Aspect-Based Sentiment Analysis in Restaurant Reviews

Abhiram Reddy Pudi
AbhiramPudi@my.unt.edu
University of North Texas
Denton, Texas, USA

Saketh Papareddy
sakethpapareddy@my.unt.edu
University of North Texas
Denton, Texas, USA

Karthik Etukuri
karthikchoudharyetukuri@my.unt.edu
University of North Texas
Denton, Texas, USA

Maddineni Naga Sampath
NagaSampathMaddineni@my.unt.edu
University of North Texas
Denton, Texas, USA

Abstract

Understanding and analyzing customer opinions at a minute level is crucial for businesses.Aspect-Based Sentiment Analysis (ABSA) aims to identify specific aspects (features or topics) referred to in the text and to determine the sentiment expressed towards each aspect.This paper explores and compares two deep learning approaches used for ABSA on restaurant reviews.The first and primary approach employs a pipeline consisting of fine-tuned transformer models (RoBERTa-base) for Aspect Term Extraction (ATE) and Aspect Sentiment Classification (ASC).These are compared with a baseline model that uses the same Transformer-based ATE but employs a different method, namely a Bidirectional Long Short-Term Memory (BiLSTM) network for ASC.We evaluate both approaches (i.e., Baseline and Pipeline) on a standard restaurant review dataset using strict end-to-end metrics.

Our results demonstrate the superior performance of the Transformer pipeline (F1-score: 0.74) compared to the LSTM baseline (F1-score: 0.65), highlighting the effectiveness of attention mechanisms and pre-training for capturing aspect-specific context in sentiment analysis.

Keywords: Aspect-Based Sentiment Analysis, Deep Learning, Transformers, RoBERTa, LSTM, Restaurant Reviews

ACM Reference Format:
Abhiram Reddy Pudi, Saketh Papareddy, Karthik Etukuri, and Maddineni Naga Sampath. 2025. Aspect-Based Sentiment Analysis in Restaurant Reviews. In . ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

In the modern age, customer reviews on platforms like Yelp, Google Maps, and TripAdvisor represent a vast and valuable source of feedback for businesses, mainly in the hospitality and service sectors. Examining these reviews allows businesses to measure customer satisfaction, identify areas needing improvement, and understand market trends. However, simply classifying an entire review as positive or negative often hides crucial details. A customer might express satisfaction with the food quality but also complain about the slow service within the same review. To capture these minute details, Aspect-Based Sentiment Analysis (ABSA) has emerged as a crucial task in Natural Language Processing (NLP).

ABSA typically involves two main sub-tasks: first, identifying the specific aspects or features being discussed (e.g., “pizza,” “waiter,” “price,” “decor”)—known as **Aspect Term Extraction (ATE)**—and second, determining the sentiment polarity expressed towards each identified aspect (e.g., positive, negative, neutral, or conflict)—known as **Aspect Sentiment Classification (ASC)** or **Aspect Polarity Classification (APC)**. Analyzing and solving the ABSA problem provides businesses with actionable insights directly linked to specific sides of their service or product.

The relevance of ABSA extends beyond individual business improvements. It also enables large-scale market analysis, competitor benchmarking, and tracking public opinion trends related to specific product features or service attributes over time. The complexity arises from the need to correctly identify not just the aspect terms, which can be multi-word expressions or implicit references, but also to associate the sentiment expressed specifically towards that aspect, often within sentences containing multiple aspects with differing polarities.

Early attempts relied on rule-based systems and sentiment lexicons, which often struggled with coverage and domain adaptation. Machine learning approaches using handcrafted features improved performance, but the emergence of deep learning, particularly Recurrent Neural Networks (RNNs) like LSTMs, and more recently, Transformer models like BERT and RoBERTa, has led to significant breakthroughs. These models can learn rich contextual representations from data, reducing the need for manual feature engineering.

This study focuses on implementing and rigorously comparing two prominent deep learning pipeline approaches for end-to-end ABSA on a restaurant review dataset. Our primary approach leverages the power of pre-trained RoBERTa models for both the ATE and ASC stages. We compare this

Conference'17, Washington, DC, USA
2025. ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

against a strong baseline that uses the identical RoBERTa model for ATE but employs a Bidirectional LSTM (BiLSTM) network specifically trained for the ASC task. By keeping the ATE component constant, we aim to isolate and quantify the impact of using a Transformer versus an LSTM for the aspect-specific sentiment classification sub-task within a standard pipeline framework.

2 Related Works

The sector of Aspect-Based Sentiment Analysis (ABSA) has been evolving significantly over the past decades, driven considerably by the use of benchmark datasets and shared tasks organized by SemEval. Various research efforts has uncovered some sub-problems within ABSA, including Aspect Term Extraction (ATE), Aspect Category Detection (ACD), and Aspect Sentiment Classification (ASC).

Early computational approaches often involved mining frequent nouns and noun phrases as candidate aspects and using sentiment lexicons (e.g., SentiWordNet) combined with linguistic rules (e.g., proximity, negation handling, dependency parsing) to identify sentiment polarity. While effective in some cases, these methods often lacked robustness and required significant manual effort for different domains or languages.

Later, the adoption of supervised machine learning techniques provided notable improvements. ATE was frequently modeled as a sequence labeling problem using algorithms like Conditional Random Fields (CRFs) or Support Vector Machines (SVMs), trained on features such as word embeddings, part-of-speech tags, and syntactic information. ASC was often treated as a standard classification task, utilizing features derived from the aspect term and its surrounding context.

The application of deep learning marked a significant change. Long Short-Term Memory networks (LSTMs), with their ability to model sequential dependencies, became a popular choice for both ATE and ASC. For ASC, specialized LSTM architectures were developed to incorporate target aspect information more directly. For instance, Target-Dependent LSTM (TD-LSTM) employed separate LSTMs to model the context before and after the aspect term. Attention mechanisms were frequently layered onto LSTMs, allowing the model to dynamically weigh the importance of different context words when determining sentiment towards a specific aspect. Convolutional Neural Networks (CNNs) were also explored, demonstrating effectiveness in capturing local contextual features relevant to sentiment.

The initiation of large pre-trained Transformer models, such as BERT, revolutionized NLP, including applications in ABSA. Fine-tuning BERT or its variants like RoBERTa has become the dominant approach for achieving state-of-the-art results in this domain. Common strategies for applying Transformers to ABSA include:

- **Pipeline approach**: Separate models are fine-tuned for ATE (token classification) and ASC (sequence-pair classification, e.g., [CLS] sentence [SEP] aspect [SEP]).
- **Joint modeling**: Simultaneously solving ATE and ASC within a single model using multi-task learning frameworks or specialized tagging schemes.
- **Generative or QA frameworks**: Framing ABSA as a sequence generation or reading comprehension task.

Our work mainly focuses on the widely used and effective pipeline strategy. We use RoBERTa, a robustly optimized variant of BERT, as the foundation for our main approach. The originality lies in the direct empirical comparison against a baseline that uses the same RoBERTa ATE component but replaces the RoBERTa ASC component with a dedicated BiLSTM ASC model. This allows for a focused evaluation of the sentiment classification stage, contrasting a modern Transformer against a representative pre-Transformer deep learning architecture within an identical extraction framework.

3 Methods

We aim to perform an end-to-end ABSA, taking a sentence as input and outputting a set of (aspect_term, sentiment_polarity) pairs. We implement this by using a two-stage pipeline which consists of Aspect Term Extraction (ATE) accompanied by Aspect Sentiment Classification (ASC).These two approaches are compared as : a full Transformer pipeline and a baseline using a Transformer for ATE and an LSTM for ASC.

3.1 Shared Aspect Term Extraction (ATE) Model

Both the baseline and the pipeline approach uses the identical ATE model to ensure a better comparison of the subsequent ASC stage. The ATE task is framed as sequence labeling.

- **Model Architecture**: We employ a pre-trained RoBERTa-base model (`roberta-base`) provided by Hugging Face. A linear layer is added on top of the final hidden-state output of the RoBERTa encoder for each token, acting as a token classifier.
- **Tagging Scheme**: The BIO tagging scheme is used. Each input token is mapped to one of three labels: `B-ASP`, `I-ASP`, or `O`.
- **Data Preparation and Fine-tuning**: Input sentences are tokenized using the RoBERTa tokenizer. Ground truth aspect terms are mapped to token spans. Fine-tuning is performed using the Hugging Face Trainer class, optimizing Cross-Entropy loss.

3.2 Approach 1: Pipeline with Transformer ASC

This approach uses RoBERTa-base for both ATE and ASC.

- **ASC Model Architecture**: A second RoBERTa-base model is fine-tuned for sequence classification. A classification head is added on top of the pooled output (typically from the [CLS] or <s> token).
- **ASC Input Formulation**: Input formatted as: `<s> sentence_text </s> </s> aspect_term_text </s>`.
- **ASC Fine-tuning**: Fine-tuned using pairs of sentences and aspect terms with Cross-Entropy loss.

3.3 Approach 2: Baseline with LSTM ASC

- **ASC Model Architecture**: The BiLSTM model includes:
  1. `nn.Embedding` layer (100d).
  2. Bidirectional `nn.LSTM` (hidden size 128 per direction).
  3. Aspect-specific pooling over hidden states.
  4. `nn.Linear` classification layer.
- **Data Preparation and Training**:
  - Tokenization using NLTK.
  - Padding to `LSTM_MAX_SEQ_LEN`.
  - Training with Cross-Entropy loss and Adam optimizer.

3.4 Implementation Details

The project was implemented in Python 3 using Google Colab with Tesla T4 GPU. Key libraries: `torch`, `transformers`, `datasets`, NLTK, pandas, `xml.etree.ElementTree`, scikit-learn, and seqeval.

4 Results

4.1 Dataset Description

The study was conducted using the `Restaurants_Train.xml` dataset, a collection of English restaurant reviews annotated for Aspect-Based Sentiment Analysis (ABSA). This dataset contains 3044 sentences overall. We performed an 80/20 split, resulting in 2435 sentences for training and 609 sentences for testing. These annotations include specific aspect terms within sentences, along with character offsets and a sentiment polarity label from the set {positive, negative, neutral, conflict}. The test split contains 445 unique ground truth (aspect term, polarity) pairs used for end-to-end evaluation.

4.2 Evaluation Metrics

Model performance is assessed using two main criteria:

- **Aspect Term Extraction (ATE):** The ability of the shared RoBERTa ATE model to correctly identify spans of aspect terms was measured using token-level Precision, Recall, and F1-score, calculated via the `seqeval` library.
- **End-to-End ABSA:** The full pipeline model (ATE followed by ASC) is evaluated using a strict matching approach, where a prediction is considered correct only if both the extracted aspect term and the sentiment polarity match exactly. Precision, Recall, and F1-score are computed under this strict criterion.

4.3 Experimental Setup

The shared RoBERTa-base ATE model was fine-tuned for 10 epochs with a learning rate of 2×10^{-5} and a batch size of 8. The pipeline's RoBERTa-base ASC model was fine-tuned for 5 epochs with a learning rate of 2×10^{-5}, a per-device batch size of 4, 2 gradient accumulation steps (effective batch size of 8), and mixed-precision (fp16) training enabled. The baseline Bi-LSTM ASC model was trained for 20 epochs using the Adam optimizer with a learning rate of 0.001 and a batch size of 16. For both ASC models, the best checkpoint based on the weighted F1-score on the validation/test set was selected for final evaluation.

4.4 Quantitative Results

ATE Performance. The RoBERTa-base model fine-tuned for ATE achieved strong performance in identifying aspect term spans, with an F1-score of 0.8912 (Precision: 0.8797, Recall: 0.9031).

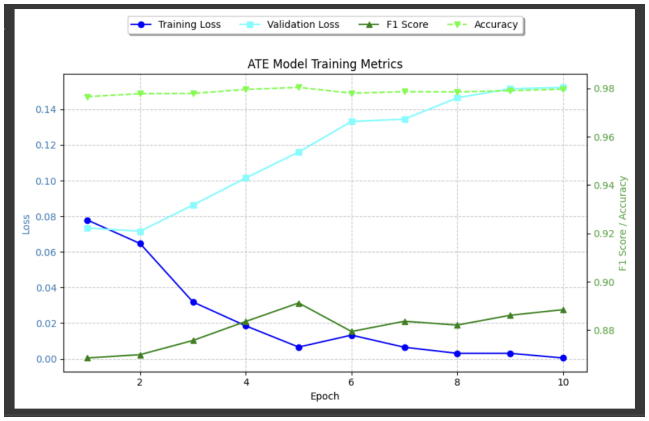

Figure 1. ATE Model Training Metrics: Training Loss, Validation Loss, F1 Score, and Accuracy.

End-to-End ABSA Performance. The comparison between the Baseline (ATE+LSTM) and Pipeline (ATE+Transformer) approaches revealed significant differences:

Approach	Precision	Recall	F1-Score
Baseline (ATE + LSTM)	0.6560	0.6472	0.6516
Pipeline (ATE + Transformer)	0.7309	0.7506	0.7406

Table 1. Strict End-to-End ABSA Matching Performance.

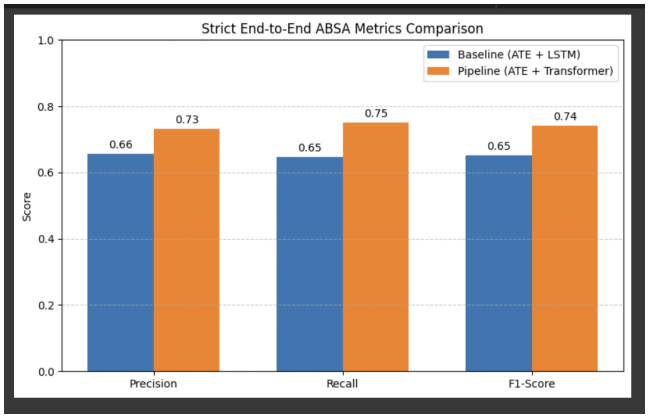

Figure 2. Strict End-to-End ABSA Metrics Comparison between Baseline and Pipeline approaches.

Training Performance. Training dynamics of the ASC models are shown below.

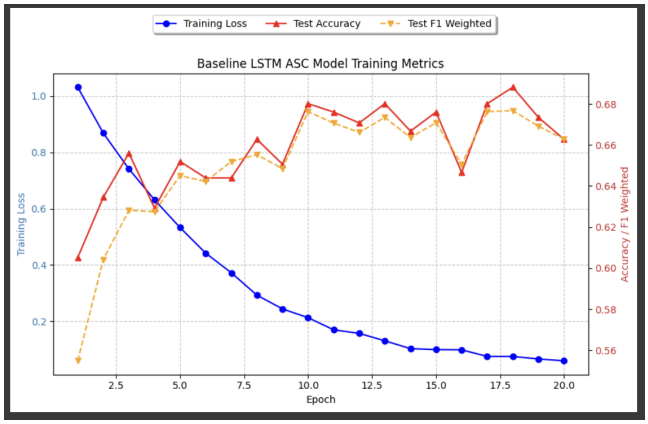

Figure 3. Training Performance of the Baseline LSTM ASC Model.

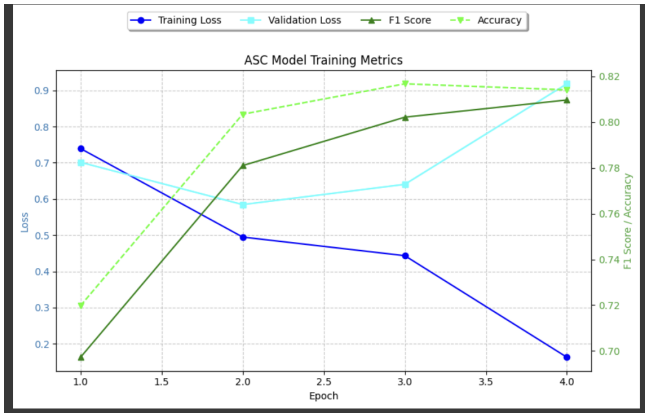

Figure 4. Training Performance of the Transformer ASC Model.

4.5 Qualitative Analysis

Examining specific examples where the Pipeline succeeded and the Baseline failed highlights the Transformer's strengths:

- **Sentence**: "The sake menu should not be overlooked!"
  - Ground Truth: ("sake menu", positive)
  - Pipeline: ("sake menu", positive) (Correct)
  - Baseline: ("sake menu", negative) (Incorrect)
- **Sentence**: "As for the bar, this is another bad idea."
  - Ground Truth: ("bar", negative)
  - Pipeline: ("bar", negative) (Correct)
  - Baseline: ("bar", neutral) (Incorrect)
- **Sentence**: "But who says Murray's is anything about service."
  - Ground Truth: ("service", neutral)
  - Pipeline: ("service", neutral) (Correct)
  - Baseline: ("service", positive) (Incorrect)

These examples illustrate the Pipeline model's superior handling of inferred sentiment (e.g., "should not be overlooked"), negations, and neutral cases compared to the LSTM-based baseline, which appears more prone to misclassification in such detailed contexts.

5 Discussion

The results clearly indicates that the pipeline model which is leveraging RoBERTa for both Aspect Term Extraction and Aspect Sentiment Classification considerably outperforms the baseline model which uses an LSTM for the sentiment classification stage. An end-to-end F1-score improvement from 0.65 to 0.74 on a strict end to end evaluation emphasises the benefits of using Transformer architectures for complex, context-dependent tasks like ASC. The shared ATE component performs well F1 $\approx$ 0.89 ,which suggests that the main performance difference arises from the ASC module's ability of interpretting sentiment towards a specific aspect within the context of the sentence. The qualitative analysis further supports this by showcasing the Transformer's superior handling of negation, indirect sentiment expression, and neutral cases when compared to the BiLSTM. This outcome aligns with expectations and findings in the broader NLP literature, where the transformer models have consistently surpassed RNNs on various benchmarks because to their self-attention mechanism, which enables for better modeling of the long-range dependencies and contextual word relationships. Pre-training on vast text collection also provides the transformers with a stronger foundation for language understanding when compared to LSTMs normally trained with randomly initialized or standard pre-trained embeddings (e.g., GloVe) on smaller, task-specific datasets. While the pipeline achieved a considerable F1-score of approximately 0.74, especially in circumstances like the strict end-to-end evaluation, there is still room for improvement. The difference between the ATE performance (approximately 0.89 F1)

and the end-to-end performance (approximately 0.74 F1) underlines the impact of error propagation. Any error in the ATE stage (missed term, incorrect boundary) makes it impossible for the ASC stage to contribute to a correct end-to-end prediction for that specific aspect. Furthermore, the ASC model itself, despite using RoBERTa, is not perfect (approximately 0.81 validation accuracy/F1 during isolated training, which likely drops when handling potentially noisy terms from ATE). The observed overfitting during ASC fine tuning (evident in validation loss trends shown in logs, even though early stopping based on F1 was used) suggests the model could benefit from further regularization or more diverse training data. The basic vagueness of language has some potential inconsistencies or difficulties present in the dataset. Future research can focus on several areas. Addressing those errors by exploring the joint ATE+ASC models can drastically yield better results by allowing the tasks to inform each other. Experimenting with larger or more domain-specific pre-trained models (e.g., models pre-trained on review corpora) can improve both ATE and ASC stages.And techniques like data augmentation could help in improving generalization in avoiding overfitting observed in the ASC model. A more detailed error analysis, perhaps breaking down performance by sentiment class (especially neutral and conflict, which are often harder) or aspect type, could pinpoint specific weaknesses. By using methods such as label smoothing and more advanced optimization strategies while fine-tuning the model can further enhance the model performance.

In conclusion, this study confirms the effectiveness of a full Transformer pipeline for ABSA compared to an LSTM-based baseline, by achieving substantial performance.Mainly by improving aspect-specific sentiment classification. While the results are promising, the analysis also highlights the remaining challenges in end-to-end ABSA and suggests avenues for future enhancements..

6 References

1. Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. *SemEval-2014 Task 4: Aspect Based Sentiment Analysis*. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 27–35, 2014.
2. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In Proceedings of NAACL-HLT, pages 4171–4186, 2019.
3. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv preprint arXiv:1907.11692, 2019.
4. Sepp Hochreiter and Jürgen Schmidhuber. *Long Short-Term Memory*. Neural Computation, 9(8):1735–1780, 1997.
5. Bing Liu. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, 2012.
6. Duyu Tang, Bing Qin, and Ting Liu. *Aspect Level Sentiment Classification with Deep Memory Network*. In Proceedings of EMNLP, pages 214–224, 2016.
7. Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. *Transformers: State-of-the-Art Natural Language Processing*. In Proceedings of EMNLP: System Demonstrations, pages 38–45, 2020.
8. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. *Attention Is All You Need*. In Advances in Neural Information Processing Systems (NeurIPS), 2017.