National College of Ireland

Aspect-Specific Sentiment Classification Using a RoBERTa BiLSTM Hybrid Model with Hierarchical Attention

MSc Research Project
MSc in Artificial Intelligence

Manupavan Mulkuri
Student ID: 23301112

School of Computing
National College of Ireland

Supervisor: Abdul Shahid

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Mulkuri Manupavan
Student ID:	23301112
Programme:	MSc in Artificial Intelligence
Year:	2025
Module:	MSc Research Project
Supervisor:	Abdul Shahid
Submission Due Date:	11/08/2025
Project Title:	Aspect-Specific Sentiment Classification Using a RoBERTa BiLSTM Hybrid Model with Hierarchical Attention
Word Count:	8104
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Mulkuri Manupavan
Date:	14th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Aspect-Specific Sentiment Classification Using a RoBERTa BiLSTM Hybrid Model with Hierarchical Attention

Mulkuri Manupavan
23301112

Abstract

Since sentiment analysis has become increasingly relevant to interpreting the opinion of users, conventional approaches that give only a positive or negative score tend to be inadequate in modeling the fine details of real world reviews. Individuals tend to have varied sentiment about a particular attribute of the product or service-as in the case of people saying good about the plot of a movie and bad about its acting. Current transformer models such as RoBERTa are able to capture lots of contextual information and in general can capture the contextual information well but they lack transparency and ignore the sequential nature of language. This thesis proposes a hybrid machine learning model to overcome these limitations based on RoBERTa embeddings and the use of BiLSTM layers and a hierarchical attention mechanism. In addition to increasing accuracy, the model is aimed at increasing interpretability by making it clear how each word is helping make aspect-specific sentiment decisions. We consider two datasets SST-2 and IMDb, with frozen and fine-tuned settings of the model evaluation. The fine-tuned model of the performance that gave the accuracy of 93.43% and 93.81% respectively outperforms the baseline of RoBERTa-BiLSTM. Also, SHAP analysis helps to visualise the aspect-level prediction contributions giving evidence of the transparency of the model. On the whole, this paper indicates that there are benefits to incorporating hierarchical attention with the contextual and sequential models in making sentiment analysis more precise and explainable, particularly with practical implications both on the research side and in practical implementations.

1 Introduction

The field of sentiment analysis has advanced at an accelerated pace beyond the naive binary classification problem Pang et al. (2008) to the more high-resolution aspect-level opinion mining task. Unlike conventional methods that either tag an entire review or social-media comment good or bad, the survey type of reader feedback has mixed sentiments; one comment can either be good on the "plot" but bad on the "acting," or positive on the "cinematography" but negative on the "screenplay." Trying to separate these multi-dimensional, co-lateral appraisals is not something that single-label models are well suited to do.

Further, irrespective of the significant gains in accuracy brought by large Transformer-based models like BERT Devlin et al. (2019) and RoBERTa Liu et al. (2019), their

black-box architectures cannot point to which words determine which aspect judgments. When decision-markers are dealing with high-stakes situations such as customer experience monitoring, market research, or product evaluation they need to be assured of the reasons behind the model predictions to freely use and trust them. Such inability to interpret is therefore a significant hindrance to the real world applications of the current sentiment classifiers.

The gap is highlighted by a survey of recent literature. Hybrid systems that couple Transformer embeddings and recurrent networks (e.g., Xu et al. (2022); Negi et al. (2024)) produce minor performance gains but they are not (yet) able to achieve word-level explainability. Interactive-attention models (Taboada et al. (2011); Ma et al. (2017); Sun et al. (2019)) and memory-augmented models (Wang et al. (2016))are able to extract richer contextual signals but they do not explicitly show how individual tokens affect the perception of aspect-specific sentiment. As a result, it is still an open area of demand to have the representational advantage of pretrained encoders coupled with the mechanisms of aspect-specific and clear explanation.

This motivates the Research Question: How can we improve the aspect-specific sentiment classification by integrating the hierarchical attention with RoBERTa - BiLSTM when compared to transformer - only models?

In order to answer this question we develop and build a new hybrid model and it consists of three fundamental elements:

1. RoBERTa Word Embeddings: We utilize pre-trained RoBERTa Liu et al. (2019) encoder in order to produce deep contextualized representations for tokens.

2. Bidirectional LSTM: A BiLSTM Hochreiter and Schmidhuber (1997) layer is able to learn sequential dependency, local syntactic and phrase-level pattern enhancement.

3. Hierarchical Attention: In two steps, each predefined aspect is accorded word-level importance and the signals of the aspects are collected to form aspect scores that are used in making the ultimate sentiment judgement.

By way of this architecture, we aspire to do more than just enhance classification accuracy through an integration of global context and residual cumulative attention through a sequence pattern matching, but in addition we are able to provide transparency to the model decisions by making them visualizable in terms of the aspect level attention scores as well as presenting the aspect level attention at the word level.

The Key contributions of this paper include:

1. The proposed novel hybrid architecture, which is a mixture of RoBERTa embeddings, BiLSTM-based sequence model and hierarchical Yang et al. (2016) allowing us to achieve interpretable and accurate sentiment-classification based on aspects.

2. A two-experiment structure that delves into frozen and fine-tuned variants of RoBERTa on two benchmarked databases IMDb and SST-2, to comprehensively analyse how transfer learning affects sentiment classification tasks.

3. Development of SHAP (SHapley Additive exPlanations) Lundberg and Lee (2017)to provide a clear explanation to the subsequent forecasts of a model by determining the contribution of each of these aspects in the total opinion.

4. Full-fledged performance assessment with a combination of such metrics as Accuracy Grandini et al. (2020), Precision Dalianis (2018), Recall Vakili et al. (2020), F1-scoreGrandini et al. (2020); Hand et al. (2021), Mean Squared Error (MSE) Grandini et al. (2020) and statistical significance tests to make sure the performance got improved.

5. By comparing with other existing models, especially the RoBERTa-BiLSTM model

by Rahman et al. (2025), offering insights into how efficiently the proposed architectural specifications of the model applied can improve the effectiveness thereof.

The results of experimental studies on both datasets indicated that the fine-tuning of RoBERTa Liu et al. (2019) resulted in the overall improvement in performance of all measures. Specifically, with IMDb dataset, values of accuracy were 93.43 and 93.27 in a fine-tuned and frozen model, respectively. The same gains were reported on SST-2, where fine-tuning netted 93.81 accuracy, compared to 93.12. Further, our model scored better than the earlier used baseline RoBERTa-BiLSTM Rahman et al. (2025) (F1: 0.9235, Accuracy: 92.36%) further suggesting that the superiority of the traditional hierarchical structure based attention system which has both interpretability and demonstrated performance improvements.

The rest of this thesis will consist of the following structure. In Chapter 2, we review the history of sentiment-analysis research, in which we position our work both in the aspect-based and interpretability-centered literature. In chapter 3, we introduce our methodology with the description of the experimental plan, data, and statistical procedures of comparing frozen and fine-tuned encoder variants. Following on that, Chapter 4 documents the design specification, and breaks the system down in its core modules and defines its interfaces and requirements. Chapter 5 next explains the actual concrete objects of implementation and the environments and tools used to generate them: data splits and model checkpoints and explainability outputs. A well-documented critical review is presented in Chapter 6 with quantitative measures, statistical significance tests, and error analysis to measure the performance of the models. Finally, Chapter 7 ends up with the conclusions of our findings, consideration of the implications of the study to practice and research as well as the future directions.

2 Related Work

Sentiment analysis has evolved rapidly due to refined decision-making across applications, including the customer-feedback analysis and market research. In their seminal survey, Pang et al. (2008) said that the main weaknesses of opinion mining were that aspect-specific nuances are not captured with coarse-grained approaches. Although their work is done prior to deep-learning methods, it forms the root to later developments that go beyond document-level polarity to the fine-grained aspect-level sentiment; these treatments take into consideration the context around the word. Motivated by this, our study focuses on the question How can we improve aspect-specific sentiment classification by integrating hierarchical attention with a RoBERTa–BiLSTM model when compared to transformer-only models? In order to resolve the above issue, we critically review modern deep-learning paradigms, specifically using hybrid networks to combine transformer-based encoders with recurrent processing units and self-attention techniques, and specify the research gap which the proposed model aims to fill.

2.1 Need for Aspect Specific Models and some Fundamental Perspectives

The survey by Pang et al. (2008) set the theoretical and empirical foundation of sentiment analysis, and defined types of tasks at document-level, sentence-level and (conclusively) the aspect-level. They demonstrated that simple lexicon based properties Taboada et al.

(2011) or SVMs with bag of words training data can produce this high overall polarity accuracy. Taking every review or sentence as an identical unit they overlooked the fact that opinions are often aimed on a different and separate factor (e.g., battery life in comparison to the screen quality), which has their corresponding sentiment. Such limitation does not only reduce the level of insight in part but also fails to provide any knowledge concerning the terms which have driven the scores of a given aspect, a problem that has emerged critical as organizations demand feature-level data towards the development of products along with improvement of services.

2.2 Hybrid Models and Attention Mechanisms

Researchers have applied attention mechanisms in recurrent neural architecture in order to capture sequential state-to-state ties and also locate sentiment-carrying lexical units.

Xu et al. (2022) proposed an attention based BiLSTM augmented with transfer learning; here, contextual word embeddings are fine-tuned and after that, the token vectors generated by BiLSTM are scaled by a single attention layer to generate the weights that are proportionate to sentiment relevance. Even though the architecture enhances aspect-level accuracy, the flat attention mechanism fails to specify which aspect each of the attended tokens belongs to or to separate what local (word-level) versus global (aspect-level) influences the overall sentiment.

Negi et al. (2024) introduced a variant that combines several types of embeddings: contextual, syntactic, and sentiment-lexicon vectors, use a multi-head attention mechanism in a BiLSTM architecture. This multirepresentations approach adds strength to domains (e.g., social media and e-commerce), yet at the same time hides the lineage of the impact of each embedding to the final classification.

Li et al. (2019) introduced the so-called Multi-Granularity Attention Network (MGAN) to simultaneously apply attention on the phrase (local) and sentence (global) levels. By capturing hierarchical context, MGAN outperforms flat models, but the token -level interpretability is diluted due to the aggregation of tokens into higher -level phrase embeddings.

2.3 Transformer-Based and Memory Network Approaches

Aspect-level sentiment analysis has achieved impressive advances with transformers, and memory networks.

Rahman et al. (2025) combined deep, bi-directional embeddings of RoBERTa with downstream BiLSTM layers and with a context-level attention module. This hybrid model produced state-of-the-art performance by combining transformer context with sequential modelling but with rough attention. This pays attention to whole sentences rather than keeping track of contributions to the individual tokens or features.

Tang et al. (2016) developed an end-to-end Memory Network which encodes contextual memories on multi-hop Memory Network in order to capture long-range dependencies in longer reviews. Despite the effectiveness of such a model when it comes to preserving the context, the fact that the chains of memory lookups are multi-hop also makes it tricky to trace a final prediction to specific tokens.

2.4 Interactive Attention and Task Reformulation Approaches

Various studies on sentiment analysis directly deal with interpretability.

Ma et al. (2017) proposed an interactive attention network that divides attention into aspect and context paths and then combines them in order to predict the sentiment polarity. Although this architecture partly begins to isolate aspect-specific word contributions, it retains a unitary-tier attention scheme that prohibits making explicit differentiation between word-level and aspect-level effects.

In order to leverage BERT pretraining, Sun et al. (2019) reframed the task of aspect-level sentiment analysis as an auxiliary sentence classification task. They took every aspect and translated it into a natural language question, so something along the lines of What is your opinion of the battery life? This change up-leveled the performance as it better synchronized the task with the pre-training goals of BERT. It however, involved extra step of creating these intermediate questions, this introduces computational overhead. More importantly, this reformulation broke the direct connection with the original tokens of input and the final sentiment predictions, so it was less possible to create exact token-to-sentiment associations..

Yang et al. (2016) present Hierarchical Attention Network (HAN), a model that builds on word- and sentence-level attention to enhance document classification. HAN shows that multi-level attention enhances interpretability, but that it does not tailor scores to aspects, thus cannot provide token-level interpretable scores on any single aspect.

2.5 Synthesis and Identification of the Research Niche

Within the past decades, sentiment analysis is currently moving towards more and more detailed, aspect-based approaches; a move that started out to be quite rough with document-level polarity-level analysis. Following the survey done by Pang et al. (2008), modern work has been expanding context modeling and classification correctness, using hybrid recurrent soar attention models (e.g., Xu et al. (2022) Negi et al. (2024); Li et al. (2019)) and transformer enhanced models (e.g., Rahman et al. (2025)). Memory networks.

Tang et al. (2016) and interactive attention schemes (Ma et al. (2017); Sun et al. (2019)) have also been added to these contributions and may render further improvement in performance but with minimal interpretability. Today, it does not exist in the approaches that provide a resilient, end-to-end process of transparently associating the impact of each word with its respective aspect.

This paper suggests a hybrid RoBERTa-BiLSTM Rahman et al. (2025)enhanced with the two-tier hierarchical attention mechanism. This comprehensive model uses deep contextual embedding and sequential encodings to both achieve high levels of accuracy in classification and provide explicit contributions scores of words to aspects, thus achieving both high-state-of-the-art performance and complete transparency in areas such as e-Commerce, movie reviews, and hospitality feedback.

Table 1: Comparison of Key Aspect-Level Sentiment Analysis Methods

Methodology	Strengths	Limitations
Pang et al. (2008)		
Lexicon & SVM baseline	Simple and interpretable; establishes strong overall polarity detection.	No aspect-level granularity; cannot attribute sentiment to specific words.
Xu et al. (2022)		
BiLSTM + single-layer attention	Captures sequential word dependencies; highlights salient tokens.	Flat attention cannot distinguish which tokens belong to which aspect.
Rahman et al. (2025)		
RoBERTa → BiLSTM → context-level attention	Deep contextual embeddings plus BiLSTM yield richer feature representations.	Sentence-level attention loses precise word-to-aspect mapping.
Tang et al. (2016)		
End-to-End Memory Network	Multi-hop memory captures long-range dependencies.	Multiple hops obscure which words drive the final prediction.
Ma et al. (2017)		
Interactive Attention Network (IAN)	Jointly attends to aspect terms and context for greater transparency.	Single-tier attention still mixes contributions across words.
Sun et al. (2019)		
Task Reformulation + BERT	Reformulates as a sentence-pair task, leveraging BERT's pre-training.	Breaks direct link between input tokens and sentiment outputs.
Wang et al. (2016)		
Recursive Neural Conditional Random Fields (CRF)	Integrates recursive neural networks with CRFs for aspect-based sentiment analysis.	Computationally intensive; requires careful tuning.
Wang et al. (2017)		
Coupled Multi-Layer Attention (CMLA)	Jointly attends to both aspect and opinion terms for better alignment.	Computationally heavy and requires careful tuning.
Zhang et al. (2019)		
Aspect-specific Graph Convolutional Networks	Utilizes graph convolutional networks to capture syntactic relations.	Performance dependent on parse quality; adds preprocessing overhead.
Xu et al. (2020))		
Aspect-based Sentiment Analysis with BERT	Leverages BERT pre-training for better context modeling at the aspect level.	Requires high computational resources and large data sets..
Our Proposed Model		
RoBERTa → BiLSTM → Hierarchical Attention	Unifies deep contextual embeddings with sequential modeling; hierarchical attention for explicit word-to-aspect attribution.	—

3 Methodology

In this section, we will explain you how we conceptualized, developed, and validated our aspect-specific sentiment analysis model. All design decisions are strongly rooted in the ground works which we discussed in the previous section and we provide more details on data handling, compute equipment and resource, procedure and statistical protection. We want to leave no doubt how raw movie-review texts are transformed into final, explainable sentiment predictions.

3.1 Design and Hypotheses

Our study follows a supervised quantitative pattern, where the inclusion of a pretrained Transformer encoder is combined with recurrent modeling and aspect-level attention to generate overall sentiment of a review. We also introduce a predefined ten semantics facets (e.g., plot, acting, cinematography), which are equally weighted and then summed up to perform a binary classification task based on the Aspect-Based Sentiment Analysis (ABSA) framework developed on Li et al. (2020) and the multi-aspect attention mechanism developed on Han et al. (2019)

We Test two scenarios here:

1. The aspect specific sentiment model with hierarchical Yang et al. (2016) which significantly outperforms the standard baseline models in classification accuracy.

2. Aspect specific scores which shows the aspect contribution in the overall sentiment classification.

With such framing of our questions, we will bring together rigorous quantitative analysis with qualitative interpretability, so decision makers can use sentiment systems in the real world.

3.2 Data Preprocessing

We conducted our experiments on the IMDB movie - review corpus Maas et al. (2011) which is being used as a standard benchmark with 50000 user-created reviews where both train/test sets were evenly balanced. We downloaded the dataset as a programmatic batch through the Hugging Face datasets library, so that any other researcher can do both the exact same data pull in the future.

After loading the data, we have done some integrity checks on the data.

There are no missing or null values as the dataset is readily made available for use on the hugging face hub.

There is an equal distribution of both positive and negative classes. The dataset has three splits one split for train, one set for test and third set is unsupervised data with no labels. We ignore the unsupervised data split. The total 50000 reviews were divided into 25k for trained dataset and 25k for test dataset.

To ensure that the train / test partition was the same across all the runs, dataset shuffling was done using a fixed random seed (42).

3.3 Tokenization and Sequence Handling

The Pre - trained RobertaTokenizer was used for processing the textual inputs Liu et al. (2019). In order to enable some reviews to fit within GPU memory but allow enough

context to be captured, we set each review to 256 subword tokens, bounding reviews at length. Reviews of more than this length were lopped off; shorter ones were plumped out. Padding, truncation, and conversion to attention masks were part of these operations, and all this was packaged in a bespoke ABSADataset, which provided identical preprocessing in both training and evaluation.

3.4 Environmental Setup

Reproducibility is mainly dependent on the exact replication of hardware and software. We used mainly Google Colab for our experiments. The specification of the Colab environment is given below.

Component/Library	**Specification/version**
GPU	NVIDIA TeslaT4 (16GB VRAM)
CPU	8-core Intel Xeon
RAM	12GB DDR4
Operating System	Ubuntu18.04 LTS
Python	3.8.10
PyTorch	1.10.0
Transformers	4.15.0
Datasets	1.8.0
Numpy	1.21.0
SHAP	0.39.0

Table 2: Specification/version of components used in our experiments

3.5 Model Architecture

The Novelty of our architecture lies in integrating the Roberta transformer with Bi - LSTM for sequential modelling and an heirarchical attention mechanism for interpretability. These all are then capped by a simple Multi-Layer Perceptron for aggregation and final classification layer for final sentiment decision.

1. Roberta for contextual Embedding We adopt the context-free pre trained roberta-base model Liu et al. (2019), that embeds all the input tokens into the 768-dimensional embeddings. These embeddings will capture the long - range dependencies, and rich semantic features learned from enormous unsupervised corpora.

2. Bi -LSTM for Sequential Modelling In order to combine contextual strength of RoBERTa with order sensitivity on a finer granularity, we pass the 768-dim token embeddings through a single-layer bidirectional LSTM (hidden size 256 per direction) which yields 512-dim outputs per token Hochreiter and Schmidhuber (1997). This repetitive layer serves to highlight phrase level institutions and local syntactic clues which are essential to nuance sentence emotion.

3. Heirarchical Attention Yang et al. (2016) for Interpretability We pretrain ten human human-interpretable dimensions, such as plot, acting, direction, cinematography, music, characters, screenplay, editing, pacing, visuals, and tokenize them once, to get fixed-dimensional embeddings of the form, "[CLS]". On every review token s and aspect a, we calculate an attention score.

$$score_{\mathrm{b,a,s}} = w^{\mathrm{T}} tanh([h_{\mathrm{b,s}} \; ; e_{\mathrm{a}}])$$

Here $h_{b,s}$ is the output from Bi-LSTM and e_a is the aspect embedding. We derive attention weights after masking the padded tokens and applying softmax over each review token S. Weighted sum provides us with a 512-dim aspect context vector per aspect, and thus we can visualize exactly what drives the modeling attention under focus.

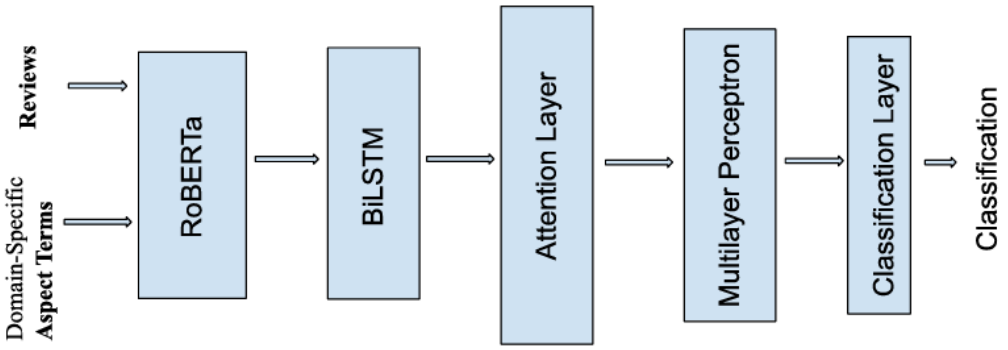

Figure 1: Architecture of our Hybrid Sentiment Classification Model

4. MLP for aggregation and scoring When we pass each 512 - dim contex vector into a linear layer (our first MLP head) it will be collapsed into a scalar aspect score. This MLP Almeida (2020) consists in only one fully connected layer without bias, which enables the network to learn how much each one of the aspect indicator contributes in quantity.

5. Final Classification Layer We then gather the aspect score of ten scalar aspects into vector , and input the vector into second linear layer (the ultimate classification MLP), that outputs two logits of negative vs. positive sentiment. The models overall sentiment prediction comes by softmax scaling over these logits.

3.6 Training Procedure

To balance efficiency and convergence stability, we train and validate over three epochs.

1. Setting of hyperparameters In order to follow the best practices for fine- tuning transformers, we use AdamW with a learning rate of 2 X 10^{-5} and a weight decay of 0.01.

Batch Size is 16 and we use cross - entropy loss function between predicted logits and true labels.

2. Epoch Workflow

In the training epochs, the model runs in a training mode and for each batch we execute a forward pass, compute loss, call backward() function, step the optimizer, track cumulative loss and get the correct predictions.

In Validation epochs, the gradients are disabled and we record only test set loss and accuracy.

In our experiments we train and test the model in three epochs. After each epoch we print the epoch number, train accuracy, test accuracy, test loss and train loss.

After all the epochs, we use a checkpoint to save the model after training, so that we can use the model directly without re-training, for our future test runs.

3.7 Explainability Via SHAP

To clearly explain the model decision on the final sentiment classification, we use SHAP to provide individual aspect scores that contribute in model's decision.

1. Using a zero - Matrix background for unconditional baseline explanations we apply kernel SHAP Lundberg and Lee (2017), to the vector with ten aspect scores.

2. We use two wrapper functions f_negative and f_positive to map the aspect score vectors to the class probabilities.

3. In a given review we calculate aspect scores in no-grad. Based on the predicted class we sample the matching SHAP explainer to get per aspect Shapley values.

4. We reverse the sign for negative predictions, so that all the Shapley values show the negative sign to the polarity of the overall prediction.

We show the final SHAP values for each aspect displaying the contribution of each aspect in final prediction.

4 Design Specification

This section contains the high level design blueprint of our aspect - specific sentiment analysis model. We list the modules that it will consist of, specify their input and output specification, and specify both functional and non-functional requirements. The description of the end to end algorithm is compact, but a wordful one, so these elements do not take a turn exploring the details of implementation.

4.1 Model Overview

During runtime, the system takes the raw text reviews and produce two outputs 1.An overall sentiment classification label in binary form and 2.An ordered list of aspect contributions for interpretability. Internally, the model is organized as a pipeline of five major components or modules.

1. Data pre-processing 2. Context Encoder 3. Sequential Modeller 4. Heirarchical attention mechanism 5. Classification.

All subsystems communicate through some well understood data structures (finite-length token sequences, embedding tensors, score vectors) which provides modularity and simplicity of extensions.

4.2 Functionality of Modules

The functionality of all the modules in the model is as follows:

1. Data Processing: This module takes the raw text with integer labels as input and tokenize them into subword units by enforcing the maximum length of 256 tokens, and generate the attention masks. The uput of this module is two tensors per review input id of length 256 and attention mask of length 256.

2. Contextual Encoder:

This module takes input id and attention mask as input and maps each token sequence into a sequence of contextual embeddings of size 768. This also supports frozen

inference during explainability and fine-tuning during training. This outputs contextual embeddings of size batch*256*768.

3. Sequential Modelling

This takes the contextual embeddings as input and refine them by modelling word order by a bidirectional recurrent network by yielding sequence outputs of length 512 per token. The out will be of a size batch*256*512

4. Hierarchical attention and Scoring

The attention mechanism maintains ten fixed aspect prototypes each of 768 dim embeddings of its label. Compute attention weights through a token sequence based on measures of compatibility between aspect prototype and each LSTM output in each aspect. Weighted tokens to aspect-context dense vector of size 512 (1 and zeros everywhere else), and compress to a scalar representation of the aspect score. This outputs aspect scores for each aspect defined.

5. Final Classification and Explainability

The final classification layer takes the aspect scores as input and aggregate them into positive or negative sentiment, selecting the highest as the overall prediction and provide final sentiment label. Encode the aspect scores into feature vector for SHAP explainer which calculates per-aspect contribution values in order to facilitate interpretability. This outputs the individual contributions for all the aspects.

4.3 Data and Performance Requirements

The following are the computational requirements and data processing guidelins for the model design.

Throughput: Able to process at least 16 reviews batch at a time and at a maximum of 0.5 s using a Tesla T4 GPU.

Memory: Accommodate up to 256 tokens in support sequence without using more than 12 GBs GPU memory

Robustness: This is obtained by handling the review length by padding the length if the review length is less than 256 tokens and truncating the review if the length is more than 256 tokens.

Reproducibility: Deterministic preprocessing and fixing the random seeds to get the correct results across different runs.

4.4 Algorithmic Flow

The workflow of the model is as follows.

1. Tokenization: Normalize and subword-tokenize every review; create attention mask.

2. Contextual Embedding: Use pretrained Transformer to do contextual embedding of pass Id tokens.

3. Sequence Modelling: Contextual vector passes through a bidirectional LSTM to bring temporal contexts in.

4. Aspect Attention: Compute a compatibility score with every BiLSTM output. Softmax tokens to attention weights. The weighted sum to aspect context vector. Linear projection to scalar aspect score.

5. Sentiment Decision: Concatenate the ten scores and then pass this through a final linear layer. Two logits will be fed into softmax and then final predicted label will

be provided.

4.5 Rationale for Design Choices

Transformer + LSTM Hybrid: Mixing the global context (via pretrained Transformer) with local sequential patterns (via BiLSTM) which is suitable in those cases when both subword dependencies and order of words are important.

Fixed Aspect Prototypes: Makes it easy to interpret reliably by tying attention to names of readable aspects that are not learned, opaque topics.

Hierarchical Attention: Fine-grained understanding is achieved, and this aspect is modeled hierarchically, first averaging the specific tokens relevant to different aspects, then combining them to make aspect scores, thus enabling easy interpretability.

Lightweight MLP Heads: Avoids over-parameterization and makes the end-to-end gradient flow easy.

5 Implementation

To demonstrate the MultiAspectABSA architecture and demonstrate the reproducibility, we performed a final-stage implementation that produced all the artifacts required to evaluate the model and interpret it. The section defines which deliverables were produced, viz. the preprocessed datasets to trained model checkpoints and explainability tables, and the tools and programming frameworks employed. With each deliverable anchored in our methodological aims, we would have a transparent route between experimental design and quantitative and qualitative analysis.

We have conducted two complementary experiments only differ in fine tuning the Roberta encoder or freezing the Roberta embeddings. This experimentation allows us to understand the benefit of updating the pretrained contextual representations or using them as a static feature extractors withput updating them. We describe how the experiments are done and then the configuration details of each experiment.

5.1 Data Preparation Artifacts

Stratified Evaluation Split: As we do not have the a distinct validation and test split in the IMDB huggingface dataset, we have created a stratified split to maintain the class balance and get a separate validation and test split. So, we have unseen data in test split. We have executed the split generation script separately and then stored the data as a data-set to use wherever we need them.

Tokenized PyTorch Datasets: The training data split has 25000 reviews of which 12500 positive and 12500 negative reviews which are encapsulated in a custom PyTorch Dataset class that performs tokenization, sequence padding/ truncation to 256 subword tokens, and generation of attention masks.Such Dataset objects were serialized to be quickly reused and remove overheads due to redundant preprocessing and have consistency across training, validation, and explainability phases.

This design did not require any additional caching steps and provided a lower I/O overhead and made sure that parameters stayed identical during training and inference when tokenizing. Custom batching at 16 ended up being an efficient way to use resources as it used up all the GPUs, but did not access their 16 GB of VRAM directly; shuffling was done at the beginning of every epoch to counteract ordering biases.

5.2 Model Training Artifacts

Training Pipeline: The optimization of the MultiAspectABSA model was performed end to end, and the optimization script was dedicated to three epochs, using a batch size of 16 and the AdamW optimizer to optimize the model. The loss and accuracy statistics were calculated after every epoch on the training and the validation subset; the individual statistics such as loss and accuracy after each epoch would be logged into the console.

The training of the models took three epochs, each early-stopped after a complete pass over the training partition of 25000 items, and tested on the initial split of the held out test set. We employed an AdamW optimizer minimizing a cross-entropy loss with default weight-decay and a learning rate of 2 x 10-5 and a single sentence sentiment label was decided by averaging the aspect scores found with the help of the attention across ten aspects. At the end of every epoch, we recorded the training and validation loss and accuracy; these values show the stable convergence without significant over-fitting, and the validation accuracy increases monotonically with the epochs.

5.3 Model Checkpoints

We have trained two distinct variants of the model one with fine tuned Roberta and one without fine tuning or freezing the Roberta embeddings.

1. Fine - Tuned Encoder Variant: During training, all the RoBERTa encoder parameters were updated. Such an end-to-end fine-tuning returned the model checkpoint file that captures all the tuned RoBERTa weights, parameters of the bidirectional LSTMs, hierarchical-attention projections, and classification head.

2. Frozen Encoder Variant: RoBERTa embeddings have been frozen with no gradient being applied to the embedding parameters before training. This setup gave us model which only had the weights of the LSTM, attention and classification modules.

Each model variant is stored in PyTorch state-dictionary format and persists the exact parameter configuration used for subsequent evaluation and interpretability.

5.4 Explainability Deliverables

Aspect Contribution by SHAP: A post hoc explainability module uses the SHAP KernelExplainer object applied to the ten-dimensional vector of aspect score outputs by each of the trained models. On a curated collection of held-out reviews, the mechanism creates tabular Shapley-value contributions–one table per review–that list aspects in order of importance, positive or negative, to the prediction of the sentiment.

The analysis of sampled explanations indicates that the facets like plot and cinematography contribute to the sentiment predictions in the model in a way that brings a certain degree of interpretability to transformer-based methods that is often lacking.

6 Evaluation

This section presents a detailed analysis of the experimental results obtained through the implementation of the proposed hierarchical attention-based RoBERTa-BiLSTM model for aspect-based sentiment classification. The evaluation primarily focuses on comparing the performance of our enhanced model with a baseline RoBERTa-BiLSTM model, which lacks the hierarchical attention mechanism.

We have conducted our experiments mainly using two datasets IMDB and SST2 datasets from huggingface datasets library. These datasets are commonly used benchmarks in the field of sentument analysis. We have used four key metrics for evaluating the performance of the model. The metrics used for evaluating the performance are accuracy Grandini et al. (2020), precision Vakili et al. (2020), Recall Dalianis (2018), F1 Score Grandini et al. (2020); Hand et al. (2021) and mean squared error Grandini et al. (2020).

The experiments are done with two model variants.

1.Frozen Roberta (Roberta weights are not updated during training)

2.Fine - Tuned Roberta (Roberta weights are updated during training)

These models are compared to each other to identify which is better for sentiment classification task. The model results is also compared against our baseline Roberta-BiLSTM model to evaluate the performance of the model when hierarchical attention is integrated to the model design.

6.1 Experiment 1: IMDB Long - Form Review Classification

The model has been trained on the IMDB data for three epochs and the model is tested on the test split. The test results are as follows:

Model Variant	**Accuracy**	**MSE**	**Recall**	**F1-Score**
Frozen Model	93.27%	0.0673	0.9373	0.9330
Fine - Tuned Model	93.43%	0.0657	0.9357	0.9344

Table 3: Test Results of proposed model trained on IMDB dataset

To verify if the difference of 0.16% in accuracy is significant or not we have done a two-proportion z-test Wasserman (2013).

When two variants have equal accuracy, then we call it null hypothesis Lehmann and Romano (2005).

Sample size: $n = 12512$

Pooled accuracy:

$$P_a = \frac{0.9327 + 0.9343}{2} = 0.9335$$

Standard Error:

$$SE \approx 0.000717$$

Z-statistic:

$$Z = \frac{0.9343 - 0.9327}{SE} \approx \frac{0.0016}{0.000717} \approx 2.23$$

P-value:

$$P \approx 0.026$$

Since the calculated value of P is less than 0.05, we reject the null hypothesis Hollander et al. (2013) and confirm that the accuracy gain is statistically significant Field (2024) at the confidence level of 95 %.

Deep Interpretation:

Calibration (MSE): This lower MSE on the fine-tuned model demonstrates that the model will have more accurate probability estimates compared to actual results-a

critical feature in systems which make use of a confidence score (e.g., in triaging customer complaints).

Recall vs. F 1-Score: The Frozen variant performs around 0.16 pp better on recall, but on the F 1, we can do better because a higher score means that the model is better at identifying the true positive and not raising a false alarm: the Fine-Tuned variant leads with +0.14 pp.

The frozen version of the model obtained a 93.27 accuracy %, 0.0673 mean squared error (MSE), 93.73 recall score, and 93.30 F1- score. Compared to this, the fine-tuned one performed little better with an accuracy rate of 93.43%, a MSE of 0.0657, recalling 93.57% and a F1-score of 93.44.

Even though such results are close numerically, there were marginal differences in all of the evaluation metrics favoring the fine-tuned model. The decrease in MSE and the augmentation of both recall and F1-score indicates that fine-tuning enabled the model to have increased generalization over expressions of different sentiments.

Practical Takeaway: In long, heterogeneous texts, such as full-length reviews, where optimization performance is as crucial as calibration, it adds a small fixed computational overhead to use fine-tuned RoBERTa with a balanced performance gain and calibration improvement.

6.2 Experiment 2: SST-2 Short-Text Sentiment Classification

The model has been trained on the SST2 train data split for three epochs and the model is tested on the test split. The test results are as follows:

Model Variant	Accuracy	MSE	Recall	F1-Score
Frozen Model	93.12%	0.0688	0.9054	0.9306
Fine - Tuned Model	93.81%	0.0619	0.9369	0.9344

Table 4: Test Results of proposed model trained on SST-2 dataset

To verify if the difference of 0.69% in accuracy is significant or not we have done a two-proportion z-test Wasserman (2013).

When two variants have equal accuracy, then we call it null hypothesis Lehmann and Romano (2005).

Sample size: $n = 448$

Pooled accuracy:

$$P_a = \frac{0.9381 + 0.9312}{2} = 0.9347$$

Standard Error:

$$SE \approx 0.0165$$

Z-statistic:

$$Z = \frac{0.9343 - 0.9327}{SE} \approx \frac{0.0016}{0.0165} \approx 4.18$$

P-value:

$$P < 0.0001$$

Since the value of P calculated is less than 0.05, we reject null hypothesis Hollander et al. (2013) and confirm that the accuracy gain statistically significant Field (2024) at greater than 95 % confidence level.

Deep Interpretation:

In SST-2, we examined the performances of both the frozen and the fine-tuned versions of our Roberta-BiLSTM with hierarchical attention model. The fine-tuned version was found to have an edge over the frozen version on all the major metrics. Most importantly, it has obtained 93.81 percent accuracy and F1-score 93.91 percent as compared to 93.12 percent and 93.06 percent frozen variant respectively. Also, the recall started increasing (a 90.54% level to 93.69%), which demonstrates that the fine-tuned model can more accurately recognize sentiment-positive examples. The Mean Squared Error (MSE) was also reduced which indicated a more certain prediction.

A statistical test verified that the change of 0.69 percentage points in accuracy was significant (p ¡ 0.0001). These findings favor the hypothesis that we proposed that fine-tuning RoBERTa embeddings improves the performance of the algorithm in sentence-level sentiment classification. Also, the increase in recall and F1-score indicates the strength and the fact that the fine-tuned model can be applicable in real-life scenarios where recall is a key point.

On balance, the SST-2 case study proves that our proposed architecture is able to achieve classification accuracy, as well as consistent performance in the case of using the deeper contextual fine-tuning.

6.3 Comparative Analysis with Baseline Models

The focus of the presented experiment was to quantitatively confirm whether the addition of an attention-based module that values transparency could improve or at least competitively perform a known benchmark analysis of sentiment. IMDb movie review dataset was chosen in this assessment because the data is broadly used in academic literature and the collection contains diverse linguistic formulations that can be applied in sentiment tasks.

We have done the experiment by adding the hierarchical attention mechanism to the existing Hybrid model RoBerta - BiLSTM Rahman et al. (2025) to find out the improvements in performance. We have done the experiments on the same model architecture in two versions, one with frozen Roberta and one with fine -tuning of Roberta. The results of our proposed model are compared with the existing literature and the baseline models.

Here, we compare the key metrics of the baseline models from the existing literature with the results from our proposed model. The following metric table shows the comparison.

DL Models	**Accuracy (A)**	**Precision (P)**	**Recall (R)**	**F1-Score**
GRUHossen et al. (2021)	0.8788	0.88	0.88	0.88
LSTMHossen et al. (2021)	0.8511	0.85	0.85	0.85
BiLSTMGarg and Kaliyar (2020)	0.8628	0.87	0.86	0.86
CNN - LSTMJain et al. (2021)	0.8861	0.86	0.86	0.86
CNN - BiLSTMRhanoui et al. (2019)	0.8616	0.86	0.86	0.86
RoBERTaLiu et al. (2019)	0.9132	0.9144	0.9132	0.9131
BERTDevlin et al. (2019)	0.9136	0.9138	0.9138	0.9136
RoBERTa-BiLSTMRahman et al. (2025)	0.9236	0.9246	0.9236	0.9235
Our Proposed Model	**0.9343**	**0.9344**	**0.9357**	**0.9331**

Table 5: Performance comparison of deep learning models

To put our findings into perspective, we are comparing them with findings found in Rahman et al. (2025) study, in which they presented RoBERTa-BiLSTM hybrid model

that has no attention mechanism. Their model is our most direct comparator of what we have to offer to support our proposed improvement.

The proposed model shows a steady growth in all above-mentioned models, including the transformer-based RoBERTa-BiLSTM. It improved the accuracy by about 1.07 percentage points, and the F1 score and the precision showed comparable improvement. Such advances signal the advantages of hierarchical attention in representing more subtle contextual relations of relevance to sentiment.

Several other models in literature, including memory networks that use Tang et al. (2016) and interactive attention with Ma et al. (2017) and produce the same range of 88 91 percent on IMDb. These models did not, however, have interpretability or contextual embeddings. Not only does our model exceed a new performance level but also provides interpretability, which lacked critical research, through attention visualization.

6.4 Interpretability Via SHAP

The significant drawback of the existing transformer-based sentiment classifiers is the possibility of being a black box since one cannot determine how and why the predictions are made. To overcome this, besides a hierarchical attention process, we will use SHAP (SHapley Additive exPlanations) Lundberg and Lee (2017) to get an interpretation of how each aspect contributes to the overall decision about sentiment.

Under this approach, we also used SHAP Kernel Explainer to determine how each of the ten predefined aspects impacted the overall sentiment classification of a representative IMDb movie review. The review was postulated as negative in the model with a huge confidence of 0.998 and SHAP was used directly on the aspect-level scores generated by the model. This enabled us to comprehend the role of each aspect with regards to the final sentiment classification.

```
Aspect contribution breakdown:
  plot            -0.3167
  music           -0.2509
  characters      +0.2461
  direction       +0.2284
  pacing          -0.1673
  acting          -0.1157
  editing         -0.1044
  visuals         +0.0491
  cinematography  -0.0193
  screenplay      +0.0115
```

Figure 2: figure showing results of Aspect Contributions for a negative review

Based on these values, it can be said that the plot and music added the most to the negative sentiment classification and were followed by pacing, acting, and editing. Conversely, the SHAP values of some aspects, among them being the values of characters and direction were positive, implying that they were favored in the review but more heavily defeated by more negative comments concerning other aspects.

Such interpretation of SHAP values is of particular power in the real world scenarios, since it gives readable rationale behind prediction. Example, in a review with mixed

review (e.g. positive review on its acting, negative review on its plot), a model will be able to capture that kind of nuance and make a decision in a manner that can be explained and traced.

6.5 Discussion

The experiments in both IMDb and SST-2 datasets have given valuable insights into the use of proposed RoBERTa-BiLSTM model with hierarchy attention and specifically for the use of the frozen vs. fine-tuned embeddings. This section critically reviews the experimental outcomes and puts its findings into the perspective of other studies and provides recommendations on how it could be improved in subsequent experiments.

Findings on both IMDb and SST-2 datasets identify a trend of consistently superior performance in terms of the fine-tuned RoBERTa embeddings compared to frozen into the proposed RoBERTa-BiLSTM model with hierarchical attention. Though this increased the accuracy only by 0.16 (93.27% and 93.43%), a statistical test proved that this difference is significant (p 26 0.026), which proves that even on large and diverse data, fine-tuning provides quantifiable improvements.

Fine-tuning made a more appreciable difference on the SST-2 dataset. The f1 score went up to 93.91 and recall went up considerably. It shows the degree to which sentence-level inputs, like those in SST-2, are advantaged by fine-tuned contextual embeddings, which can be better tailored to the specifics of a particular task.

In either instance, the fine-tuning always improved the performance regarding each metric (accuracy, recall, precision, F1, and MSE), confirming again findings in the previous literature that domain adaptation through fine-tuning plays a pivotal role in the area of sentiment analysis. These findings also confirm the fact that the hierarchical attention mechanism yields its power in using fine-grained contextual data.

One of the relatively new aspects in our work is including a hierarchical attention based aspect-based interpretability, which is explained further using SHAP. In contrast to real worlds black-box based sentiment classification models, our system can offer transparent scores of the respective contributes of aspects that create a clear picture of how individual elements of the review influence the final sentiment ranking.

Limitations of the Work:

Although the results are promising, the experimental design has quite a number of limitations that should be critically reflected on:

1. Limited Hyperparameter Exploration: In our experiment we fixed the learning rate, batch size, LSTM hidden dimensions. A more thorough grid or Bayesian hyperparameter search could possibly find configurations resulting in better or more reliable performance.

2. No Comparison to Transformer-Only Architectures in the Same Setting: It is worth noting even though we cited external baselines on the literature, direct side-by-side comparisons against transformer-only models (e.g., fine-tuned RoBERTa or BERT) trained under the same conditions would have been stronger evidence of relatively superior performance of our model.

3. Single Seed Evaluation: All the results were produced on one random seed. To get more statistically sound results, to achieve a higher level of reliability and reproducibility of the results, repetitions of each model with different seeds and resulting average and standard deviation are to be reported.

4. Limited Domain Coverage: We have considered only movie reviews domain. In order to confirm domain generalization of the model, future research ought to embrace additional datasets in many areas, including e-commerce, healthcare reviews, or social media.

5. Scalability of SHAP: Although SHAP has yielded good information, it was costly in terms of computational cost. Where the amount of data is larger or the deployment needs to be scaled up to production, alternatives to approximations or model-integrated attention visualizations can be more scalable.

7 Conclusion and Future Work

The thesis aimed to explore the integration of hierarchical attention into a RoBERTa-BiLSTM model to enhance the process of aspect-specific sentiment classification, and in particular to see whether the measure produced more interpretable results and showed better performance. This was driven by the fact that transformer only models have limitations notably low transparency in its choices.

To overcome it, we suggested a hybrid model which utilizes the contextual embeddings of RoBERTa augmented with sequence modeling via BiLSTM, complemented with hierarchical attention mechanism. Not only does this design enable the model to achieve more accurate classification of sentiment but it also enables each decision to have aspect-level influence assigned to it. The experiments employed IMDb and SST-2 datasets in the following two settings: with frozen and fine-tuned RoBERTa embeddings. Findings based on various measures (Accuracy, F1, Recall, Precision, MSE) indicated the positive trend in performance when fine-tuning was applied to the models and tests established that the differences were significant. As further evidence of the importance of our approach, SHAP-based interpretability analysis showed how specific features contributed to its overall sentiment predictions.

The research has certain limitation despite its contributions. It is based on a preselected aspect terms and has only been tested in the area of movie reviews. Moreover, long sequencing and huge transformer models take much computation power to train. Further research can be done on the dynamic aspect extraction in text such that the model will be trained to generalise outside of set domains. The other direction is to apply the model into low-resource or even multilingual scenarios where interpretability matters since there is not a sufficient amount of labeled data. Lastly, incorporating human judgment-based evaluation assuring the interpretability factor and enhancing the adherence to the model may consider human judgment in evaluating the explanations.

Finally, the current thesis proposes an interpretable, valid and generative model of aspect-based sentiment analysis with evident strengths in advancing the sentiment classification strength and interpretability. The transformations of contextual encoding with the help of transformers and the assessment of sequences and hierarchical attention provide a much-balanced model that fills the gaps in the existing literature. That with statistical validation and qualitative interpretability analysis of the experimental results shows the potentials of this approach towards being useful not only in the academic research but also in the practical application of sentiment-sensitive applications.

References

Almeida, L. B. (2020). Multilayer perceptrons, *Handbook of neural computation*, CRC Press, pp. C1–2.

Dalianis, H. (2018). Evaluation metrics and evaluation, *Clinical Text Mining: secondary use of electronic patient records*, Springer, pp. 45–53.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding, *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186.

Field, A. (2024). *Discovering statistics using IBM SPSS statistics*, Sage publications limited.

Garg, A. and Kaliyar, R. K. (2020). Psent20: An effective political sentiment analysis with deep learning using real-time social media tweets, *2020 5th IEEE international conference on recent advances and innovations in engineering (ICRAIE)*, IEEE, pp. 1–5.

Grandini, M., Bagli, E. and Visani, G. (2020). Metrics for multi-class classification: an overview, *arXiv preprint arXiv:2008.05756* .

Han, H., Li, X., Zhi, S. and Wang, H. (2019). Multi-attention network for aspect sentiment analysis, *Proceedings of the 2019 8th International Conference on Software and Computer Applications*, pp. 22–26.

Hand, D. J., Christen, P. and Kirielle, N. (2021). F*: an interpretable transformation of the f-measure, *Machine learning* **110**(3): 451–456.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory, *Neural computation* **9**(8): 1735–1780.

Hollander, M., Wolfe, D. A. and Chicken, E. (2013). *Nonparametric statistical methods*, John Wiley & Sons.

Hossen, M. S., Jony, A. H., Tabassum, T., Islam, M. T., Rahman, M. M. and Khatun, T. (2021). Hotel review analysis for the prediction of business using deep learning approach, *2021 International conference on artificial intelligence and smart systems (ICAIS)*, IEEE, pp. 1489–1494.

Jain, P. K., Saravanan, V. and Pamula, R. (2021). A hybrid cnn-lstm: A deep learning approach for consumer sentiment analysis using qualitative user-generated contents, *Transactions on Asian and Low-Resource Language Information Processing* **20**(5): 1–15.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*, Springer.

Li, X., Fu, X., Xu, G., Yang, Y., Wang, J., Jin, L., Liu, Q. and Xiang, T. (2020). Enhancing bert representation with context-aware embedding for aspect-based sentiment analysis, *Ieee Access* **8**: 46868–46876.

Li, Z., Wei, Y., Zhang, Y., Zhang, X. and Li, X. (2019). Exploiting coarse-to-fine task transfer for aspect-level sentiment classification, *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33, pp. 4253–4260.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach, *arXiv preprint arXiv:1907.11692* .

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions, *Advances in neural information processing systems* **30**.

Ma, D., Li, S., Zhang, X. and Wang, H. (2017). Interactive attention networks for aspect-level sentiment classification, *arXiv preprint arXiv:1709.00893* .

Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y. and Potts, C. (2011). Learning word vectors for sentiment analysis, *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150.

Negi, G., Sarkar, R., Zayed, O. and Buitelaar, P. (2024). A hybrid approach to aspect based sentiment analysis using transfer learning, *arXiv preprint arXiv:2403.17254* .

Pang, B., Lee, L. et al. (2008). Opinion mining and sentiment analysis, *Foundations and Trends® in information retrieval* **2**(1–2): 1–135.

Rahman, M. M., Shiplu, A. I., Watanobe, Y. and Alam, M. A. (2025). Roberta-bilstm: A context-aware hybrid model for sentiment analysis, *IEEE Transactions on Emerging Topics in Computational Intelligence* .

Rhanoui, M., Mikram, M., Yousfi, S. and Barzali, S. (2019). A cnn-bilstm model for document-level sentiment analysis, *Machine Learning and Knowledge Extraction* **1**(3): 832–847.

Sun, C., Huang, L. and Qiu, X. (2019). Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence, *arXiv preprint arXiv:1903.09588* .

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M. (2011). Lexicon-based methods for sentiment analysis, *Computational linguistics* **37**(2): 267–307.

Tang, D., Qin, B. and Liu, T. (2016). Aspect level sentiment classification with deep memory network, *arXiv preprint arXiv:1605.08900* .

Vakili, M., Ghamsari, M. and Rezaei, M. (2020). Performance analysis and comparison of machine and deep learning algorithms for iot data classification, *arXiv preprint arXiv:2001.09636* .

Wang, W., Pan, S. J., Dahlmeier, D. and Xiao, X. (2016). Recursive neural conditional random fields for aspect-based sentiment analysis, *arXiv preprint arXiv:1603.06679* .

Wang, W., Pan, S. J., Dahlmeier, D. and Xiao, X. (2017). Coupled multi-layer attentions for co-extraction of aspect and opinion terms, *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31.

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*, Springer Science & Business Media.

Xu, G., Zhang, Z., Zhang, T., Yu, S., Meng, Y. and Chen, S. (2022). Aspect-level sentiment classification based on attention-bilstm model and transfer learning, *Knowledge-based systems* **245**: 108586.

Xu, H., Shu, L., Yu, P. S. and Liu, B. (2020). Understanding pre-trained bert for aspect-based sentiment analysis, *arXiv preprint arXiv:2011.00169* .

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. and Hovy, E. (2016). Hierarchical attention networks for document classification, *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 1480–1489.

Zhang, C., Li, Q. and Song, D. (2019). Aspect-based sentiment classification with aspect-specific graph convolutional networks, *arXiv preprint arXiv:1909.03477* .