasets for abstractive summarization, domain adaptation, and rating-based sentiment classification of airline reviews. The datasets have been collected from the Skytrax reviews portal (https://www.airlinequality.com/) using Python library packages.

5.1. Dataset for abstractive summarization

This dataset contains review-summary pairs. The shape of the dataset is represented by a tuple (500,2) where there are 500 rows and 2 columns. Fig. 6 shows the review size and summary size distribution across the abstractive summarization dataset. The longest review is around 600 words. Most of the reviews are under 200 words. Almost all the summaries are under 100 words. The mean compression ratio between reviews and summaries is 26.3 %.

Fig. 7 shows the most common words found in reviews and abstractive summaries. The stop words are excluded before generating

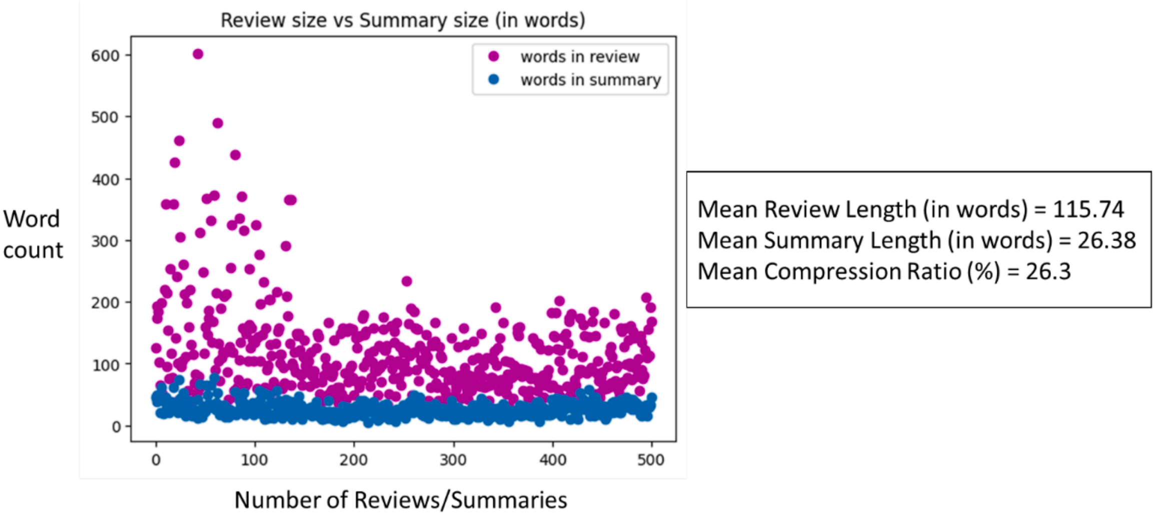


Fig. 6. Review size and summary size distribution in Abstractive Summarization dataset.

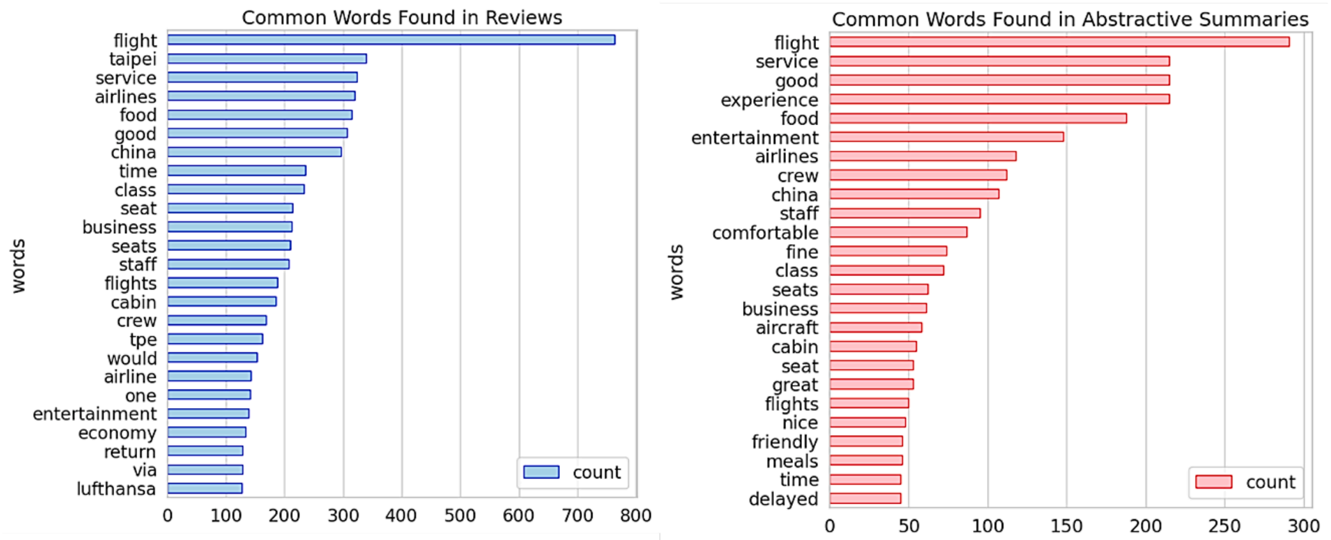


Fig. 7. Common words found in Reviews and Summaries.

these plots. It can be noted from these graphs that many words are overlapping between reviews and summaries indicating that the summaries cover most of the topics discussed in reviews.

Figs. 8 and 9 present the network of bigrams and trigrams found in the abstractive summarization dataset. These network diagrams are useful for exploring relationships between co-occurring words.

5.2. Dataset for domain adaptation

The dataset for domain adaptive training consists of review and review title pairs. It has 7079 data entries and two columns. Fig. 10 shows statistics of this dataset, such as character distribution, word distribution, and average word length across customer reviews and review titles.

5.3. Sentiment classification dataset

The sentiment classification dataset consists of five columns; review, title, rating, recommended, and sentiment. The sentiment column is manually annotated based on customer rating value. Fig. 11 reveals the sentiment class distribution as well as the token count density distribution across the dataset. Most samples belong to the negative class. There are about 700 reviews representing the negative class, 300 reviews for the positive class, and 100 for the neutral class.

Figs. 12 and 13 present word clouds of bigrams and trigrams for reviews exhibiting positive and negative sentiments. The bigrams and trigrams provide information and context for customers' positive and negative sentiments. For example, in the bigram word clouds, positive sentiment mostly relates to 'business class and 'cabin crew', while negative sentiment mostly relates to 'customer service' and 'Air Canada'. An interesting observation is the bigram 'Air Canada' that appears in both positive and negative word clouds. However, the bigram frequency for 'Air Canada' on the negative side is far more as compared to the positive one, signaling that most customers express negative opinions on 'Air Canada' rather than positive.

Fig. 14 presents a pie chart showing statistics of airline customers who 'recommended' versus 'not recommended' certain airlines. The pie chart indicates that most of the customers 746/1100 do not recommend their experienced product. Only 354/1100 customer reviews give a 'yes' to the recommendation.

Fig. 15 shows a bar plot representing customer rating distribution across the dataset. Most reviewers give a rating of 1/10. The rating distribution also explains the majority of 'negative sentiments' and 'not

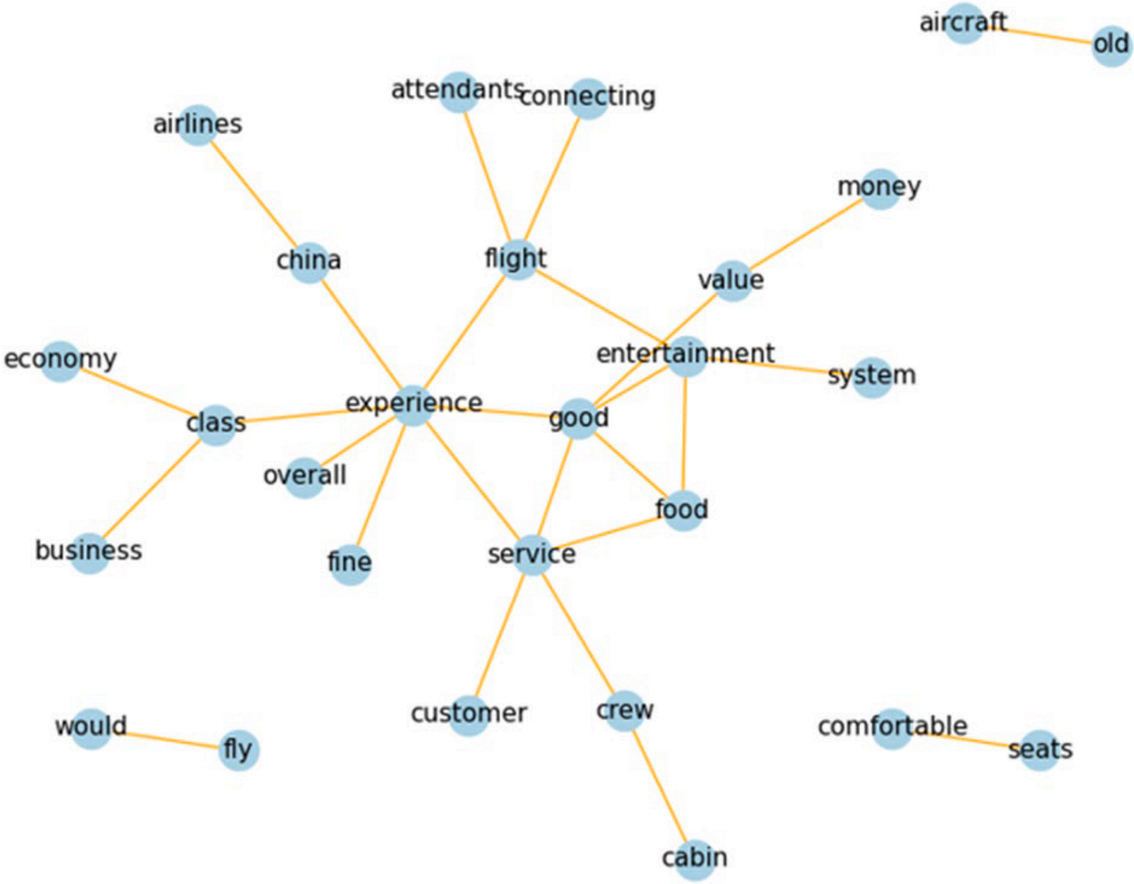


Fig. 8. Bigram network in abstractive summarization dataset.

recommended' in the review dataset.

Fig. 16 shows a contour plot representing the bivariate probability distribution for customer rating and sentiment with recommendation value mapping. The graph clearly explains the relationship between customer rating, sentiment, and recommendation value. The lower left region that corresponds to negative sentiment and low rating values shows a recommendation value of 0 (no) while the upper right region with positive sentiment and high rating values indicates a recommendation value of 1 (yes). However, in reviews with neutral sentiment or middle rating values, both recommendation values (yes & no) are present. It can also be observed that regions with negative recommendations are denser as compared to regions with positive recommendations. The same condition is there when the sentiment is neutral.

The correlation matrix in Fig. 17 represents the correlation patterns between customer rating, recommendation value, and sentiment. From the correlation matrix, rating and sentiment exhibit a correlation value of 0.96. The correlation between recommendation value and sentiment is 0.91 while between rating and recommendation value, the correlation is 0.89. All the variables show a significant relationship and a close association with one another.

Figs. 18 and 19 illustrate bubble charts of trigrams in reviews with positive recommendations and those with negative recommendations. The size of the bubble represents the frequency of the trigram. These charts help analyze and relate information that led to positive or negative recommendation decisions by the customer

In Figs. 18 and 19, the major trigrams that reflect positive recommendation are 'business class passengers', 'business class product', 'long haul flight' while major trigrams reflecting negative recommendation are 'worst customer service', 'worst airline ever' and 'missed connecting flight'.

6. Abstractive summarization framework

6.1. Task formulation

We model the abstractive summarization of airline reviews as a sequence-to-sequence text generation task. Given a review R consisting of a sequence of n words $R = \{r_1,\ r_2,\ r_3,\ \ldots.,\ r_n\}$, the abstractive summarization model generates a concise summary of m words for the review as $S = \{s_1,\ s_2,\ s_3,\ \ldots.,\ s_m\}$ where $m < n$ by finding a mapping from $R \rightarrow S$. Moreover, we have a language model L pretrained on a source domain d_p and we have a target domain consisting of airline reviews d_r. With varying source and target domains $(d_p \neq d_r)$ and similar source and target tasks, the summarization model learns to adapt and transfer knowledge effectively from the source domain to the target domain.

6.2. The proposed model

This section introduces our proposed model. The backbone architecture implementing transfer learning in our framework is the pretrained PEGASUS language model. PEGASUS is based on a standard Transformer encoder-decoder architecture. We used PEGASUS$_{\text{LARGE}}$ in our experiments which has 16 encoder-decoder transformer blocks, a hidden layer of 1024 neurons, and a feedforward layer size of 4096. Our model goes through two learning stages. The first stage is for domain adaptation purposes and implements intermediate finetuning of the PEGASUS model on the review title generation task. The second stage further trains the model on the final task of an abstractive summary generation. The model is presented in Fig. 20.

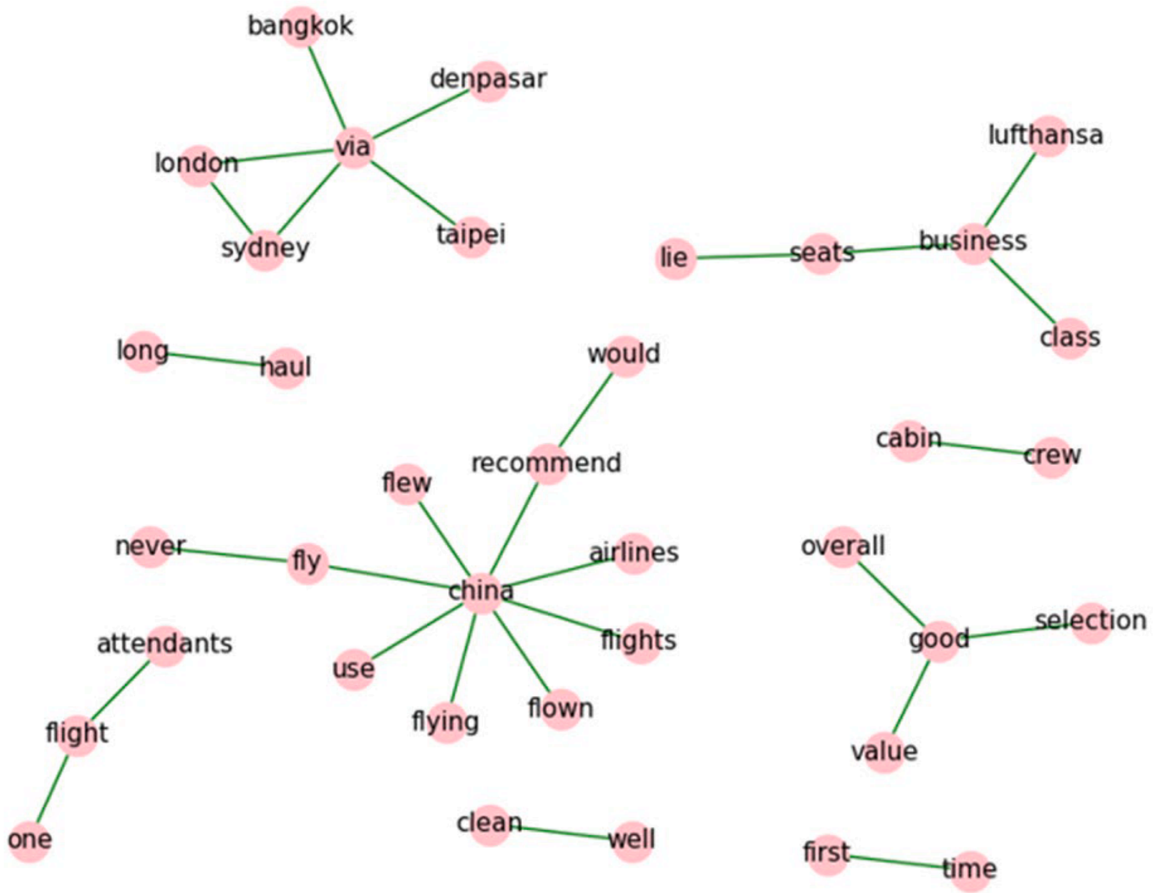


Fig. 9. Trigram network in abstractive summarization dataset.

6.2.1. *The encoder function*

The encoder block produces a contextualized encoded sequence from the input sequence as:

$$g_\theta : R_{1:n} \rightarrow \overline{R}_{1:n}. \tag{1}$$

Each encoder has a bidirectional self-attention mechanism to relate each input vector with other vectors thereby converting each input vector into its contextual representation. For example, the input vector r_j is transformed into a contextualized representation $r_j^{'}$ where $j \in \{1, \ldots, n\}$. The bidirectional self-attention projects each vector of the encoder input sequence to a key (k), query (q), and value (v) vector using the corresponding weight matrices w_k, w_q, and w_v as:

$$k_i = \omega_k r_i, \tag{2}$$

$$q_i = \omega_q r_i, \tag{3}$$

$$v_i = \omega_v r_i \tag{4}$$

The importance of the value vector is determined by comparing each of the query vectors to all the key vectors. The high similarity between the query and key vector gives more weight to the value vector. After applying the self-attention, the encoder output is computed as:

$$R^{'}_{1:n} = softmax\left(Q_{1:n} K^T_{1:n}\right) V_{1:n} + R_{1:n} \tag{5}$$

The encoder component consists of multiple encoders, in this case 16. The representation created by the first encoder is passed on to the following encoders which further refine the contextual information in the encoded representation until the final encoded representation $\overline{R}_{1:n}.$ is obtained.

6.2.2. *The decoder function*

The decoder takes the contextualized sequence from the encoder and determines the conditional probability distribution of the target output as a product of the conditional distributions of the target vector given the previous target vectors and the encoder representation.

$$p_\theta(S_{1:m}|\overline{R}_{1:n}) = \prod_{i=1}^{m} p_\theta(s_i|S_{0:i-1}, \overline{R}_{1:n}) \tag{6}$$

The decoder is a stack of decoder blocks followed by a language modeling head. To obtain a quality representation of the next target vector, each decoder consists of a unidirectional self-attention and an encoder-decoder cross-attention. The unidirectional self-attention relates each input vector $s_j^{'}$with all the previous input vectors $s_i^{'}$ and is expressed as:

$$s_i^{''} = V_{0:i} * softmax\left(K^T_{0:i} * Q_i\right) + s_i^{'} \tag{7}$$

The encoder-decoder attention relates each of its inputs with all the contextualized vectors from the encoder and is computed as:

$$s_i^{'''} = V_{1:n} * softmax\left(K^T_{1:n} * Q_i\right) + s_i^{''} \tag{8}$$

Each decoder in the decoder stack maps the encoded sequence $\overline{R}_{1:n}$ and the target vector sequence $S_{0:i-1}$ to an encoded sequence of target vectors $\overline{S}_{0:i-1}$ which is then mapped to the logit vector sequence $L_{1:n} = l_1$, l_2, …., l_n. The dimension of the output logit vector is that of the size of the vocabulary. The SoftMax function transforms each logit vector to a conditional probability distribution. The conditional probabilities of all

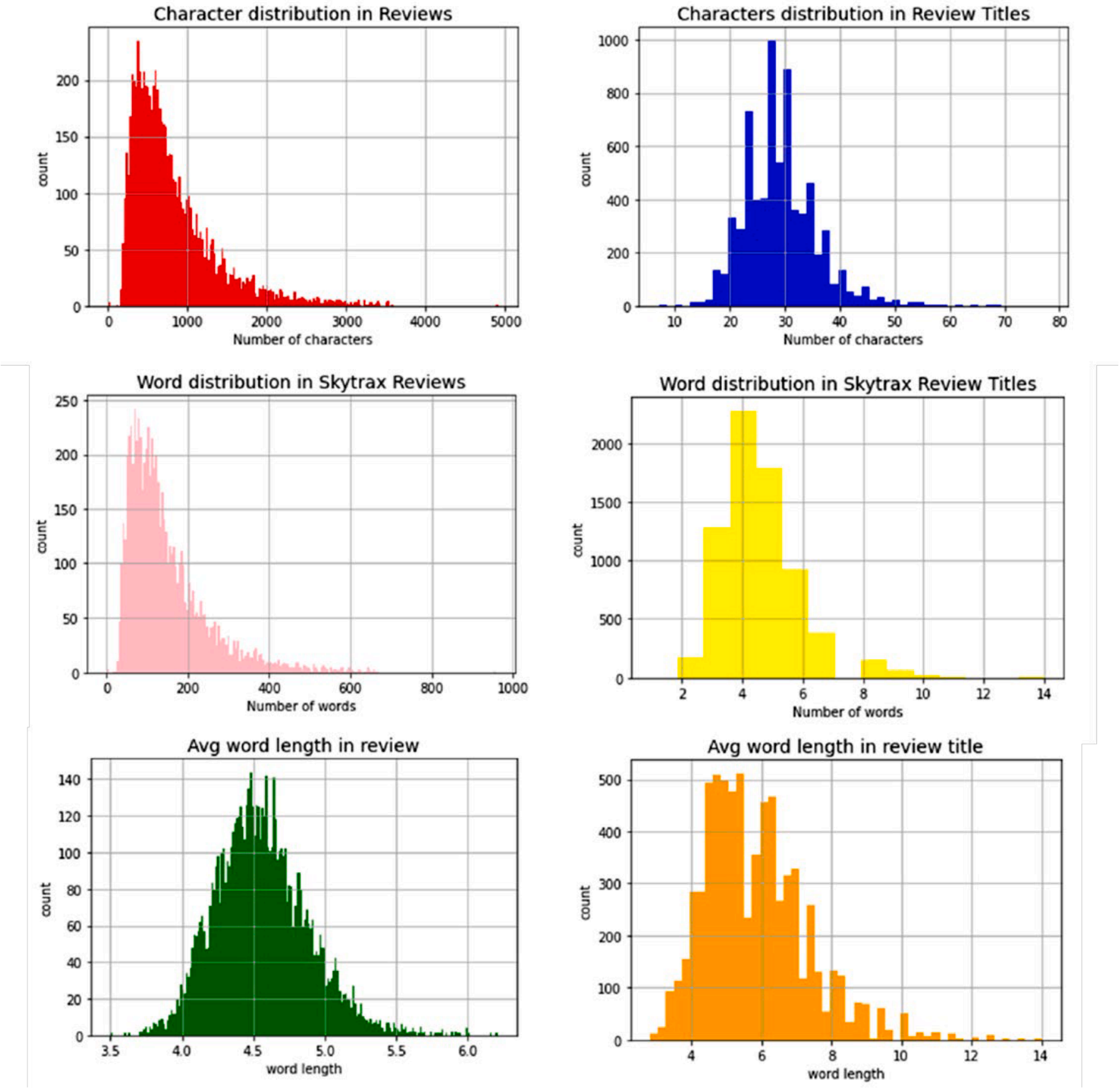


Fig. 10. Statistics of the domain-adaptive training dataset (review-title pairs).

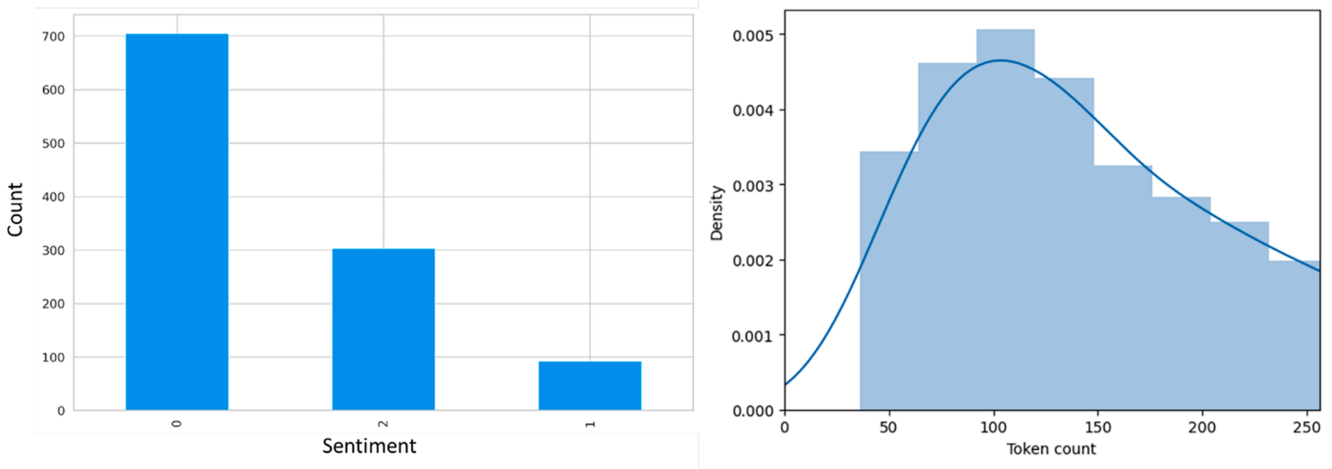


Fig. 11. Sentiment class distribution and token count density distribution in sentiment classification dataset.

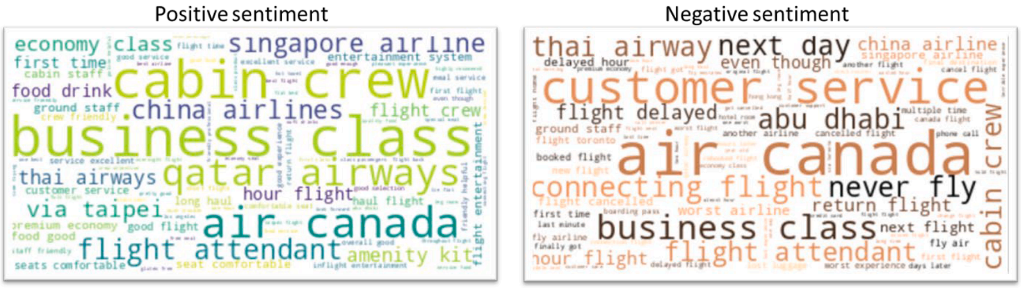


Fig. 12. Word cloud of bigrams representing positive and negative sentiment.

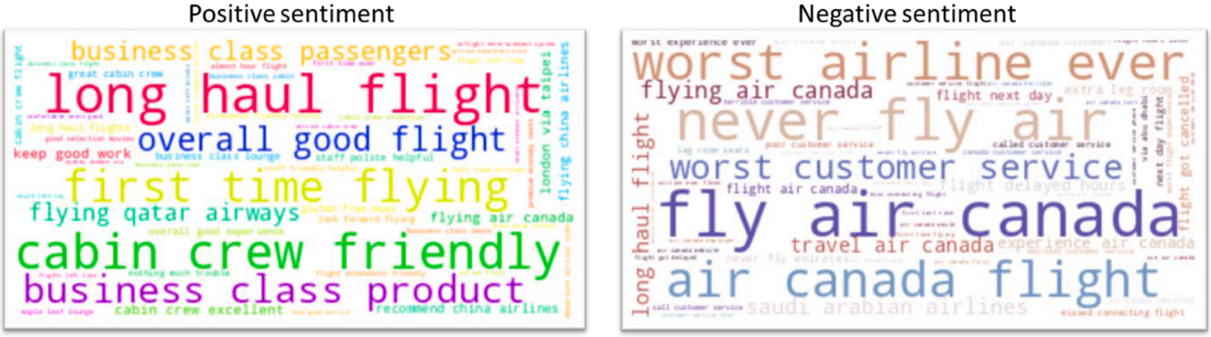


Fig. 13. Word cloud of trigrams representing the positive and negative sentiment.

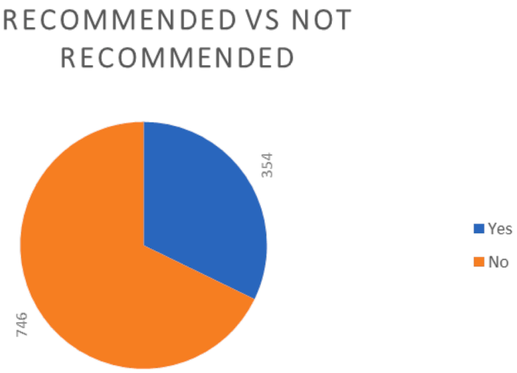


Fig. 14. Recommendation value distribution across the sentiment classification dataset.

the target vectors are multiplied to reveal the final conditional probability distribution of the target sequence.

6.3. Experiment details

The abstractive summarization model has two finetuning stages. During the first stage of finetuning, we tuned PEGASUS$_{\text{LARGE}}$ on the domain-specific data (airline reviews and titles) using a learning rate of 5×10^{-5}. We specified a learning rate scheduler of type ‘linear’. We used the Adam optimizer with exponential decay rates β_1 and β_2 as 0.9 and 0.99, and epsilon value as 1×10^{-8} respectively. We maintained a training batch size of 1 and an evaluation batch size of 8. We continued finetuning up to 150,000 steps and 30 epochs. For the second stage/final tuning, we used our intermediate-tuned model as the base model and continued training on the final task of abstractive summary generation with the same set of hyperparameters as in the intermediate tuning stage but now we trained for 5 epochs and 2000 steps. Till this point, we were able to achieve a balanced training and validation loss reduction.

To investigate the impact of two-stage tuning, we also obtained a model with vanilla or standard finetuning of the PEGASUS-XSUM model on an abstractive summarization dataset of airline reviews. We continued training for 5 epochs and 2000 steps. The training and validation losses for models with standard finetuning and two-stage tuning are illustrated in Fig. 21.

The variation in loss reduction between models at a few time steps is presented in Table 3. It is worth notable that the model with two-stage tuning for domain adaptation reduces loss efficiently on the final task as compared to the standard model, thus reducing the training time, and increasing the training efficiency.

6.4. Evaluation

The evaluation of models is conducted using ROUGE (Recall Oriented Understudy for Gisting Evaluation) scores (Lin, 2004), the BERT Score metric (Zhang et al., 2020), BLEU (BiLinngual Evaluation Understudy) score (Papineni et al., 2001) and human evaluation. The ROUGE metric is the well-known metric for the evaluation of automatic summarization and is based on the concept of recall and computes n-gram overlap between the model summary and one/more reference summaries. We have used ROUGE-1, ROUGE-2, and ROUGE-L for assessing our models’ performance. ROUGE-1 (R1) measures the unigram overlap between the model/candidate summary and the ground truth/reference summary. ROUGE-2 (R2) determines the bi-gram overlap between the model summary and the ground truth. ROUGE-L is based on LCS (Longest Common Subsequence) and computes the longest overlapping sequence of words between the model summary and the gold reference. BERT Score is based on BERT contextual embeddings and measures the similarity of each token in the model summary with each token in the reference summary. However, unlike ROUGE, that measures the surface level n-gram overlap, the BERT Score measures the sematic similarity using each token’s contextual embeddings. In literature, BLEU has been used by some researchers to complement the ROUGE metric for evaluation of the quality of the model generated summaries against gold references. The summarization models presented in this research also use multiple metrics including BERT Score,

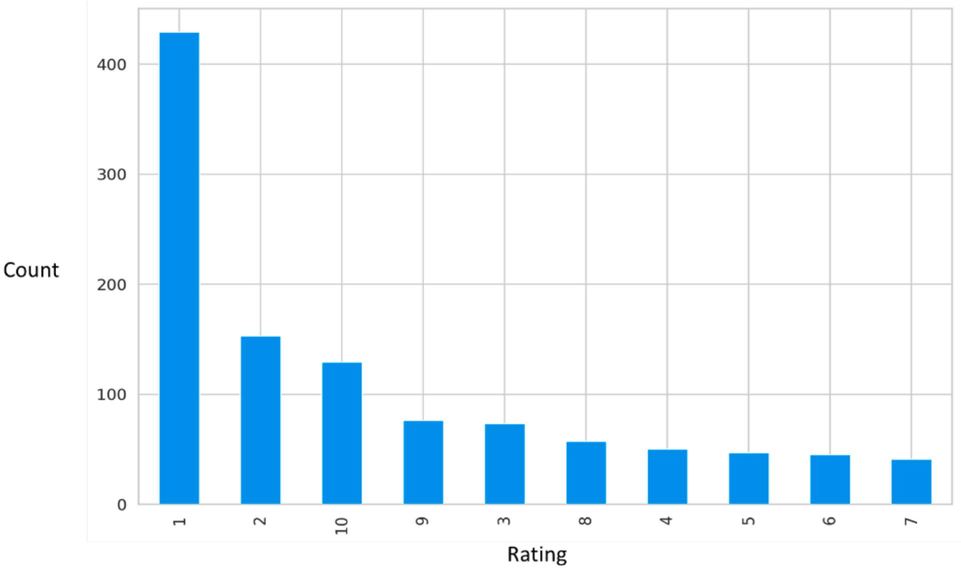


Fig. 15. Customer rating distribution across sentiment classification dataset.

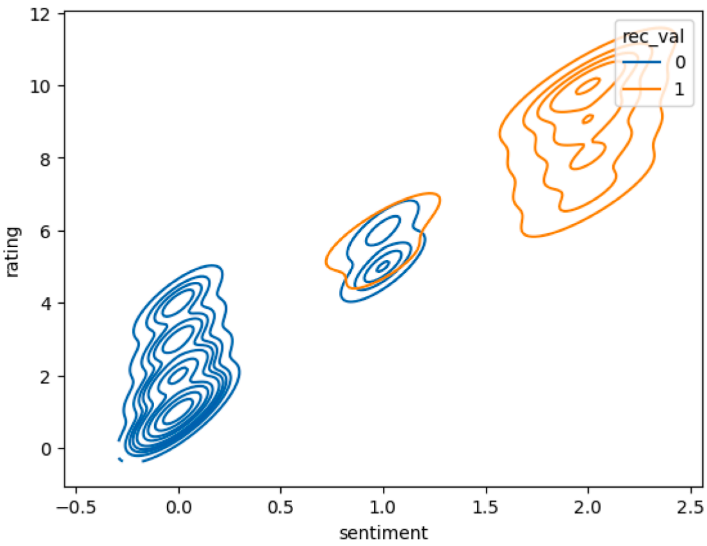


Fig. 16. Bivariate kernel density plot of customer rating and sentiment class with recommendation value mapping.

ROUGE, and BLEU to obtain comprehensive performance evaluation of the models.

6.4.1. *Automatic evaluation*

The abstractive summarization model has been evaluated from two perspectives using automatic metrics:

1. First, we have proposed PLM based approach using finetuning PEGASUS language model for abstractive summarization of airline reviews in a limited data scenario. To demonstrate the usefulness of the proposed approach for the kind of airline reviews dataset used, it is also compared with non-pretraining approach, that is the standard transformer model. Furthermore, the evaluation metrics ROUGE and BERT Score are complemented with the BLEU (BiLingual Evaluation Understudy) metric for a more comprehensive evaluation of the models. The results are presented in Table 4.

The evaluation results reveal that two-stage finetuning model surpass other models in terms of most of the metric scores. For the standard transformer model, the scores are significantly lower as compared to the transfer learning-based models on our airline reviews dataset. The reason for this performance is the smaller number of samples in the training dataset. Because we are training a transformer encoder-decoder architecture from scratch and to learn and perform better on the target task, it needs a larger training dataset. Due to limited training data, the summaries predicted with the transformer model were inconsistent, repetitive, barely informative, and lack meaning. On the other hand, the summaries produced with transfer learning models were better at fluency, coherence, informativeness, and exhibited non-redundancy. The reason is that pretrained language models possess a deeper language understanding ability including grammar, syntax, semantics, and relationships between the textual components. Finetuning these models on the target task data enhances their capabilities on the target task resulting in a good performance at the final task.

2. Second, we have proposed two-stage finetuning as a domain adaptation strategy for abstractive summarization of airline reviews. We have compared our proposed approach to alternative methodologies

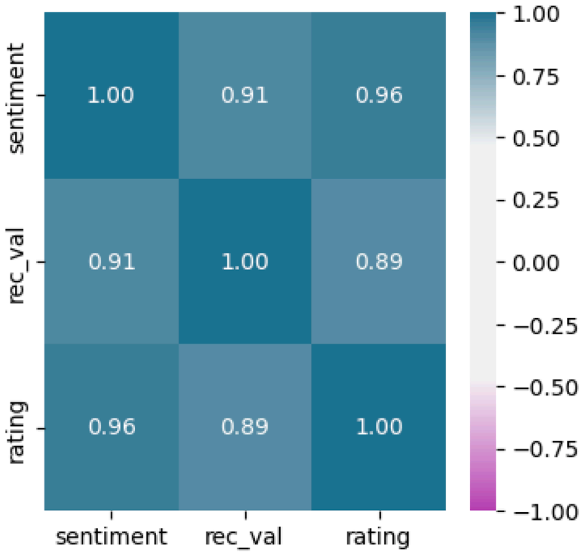


Fig. 17. The correlation heatmap between rating, recommendation value (rec_val), and sentiment.

existing in literature for domain adaptation in abstractive summarization. For this purpose, we have compared our works with several works ((Yu et al., 2021), (Bražinskas et al., 2022), and (Hoang et al., 2019)) implementing the domain adaptation for abstractive summarization. Table 6 presents a comparison of various domain adaptation techniques with our proposed two-stage finetuning approach.

Referring to Table 6, AdaptSum (Yu et al., 2021) is a low-resource abstractive text summarization model comprising multiple target domains that implements domain adaptation using three methods: SDPT (Source Domain Pretraining), DAPT (Domain Adaptive Pretraining), and TAPT (Task Adaptive Pretraining) with and without the use of RecAdam (Chen et al., 2020) optimizer. AdaSum (Bražinskas et al., 2022) is a few-shot opinion summarization model. It proposes self-supervised pretraining of the BART (Lewis et al., 2020) language model on customer reviews using adapters to introduce in-domain knowledge. The work of Hoang et al. (2019) is an abstractive summarizer with efficient adapting capability. It proposes domain adaptive training to adapt the transformer based GPT language model to the newswire text data. Our approach to domain adaptation uses two-stage finetuning, where we first finetuned the PEGASUS model on an intermediate task of title generation and then finally tuned the model on abstractive summarization task.

In Table 6, the models are compared in terms of improvement in ROUGE scores obtained with different approaches to domain adaptation. The improvement is computed by taking the difference between the ROUGE scores obtained with standard finetuning versus ROUGE scores obtained with various domain adaptation methods. As noted, our two-stage tuning model achieves an increment of 2.7 R2 points which is the best R2 improvement score as compared to models using alternative domain adaptation techniques. These results support the effectiveness of the two-stage finetuning approach for domain adaptation, reduced overfitting, and better knowledge transfer, leading to improved final task performance as compared to the pretraining-based domain adaptation approaches e.g., SDPT, DAPT, TAPT, etc., that come with a high computational expense and training costs.

6.4.2. *Human evaluation*

The summaries from the models with vanilla finetuning and two-stage finetuning have been evaluated manually using the Best Worst Scaling (BWS) method as in (Bražinskas et al., 2020). Since the evaluation set was small, two people (language experts) were employed to evaluate the summaries. The summaries were assessed against reference summaries as well as original reviews. The evaluation criteria included summaries assessment based on fluency, coherence, non-redundancy, informativeness, and sentiment. Fluency measures whether the summary is well-understandable, grammatically correct, and, easy to read. Coherence inspects the summary structure and organization of information. Non-redundancy determines whether there is any redundant information in the summary. Informativeness measures the information coverage of the summary when estimated against the reference summary and the original review. Sentiment defines how well the summary expresses the sentiment of the original review.

The scores for each criterion are computed based on the difference between the percentage of times a model was selected as the first best

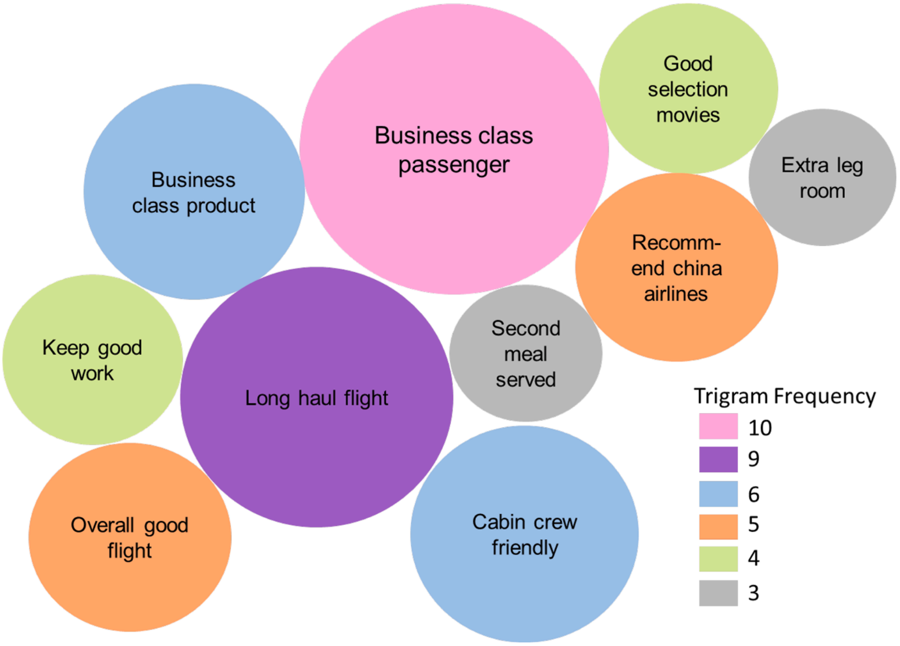


Fig. 18. Bubble chart of trigrams in reviews with a positive recommendation.

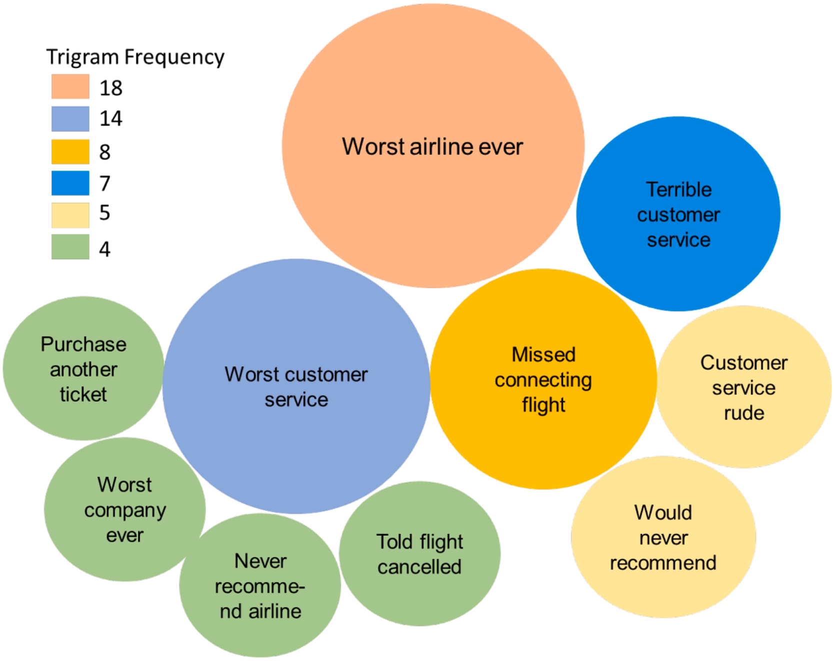


Fig. 19. Bubble chart of trigrams in reviews with a negative recommendation.

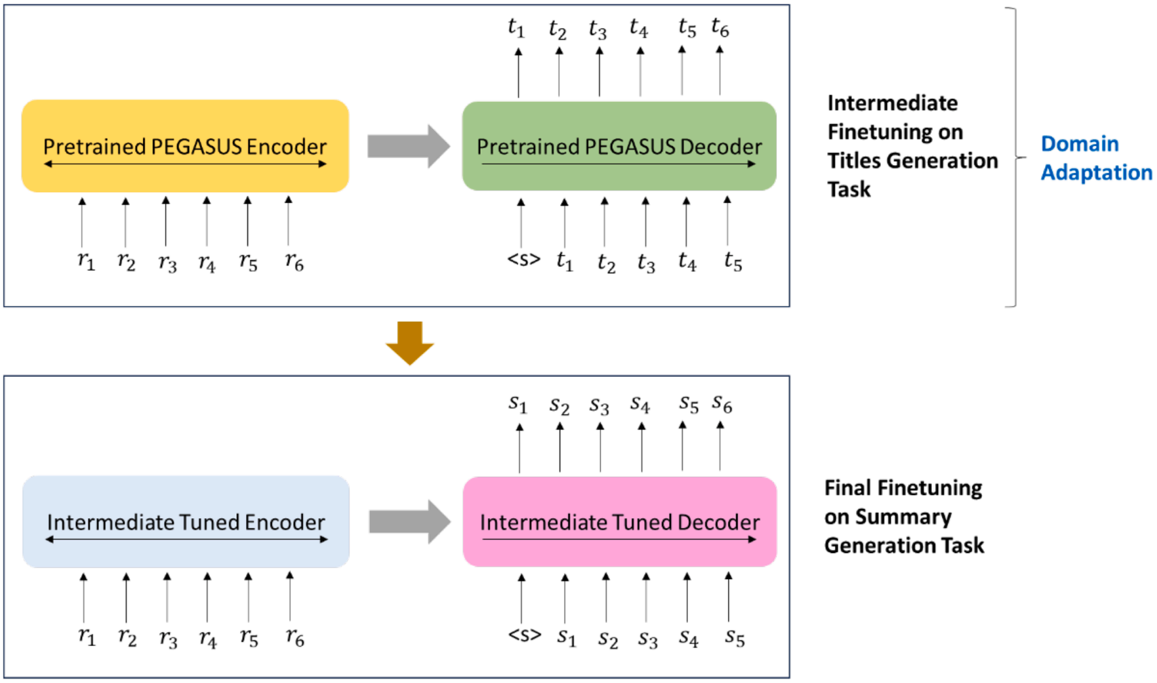

t_1 t_2 t_3 t_4 t_5 t_6
Pretrained PEGASUS Encoder
Pretrained PEGASUS Decoder
r_1 r_2 r_3 r_4 r_5 r_6
<s> t_1 t_2 t_3 t_4 t_5
Intermediate Finetuning on Titles Generation Task
Domain Adaptation
s_1 s_2 s_3 s_4 s_5 s_6
Intermediate Tuned Encoder
Intermediate Tuned Decoder
r_1 r_2 r_3 r_4 r_5 r_6
<s> s_1 s_2 s_3 s_4 s_5
Final Finetuning on Summary Generation Task

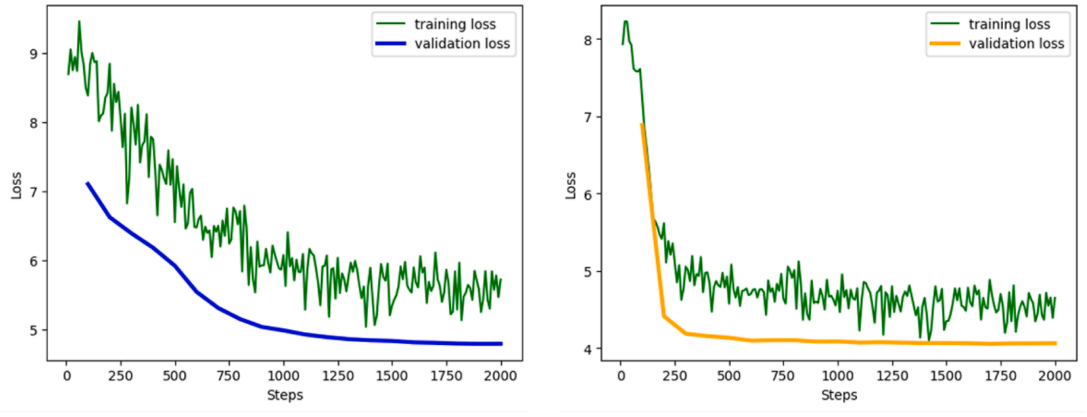
Fig. 20. Framework for abstractive summarization of airline reviews.

and the percentage of times it was selected as the second best. The scores range between −0.1 and +0.1 with positive scores inclining toward 1st best and negative scores representing the second best. The results are presented in Table 5. It is noted that the two-stage tuned model outperformed the standard model for each criterion. The results are also in line with automatic evaluation results.

7. Sentiment classification framework

7.1. Task formulation

We model rating-based sentiment prediction of airline reviews as a multiclass classification problem. Given a review R consisting of a sequence of n words $R = \{r_1, r_2, r_3, \ldots., r_n\}$ and a set of customer ratings over L different kinds of sentiment classes ($L \in \{0, 1, 2\}$), the sentiment classification model predicts an overall sentiment score based on customer rating. The sentiment categories over customer ratings are predefined where class 0 indicates negative sentiment, class 1 specifies



Standard finetuning

Two-stage finetuning

Fig. 21. Training & validation losses for model with standard finetuning vs two-stage finetuning.

Table 3
Variation in loss reduction – standard finetuning vs two-stage finetuning.

step	Model (vanilla finetuning) loss	Model (two-stage tuning) loss
50	8.7411	7.9192
250	8.0597	5.0974
750	6.2406	4.6374
990	5.8669	4.5501
1500	5.3431	4.3464

Table 4
Results of automatic evaluation – ROUGE scores and BERT score.

Model	Rouge Scores R1/R2/RL	BERT Score	BLEU Score
Two-stage finetuning	30.6/9.9/24.9	0.54	0.094
Standard finetuning	32.6/7.2/24.1	0.54	0.068
Standard Transformer	20.3/1.7/17.9	0.51	0.022

neutral sentiment, and class 2 denotes positive sentiment.

7.2. *The proposed model*

Our model for rating-based sentiment classification of airline reviews is built on top of the pretrained BERT language model (Fig. 22). BERT (Bidirectional Encoder Representations from Transformers) model is pretrained in an unsupervised manner on large text corpora using Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) pretraining objectives to obtain deep bidirectional (left and right) contextual language representations. Our architecture includes a dropout layer for regularization purposes and a SoftMax classifier as an output layer on top of the BERT language model.

7.2.1. *Preprocessing*

- The text is tokenized using the Bert tokenizer that converts sentences to words and sub-words with a vocabulary from the WordPiece algorithm.
- BERT-specific tokens including [CLS] and [SEP] are inserted at the beginning and the end of the sentences.
- Since Bert accepts constant-sized inputs, we specified a maximum length and introduced padding.
- The attention masks are created with padded tokens and original tokens.

Table 5
A comparison of various domain adaptation techniques with two-stage finetuning.

Domain	Base PLM	Model	R1/R2/RL	Improvement R1/R2/RL
Movie Reviews	BART	AdaptSum stand. (Yu et al., 2021)	25.13/9.22/20.04	—
		AdaptSum SDPT	26.06/10.27/20.91	0.93/1.05/0.87
		AdaptSum DAPT	25.78/9.84/20.69	0.65/0.62/0.65
		AdaptSum TAPT	25.65/9.13/20.45	0.52/0.00/0.41
Amazon	BART	AdaSum full (Bražinskas et al., 2022)	37.22/9.17/23.51	—
		Adasum 5 %+L1O	39.78/10.80/25.55	2.56/1.63/2.04
YELP		AdaSum full	37.40/10.27/23.76	—
		Adasum 5 %+L1O	38.82/11.75/25.14	1.42/1.48/1.38
CNN/Daily mail	GPT	T-LM (ETT) (Hoang et al., 2019)	36.82/16.04/34.03	—
		T-LM (DAT+ETT)	38.00/17.13/35.20	1.18/1.09/1.17
		T-SM (ETT)	37.81/16.82/34.87	—
		T-SM (DAT+ETT)	38.34/17.34/35.44	0.53/0.52/0.57
Airline Reviews (current work)	PEGASUS	Standard finetuning	32.6/7.2/24.1	—
		Two-stage finetuning	30.6/9.9/24.9	0.00/**2.70**/0.80

7.2.2. *BERT embeddings*

BERT input is represented as embeddings. Embeddings are vectors that hold the semantics of a word. BERT input embeddings are a combination of position, segment, and token embeddings as seen in Fig. 23. Position embeddings convey information about the position of a word within a sentence. The segment embeddings represent whether a

Table 6
Results of human evaluation.

Model	Fluency	Coherence	Non-redundancy	Informativeness	Sentiment
Standard finetuning	−0.01	−0.1	−0.05	−0.03	−0.02
Two-stage finetuning	+0.01	+0.1	+0.05	+0.03	+0.02

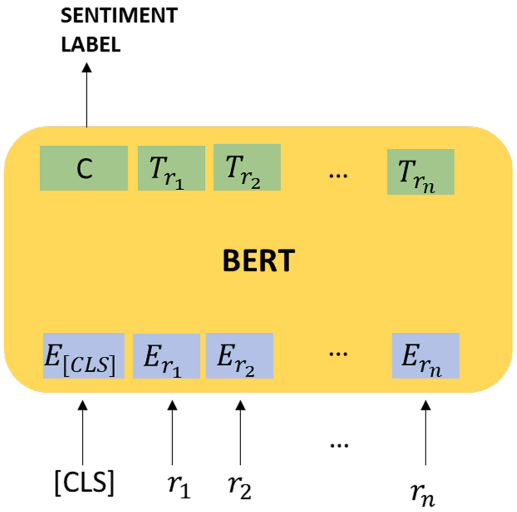


Fig. 22. The rating-based sentiment classification framework.

particular word belongs to which segment of the input, whether it belongs to the sentence ‘A’ or sentence ‘B’. The token embeddings are pretrained embeddings for the input words and are obtained through WordPiece tokenization. All the embeddings are added to create a meaningful input representation for the BERT model.

7.2.3. The BERT architecture

The BERT model is a stack of bidirectional Transformer encoders. Each encoder consists of multi-head attention and feedforward neural networks. There is a residual block around each of the attention and feedforward layers. The residual connections make the model easy to train and optimize. The residual blocks are followed by the normalization layer.

The attention mechanism in the BERT model is scaled dot product attention. The query and key vectors have a dimension d_k and value vectors have a dimension d_v. The attention function is computed as the SoftMax of the dot product of the query with all the keys scaled by 1 $/\sqrt{d_k}$ to give weights to the value vectors.

$$attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) * V \tag{9}$$

The multi-headed attention consists of multiple attention layers running in parallel and is given as,

$$multihead.attn(Q, K, V) = concat(head_1, \ldots., \ head_h)W^O \tag{10}$$

$$head_i = attention\left(QW_i^Q, \ KW_i^K, VW_i^V\right) \tag{11}$$

7.2.4. Dropout

We added a dropout layer to prevent the overfitting problem. We set the probability factor p to 0.3. p represents the probability of a neuron being dropped off or eliminated from the neural network layers.

7.2.5. The classification layer

The final classification layer is a fully connected feedforward network with a SoftMax activation function. The SoftMax function computes the relative probabilities of the input review for the three sentiment classes in a manner that the sum of all probabilities is 1. The node with the maximum probability represents the predicted class label. For $i = (1, \ldots.., k)$ and logit output, $z = (z_1, \ldots.., z_k) \in \mathbb{R}^k$, the SoftMax function is given as,

$$softmax(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \tag{12}$$

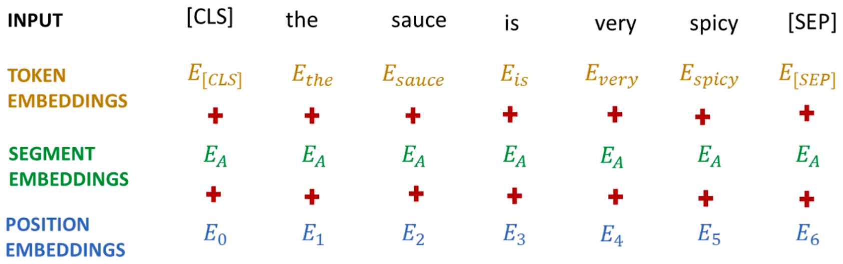
7.3. Experiment & evaluation

The sentiment classification model has a hidden layer size of 768. To deal with fixed-length sequences, we specified a maximum length of 400 tokens. We finetuned the BERT model for 10 epochs with a batch size of 16 and a learning rate of 5 $\times 10^{-5}$. We used the AdamW optimizer and computed the loss function as cross-entropy loss.

The evaluation of the model has been conducted once via a random train-valid-test split as well as using k-fold (5-fold) cross validation. With k-fold cross validation, the dataset is divided into k smaller sets. The model is trained using k-1 folds and validated on the remaining part. The process is repeated for k-splits of the data. During each split, the model is validated on a different fold of the data. We used the 5-fold stratified cross validation process that is a variation of the conventional k-fold cross validation suitable for classification models with imbalanced datasets. The basic process is the same as that of k-fold cross validation. The difference is in the sampling process where the samples for each fold are produced such that the ratio of original dataset class distribution is maintained across each fold of the data. The stratified 5-fold cross validation sampling on the sentiment classification dataset is illustrated in Fig. 24.

7.3.1. Evaluation

The model is evaluated using metrics such as accuracy, precision, recall, F1-score, and confusion matrix. The metrics are defined below.



INPUT	[CLS]	the	sauce	is	very	spicy	[SEP]
TOKEN EMBEDDINGS	$E_{[CLS]}$	E_{the}	E_{sauce}	E_{is}	E_{very}	E_{spicy}	$E_{[SEP]}$
	+	+	+	+	+	+	+
SEGMENT EMBEDDINGS	E_A	E_A	E_A	E_A	E_A	E_A	E_A
	+	+	+	+	+	+	+
POSITION EMBEDDINGS	E_0	E_1	E_2	E_3	E_4	E_5	E_6

Fig. 23. BERT Embeddings.

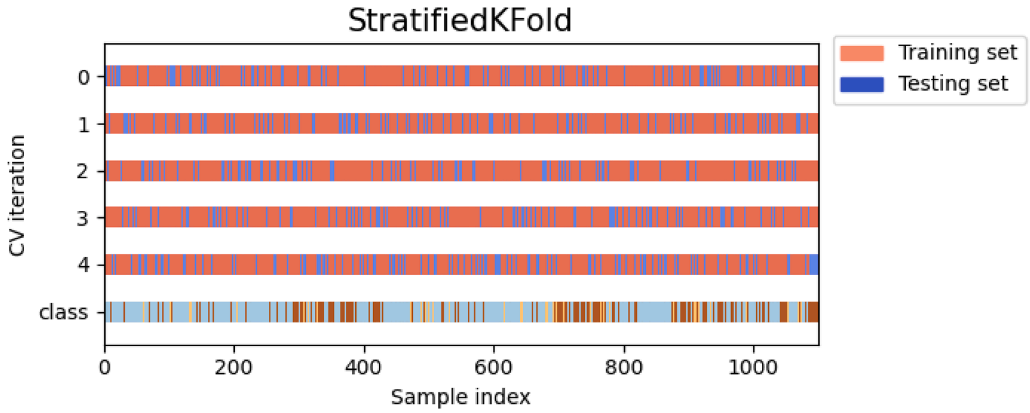


Fig. 24. Visualization of Stratified k-fold (5-fold) cross-validation on the sentiment classification dataset.

$$Accuracy = \frac{Number\ of\ correct\ class\ predictions}{Total\ number\ of\ predictions} \tag{13}$$

$$Precision = \frac{True\ positive\ class\ predictions}{Total\ positive\ class\ predictions} \tag{14}$$

$$Recall = \frac{True\ positive\ class\ predictions}{Total\ actual\ positive\ predictions} \tag{15}$$

$$F1\ score = 2\left(\frac{precision\ x\ recall}{precision + recall}\right) \tag{16}$$

A confusion matrix is an $m \times m$ matrix, where m is the number of target classes. It provides information on model classification performance by showcasing the number of true predictions and false predictions on the target classes.

Table 7 presents the results of rating-based sentiment classification model evaluation via random train/validation/test split with the ratio 80:10:10. Despite a highly imbalanced dataset with a limited number of samples, the model has achieved an overall 89 % accuracy with a very good performance on the classification of positive and negative classes. The precision, recall, and F1-score for negative, neutral, and positive classes are shown in Table 7. The scores are low for the ‘neutral’ class having the smallest number of samples.

Fig. 25 illustrates the confusion matrix obtained with the rating-based sentiment classifier. Out of 66 total predictions for the ‘negative’ class, 5 reviews were misclassified as ‘neutral’. For the ‘neutral’ class, there are 3 mispredictions out of a total of 8 predictions. The ‘positive’ class missed 4 shots out of a total of 36. Overall, there are 12 misclassifications out of a total of 110 samples in the test set.

The classification accuracy computed across each fold in the cross-validation process is shown in Table 8. The minimum model accuracy is 85.45 % which is computed across the first fold. The maximum model accuracy is 100 % which is achieved across 4th fold of the data. The average model accuracy using cross-validation is computed as 95.36 %. The confusion matrices generated across the 5-fold cross-validation process is illustrated in Figs. 26, 27, and 28.

Based on Fig. 26, the 1st fold confusion matrix shows 11 misclassifications for negative class, 12 missed shots for the neutral class, and 9 wrong predictions for the positive class. It can be observed from the 2nd fold confusion matrix that there are 7 misclassifications for negative class, 6 mis-predictions for the neutral class, and 2 invalid predictions for the positive class.

It can be noted from Fig. 27 that for the 3rd fold confusion matrix, there are no misclassifications for the negative class, still there are 2 missed shots for the neutral class, and 1 wrong prediction for the positive class. For the 4th fold confusion matrix, all predictions are correct. The confusion matrix for 5th fold in Fig. 28 shows only one wrong prediction for the neutral class.

Table 9 compares the performance of sentiment classification in terms of overall model accuracy with existing related works of (Kang et al., 2022), (Tan et al., 2022), and (Hasib et al., 2021). All three works are based on BERT or its variants to classify sentiments on airline reviews. Our model achieves 89 % (via random train-test split) and 95.36 % (via 5-fold cross-validation) accuracy as compared to 86 %, 85.89 %, and 83 % accuracy obtained by the existing works. The results of the rating-based sentiment prediction model for airline reviews reveal the effectiveness of using the ‘customer rating’ outcome for classifying airline customer sentiments at the review level as compared to other conventional methods for sentiment category labeling. The results also

Table 7
Results of the rating-based sentiment classification model.

	Precision	Recall	F1-score
Class 0 – Negative	0.95	0.92	0.94
Class 1 – Neutral	0.38	0.62	0.48
Class 2 – Positive	0.97	0.89	0.93
Accuracy			0.89
Macro Average	0.77	0.81	0.78
Weighted Average	0.92	0.89	0.90

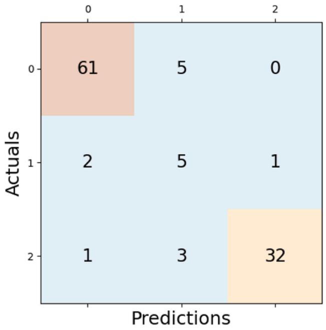


Fig. 25. The confusion matrix for classification performance for one-time evaluation.

Table 8
The classification accuracy across k-folds.

Fold	Accuracy (%)
1st fold	85.45 %
2nd fold	93.18 %
3rd fold	98.64 %
4th fold	100 %
5th fold	99.54 %
Average accuracy	95.36 %

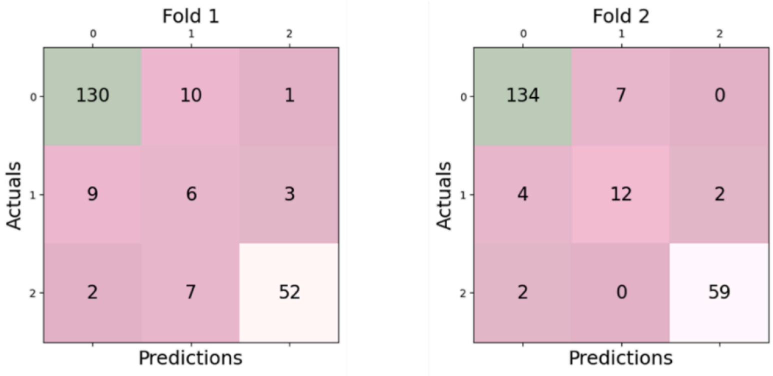


Fig. 26. The confusion matrices for fold 1 and fold 2.

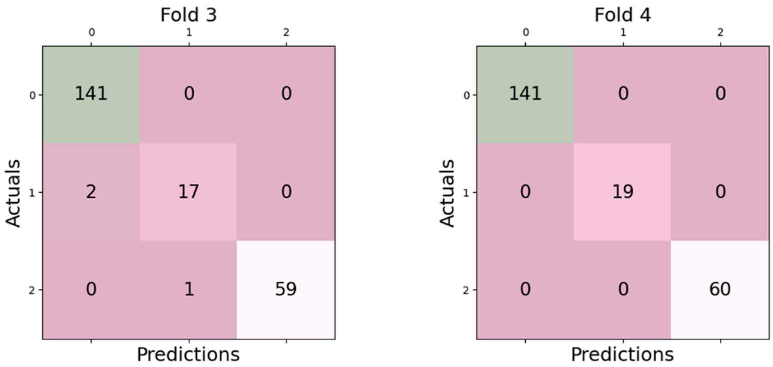


Fig. 27. The confusion matrices for fold 3 and fold 4.

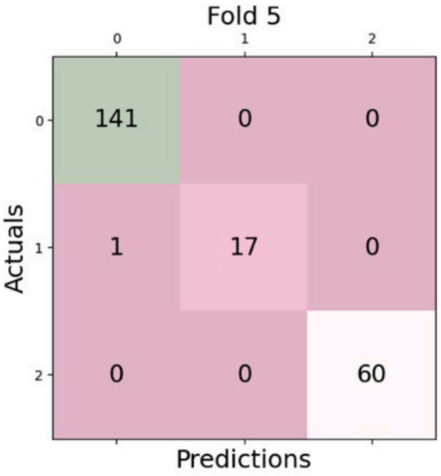


Fig. 28. The confusion matrix for Fold 5.

support the efficacy of the BERT language model with its ability to effectively capture the context and nuances in textual data for handling the sentiment classification of airline reviews.

As part of our transfer learning-based sentiment classification model assessment, we have conducted our rating-based sentiment classification experiment using other techniques including unsupervised approaches like VADER (Valence Aware Dictionary for Sentiment Reasoning) and BERT pipeline, standard machine learning approaches like SVM (Support Vector Machines) and Decision Trees, and deep learning models like RNN (Recurrent Neural Network) and LSTM (Long Short-Term Memory). The results are presented in Table 10.

Table 9
The comparison of rating-based sentiment classification with existing works.

Model	Accuracy	Dataset
Sentiment analysis on Malaysian airlines using BERT (Kang et al., 2022)	86 %	Malaysian airlines (14,000 samples)
Roberta-LSTM (without data augmentation) (Tan et al., 2022)	85.89 %	Twitter US airline sentiment dataset (10,000 tweets)
Sentiment classification on Bangladesh airline service using BERT (Hasib et al., 2021)	83 %	Bangladesh Airlines (1047 reviews)
Rating-based sentiment prediction with BERT (current work)	89 % (random train-test split) 95.36 % (5-folds cross-validation)	Airline reviews dataset (1100 samples)

Based on Table 10, the accuracy scores reflect that transfer learning using finetuned BERT shows the best classification performance on our dataset. These results can be interpreted from various perspectives including model architecture, structure and complexity, feature size, feature engineering, language representation, number of parameters,

Table 10
A Comparison of multiple approaches on rating-based sentiment analysis task.

Technique	Approach	Accuracy
Unsupervised	VADER	77.53 %
	BERT without finetuning	85.07 %
Standard Machine Learning (ML)	SVM	84.00 %
	Decision Tree	74.00 %
Deep Learning (DL)	RNN	74.55 %
	LSTM	80.91 %
Deep Transfer Learning	BERT (random train-test split)	89.00 %
	BERT (5-fold cross-validation)	95.36 %

parameter optimization techniques, amount of training data, algorithm training, and computational requirements. BERT with its large feature size, huge number of trainable parameters, and its ability to encode linguistic knowledge through contextualized language representations outperforms other models when finetuned even on a limited dataset. The standard machine learning models are less complicated as compared to deep learning or transfer learning models. With standard machine learning models, SVM shows better performance as compared to the decision trees approach as well as surpasses deep learning models. The problem with deep learning models like RNN and LSTM is that these methods generally do not work well in limited data situations and need huge amounts of training data and longer training times to increase the model accuracy. SVM, on the other hand, doesn't need huge amounts of training data and is trained faster as compared to other deep learning models. However, it still requires careful feature engineering to turn the textual data into features. From the computational cost dimension, the use of BERT is compute-intensive requiring GPU, large RAM, and storage resources while the use of SVM with moderate computational requirements, good feature engineering, and at the expense of somewhat reduced accuracy compared to BERT is cost-efficient. The transfer learning paradigm, however, offers ease in feature creation and hyperparameter selection while saving human time and effort and offering better scalability across various applications in real-world practices.

8. Discussion

This Sections discusses the contributions of this research to the literature as well as the practical implications. It also discusses some limitations of this study and recommends some future directions to follow.

8.1. Contributions to literature

This study enhances the literature on information management and processing of customer reviews across the civil aviation industry. It contributes towards enhancing the understanding of two important NLP tasks, abstractive summarization and sentiment classification using deep transfer learning. It has advanced the progress in these research areas by enhancing the capabilities of these models via improving the performance of these models for the airline customer opinion domain. Most of the existing research is focussed on solving one of these tasks for customer opinions. Still, there exists a few models in the literature that perform these two tasks jointly, but these models use methods that are still away from state of the art in NLP.

From an information management perspective, the use of deep transfer learning technology may be viewed as a junction between abductive, inductive, and deductive reasoning in the context of data science and artificial intelligence. On the inductive reasoning dimension, this research proposes models for abstractive summarization and rating-based sentiment classification for airline customer reviews that generalize well on the airline reviews domain.

Using deductive reasoning approach, the current study concludes that the two-stage finetuning approach assists the abstractive summarization model in adapting effectively to the airline review domain, thereby improving model performance on the target task of airline review summary generation. The proposed finetuning-based approach to domain adaptation differs from the previous research on domain adaptation for abstractive summarization in terms of computational resource utilization because most existing domain adaptation approaches for abstractive summarization rely on pretraining language models on target domain tasks or data, which is computationally expensive due to hardware requirements, large storage resources, and long training times. While for rating-based sentiment classification models, the research concludes that using sentiment supporting data such as 'customer rating' and 'recommendation value' for an overall sentiment analysis of multi-aspect reviews such as airline reviews improves the accuracy of sentiment classification tasks for airline reviews when compared to existing conventional techniques. Future research should investigate other signals, such as emoticons, along with the proposed customer attributes to further improve the accuracy of sentiment prediction.

For the abductive reasoning aspect of the utilization of the deep learning technology for information management of airline customer reviews, the research infers from the existing literature on PLMs that when pre-trained language models are finetuned on downstream tasks and domains, there usually occurs a problem known as the 'domain shift'. The domain shift issue occurs when data distributions for the source and target domains do not match. In this case, the source domain comprises of the domains including news, science, short stories, instructions, emails, patents, etc., particularly the domains on which PEGASUS language model has been pretrained. The proposed abstractive summarization dataset belongs to customer reviews domain (target domain) particularly for airline customers. It has been hypothesized that domain shift issue likely degrades the performance of the abstractive summarization model for airline customer review domain. The present research addresses the domain shift issue for airline reviews domain by proposing two-stage finetuning as a domain adaptation strategy for abstractive summarization of airline reviews, resulting in better performance of the abstractive summarization model. For rating-based sentiment classification task, it was figured out from the proposed sentiment classification dataset that airline reviews are multi-aspect reviews and consists of intricate narratives. It was also presumed that sentiment supporting data like customer rating and recommendation value would better predict an overall sentiment of multi-aspect reviews (airline customer reviews) as compared to the conventional sentiment labelling techniques. The presumption was supported by investigating the correlation between rating and recommendation value signals that very found to be highly correlated. The positive correlation between the mentioned customer attributes was also inferred from the existing literature. The task was modelled using BERT language model and the performance turned out to be better as compared to existing models despite the dataset being highly imbalanced in terms of positive, negative, and neutral sentiment classes.

Overall, the present study adds up to the interdisciplinary research where NLP (abstractive summarization & sentiment classification), deep transfer learning, information management, e-commerce, and civil aviation sector meet at a junction.

8.2. Practical implications

The research has a direct practical application across the aviation industry. It can be applied to real-time airline review processing, or the models can be integrated into airline customer service platforms.

From the practical perspective, opinion summarization and sentiment classification play an important role in improving the management and processing of customer reviews on air travel review websites, making it convenient for industry practitioners to respond, address issues, and continuously enhance their services, leading to higher levels of customer satisfaction and loyalty.

The research also contributes publicly accessible datasets that can benefit other researchers in advancing the research, development, and innovation. Public datasets can also assist industry personnel for data-driven decision making or to inform strategic planning or marketing research.

8.3. Limitations and directions for future work

For abstractive summarization and rating-based sentiment classification, this research particularly focuses on the airline reviews domain, which has obvious benefits for the aviation sector but limits the applicability of this research on domains and sectors other than the aviation sector. As a direction for future research, it would be useful to conduct this study on other domains and compare the results.

For the rating-based sentiment classification task, the dataset collected and used in this research suffers from the class imbalance issue. The class imbalance poses a risk to the trained model to be biased toward major classes. This situation likely leads to low performance of the classification model on minor classes. For future direction, it would be beneficial to explore data augmentation techniques to deal with class imbalance issues.

To verify the effectiveness of the two-stage finetuning approach for domain adaptation, it is suggested to compare it with other current alternative methodologies. The future research should also focus on how the proposed models can be used in real-world situations.

9. Conclusion

In this research, we proposed models for abstractive summarization and rating-based sentiment prediction of airline reviews using pre-trained PEGASUS and BERT language models. For abstractive summarization, we used a two-stage finetuning approach to help our model better adapt to the target abstractive summarization task on the airline reviews dataset. We evaluated our model with automatic and manual evaluation methods. We demonstrated that the two-stage finetuning approach works better as compared to the standard finetuning approach for varying domain data. The rating-based sentiment classification model is based on pretrained BERT architecture. Despite its simplicity in architecture, the model shows an overall accuracy of 89 % (random train-test split) and 95 % (via cross-validation). We discussed the contributions of our work to the literature, practical scenarios, and industry. We also discussed limitations of our research and presented some directions for future work. Along with multiple future work recommendations from this study, the abstractive summarization and sentiment classification frameworks can be integrated into a unified framework using a multitask learning (Zhang et al., 2022) approach or prompt tuning a single Large Language Model (LLM) on these tasks.

CRediT authorship contribution statement

Ayesha Ayub Syed: Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Ford Lumban Gaol:** Resources, Investigation, Conceptualization. **Alfred Boediman:** Writing – review & editing, Validation, Conceptualization. **Widodo Budiharto:** Supervision, Methodology, Conceptualization.

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.

References

Alduailej, A., & Alothaim, A. (2022). AraXLNet: Pre-trained language model for sentiment analysis of Arabic. *Journal of Big Data, 9*(1). https://doi.org/10.1186/s40537-022-00625-z

Al-Natour, S., & Turetken, O. (2020). A comparative assessment of sentiment analysis and star ratings for consumer reviews. *International Journal of Information Management, 54*. https://doi.org/10.1016/j.ijinfomgt.2020.102132

Baniya, R., Dogru-Dastan, H., & Thapa, B. (2021). Visitors' experience at Angkor Wat, Cambodia: Evidence from sentiment and topic analysis. *Journal of Heritage Tourism, 16*(6), 632–645. https://doi.org/10.1080/1743873X.2020.1833892

Bigne, E., Ruiz, C., Cuenca, A., Perez, C., & Garcia, A. (2021). What drives the helpfulness of online reviews? A deep learning study of sentiment analysis, pictorial content and reviewer expertise for mature destinations. *Journal of Destination Marketing and Management, 20*, Article 100570. https://doi.org/10.1016/j.jdmm.2021.100570

Bigne, E., Ruiz, C., Perez-Cabañero, C., & Cuenca, A. (2023). Are customer star ratings and sentiments aligned? A deep learning study of the customer service experience in tourism destinations. In *Service business, 17*. Springer Berlin Heidelberg. https://doi.org/10.1007/s11628-023-00524-0

Bordoloi, M., & Biswas, S. K. (2023). Sentiment analysis: A survey on design framework, applications and future scopes. *Artificial intelligence review*. Springer Netherlands. https://doi.org/10.1007/s10462-023-10442-2

Bražinskas, A., Lapata, M., & Titov, I. (2020). Few-shot learning for opinion summarization. In *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (pp. 4119–4135). https://doi.org/10.18653/v1/2020.emnlp-main.337

Brazinskas, A., Nallapati, R., Bansal, M., & Dreyer, M. (2022). *Efficient few-shot fine-tuning for opinion summarization*. 1509–1523. https://doi.org/10.18653/v1/2022.findings-naacl.113.

Bražinskas, A., Nallapati, R., Bansal, M., & Dreyer, M. (2022). Efficient few-shot fine-tuning for opinion summarization. In *Findings of the Association for Computational Linguistics: NAACL 2022 - Findings* (pp. 1509–1523). https://doi.org/10.18653/v1/2022.findings-naacl.113

Cambria, E., Poria, S., Gelbukh, A., Nacional, I. P., & Thelwall, M. (2017). Affective computing and sentiment analysis sentiment analysis is a big suitcase. *Ieee Intelligent Systems*.

Chamekh, A., Mahfoudh, M., & Forestier, G. (2022). Sentiment analysis based on deep learning in E-commerce. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 13369 LNAI* (pp. 498–507). https://doi.org/10.1007/978-3-031-10986-7_40

Chatterjee, S. (2019). Explaining customer ratings and recommendations by combining qualitative and quantitative user generated contents. *Decision Support Systems, 119* (November 2018), 14–22. https://doi.org/10.1016/j.dss.2019.02.008

Chen, S., Hou, Y., Cui, Y., Che, W., Liu, T., & Yu, X. (2020). Recall and learn: Fine-tuning deep pretrained language models with less forgetting. In *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (pp. 7870–7881). https://doi.org/10.18653/v1/2020.emnlp-main.634

Chintalapudi, N., Battineni, G., Canio, M. D., Sagaro, G. G., & Amenta, F. (2021). Text mining with sentiment analysis on seafarers' medical documents. *International Journal of Information Management Data Insights, 1*(1), Article 100005. https://doi.org/10.1016/j.jjimei.2020.100005

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *, 1. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* (pp. 4171–4186).

Guo, Y., Shi, H., Kumar, A., Grauman, K., Rosing, T., & Feris, R. (2019). Spottune: Transfer learning through adaptive fine-tuning. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019-June* (pp. 4800–4809). https://doi.org/10.1109/CVPR.2019.00494

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Zhang, L., Han, W., Huang, M., Jin, Q., Lan, Y., Liu, Y., Liu, Z., Lu, Z., Qiu, X., Song, R., Tang, J., Wen, J.-R., … Zhu, J. (2021). Pre-trained models: Past, present and future. *AI Open*. https://doi.org/10.1016/j.aiopen.2021.08.002

Hasib, K. M., Towhid, N. A., & Alam, M. G. R. (2021). Online review based sentiment classification on Bangladesh airline service using supervised learning. In *2021 5th International Conference on Electrical Engineering and Information and Communication Technology, ICEEICT 2021*. https://doi.org/10.1109/ICEEICT53905.2021.9667818

Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., & Yeganegi, M. R. (2020). Text mining in big data analytics. *Big Data and Cognitive Computing, 4*(1), 1–34. https://doi.org/10.3390/bdcc4010001

Hoang, A., Bosselut, A., Celikyilmaz, A., & Choi, Y. (2019). Efficient adaptation of pretrained transformers for abstractive summarization. http://arxiv.org/abs/1906.00138.

Iddrisu, A. M., Mensah, S., Boafo, F., Yeluripati, G. R., & Kudjo, P. (2023). A sentiment analysis framework to classify instances of sarcastic sentiments within the aviation sector. *International Journal of Information Management Data Insights, 3*(2), Article 100180. https://doi.org/10.1016/j.jjimei.2023.100180

Jain, S., Tang, G., & Chi, L.S. (2021). *MRCBert: A machine reading comprehensionapproach for unsupervised summarization*. 1–15. http://arxiv.org/abs/2105.00239.

Kang, H. W., Chye, K. K., Ong, Z. Y., & Tan, C. W. (2022). Sentiment analysis on Malaysian airlines with BERT. *The Journal of The Institution of Engineers, Malaysia, 82* (3). https://doi.org/10.54552/v82i3.98

Kant, N., Puri, R., Yakovenko, N., & Catanzaro, B. (2018). *Practical text classification with large pre-trained language models*. http://arxiv.org/abs/1812.01207.

Kar, A. K., Angelopoulos, S., & Rao, H. R. (2023). Guest Editorial: Big data-driven theory building: Philosophies, guiding principles, and common traps. *International Journal of Information Management, 71*(April). https://doi.org/10.1016/j.ijinfomgt.2023.102661

Katwe, P. K., Khamparia, A., Gupta, D., & Dutta, A. K. (2023). Methodical systematic review of abstractive summarization and natural language processing models for biomedical health informatics: Approaches, metrics and challenges. *ACM Transactions on Asian and Low-Resource Language Information Processing*. https://doi.org/10.1145/3600230

Kumar, S., Kar, A. K., & Ilavarasan, P. V. (2021). Applications of text mining in services management: A systematic literature review. *International Journal of Information Management Data Insights, 1*(1), Article 100008. https://doi.org/10.1016/j.jjimei.2021.100008

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7871–7880). https://doi.org/10.18653/v1/2020.acl-main.703

Ligthart, A., Catal, C., & Tekinerdogan, B. (2021). Systematic reviews in sentiment analysis: A tertiary study. In *Artificial intelligence review, 54*. Springer Netherlands. https://doi.org/10.1007/s10462-021-09973-3

Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. In *, 1. Proceedings of the Workshop on Text Summarization Branches out (WAS 2004)* (pp. 25–26). papers2://publication/uuid/5DDA0BB8-E59F-44C1-88E6-2AD316DAEF85.

Lu, L., Mitra, A., Wang, Y.-Y., Wang, Y., & Xu, P. (2022). Use of electronic word of mouth as quality metrics: A comparison of airline reviews on Twitter and Skytrax. In *Proceedings of the 55th Hawaii International Conference on System Sciences* (pp. 1349–1357). https://doi.org/10.24251/hicss.2022.165

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations ofwords and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 1–9.

Min, B., Ross, H., Sulem, E., Ben Veyseh, A.P., Nguyen, T.H., Sainz, O., Agirre, E., Heinz, I., & Roth, D. (2021). *Recent advances in natural language processing via large pre-trained language models: A survey*. http://arxiv.org/abs/2111.01243.

Munikar, M., Shakya, S., & Shrestha, A. (2019). Fine-grained sentiment classification using BERT. In *International Conference on Artificial Intelligence for Transforming Business and Society, AITB 2019* (pp. 2–5). https://doi.org/10.1109/AITB48515.2019.8947435

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). BLEU: A method for automatic evaluation of machine translation. *Acl, July*, 311–318. https://doi.org/10.3115/1073083.1073135

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). https://doi.org/10.1080/02688697.2017.1354122

Qiu, X. P., Sun, T. X., Xu, Y. G., Shao, Y. F., Dai, N., & Huang, X. J. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences, 63*(10), 1872–1897. https://doi.org/10.1007/s11431-020-1647-3

Ramponi, A., & Plank, B. (2021). *Neural unsupervised domain adaptation in NLP—A survey*. 6838–6855. https://doi.org/10.18653/v1/2020.coling-main.603.

Sciforce. (2019). *Towards automatic summarization. Part 2. Abstractive methods.* https://medium.com/sciforce/towards-automatic-summarization-part-2-abstractive-methods-c424386a65ea#id_token=eyJhbGciOiJSUzI1NiIsImtpZCI6ImEzYmRiZmRlZGUzYmFiYjI2NTFhZmNhMjY3OGRkZThjMGIzNWRmNzYiLCJ0eXAiOiJKV1QifQ.eyJpc3MiOiJodHRwczovL2FjY291bnRzLmdvb2dsZS5.

Setiyawan, A., Wijayanto, A. W., & Youshi, H. (2021). Extracting consumer opinion on Indonesian E-commerce: A rating evaluation and Lexicon-based sentiment analysis. In *Proceedings of 2021 International Conference on Data Science and Official Statistics (ICDSOS)* (pp. 1–11). https://proceedings.stis.ac.id/icdsos/article/view/22.

Shobana, J., & Murali, M. (2021). Abstractive review summarization based on improved attention mechanism with pointer generator network model. *Webology, 22*(1), 77–91. https://doi.org/10.14704/WEB/V18I1/WEB18028

Syed, A. A., Gaol, F. L., Boediman, A., Matsuo, T., & Budiharto, W. (2022). A survey of abstractive text summarization utilising pretrained language models. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 13757 LNAI* (pp. 532–544). https://doi.org/10.1007/978-3-031-21743-2_42

Syed, A. A., Gaol, F. L., Boediman, A., Matsuo, T., & Budiharto, W. (2023). A data package for abstractive opinion summarization, title generation, and rating-based sentiment prediction for airline reviews. *Data in Brief, 50*, Article 109535. https://doi.org/10.1016/j.dib.2023.109535

Syed, A. A., Gaol, F. L., & Matsuo, T. (2021). A survey of the state-of-the-art models in neural abstractive text summarization. *IEEE access : practical innovations, open solutions, 9*, 13248–13265. https://doi.org/10.1109/ACCESS.2021.3052783

Tan, K. L., Lee, C. P., Anbananthen, K. S. M., & Lim, K. M. (2022). RoBERTa-LSTM: A hybrid model for sentiment analysis with transformer and recurrent neural network. *IEEE access : practical innovations, open solutions, 10*, 21517–21525. https://doi.org/10.1109/ACCESS.2022.3152828

Tan, K. L., Lee, C. P., & Lim, K. M. (2023). A survey of sentiment analysis: Approaches, datasets, and future research. *Applied Sciences (Switzerland), 13*(7). https://doi.org/10.3390/app13074550

Ullah, M. A., Marium, S. M., Begum, S. A., & Dipa, N. S. (2020). An algorithm and method for sentiment analysis using the text and emoticon. *ICT Express, 6*(4), 357–360. https://doi.org/10.1016/j.icte.2020.07.003

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 2017-Decem*(Nips), 5999–6009.

Weiss, K., Khoshgoftaar, T. M., & Wang, D. D. (2016). A survey of transfer learning. In *Journal of big data, 3*. Springer International Publishing. https://doi.org/10.1186/s40537-016-0043-6

Xu, X., Wang, Y., Zhu, Q., & Zhuang, Y. (2024). Time matters: Investigating the asymmetric reflection of online reviews on customer satisfaction and recommendation across temporal lenses. *International Journal of Information Management, 75*. https://doi.org/10.1016/j.ijinfomgt.2023.102733

Yu, T., Liu, Z., & Fung, P. (2021). *AdaptSum: Towards low-resource domain adaptation for abstractive summarization*. 5892–5904. https://doi.org/10.18653/v1/2021.naacl-main.471.

Zhang, J., Zhao, Y., Saleh, M., & Liu, P. J. (2020a). PEGASUS: Pre-Training with extracted gap-sentences for abstractive summarization. In *37th International Conference on Machine Learning, ICML 2020, PartF16814* (pp. 11265–11276).

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020b). Bertscore: Evaluating text generation with BERT. In *ICLR 2020* (pp. 1–43).

Zhang, Z., Yu, W., Yu, M., Guo, Z., & Jiang, M. (2022). *A survey of multi-task learning in natural language processing: Regarding task relatedness and training methods*. http://arxiv.org/abs/2204.03508.

Zhu, L., Lin, Y., & Cheng, M. (2020). Sentiment and guest satisfaction with peer-to-peer accommodation: When are online ratings more trustworthy? *International Journal of Hospitality Management, 86*. https://doi.org/10.1016/j.ijhm.2019.102369

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE, 109*(1), 43–76. https://doi.org/10.1109/JPROC.2020.3004555