University of Bonn

CAISA LAB

Introduction to Natural Language Processing

(Winter Semester 2023/2024)

Default Project

Final Report of Team 2

Bites and Bytes: Advanced Techniques in Aspect-Based Sentiment Classification for Restaurants and Laptops

Group Members:

Nikita Bondarenko	3460192
Simon Welz	3233568
Jon Breid	3067704
Aditya Vijay Jogalekar	3189939

Lecturer: Prof. Dr. Lucie Flek
Course coordinator: Vahid Sadiri Javadi

CAISA Lab
https://caisa-lab.github.io/

UNIVERSITÄT BONN

February 11, 2024

Contents

1. Introduction

In an era increasingly dominated by digital platforms, online reviews have become a key driver of consumer choice. The ability to detect and understand the sentiments embedded in these reviews is essential for businesses seeking to align themselves with customer preferences. Aspect based sentiment analysis is an advanced facet of sentiment analysis. It focuses on identifying and interpreting the sentiment associated with specific aspects within textual data, providing a more refined understanding of customer opinions. Traditional sentiment analysis provides a broad overview of sentiment, but often misses the nuances in text. For example, a review that says ”The new laptop model is decent, but it's quite expensive” might receive a neutral overall sentiment score. While this approach is efficient for measuring overall sentiment, it overlooks the complexity and multi-layered sentiments in such statements.

In contrast, Aspect-Based Sentiment Analysis (ABSA) provides a more nuanced analysis. It would break down the laptop review above into sentiment about specific aspects: positive sentiment about the laptop's quality and negative sentiment about its price. This granular approach provides a more accurate picture of customer opinion on different facets of a product.

The primary goal of this project is to develop a sophisticated model capable of accurately identifying and classifying sentiments related to specific aspects of laptop and restaurant reviews, thereby addressing the inherent ambiguity and complexity of user-generated content.

The guiding research question is: ”How effectively can aspect-based sentiment analysis models classify sentiment in restaurant and laptop reviews, and what are the key factors influencing their accuracy?”

The significance of this research goes beyond the technical development of a sentiment analysis model. With the proliferation of online platforms, customer reviews have become a gold mine of data for businesses. By using ABSA, companies can gain a deeper understanding of customer needs and preferences, allowing them to tailor their products, services and marketing strategies more effectively.

A major challenge in sentiment analysis is the inherent ambiguity of the sentiments expressed in user-generated content. The nuances of language, such as irony, mixed emotions and varying degrees of sentiment intensity, add layers of complexity to accurate sentiment classification.

2. Literature Review:

Sentiment analysis has undergone a significant evolution, moving from simple sentiment classification to complex aspect-based approaches. The field's journey reflects advances in both linguistic analysis and computational methods, reshaping the way textual sentiment is understood and analysed.

Initially, sentiment analysis was primarily concerned with categorising entire texts into positive, negative or neutral sentiments. Fundamental studies by researchers such as Pang and Lee [1] were crucial in this phase, providing insights into the overall sentiment of texts ranging from product reviews to social media posts. This era laid the groundwork for understanding the broader emotional context of written language.

The advent of ABSA marked a significant shift in the field, focusing on sentiments associated with specific aspects within texts. The pioneering work of Liu and Zhang [2] introduced methods for extracting aspect terms and assessing the sentiments associated with them, providing a more detailed analysis that is crucial for understanding complex customer feedback.

As ABSA evolved, the integration of machine learning algorithms with linguistic rules began to gain prominence. The work of Kiritchenko et al. [3] exemplified the effectiveness of hybrid models in capturing complex linguistic features, such as sarcasm and implicit sentiment, and improving the precision and contextual accuracy of sentiment analysis.

The incorporation of deep learning models such as RNNs, LSTMs and CNNs represented a methodological leap in ABSA. The ability of these models to capture the sequential and hierarchical nature of language [4] significantly improved the contextual understanding of sentiment within texts.

The introduction of pre-trained language models, in particular BERT and its variants, has revolutionised ABSA. These models, trained on large corpora, excel at understanding complex language contexts [5], making them highly effective at identifying sentiment contexts around aspect terms.

The inclusion of attention mechanisms in ABSA models was a major advance, allowing these models to focus on text segments that are critical for aspect sentiment classification [6]. This approach is critical for isolating sentiment associated with specific aspects, leading to more accurate predictions.

Finally, the application of transfer learning and the fine-tuning of pre-trained models for specific ABSA tasks has become a standard approach [7]. This adaptation enhances the general language understanding capabilities of the models, tailoring them to the specific requirements of ABSA tasks and often leading to significant improvements in accuracy and efficiency.

3. Data Exploration and Preprocessing:

The following section details the dataset used for training and evaluation, and the preprocessing of the data to prepare it for training and testing.

3..1 Dataset

The datasets used for this research come from the SemEval-2014 Task 4 challenge[8] and the Semeval-2016 task 5 challenge[9], which focus on aspect-based sentiment analysis (ABSA). These datasets cover two main domains: restaurant reviews and laptop reviews.

1. **Restaurant Reviews:**
    - The restaurant review dataset was partially derived from the dataset created by Ganu et al. (2009), which provided a basic structure for aspect categories and sentiment annotations
    - Additional reviews were collected from popular online platforms where customers commonly leave feedback about their dining experiences.
2. **Laptop Reviews**:
    - The dataset of laptop reviews was compiled from various online retail platforms. These platforms are hotspots for consumers to express their opinions on technology products, providing a diverse range of viewpoints and aspects mentioned.

3..1.1 Data Collection Methodology

The data collection process for our Aspect Based Sentiment Analysis study was carefully designed to capture a wide range of sentiments, aspects and customer viewpoints from online reviews. This process is divided into two main phases: Review Extraction and Aspect and Sentiment Annotation.

In the first phase reviews are systematically extracted form different source platforms. These platforms are selected to ensure a high diversity in the collected data. A wide range of customer opinions, covering different sentiments and aspects mentioned in the reviews should be covered. Manual and automated methods to extract reviews are used. Web scaping tools and APIs allow to efficiently collect large amounts of data. Reviws that may have been missed by the tools are manually extracted to ensure thorough coverage. Extracted reviews underwent a pre-processing step, essential to standardise the format of the reviews and to make them suitable for analysis. The data is cleaned by remiving irrelevant information such as advertisements, non-textual elements.

In the second phase the collected data is carefully annotated by expert annotators. Annotators are specifically trained in aspect-based sentiment analysis. In each review aspect terms together with their corresponding sentimens are identified. To ensure quality and consistency of the annotations, the process was governed by strict guidlines helping to maintain a high labelling accuracy.

3..1.2 General Dataset Characteristics

Dataset	**Training Sentences**	**Testing Sentences**	**Total Sentences**
Restaurant Reviews	3566	1117	4683
Laptop Reviews	2252	635	2887

Table 1: Dataset Size and Composition

Size and Composition: Each dataset, as detailed in Table 1, consists of sentences that have been meticulously tagged with aspect terms and their corresponding sentiment labels. These labels cover a spectrum of sentiments - positive, negative, neutral - providing a comprehensive perspective on customer feedback.

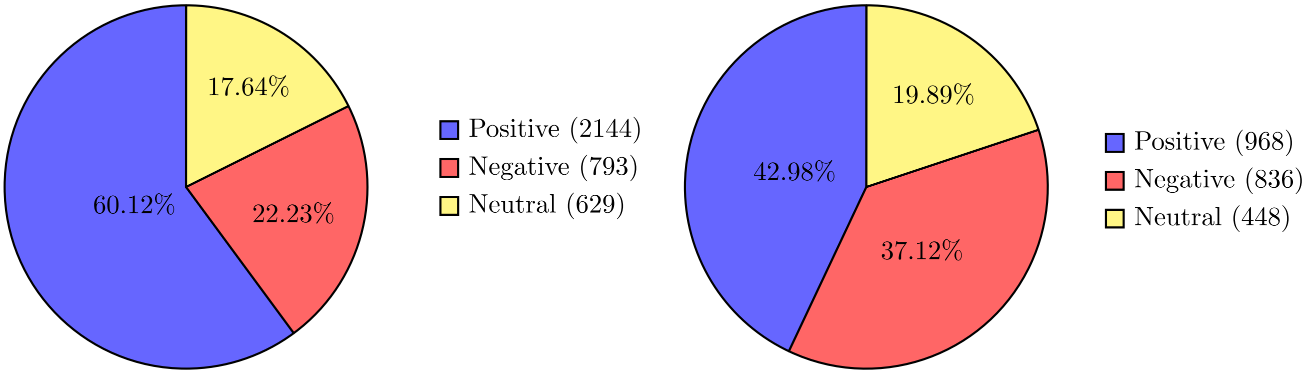

Figure 1: Sentiment distribution in Restaurant and Laptop Reviews.

Sentiment Distribution: The sentiment distributions shown in Figure 1 in the restaurant and laptop reviews datasets have different characteristics, reflecting different consumer behaviour in these sectors.

The dataset of restaurant reviews is heavily skewed towards positive sentiments. This trend suggests that customers are more likely to share positive experiences in the context of dining out. The prevalence of positive reviews in this dataset underlines the tendency of diners to post positive feedback about their culinary experiences.

In contrast, the laptop reviews dataset shows a more balanced distribution between positive and negative sentiments, with a comparatively lower occurrence of neutral sentiments. This balance suggests that technology customers are equally likely to express both satisfaction and dissatisfaction with their purchases, providing a more balanced representation of opinions.

This difference in sentiment distribution between the two datasets presents unique challenges and considerations for our aspect-based sentiment analysis model. It highlights the importance of tailoring the model to effectively interpret and analyse the specific sentiment dynamics of each domain.

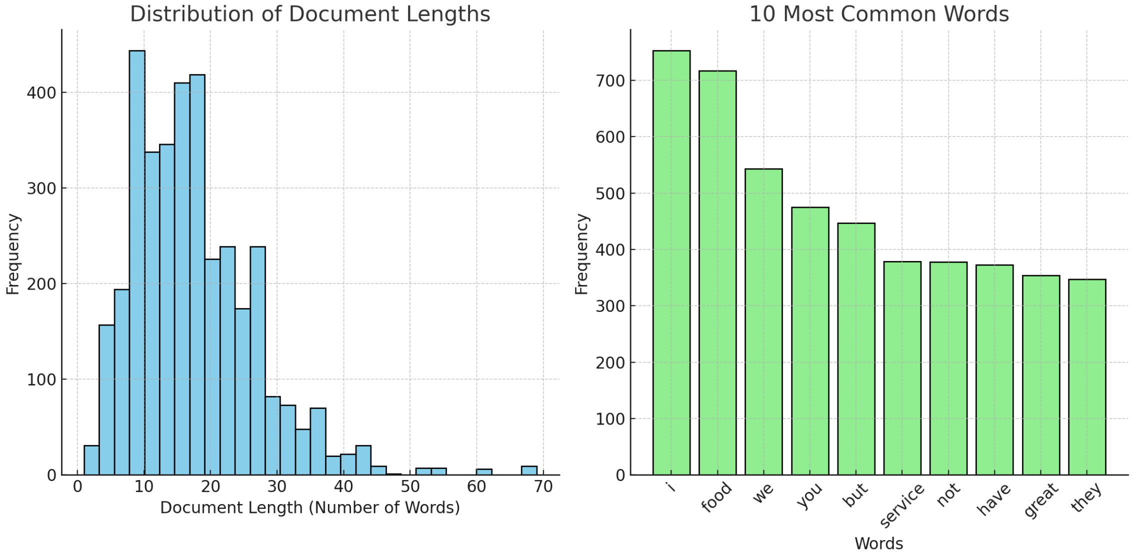

Figure 2: Statistics of the Restaurant Dataset.

Statistics of the Datasets: Most of the documents consist of short to medium length sentences, suggesting that reviewers are often expressing focused opinions about specific aspects of their dining experience. The frequent use of personal pronouns such as 'I', 'we' and 'you' underlines the personal nature of these reviews, framing sentiments within personal experiences. Words such as 'great' and 'not' serve as direct sentiment indicators, essential for training models to recognise positive and negative sentiments. It's important to focus on accurately interpreting sentiments that are explicitly expressed within the limited textual context of short, personal narratives.

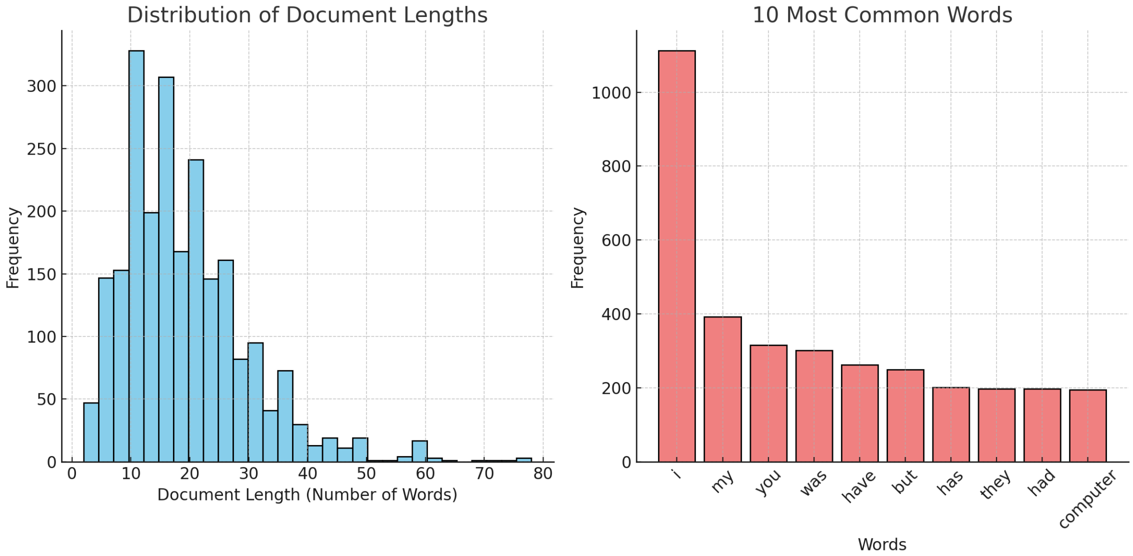

Figure 3: Statistics of the Laptop Dataset.

In laptop reviews, the frequency of personal pronouns and ownership terms ('my') highlights the importance of personal experience and individual evaluation of laptop features. Sentiments such as 'not', 'very' and 'but' suggest that reviewers often express clear sentiments about specific aspects of laptops, which could include performance, design, usability or customer service. The concise nature of the reviews means that each sentence is likely to contain a focused opinion about a specific feature, making the task of sentiment analysis easier, but requiring the model to accurately understand and attribute sentiment within a limited context.

3..2 Data Preprocessing

The pre-processing of the dataset is a multifaceted process that is crucial in preparing the textual data for efficient and effective analysis by the DeBERTa[10] model. Firstly, **data cleaning** is carried out to ensure the purity and uniformity of the dataset. This step involves removing numerical digits, converting all text to lower case and removing all extraneous symbols, thereby standardising the textual content.

Then, in order to increase the model's focus on relevant aspects, each sentence is prefixed with its corresponding **aspect term**, followed by the separator **'[SEP]'**. This clear method of separation is used to signal to the DeBERTa model to draw attention specifically to the aspect terms in relation to the rest of the sentence. Such contextual highlighting is crucial for aspect-based sentiment analysis, as it guides the model to focus on the most relevant parts of the text.

The final pre-processing step is **tokenisation** using the **DeBERTa Tokenizer**. This tokenizer adapts the text to the requirements of the DeBERTa model. It splits the text into tokens and transforms it into a format that the model can process. The tokeniser includes special tokens such as [CLS] and [SEP], which are critical for the model to understand the start, end and separation of contextual elements within the text. It also generates an attention mask, which is crucial for controlling the model's attention mechanism. This mask enables the DeBERTa model to focus on the meaningful content within each tokenised sequence and ignore the padded areas, thereby increasing the efficiency and accuracy of sentiment analysis. These tokenisation steps ensure that the text data is optimally formatted and prepared for processing by the DeBERTa model, enabling more nuanced and accurate aspect-based sentiment analysis.

4. Methodology

The following section presents the methods used to pre-process and augment the dataset and the model architecture used.

4..1 Data Augmentation

To address the imbalance in sentiment distribution within our dataset, we used data augmentation techniques. This process is crucial to increase the diversity and volume of the training data, especially for underrepresented sentiment classes.

We used the **'nlpaug'** library[11], a versatile tool for natural language processing data augmentation. Our augmentation strategy included

1. **Synonym Replacement:** This involved replacing words in sentences with their synonyms, preserving the original meaning while diversifying the linguistic expression.

2. **Word Swapping:** We randomly swapped words within a sentence. This technique added variability to the dataset without changing the core meanings of the sentiments.

3. **Backtranslation:** Sentences were translated into another language and then back into the original language. This approach often results in paraphrased content, which adds to the richness of the dataset.

These augmentation methods were instrumental in creating a more balanced dataset, providing our model with a wider range of expressions to learn from, and thus improving its ability to accurately classify sentiments.

4..2 DeBERTa Model Architecture

For our aspect-based sentiment analysis, we chose the pre-trained DeBERTa (Decoding enhanced BERT with disentangled attention) model. This model, already trained on large corpora of diverse text data, provided a solid foundation for our ABSA tasks. By fine-tuning this pre-trained model on our specific dataset, we were able to leverage its inherent understanding of linguistic nuances while tailoring it to the unique requirements of sentiment analysis of restaurant and laptop reviews. DeBERTa is known for its effectiveness in NLP tasks due to its innovative architecture.
The DeBERTa model introduces several important innovations:

1. **Disentangled Attention Mechanism:** Unlike traditional BERT, DeBERTa separates content and positional attention, allowing for a more accurate understanding of word context and relationships.

2. **Enhanced Mask Decoder:** Uses an enhanced mask decoder to better predict masked tokens, which is critical for understanding the nuances of sentiment analysis.

3. **Integration with Aspect Information:** In our implementation, the model has been adapted to use aspect-specific information, making it particularly suitable for ABSA tasks.

DeBERTa's advanced attention mechanism and superior contextual understanding make it uniquely suited to disentangling the complex sentiment expressions in our dataset. Its ability to process and analyse aspect-annotated sentences enables nuanced sentiment classification, meeting the complex requirements of aspect-based sentiment analysis.

4..2.1 Comparison with Other Models

In our quest to find the most effective model for aspect-based sentiment analysis, we also experimented with several other models, including BERT[12], RoBERTa[13], SVM[14] and Logistic Regression. Each of these models brought unique strengths and perspectives to the task. However, after rigorous testing and evaluation, it was the DeBERTa model that consistently outperformed the others in terms of accuracy and efficiency in handling the complexity of our dataset.

4..3 Discussion on Libraries and Frameworks

4..3.1 PyTorch for Training

PyTorch[15] is a cornerstone in the field of deep learning due to its flexibility and dynamic computational graph. In our project:

- **Use:** PyTorch was the primary framework used to train our DeBERTa model. Its dynamic nature allowed for more intuitive coding of complex models and experiments.
- **Benefits:** With PyTorch, we benefited from an efficient workflow and accelerated computations thanks to its GPU support. This made iterating over different model architectures and hyperparameters much more efficient.

4..3.2 Scikit-learn for evaluation

Scikit-learn (sklearn)[16], a versatile Python library, was instrumental in the model evaluation phase.

- **Role:** Sklearn provided us with a variety of performance metrics and tools to evaluate our trained models. These included confusion matrices, accuracy, precision, recall and F1 scores.
- **Benefits:** The comprehensive suite of tools and straightforward API made it easy to implement and compare the performance of different models.

4..3.3 nlpaug for data augmentation

As mentioned above, nlpaug played a key role in augmenting our dataset to address class imbalances and improve model training.

- **Use:** It allowed us to apply various augmentation techniques, such as synonym replacement and back translation, which enriched our dataset and improved the robustness of the model.

4..3.4 Transformers library for pre-trained models

Finally, Hugging Face's Transformers library[17] was crucial for accessing state-of-the-art pre-trained models.

- **Integration:** This library allowed us to easily implement and fine-tune models such as DeBERTa, BERT and RoBERTa[13], providing a high starting point for our training process.
- **Versatility:** It offered a wide range of pre-trained models and supported seamless integration with PyTorch, simplifying our experimentation and model development process.

5. Experimental Setup

The following section presents the setup and parameters for the models and training. Our approach started with conventional models and gradually progressed to state-of-the-art pre-trained models, with an emphasis on fine-tuning the DeBERTa model to suit our specific task.

We adhered to the train-test split ratio defined in the accompanying paper of the original dataset, thus maintaining the integrity of the comparative analysis between different models [8].

5..1 Baseline Models and Perfomance

Our research began with the implementation of Support Vector Machines (SVM) and Logistic Regression models, coupled with TF-IDF vectorisation, to establish a performance baseline. These models were chosen because of their proven track record in text classification tasks, due to their ability to handle high dimensional data and interpret linear relationships within text. $Achievinganaccuracyof71\%ontherestaurantdatasetand65\%onthelaptopwiththesebaselinemodelsp$

5..2 Adoption of Pretrained Models

The introduction of BERT (Bidirectional Encoder Representations from Transformers) marked our first foray into pre-trained models. BERT's deep bidirectional training and attention mechanism allows it to capture context more effectively than traditional models, making it particularly suited to sentiment analysis. Although BERT was a significant improvement on the baseline models, we believed that further improvements were possible through the use of more advanced architectures.

The introduction of DeBERTa (Decoding-enhanced BERT with Disentangled Attention) into our experimental setup was a pivotal moment. DeBERTa improves on BERT and its variants by incorporating a disentangled attention mechanism and improving the representation of words through the use of absolute position embeddings. To adapt DeBERTa to our task of classifying sentiments into three distinct categories - positive, neutral and negative - we focused on fine-tuning the top layer of the model. This adjustment allowed the model to effectively discriminate between the subtle tonal variations inherent in the sentiment labels of our dataset.

5..3 Hyperparameter Optimisation and Learning Rate Scheduling

The pursuit of optimal performance in our sentiment classification models required a rigorous hyperparameter optimisation process. Recognising the complexity and high dimensionality of the hyperparameter space, we deployed Optuna[18], a state-of-the-art hyperparameter optimization framework, to streamline and enhance this process.

Optuna is designed to automate the search for the best hyperparameters by using efficient sampling methods, such as the Tree-structured Parzen Estimator, and a novel pruning technique that discards unpromising trials early in the training process. This approach significantly reduces the computational resources and time required for optimisation, while ensuring a thorough exploration of the hyperparameter space.

In our experiments, we used Optuna to optimise the learning rate and the number of epochs. Using an objective function to guide the training and evaluation phases of the model, we sought to improve tracking accuracy. However, after 15 attempts by Optuna to adjust these hyperparameters, our initial configuration proved to be the best choice.

- **Learning Rate:** $2 \cdot 10^{-5}$
- **Number of Epochs:** 15

A **batch size** of 16 is chosen to balance the trade-off between computational efficiency and stability of the gradient updates.

Our approach to learning rate scheduling departed from automated optimisation methods. Instead, we conducted manual experiments with different learning rate schedulers, including linear warm-up with linear scheduling, exponential, and reduce on plateau, to empirically determine their impact on model performance. This hands-on methodology allowed us to directly assess the effectiveness of each scheduler, leading to the identification of the most appropriate scheduler for our task. **The best scheduler is to reduce on a plateau or exponentially**

5..4 Data Augmentation Techniques

In our efforts to mitigate the effects of class imbalance on model performance, we explored various data augmentation strategies to expand our training dataset. Initially, we experimented with techniques such as synonym substitution and word swapping, but these approaches failed to produce meaningful improvements in model results. We then turned to back translation. Despite its potential to introduce syntactic diversity into the dataset, back-translation did not provide the expected improvement in model performance.

6. Evaluation and Results

6..1 Evaluation Metrics

The performance of our aspect-based sentiment classification model was evaluated using a comprehensive set of evaluation metrics, including accuracy, precision, recall and the F1 score. These metrics - accuracy, precision, recall and F1 score - provide a comprehensive evaluation of the model's performance, capturing not only its overall accuracy, but also its detailed ability to discriminate between sentiment classes with discriminating precision and recall. The formulas for these metrics are as follows:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{3}$$

$$F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precisionrecall}} \tag{4}$$

In these equations, TP (true positives) represents the number of correctly identified positive cases, TN (true negatives) the number of correctly identified negative cases, FP (False Positives) the number of negative cases mislabelled as positive, and FN (False Negatives) the number of positive cases mislabelled as negative. This comprehensive set of metrics not only quantifies the overall accuracy of the model, but also breaks down its performance in identifying each sentiment class, providing a balanced and detailed evaluation.

6..2 Results

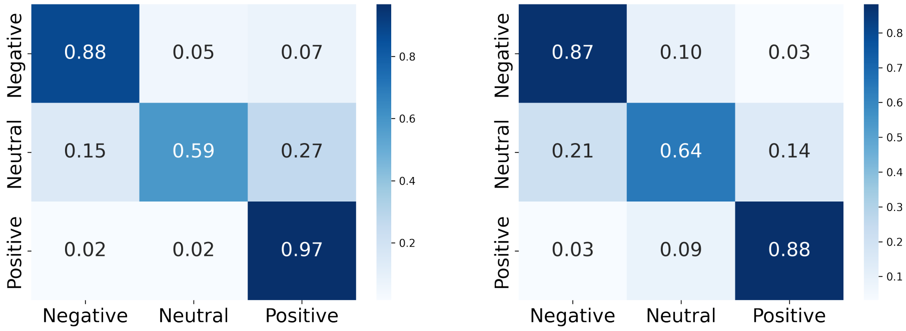

Figure 4: Normalized Confusion Matrix of **Restaurant (Left)** and **Laptop (Right)** Dataset

The quantitative results for both datasets reveal significant findings. The model is particularly good at identifying positive reviews in this domain, as evidenced by the high true positive rate for accurately identifying positive sentiments in the restaurant dataset, as shown in the normalised confusion matrix for the restaurant review dataset in Figure 1. Although the true positive rate of 0.59 indicates a low level of confusion with positive cases, the model struggles with neutral sentiments. The true positive rate for negative sentiment is 0.88, indicating high accuracy.

In contrast, the laptop review dataset shows a more balanced performance across sentiments. In particular, negative and positive sentiments are equally balanced. Neutral sentiment is, as in the restaurant dataset, lower than the true positive rate for positive and negative, but the gap is smaller than in the restaurant domain.

These results suggest that while the model is generally effective in identifying sentiment across the datasets, it is more challenged by neutral sentiment, particularly in the Restaurant dataset. The strength in detecting positive and negative sentiment in both domains suggests domain-specific performance characteristics.

	positive	neutral	negative
precision	0.742	0.899	0.842
recall	0.968	0.587	0.878
F1	0.840	0.710	0.859

Table 2: Precision and recall achieved on the **Restaurant** dataset with DeBERTa.

	positive	neutral	negative
precision	0.835	0.781	0.781
recall	0.882	0.645	0.873
F1	0.859	0.707	0.824

Table 3: Precision and recall achieved on the **Laptop** dataset with DeBERTa.

Input instance	Prediction	Truth
cake [SEP] the icing made this cake it was fluffy not ultra sweet creamy and light	*negative*	*positive*
food [SEP] since i cook for a living im very fussy about the food i eat in restaurants	*negative*	*neutral*
quality of the meat [SEP] the quality of the meat was on par with your local grocery store	*positive*	*negative*
calamari [SEP] for an appetizer their calamari is a winner	*positive*	*positive*
courses [SEP] it took about hours to be served our courses	*neutral*	*neutral*
slices [SEP] finally let into the store at a time to buy expensive slices from a harried staff	*negative*	*negative*

Figure 5: Samples of instances where the sentiment is predicted with DeBERTa of the restaurant dataset.

Input instance	Prediction	Truth
backlit keys [SEP] the only thing I miss is that my old alienware laptop had backlit keys.	*negative*	*positive*
HDMI port [SEP] no HDMI port	*negative*	*neutral*
i5 [SEP] not as fast as I would have expected for an i5	*positive*	*negative*
Set up [SEP] set up was a breeze.	*positive*	*positive*
Mac OS X 10.8 'Mountain Lion' [SEP] I then upgraded to Mac OS X Mountain Lion.	*neutral*	*neutral*
keyboard [SEP] it feels cheap, the keyboard is not very sensitive.	*negative*	*negative*

Figure 6: Samples of instances where the sentiment is predicted with DeBERTa of the laptop dataset.

Figures 5 and 6 present selected sentences from the restaurant and laptop datasets respectively, showing the predictions made by the DeBERTa model alongside the actual sentiment labels. Each table is organised to highlight examples of both successful predictions and misclassifications across positive, neutral and negative sentiments. The examples are chosen to illustrate the performance of the model in different contexts, with the input instance being a key aspect followed by the review text expressing the sentiment towards that aspect.

6..2.1 Comparison with Other Models

Model	Restaurant Accuracy (%)	Laptop Accuracy (%)
DeBERTa	88.5	81.73
RoBERTa	86.43	80.72
Best Model of [8]	80.95	70.48
Logistic Regression	71.00	65.00
Baseline of [8]	64.28	51.37
LSA-X-DeBERTa[19]	90.33	86.21

Table 4: Accuracy achived with different models.

DeBERTa proves to be a robust performer, achieving accuracies of 88.5% for restaurant reviews and 81.73% for laptop reviews, due to its advanced mechanism for fine-tuning contextual understanding of words. RoBERTa has lower accuracies of 86.43 % and 80.72 % respectively, reflecting its earlier generation technology. The best model from [8], which is the model from DCU uses an SVM classifier with features such as n-grams and parse trees along with publicly available sentiment lexicons, shows decent performance but doesn't reach the heights of DeBERTa. Logistic regression lags behind with 71% for Restaurant and 65% for Laptop, demonstrating the challenge that simpler models face in capturing the nuances of natural language. The baseline model of [8] assigns a sentiment polarity to an aspect term based on its frequency in similar training sentences, as measured by the Dice coefficient. If the aspect term is new, it assigns the most common polarity from the training set. This simple approach based on statistical frequency and word overlap is the baseline with an accuracy of 64.28% for the restaurant and 51.37% for the laptop dataset.

The new LSA-X-DeBERTa [19] model outperforms DeBERTa by utilising a Local Sentiment Aggregation (LSA) mechanism that enhances aspect sentiment coherence through a differentiated weighted sentiment aggregation approach. This sophisticated method allows the model to strategically weight adjacent sentiment features, resulting in a remarkable jump in accuracy to 90.33% and 86.21% for the restaurant and laptop datasets, respectively.

7. Analysis and Discussion

Models have been successfully trained using different architectures. The DeBERTa model learned the sentiments in the sentences with respect to the given aspects. The DeBERTa model achieved results reasonably close to those of LSA-X-DeBERTa, which currently has one of the best results on the datasets used. This shows that DeBERTa models can be successfully fine-tuned on sentiment classification datasets from specific domains, such as restaurant or laptop reviews.

Looking more closely at the model's results, it is easy to see that the model is more adept at correctly predicting positive and negative sentiment than neutral sentiment. This observation may be attributed to several factors. A possible reason for this could be that the model's training data contains fewer instances of neutral sentiments, resulting in less learning about this particular category. A smaller number of examples may result in the model being less adept at recognising the nuanced language that often characterises neutral sentiments. Moreover, even the annotators - a linguistic expert and a graduate student - encountered difficulties in consistently categorizing sentiments. This indicates an inherent complexity in classifying sentiments that are neither clearly positive nor negative, as the neutrality can be subjective and context-dependent. The examples in Figure 6 and Figure 5 suggest that neutral sentiments often contain elements of both positive and negative sentiments, or lack strong sentiment indicators. For example, phrases such as "not as fast as I would have expected" or "feals cheap" might hover between a negative and a neutral sentiment, depending on the context and the expectation of the reader. Sentences that express sentiments in a more comparative or context-dependent manner, such as 'the quality of the meat was on a par with your local grocery store', can be particularly challenging. The sentiment here could be interpreted differently based on individual expectations of the quality of the grocery store, leading to a positive, negative or neutral classification. In addition, neutral sentences often contain subtleties, implying a feeling without explicit affective words. For example, "set up was a breeze" is positive because it implies that it was easy, but it may not be as straightforwardly positive as other sentiments expressed in more emphatic language. Finally, neutral sentiments are frequently at the line between positive and negative, making it difficult for the model to assign a definitive sentiment. This is compounded by linguistic nuances and the use of mitigating phrases that can temper or intensify the sentiment expressed, which the model may not capture effectively.

When analysing the difference in performance between the restaurant and laptop datasets, it is evident that the restaurant dataset yields better results. There are several possible explanations for this. A greater number of training instances generally results in a model acquiring a richer and more varied representation of language, which is the main advantage of the restaurant dataset. Furthermore, the language used in restaurant reviews is often less technical and more general than the jargon and technical terminology used in laptop reviews. The pretraining corpus for models such as DeBERTa is likely to have a higher representation of general-domain text, such as restaurant reviews, rather than specialised technical content. In addition, restaurant reviews may also exhibit more explicit sentiment expressions, making it easier for models to learn and predict sentiment. The inherently different structures and vocabularies of these domains suggest that domain-specific adaptations in pretraining or fine-tuning might be essential to improve performance on more technical datasets. Another contributing factor could be the subjectivity in technical reviews, where opinions may be less about the inherent qualities of the product and more about the user's expectations or specific use-cases, leading to more subtle sentiment expressions that are more challenging for models to understand.
With these challenges in mind, our results can be applied to specific real-world scenarios, such as the following examples.

Customer Feedback analysis: Companies in the hospitality and electronics sectors can use such models to analyse customer feedback from a variety of sources, such as online reviews, surveys or social media posts. This analysis can help identify areas of strength and weakness, allowing companies to make informed decisions about product improvements, service enhancements or addressing specific customer concerns.

Product and Service Development: By understanding the sentiments expressed in reviews, companies can gain insight into which features or aspects of their products and services are most valued by customers and which need improvement. This information can guide product development, service offerings and marketing strategies to better meet customer needs and preferences.

Customer Support: Automated sentiment analysis can improve customer support by identifying and prioritising reviews or feedback that express negative sentiment, allowing companies to respond more quickly to dissatisfied customers. Positive feedback can also be used for marketing purposes or to recognise and reward employees who contribute to customer satisfaction.

Personalisation and recommendation systems: E-commerce platforms and content providers can use sentiment analysis to better understand user preferences and tailor recommendations or advertising to individual users based on the sentiments expressed in their reviews or feedback.

8. Conclusion

In conclusion, this project has delved into the nuanced field of aspect-based sentiment analysis with a focus on laptop and restaurant reviews, providing valuable insights into the challenges and achievements of sentiment classification models. Our results suggest that the models are more adept at discriminating between positive and negative sentiment than neutral sentiment. This disparity can be attributed to several factors, including the underrepresentation of neutral sentiment in the training data and the inherent difficulty in categorising sentiment that is not clearly positive or negative due to its subtlety and context dependency. The analysis highlighted the challenge posed by neutral sentiments, which often contain elements of both positive and negative sentiments, or lack strong sentiment indicators, making them more difficult for models to accurately classify.

Furthermore, the comparative analysis between the restaurant and laptop datasets highlighted the different performance of the models in different domains. The restaurant dataset, with its larger volume of training instances and less technical language, enabled the models to perform better than the more specialised and jargon-heavy laptop reviews. This suggests that domain-specific language and structure have a significant impact on model performance, indicating the need for domain-specific adaptations in pre-training or fine-tuning processes.

In response to the guiding research question, the project demonstrated that aspect-based sentiment analysis models can effectively classify sentiments in restaurant and laptop reviews, albeit with varying degrees of accuracy influenced by factors such as dataset size, domain specificity, and the inherent complexity of neutral sentiments. While the models show commendable accuracy in identifying clear positive and negative sentiments, their performance on neutral sentiments and in more technical domains such as laptop reviews presents areas for future improvement. Thus, while the research question is largely addressed, the nuanced challenges identified invite further exploration and refinement of sentiment analysis models to improve their adaptability and accuracy across domains.

9. Future Work

The exploration of aspect-based sentiment analysis has laid a solid foundation and revealed several directions for future research and improvement. In order to build on the current findings and address the identified limitations, the following areas offer promising opportunities for further investigation:

Enhanced representation of neutral sentiments: Given the challenges of accurately classifying neutral sentiments, future work could focus on enriching the training datasets with a wider range of neutral examples. This could involve developing more sophisticated data collection and annotation strategies to capture the nuanced language that characterises neutral sentiments, thereby improving the models' ability to accurately detect and classify these sentiments.

Domain-specific model tuning: The difference in model performance between restaurant and laptop reviews highlights the impact of domain specificity on sentiment analysis. Future research could explore domain-adaptive pre-training or fine-tuning techniques that tailor models more closely to the linguistic and structural characteristics of specific domains, particularly those involving technical language or jargon.

Advanced Language Understanding: The inherent complexity of sentiment analysis, particularly in detecting subtle or contextual sentiments, requires models with deeper language understanding capabilities. Investigating advanced NLP techniques and architectures that better capture the semantic subtleties and contextual cues within text could significantly improve sentiment classification, especially for sentences that cross the boundary between sentiment categories.

Cross-domain sentiment analysis: Investigating the transferability of models trained on one domain (e.g. restaurant reviews) to another domain (e.g. laptop reviews) could provide insights into the generalisability of aspect-based sentiment analysis models. This line of research could include cross-domain adaptation strategies and exploring the ability of models to use learned sentiment indicators in different contexts.

References

[1] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[2] Bing Liu. *Sentiment Analysis and Opinion Mining*, volume 5 of *Synthesis Lectures on Human Language Technologies*. Morgan Claypool Publishers, 2012.

[3] Svetlana Kiritchenko, Xiaodan Zhu, and Saif M. Mohammad. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762, 2014.

[4] Yunfei Long, Lu Qin, Rong Xiang, Minglei Li, and Chu-Ren Huang. Deep learning for sentiment analysis : A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(4):e1353, 2020.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, 2019.

[6] Zhengyang Xu, Chengyang Huang, and Pengfei Zhu. Gated convolutional neural network for aspect-based sentiment analysis. In *Proceedings of the 2019 International Conference on Asian Language Processing*, pages 77–82, 2019.

[7] Sebastian Ruder. An overview of transfer learning in natural language processing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 15–18, 2019.

[8] Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. SemEval-2014 task 4: Aspect based sentiment analysis. In Preslav Nakov and Torsten Zesch, editors, *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland, August 2014. Association for Computational Linguistics.

[9] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. SemEval-2016 task 5: Aspect based sentiment analysis. In Steven Bethard, Marine Carpuat, Daniel Cer, David Jurgens, Preslav Nakov, and Torsten Zesch, editors, *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California, June 2016. Association for Computational Linguistics.

[10] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention, 2021.

[11] Edward Ma. Nlp augmentation. https://github.com/makcedward/nlpaug, 2019.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[13] 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, 2019.

[14] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press, 1999.

[15] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.

[16] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[17] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020.

[18] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework, 2019.

[19] Heng Yang and Ke Li. Improving implicit sentiment learning via local sentiment aggregation, 2023.