A MULTI-SCALE APPROACH TO FINE-GRAINED SENTIMENT ANALYSIS USING DEBERTAV3

Lana Do and Tehmina Amjad

Khoury College of Computer Sciences, Northeastern University,
Silicon Valley Campus, California, USA

ABSTRACT

Fine-grained sentiment analysis captures subtle emotional tones in text, offering insights beyond positive and negative classifications. It helps users make informed decisions by revealing nuanced opinions and sentiment intensities in textual data. This paper introduces Sentiment-Enhanced Fine-Tuned DeBERTaV3 (FiTSent DeBERTaV3), a classification model designed for both sentence-level and document-level sentiment analysis. Built upon the DeBERTaV3 architecture, our model incorporates tailored fine-tuning strategies to address the unique characteristics of each dataset. On the Stanford Sentiment Treebank (SST5), fine-tuning addresses shorter, nuanced texts, while for Yelp Reviews, strategies are adapted for longer, narrative-style reviews. Additionally, the use of attention pooling allows the model to prioritize sentimentcritical tokens, enhancing its ability to capture subtle sentiment distinctions. FiTSent DeBERTaV3 achieved competitive performance, outperforming baselines on both tasks. These results highlight the effectiveness of our approach and its versatility in handling datasets with varying lengths and complexities, which have not been jointly evaluated before.

KEYWORDS

Fine-Grained Sentiment Analysis, DeBERTaV3, Dataset-specific Fine-Tuning, Sentiment-Focused Attention Pooling, Sentence-level analysis & Document-level analysis

1. INTRODUCTION

As the internet continues to grow, people tend to share their thoughts, opinions, and experiences about several entities that they buy, watch, or know about. Not only do people like to share their reviews, but they also like to know the opinion of other people who review similar products. Due to the diverse nature and complexity of the reviews, people sometimes find it difficult to get an insight on sentiments and trends in a wider community [1]. Sentiment analysis is a natural language processing (NLP) technique that aims to identify the emotional tone in the body of text automatically and classify it as positive, negative or neutral. With the rapid growth of social media and review platforms, sentiment analysis plays a vital role in analyzing user feedback and extracting insights, as discussed in prior work focusing on user ratings across domains [2]. Most sentiment classifiers are binary (positive or negative) or coarse-grained with a neutral class [3]. While useful in large-scale applications, these approaches often fail to capture the full spectrum of emotions and nuances in human expression [4]. For example, dual-polarity sentiments like "The beautiful, unusual music is this film's chief draw, but its dreaminess may lull you to sleep" can confuse binary classifiers, leading to the loss of valuable information and incorrect predictions. Fine-grained sentiment analysis addresses these limitations by categorizing

David C. Wyld et al. (Eds): SOFTPA, NATAP, ARIN, CSITA, NMCO, CONEDU, ISPR – 2025
pp. 11-29, 2025. CS & IT - CSCP 2025 DOI: 10.5121/csit.2025.150602

sentiments into more specific classes, such as very negative, negative, neutral, positive, and very positive, to better capture subtleties in text.

Despite its importance, fine-grained sentiment analysis remains a challenge. Dividing sentiments into multiple classes often leads to misclassification of intensity due to the inherent complexity of human language and its subtle emotional cues. Pretrained models like Bidirectional Encoder Representation from Transformers (BERT) have achieved remarkable success in NLP tasks by leveraging bidirectional architectures to capture contextual meaning. Improved variants of BERT, such as Robustly Optimized BERT Pretraining Approach (RoBERTa), have shown promising results in sentiment analysis by demonstrating more robust performance as compared to BERT on datasets like SST-5 [5].

A recent transformer-based model that builds upon the strength of BERT and RoBERTa - Decoding-enhanced BERT with disentangled attention (DeBERTa), employs disentangled attention, an enhanced mask decoder, and virtual adversarial training to improve pretraining efficiency and has shown to boost performance on Natural Language Understanding (NLU) and Natural Language Generation (NLG) task [6]. DeBERTaV3 further enhances the original DeBERTa model with replaced token detection and gradient disentangled embeddings sharing, improving efficiency and representation quality [7]. In this study, we proposed fine-tuning DeBERTaV3 and tested on two datasets: SST-5, a benchmark for sentiment classification, and the Fine-Grained Yelp Reviews, to evaluate its performance across different text styles. The results show that fine-tuned DeBERTaV3 showed improvement over other BERT variants in this task.

2. RELATED WORKS

The proliferation of the internet and social media platforms has significantly increased the sharing of opinions, reviews, and sentiments online, making data more abundant and accessible while driving interest in sentiment analysis research. For example, from 2008 to 2022, publications mentioning “sentiment analysis in social networks” grew at an annual rate of 34%, highlighting its rapid growth as a research area [8]. For sentiment analysis, research has progressed from traditional machine learning and deep learning techniques to advanced transformer-based architectures, which have set a new benchmark for fine-grained sentiment analysis.

2.1. Machine Learning and Deep Learning Approaches

Traditional machine learning models such as Support Vector Machines (SVM) and Naive Bayes have been widely applied to sentiment classification tasks. SVM achieves 99.5% accuracy on a dataset of Tweets about homosexual topics [9], and Naive Bayes reach 85% accuracy on the Sentiment140 dataset [10]. These results are impressive but only tested on binary or three-class sentiment dataset and have not been tested on more nuanced, multi-class sentiment tasks.

Deep learning approaches marked a significant improvement in sentiment analysis. In these methods, input data is first encoded using pretrained embeddings such as GloVe and word2vec, which then are fed into deep learning models such as recurrent neural networks (RNNs), long short-term memory networks (LSTM), or gated recurrent units (GRUs) for representation learning and classification [4]. Convolutional Neural Networks (CNNs), initially developed for image classification, achieves an impressive accuracy of 99.33% on IMDb binary sentiment classification [11] but only 53.4% on the more complex SST-5 dataset [12].

2.2. Hierarchical Attention Networks

Hierarchical Attention Networks (HAN) were introduced to address the challenges posed by multi-sentence, document-level sentiment analysis. Using word-level and sentence-level attention mechanisms, HAN effectively captures dependencies within and across sentences. It achieves impressive results on two five-class datasets: 71% accuracy on Yelp-2015 reviews and 63.6% accuracy on Amazon product reviews [13]. Building on this research, a hybrid model of Hierarchical Attention Networks and Neural Networks improves the accuracy of the fine-grained Yelp-2015 reviews dataset to 73.8% [14].

2.3. Transformer-Based Models

Transformer-based models such as BERT have set a new benchmark in NLP, excelling at tasks that require context-aware comprehension. For document-level sentiment analysis in datasets, Big Bird, an extension of the transformer architecture, employs a sparse attention mechanism to effectively capture long-range dependencies while maintaining computational performance. It achieves a good result of 72.15% accuracy on the Yelp Reviews dataset, demonstrating its ability to handle extended text inputs [15]. Another approach, BERT incorporated with Intermediate Task Pre-Training (ITPT) and Fine-Tuning (FiT), adapted the model to domain-specific tasks, achieving an accuracy of 70.58% on Yelp Reviews [16]. These approaches highlight the potential for further improvements in fine-tuning techniques and exploration of other alternative transformer architectures.

For shorter sentiment reviews, such as those in the SST-5 dataset, fine-tuned transformer- based models have shown promising results. Fine-tuned BERTLARGE achieves a best test accuracy of 55.5%, while RoBERTaLARGE improved on this with an accuracy of 58.2% [5]. Heinsen Routing and RoBERTaLARGE introduce dynamic routing that prioritize sentiment-critical tokens while reducing noise [17]. This approach enhances RoBERTaLARGE's ability to discern subtle sentient differences in fine-grained sentiment tasks. The model cur- rently achieves the highest accuracy on the SST-5 dataset, with a score of 59.8%. Another approach, using RoBERTaLARGE combined with Self-Explaining mechanisms, integrates auxiliary explanation techniques to enhance interpretability and fine-grained classification [18].

Decoding-enhanced BERT with Disentangled Attention (DeBERTa) [6], a recently developed and evaluated transformer model, significantly improves upon BERT and RoBERTa with three key components:

(1) Disentangled Attention separates content and relative position embeddings (Figure 1).

Attention weights are computed across content-to-content and content-to-position interactions while excluding position-to-position interactions for efficiency. This disentanglement allows DeBERTa to model relationships more precisely, especially in long input sequences, where positional relations are critical.

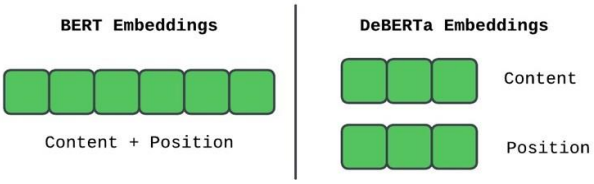

Figure 1. Embedding representation in BERT and DeBERTa. DeBERTa separates content and position embeddings for better contextual understanding.

(2) Enhanced Mask Decoder incorporates absolute positional embeddings during the decoding stage rather than in input embeddings as in BERT (Figure 2). This design allows DeBERTa to capture relative positioning throughout the Transformer layers and benefits from absolute positioning for final predictions.

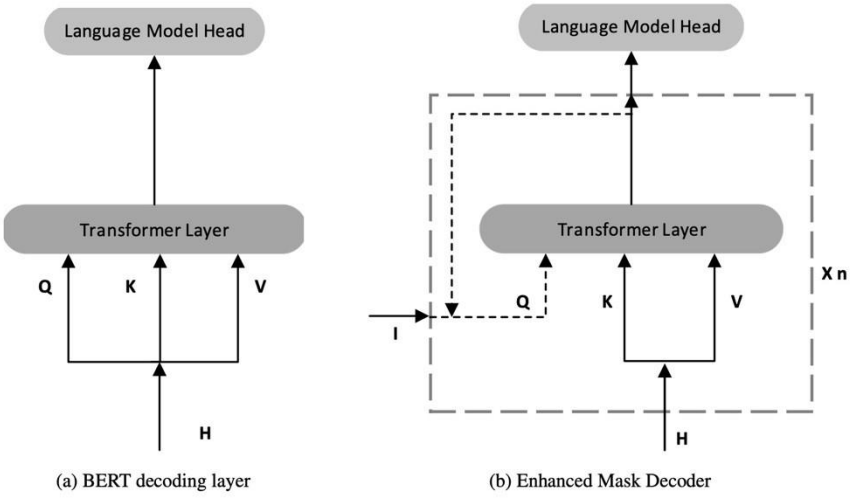

Figure 2. Decoding process in BERT vs. DeBERTa. DeBERTa applies absolute positional embeddings at the decoding stage, improving representation. Adapted from He et al. (2021) [6].

(3) Virtual Adversarial Training introduces small noise to normalized input word embeddings to enhances the model's generalization.

DeBERTaV3 further enhances the original DeBERTa model with advanced pretraining techniques, achieving improvements in efficiency and contextual representation [7]. On the General Language Understanding Evaluation, a comprehensive NLP evaluation suite, DeBERTaV3$_{\text{LARGE}}$ outperformed other models, surpassing previous state-of-the-art results by 1.37% in average score [7]. The progression from machine learning to transformer-based models like DeBERTa and its successor DeBERTaV3 highlights advances in sentiment analysis, setting the foundation to explore new architectures and fine-tuning strategies tailored to specific datasets in this research.

3. PROPOSED METHOD

In this research, DeBERTaV3 is selected for fine-tuning due to its advanced architecture and superior performance on NLP tasks. Critical to DeBERTa's performance are two key innovations: Replaced Token Detection and Gradient-Disentangle Embeddings Sharing which

improve the efficiency and quality of the model's pretraining process [7]. These techniques build upon the limitations of the previously used Masked Language Modeling approach.

(1) Masked Language Modeling (MLM)

MLM is a common pretraining objective for transformers like BERT. It masks random tokens in input sequences, training the model to predict them using context. While MLM helps build contextual understanding, it only optimizes masked tokens, reducing training efficiency and increasing computation for unmasked tokens.

(2) Replaced Token Detection (RTD)

RTD addresses the inefficiencies in MLM by replacing it as the primary pretraining task in DeBERTaV3. First introduced in Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA), RTD uses a generator replaces certain tokens in the input with plausible alternatives, and a discriminator learns to identify whether each token is original or replaced [19]. Unlike MLM, RTD allows the model to learn from all tokens in the input, significantly improving data efficiency and training speed, even in scenarios with limited training data. However, MLM remains essential for training the generator, ensuring high-quality token replacements.

(3) Gradient-Disentangled Embeddings Sharing (GDES)

In earlier models like ELECTRA, embedding sharing between the generator and discriminator caused inefficiencies due to conflicting MLM and RTD optimization goals. DeBERTaV3 introduces GDES to separate their gradient updates while retaining shared embeddings. The generator updates embeddings via MLM loss, while RTD optimizes the discriminator, improving efficiency and convergence without degrading embedding quality.Fig. 3 outlines a high-level architecture of DeBERTaV3.

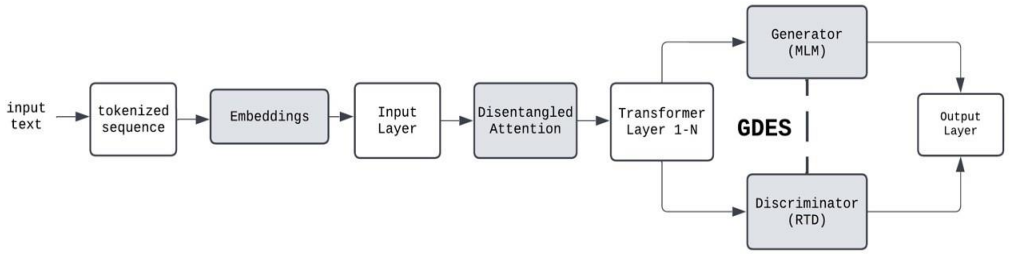

Figure 3. Illustration of Gradient Disentangled Embedding Sharing (GDES) between the Masked Language Model (MLM) and the Replaced Token Detection (RTD) objectives in DeBERTaV3.

Table 1 presents the configurations of different DeBERTaV3 variants, with $\text{DeBERTaV3}_{\text{LARGE}}$ chosen for fine-tuning due to its increased capacity, including a larger hidden size, more attention heads, and additional transformer layers**.** These architectural improvements enhance the model's ability to capture nuanced sentiment distinctions, making it well-suited for finegrained sentiment analysis across five sentiment classes. However, the increased model size introduces higher computational demands, requiring greater memory and processing power. While smaller variants offer improved efficiency, they may lack the representational capacity necessary for fine-grained sentiment classification. The trade-off between performance and computational feasibility is considered, with $\text{DeBERTaV3}_{\text{LARGE}}$ selected to maximize accuracy despite its higher resource

requirements. Due to these computational challenges, specific training strategies are employed to improve efficiency, including optimized batch processing, gradient accumulation, and mixed-precision training.

Table 1. Configurations of different variants of DeBERTaV3 [7]

Model	Hidden Size	Attention Heads	Layers	Parameters
DeBERTaV3SMALL	768	12	6	86M
$\text{DeBERTaV3}_{\text{BASE}}$	768	12	12	140M
$\text{DeBERTaV3}_{\text{LARGE}}$	1024	16	24	350M

Figure 4 outlines the proposed fine-tuned DeBERTaV3 process. First, the input text is preprocessed to ensure consistency. This includes removing special characters, punctuation, and HTML tags, expanding contractions, and normalizing the text by converting to lowercase and tokenizing sentences. The preprocessed text is then tokenized using $\text{DeBERTaV3}_{\text{LARGE}}$ tokenizer, which converts text into sub-word tokens while adding special tokens such as [CLS] for classification and [SEP] for separating sequences. The tokenized sequences are passed through the pretrained $\text{DeBERTaV3}_{\text{LARGE}}$ model which computes contextual embeddings. The Disentangled Attention Mechanism separates content and relative position information in the embeddings to model both semantic and relationship dependencies effectively.

A key contribution of our approach is the introduction of an attention pooling layer, which enhances sentiment classification by selectively weighting token embeddings, prioritizing informative tokens over simple pooling methods like averaging. The attention network transforms token representations through a linear layer, normalization, and GELU activation before generating scalar attention scores, which are normalized via softmax. These scores compute a weighted sum of token embeddings, which is then concatenated with the [CLS] token to capture both local and global context. This approach improves computational efficiency by reducing redundant computations, operating in $O(n)$ per batchrather than$O(n^2)$ as in self-attention, and enabling parallel execution on GPUs. Additionally, intermediate dimensionality reduction lowers memory overhead, making the model more efficient while maintaining expressiveness.

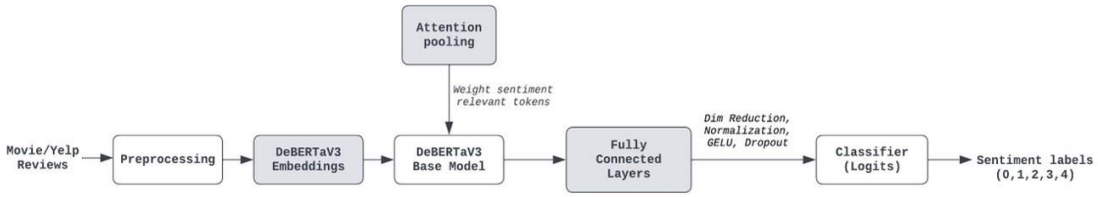

Figure 4. Architecture of Sentiment-Enhanced Fine-tuned DeBERTaV3 (FiTSent-DeBERTaV3), where DeBERTaV3 is enhanced with an attention pooling mechanism to prioritize sentiment-critical features for fine-grained sentiment analysis.

4. EXPERIMENTS AND RESULTS

4.1. Datasets

Our proposed approach is evaluated on two datasets: SST-5 and Yelp Reviews. SST-5 serves as a benchmark for fine-grained, sentence-level sentiment analysis, while Yelp Reviews supports the evaluation of document-level text classification. Due to computational constraints, we sampled a

portion of the Yelp Reviews, preserving the original distribution to ensure comparability with the full dataset.

Table 2 summarizes key statistics of SST-5 and Yelp Reviews, including dataset size and sentence/word counts. Table 3 details class distributions, showing review counts and proportions for each sentiment class, while Figure 5 visualizes these distributions. Detailed examples for each sentiment class in SST-5 and Yelp Reviews can be found in Appendix A.

Table 2. Dataset statistics: #s denotes the number of sentences, and #w denotes the number of words per entry

Dataset	**Size**	**Avg #s**	**Max #s**	**Avg #w**	**Max #w**
SST-5	11,855	1.21	12	19.17	56
Sampled Yelp Reviews	100,000	9.04	351	104.20	1013
Full Yelp Reviews	1,569,254	9.05	2101	104.06	1056

Table 3. Class-wise Counts and Proportions of SST-5, Sampled Yelp Reviews, and Original Yelp Reviews. Class labels 1 through 5 represent sentiments ranging from Very Negative (1) to Very Positive (5)

Dataset	**Class 1 (Count, %)**	**Class 2 (Count, %)**	**Class 3 (Count, %)**	**Class 4 (Count, %)**	**Class 5 (Count, %)**
SST-5	1510 (~13%)	3140 (~26%)	2242 (~19%)	3111 (~26%)	1852 (~16%)
Sampled Yelp	17,246 (~17%)	8331 (~8%)	9272 (~9%)	19,419 (~20%)	45,732 (~46%)
Full Yelp	270,626 (~17%)	130,744 (~8%)	145,506 (~9%)	304,736 (~20%)	717,562 (~46%)

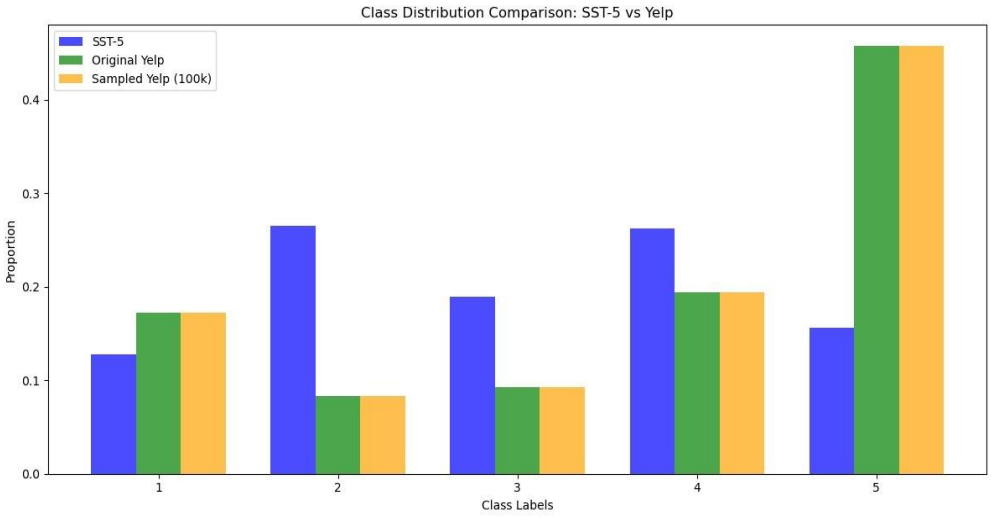

Figure 5. Visual Comparison of class distributions across SST-5, Sampled Yelp Reviews, and Original Yelp Reviews datasets. Class labels 1 through 5 represent sentiments ranging from Very Negative (1) to Very Positive (5).

4.1.1. Stanford Sentiment Treebank (SST-5)

The initial dataset, consisting of 10,662 movie review excerpts from rottentomatoes.com, was first collected and published by Wu and Pao [20]. It was later expanded to 11,855 single sentences to evaluate the Recursive Neural Tensor Network (RNTN) model's ability to understand sentence structure and sentiment [21]. Sentences were parsed into sub-phrases, which were annotated with sentiment labels via Amazon Mechanical Turk. The sentiment of each full sentence was then determined based on the annotations of its constituent phrases. The dataset classifies sentiment into five categories: Very Negative, Negative, Neutral, Positive, and Very Positive.

4.1.2. Yelp Reviews

The Yelps Reviews dataset, sourced from the Yelp Dataset Challenge in 2015, features reviews rated on a five-point scale from 1 to 5 (higher is better). Due to computational limitations, we sampled 100,000 data points from the original 1,569,254 reviews. The sample preserves the original class distribution, ensuring the proportion of each rating matches its occurrence in the full dataset, as shown in Table 3 and Figure 5. We can see the natural class imbalance in both the sampled and full datasets, where class 5 dominates (approximately 46%), and class 2 is the least frequent (around 8%). Additionally, Table 2 also shows that the statistics on the original dataset and our sampled datasets are very comparable in terms of sentence lengths and word counts.

4.1.3. Comparative Analysis

From Table 3 and Figure 5, SST-5 dataset exhibits a more balanced distribution across its five sentiment classes, with each class contributing between 13% and 26% of the data. This balance provides a controlled setting for evaluating models on fine-grained sentiment distinctions. By comparison, the Full and Sampled Yelp reviews are characterized by a skewed distribution. For example, a substantial portion of reviews falling into the Class 4 (Positive) and Class 5 (Very Positive), while Class 3 (Neutral) and Class 2 (Negative) sentiments are notably less frequently. This imbalance reflects the natural tendency in real-word review datasets, where user feedback often skews toward more extreme sentiments.

Table 2 and Appendix A highlight additional key differences between the datasets. SST-5 consists of shorter, sentence-level reviews averaging 1.2 sentences and 19 words per entry. These reviews often use nuanced, sophisticated language to convey sentiments, making subtle distinctions, such as between Class 3 (Neutral) vs Class 4 (Positive) or Class 2 (Negative) particularly challenging to detect. In contrast, Yelp reviews feature significantly longer reviews, averaging 9 sentences and 104 words per entry. The language in Yelp reviews is conversational and descriptive, often following a narrative structure and incorporating specific experiences, opinions, or events with explicit sentiment cues.

The variability in language and class distribution tests different aspects of model performance. The explicit cues in Yelp Reviews can simplify sentiment classification for dominant classes, but the skewed distribution necessitates careful evaluation to ensure the model performs well across all sentiment classes, particularly the underrepresented ones. On the other hand, SST-5's balanced distribution and nuanced language require a model capable of handling fine-grained sentiment distinctions.

Given these differences, baseline models align with the characteristics of each dataset. Models such as Heinsen Routing and RoBERTa$_{\text{LARGE}}$ [17], RoBERTa$_{\text{LARGE}}$ and Self Explaining [18] are designed to handle SST-5's shorter reviews and their focus on subtle sentiment distinctions. On

the other hand, hierarchical and document-level models such as HAN [13] and BigBird [15] are tailored specifically for Yelp's longer reviews. Despite the variance in baseline models, our proposed model is designed to perform well on both tasks, demonstrating its generalizability across sentence-level and document-level sentiment analysis.

4.2. Model and Training Configurations

The SST-5 dataset is split into 8,544 training samples, 1,101 validation samples, and 2,210 test samples, while the Sampled Yelp Reviews dataset consists of 72,000 training samples, 8,000 validation samples, and 20,000 test samples. Input text is preprocessed and tokenized using the pretrained DeBERTaV3 tokenizer. The maximum sequence length is set based on the 95th percentile of review lengths—128 tokens for SST-5 and 512 tokens for Yelp—to minimize truncation while padding shorter texts for uniform batch sizes. Efficient GPU utilization is ensured through PyTorch Dataset and DataLoader, multi-threading, pinned memory, and shuffled training batches.

Fine-tuning configurations differ based on dataset characteristics. Yelp Reviews, containing longer and more descriptive texts, uses a base learning rate of 8×10^{-6} and a higher rate of of 1×10^{-5} for the pooling and classifier layers. Gradient accumulation with a step size of 4 results in an effective batch size of 64, optimizing computational efficiency. A shorter training schedule of three epochs is applied, with label smoothing (0.1) to mitigate overconfidence and dropout (0.1) to enhance generalization. SST-5 fine-tuning is tailored for shorter, nuanced movie reviews. A learning rate of 1×10^{-5} for the base model and 5×10^{-5} for other layers are used with an extended training schedule of 12 epochs to capture subtle sentiment distinctions. While the dataset has a more balanced label distribution, sentiment classes are harder to distinguish due to complex language. To address this, dynamically computed class weights in cross-entropy loss ensure that underrepresented classes receive appropriate attention during training. This weighting mechanism helps the model better capture subtle emotional expressions and improve classification of less frequent sentiment classes.

Table 4. summarizes the key hyperparameters used for fine-tuning DeBERTaV3 on both datasets. Both datasets are trained using mixed precision with GradScaler for efficient GPU utilization. Learning rate schedules are adapted for each dataset, with a linear warmup applied to Yelp Reviews and a cosine scheduler for SST-5 for smooth convergence. Validation metrics, including loss and accuracy, are evaluated at the end of each epoch, with the best model weights saved for subsequent evaluation. Unlike traditional sentiment classification models that use uniform learning rates, fixed schedules, and standard loss functions**,** this fine-tuning strategy integrates adaptive learning rates, label smoothing for imbalanced distributions, class-weighted loss for nuanced sentiment categories, and dataset-specific scheduling. These optimizations collectively improve the model's ability to generalize across long-form descriptive reviews (Yelp) and short, nuanced expressions (SST-5), ensuring both computational efficiency and performance robustness.

Table 4. Hyperparameter Configurations for Fine-Tuning DeBERTaV3 on SST-5 and Sampled Yelp Reviews datasets.

Hyperparameter	**SST-5**	**Sampled Yelp Reviews**
Max length	128	512
Batch Size	16	16 (Effective: 64)
Gradient Accumulation Steps	-	4
Learning Rate (Base Model)	1e-5	8e-6
Learning Rate (Custom Layers)	5e-5	1e-5
Weight Decay	0.05	0.01
Dropout Rate	0.3	0.2
Scheduler	Cosine	Linear
Epochs	12	3

4.3. Results and Discussion

4.3.1. Metrics Overview

To evaluate model performance, we employ four key metrics: accuracy, precision, recall and F1score, each contributing distinct insights into the model's capabilities in fine-grained sentiment analysis:

(1) Accuracy measures the proportion of correct classified predictions over the total number of samples. It provides a straightforward indication of overall model performance.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$$

While accuracy is often the most intuitive metric, it can be misleading for imbalanced datasets like Yelp Reviews, where dominant class, for example class 5 (Very Positive) may inflate overall accuracy.

(2) Precision evaluates the proportion of correct positive predictions among all cases predicted as positive.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

In fine-grained sentiment analysis, where distinguishing between subtle sentiment classes is critical, high precision ensures the model assigns a sentiment label only when it is highly confident. For example, a high precision score means that model correctly identifies Positive reviews without misclassifying Neutral or Very Positive reviews as Positive.

(3) Recall, or sensitivity, measures the proportion of actual positives that the model correctly identifies.

$$Precision = \frac{True \quad Positives}{True \quad Positives + False \quad Positives}$$

Recall is vital in capturing underrepresented sentiments in imbalanced datasets, like Class 2 (Negative) or Class 3 (Neutral) in Yelp reviews. High recall ensures that subtle or less frequent sentiments are not overlooked. Missing critical sentiments (false negatives) can result in an incomplete understanding of user sentiment.

(4) F1-score combines precision and recall into a single harmonic mean to balance their tradeoff:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

F1-Score provides a balanced view as it combines both abilities to avoid false positives (precision) and false negatives (recall) into a single value. Unlike the arithmetic mean, the harmonic mean emphasizes the lower of the two values (precision or recall). This ensures that the F1-score only becomes high when both Precision and Recall are high, making it a robust metric for evaluating model performance in tasks like fine-grained sentiment analysis.

4.3.2. Performance on SST-5

We evaluated FiTSent DeBERTaV3 against several established baselines on SST-5, as shown in Table 5. We report accuracy as it is the primary metric used in existing baselines. While precision, recall, and F1-score are available for our model, they are omitted here to focus on direct comparisons with baselines.

Our model achieved the highest accuracy of 60.27%, outperforming all other methods on the SST- 5 dataset. It surpasses the best-performing baseline, Heinsen Routing with $RoBERTa_{LARGE}$ by 0.47%. Compared to other variants of BERT, including $RoBERTa_{LARGE}$ and fine-tuned $BERT_{LARGE}$, FiTSent DeBERTaV3 shows improvement of 2.07% and 4.77% respectively. These results highlight the strength of DeBERTaV3 architecture for sentiment classification tasks. FiTSent DeBERTaV3 also outperforms $RoBERTa_{LARGE}$ and Self-Explaining by 1.17%. While both approaches utilize advanced techniques to improve performances, this result shows that DeBERTaV3 with attention pooling is better suited for fine-grained sentiment analysis.

Table 5. Comparison of FiTSent DeBERTaV3 with Established Baselines on SST-5

Methods	**Accuracy (%)**
Fine-tuned $BERT_{LARGE}$	55.4
$RoBERTa_{LARGE}$	58.2
$RoBERTa_{LARGE}$ and Self-Explaining	59.1
Heinsen Routing and $RoBERTa_{LARGE}$	59.8
Proposed Model (FiTSent DeBERTaV3)	**60.27**

To better illustrate how attention pooling contributes to the model's performance, Figure 6 provides examples of how the model assigns attention weights to sentiment-related tokens in the

input text. The attention mechanism effectively identifies key tokens that contribute to the sentiment of the sentence, such as “professional” (1.00) and “plodding” (1.00). This targeted focus on sentiment-bearing tokens helps the model capture the subtle nuances required for fine-grained sentiment analysis in complex texts.

Figure 6. Attention heatmap on SST-5 reviews. Label 2 denotes Neutral, and Label 3 denotes Positive. Red highlights sentiment-critical words with higher attention scores.

4.3.3. Performance on Yelp Reviews

Table 6 summarizes the accuracy reported for several established baseline methods on either the full Yelp Reviews dataset or subsets of it. For a fair comparison, we also implemented and tested these baseline methods on our Sampled Yelp Reviews dataset, consisting of 100,000 data points. Since these models were implemented on our sampled dataset, we were able to calculate additional metrics—precision, recall, and F1-score—presented in Table 7.Including these metrics is both practical and beneficial, as they provide deeper insights into the model’s ability to handle the class imbalance in the Yelp Reviews dataset. Precision and recall are particularly important for analyzing how well the model balances predictions across dominant and underrepresented classes, making them critical for evaluating sentiment classification in real-world reviews.

Table 6. Established Baselines on Full Yelp Reviews

Methods	**Accuracy (%)**
BERT-ITPT-FiT	70.58
BigBird	72.16
HAHNN	73.08

Table 7. Performance of FiTSent DeBERTaV3 and Baselines on Sampled Yelp Reviews. All metrics are in %

Methods	**Accuracy**	**Precision**	**Recall**	**F1-score**
BERT-ITPT-FiT	70.68	69.26	70.68	69.81
BigBird	77.19	70.23	68.75	69.38
HAHNN	70.09	67.22	70.03	67.83
Proposed Model (FiTSent DeBERTaV3)	**78.71**	**77.91**	**78.71**	**78.17**

The baseline models—BERT-ITPT-FiT, BigBird, and HAHNN—demonstrated their ability to handle long, context-rich reviews in the original Yelp Reviews dataset. HAHNN achieved the highest accuracy by utilizing its hybrid neural networks and hierarchical attention mechanism to capture dependencies through both word-level and sentence-level attention. BigBird, optimized for long-range input, followed with 72.16%, while BERT-ITPT-FiT achieved 70.58%, benefiting from its robust pretraining for general text understanding.

On the Sampled Yelp Reviews dataset, we observed notable changes in their performances. BigBird improved significantly, reaching 77.91%, approaching FiTSent DeBERTaV3's 78.71%. This likely reflects the benefits of a smaller, less noisy dataset for BigBird's long-range input optimization. However, its lower precision (70.23%) and recall (68.75%) compared to FiTSent DeBERTaV3 (77.91% and 78.71%, respectively) indicate challenges in identifying underrepresented classes and minimizing false positives. HAHNN decreased to 70.09%, possibly due to less training data limiting its ability to generalize hierarchical dependencies. BERT-ITPTFiT remained stable, with a slight increase 70.68%, reflecting its consistency across changes in dataset size. These models' precision and recall metrics also fell below those of FiTSent DeBERTaV3, reflecting their limitations in addressing class imbalances and capturing subtle sentiment cues.

FiTSent DeBERTaV3's balanced performance across accuracy, precision, recall, and F1-score demonstrates its strength in capturing nuanced sentiment distinctions while maintaining robustness across all sentiment classes. These results underscore the benefits of fine-tuning advanced pretrained models like DeBERTaV3 with strategies tailored to specific datasets, enabling effective handling of complex, context-rich reviews.

4.3.4. Cross-Dataset Performance Comparison

Building on the comparisons of the proposed model with baselines for each dataset individually, we now evaluate the model's performance across the two datasets to highlight its adaptability to diverse sentiment analysis tasks. Table 8 summarizes the results using following metrics: accuracy, precision, recall, and F1-score.

Table 8. Performance of the Proposed Model on SST-5 and Sampled Yelp Reviews. All metrics are in %

Dataset	Accuracy	Precision	Recall	F1-score
SST-5	60.27	60.26	60.27	60.10
Yelp Reviews	78.71	77.91	78.71	78.17

Performance Across Datasets

The results show that the proposed model achieves consistent performance across all metricsaccuracy, precision, recall, and F1-score - on both SST-5 and Sampled Yelp Reviews. These metrics underscore the model's ability to address the unique challenges posed by each dataset. For SST-5, the balanced precision, recall, and F1-score highlight the model's strength in identifying nuanced sentiments. By successfully distinguishing closely related sentiment classes, such as Neutral versus Positive or Negative versus Very Negative, the model demonstrates its ability to capture subtle contextual and figurative language cues.

For Yelp Reviews, the high precision, recall, and F1-score reflect the model's capability to address class imbalance effectively. Precision ensures that the model avoids false positives in dominant classes, while recall demonstrates its effectiveness in identifying underrepresented

classes. This balance allows the model to maintain robustness across the full range of sentiment labels, particularly in long and context-rich reviews. Overall, the strong performance on both datasets demonstrates the model's adaptability to sentence-level fine-grained sentiment distinctions in SST-5 and document-level classification with class imbalance in Yelp Reviews.

Confusion Matrix Insights

To further understand the model's performances, we analyzed the confusion matrices for SST-5 and Sampled Yelp Reviews as seen in Figure 7.

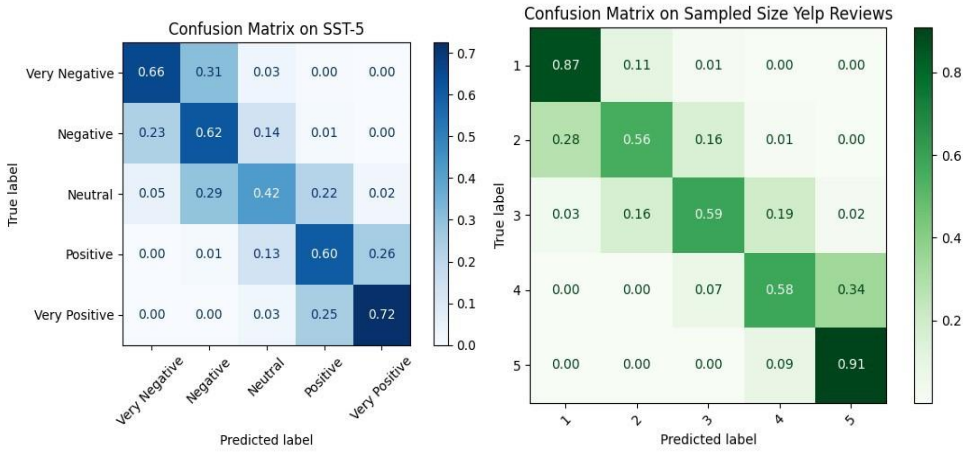

Figure 7. Confusion matrices for SST-5 and Sampled Yelp Reviews datasets, showing strong accuracy in polar classes but challenges in mid-range sentiments.

The model excels in identifying polar sentiment classes in both datasets. For SST-5, it achieves 66% accuracy for Very Negative and 72% for Very Positive. Similarly, in the Yelp Reviews dataset, the model achieves 87% accuracy for 1-star reviews and 91% for 5-star reviews. These results highlight the model's ability to effectively classify sentiments with clear emotional and linguistic cues, demonstrating stronger performance in Yelp Reviews due to the straightforward language and availability of richer contextual information.

However, the model struggles with mid-range sentiment classes in both datasets. In SST-5, the Neutral class achieves only 42% accuracy, with 26% of Neutral samples misclassified as Somewhat Negative and 25% as Somewhat Positive, indicating difficulty in distinguishing between moderately positive and negative sentiments. Similarly, Somewhat Positive (68% accuracy) is frequently misclassified as Positive (21%), while Somewhat Negative (69% accuracy) is often predicted as Negative (17%). These errors suggest that adjacent sentiment classes share overlapping linguistic patterns, making it challenging to draw clear distinctions. A similar trend is observed in Yelp Reviews, where 3-star ratings (59% accuracy) are often misclassified into 2-star (19%) and 4-star (18%) categories. This pattern suggests that reviews expressing mixed sentiments or nuanced opinions are harder to classify than those with strong emotional polarity. The high misclassification rates in mid-range classes likely stem from the gradual sentiment transitions in real-world text, where phrases do not always express clear positivity or negativity but instead exhibit a blend of both. Furthermore, mid-range sentiments often rely on context-dependent expressions, sarcasm, or nuanced word choices that the model may not effectively capture. Since the dataset includes diverse writing styles, certain expressions with ambiguous emotional tones may lead the model to favor adjacent classes.

To address these challenges, several improvements can be explored. First, incorporating contrastive learning could help the model better differentiate subtle sentiment variations by learning fine-grained distinctions between adjacent classes. Second, reinforcing ordinal classification through additional constraints on the output space could ensure that predictions follow the natural ordering of sentiment labels. Third, enhancing contextual embeddings by leveraging domain-specific pretraining on sentiment-rich corpora may improve the model's ability to discern nuanced sentiment expressions. Lastly, data augmentation techniques, such as paraphrasing or sentiment perturbation, could enrich the training data and expose the model to a broader spectrum of sentiment expressions, potentially reducing misclassifications in mid-range classes. By addressing these limitations, the model could improve its ability to capture nuanced sentiment distinctions, leading to more reliable fine-grained sentiment classification, particularly in cases where sentiment boundaries are less clearly defined.

Performance Gap Across Text Lengths and Contexts

The model demonstrates strong performance across both datasets, with higher metrics observe on the longer, context-right Yelp Reviews dataset. This can be attributed to several factors. First, longer texts provide more explicit sentiment cues and context, enabling the model to better identify and classify sentiments accurately. On the other hand, SST-5 relies on shorter, singlesentence reviews that often use subtle or figurative language, making sentiment distinctions more challenging.

Second, the smaller size of the SST-5 dataset further limits the diversity of training samples, reducing the model's ability to generalize effectively across all sentiment classes. With fewer examples, the model struggles to learn nuanced patterns that differentiate between closely related sentiment classes. In contrast, the larger Yelp Reviews dataset offers more diverse examples, contributing to more robust training and better generalization.

These results highlight the model's capability to handle both short and long texts, while demonstrating that the inherent characteristics of the dataset – text length, linguistics complexity, and dataset size – significantly influence its performance outcomes.

5. CONCLUSIONS

In this study, we presented FiTSent DeBERTaV3, a model fine-tuned for both sentence-level and document-level sentiment classification tasks. Through comprehensive evaluations on SST-5 and sampled Yelp Reviews datasets, our model demonstrated significant performance improvements over established baselines, with our model achieving the highest metrics on both datasets. On SST-5, FiTSent DeBERTaV3 achieved moderate gains, reflecting the inherent challenges of finegrained sentiment analysis in shorter, subtle texts. On the Sampled Yelp Reviews dataset, the model showcased its robustness and adaptability by effectively processing long and narrativestyle reviews. Importantly, our model maintained balanced metrics across both datasets, achieving consistent performance in terms of accuracy, precision, recall and F1-score.

This highlights its ability to generalize effectively across tasks with different text lengths and complexities. These results highlight the advantages of leveraging the DeBERTaV3 architecture, which combines advanced pretraining techniques with efficient contextual representation. It also emphasizes the importance of tailored fine-tuning strategies and attention mechanisms in improving sentiment classification performance. While FiTSent DeBERTaV3 demonstrated strong results, particularly in handling diverse text structures, the moderate gains on SST-5 and challenges with mid-range classes across both datasets suggest areas for further improvement.

To enhance FiTSent DeBERTaV3's performance on SST-5, we plan to explore domain-specific pretraining using movie review datasets such as IMDB and Rotten Tomatoes. This will expose the model to language patterns specific to film critics, potentially improving its ability to understand nuanced sentiment in SST-5. While DeBERTaV3-large has already undergone extensive pretraining, targeted domain adaptation could still offer performance gains and is worth investigating. Given the limited size of SST-5, we will also explore training strategies that improve generalization. One approach is cross-fold training, where the model is trained on multiple overlapping subsets of the data to maximize exposure to diverse examples. Additionally, contrastive learning techniques may help the model distinguish subtle sentiment variations by learning relationships between different sentiment classes. For scalability on the full Yelp Reviews dataset, efficient training methods will be explored to handle its large size. Progressive training, where initial training is conducted on smaller subsets before fine-tuning on the full dataset, may help manage computational costs more effectively.

To assess adaptability to informal, context-dependent text, evaluations will extend to social media datasets such as Twitter and Reddit, where sentiment is often conveyed through slang, emojis, and sarcasm. Fine-tuning on these datasets may improve the model's ability to handle short-form, conversational sentiment, making it more effective for real-time sentiment monitoring and content analysis. Inspired by related work on integrating sentiment and content relevance in usergenerated data, we plan to extend the application of our model to datasets like Yahoo Answers [2]. Such use cases highlight the broader potential of sentiment analysis to evaluate user satisfaction, intent, and engagement across diverse platforms. These research directions aim to enhance the performance and adaptability of FiTSent DeBERTaV3, addressing its current limitations while contributing to advancements in sentiment analysis for real-world applications.

ACKNOWLEDGEMENTS

We would like to express our gratitude to Khoury College of Computer Sciences for providing resources and support necessary to conduct this study. We extend our appreciation to our peers and colleagues for their constructive discussions and moral support, which have greatly contributed to the completion of this paper.

REFERENCES

[1] Tiquan Gu, Zhenzhen He, Hui Zhao, Min Li, and Di Ying. Aspect-based sentiment analysis with multigranularity information mining and sentiment hint. *Expert Systems with Applications*, 252:124104, 10 2024.

[2] Ameen Banjar, Awais Shaheen, Tehmina Amjad, Riad Alharbey, and Ali Daud. Users' satisfactionbased ranking for Yahoo Answers. *Multimedia Tools and Applications*, 83(28):71265–71284, 8 2024.

[3] Kumar Ravi and Vadlamani Ravi. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, 89, 2015.

[4] Kian Long Tan, Chin Poo Lee, and Kian Ming Lim. A Survey of Sentiment Analysis: Approaches,Datasets, and Future Research. *Applied Sciences (Switzerland)*, 13(7):4550, 4 2023.

[5] Jiaxi Li. Fine-Grained Sentiment Analysis with a Fine-Tuned BERT and an Improved PreTrainingBERT. In *2023 IEEE International Conference on Image Processing and Computer Applications, ICIPCA 2023*, 2023

[6] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DEBERTA: DECODINGENHANCED BERT WITH DISENTANGLED ATTENTION. In *ICLR 2021 - 9th InternationalConference on Learning Representations*, 2021.

[7] Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRAStyle Pre-Training with Gradient-Disentangled Embedding Sharing. *11th International Conference on Learning Representations, ICLR 2023*, 11 2021.

[8] Margarita Rodr´ıguez-Iba´nez, Antonio Casa´nez-Ventura, F´elix Castejo´n-Mateos, and Pedro ManuelCuenca-Jim´enez. A review on sentiment analysis from social media platforms. *Expert Systems with Applications*, 223:119862, 8 2023.

[9] Umroh Makhmudah, Saiful Bukhori, Januar Adi Putra, and Bratasena Anggabayu Bhirawa Yudha.Sentiment Analysis of Indonesian Homosexual Tweets Using Support Vector Machine Method. *Proceedings - 2019 International Conference on Computer Science, Information Technology, and Electrical Engineering, ICOMITEE 2019*, pages 183–186, 10 2019.

[10] Satuluri Vanaja and Meena Belwal. Aspect-Level Sentiment Analysis on ECommerce Data. *Proceedings of the International Conference on Inventive Research in Computing Applications, ICIRCA 2018*, pages 1275–1279, 12 2018.

[11] Tanushree Dholpuria, Y. K. Rana, and Chetan Agrawal. A sentiment analysis approach through deep learning for a movie review. *Proceedings - 2018 8th International Conference on Communication Systems and Network Technologies, CSNT 2018*, pages 173–181, 11 2018.

[12] Yi Yang. Convolutional neural networks with recurrent neural filters. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, 2018.

[13] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 Proceedings of the Conference*, 2016.

[14] Jader Abreu, Luis Fred, David Macˆedo, and Cleber Zanchettin. Hierarchical Attentional Hybrid NeuralNetworks for Document Classification. In *Lecture Notes in Computer Science (including subseries Lecture*

[15] Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon,Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big Bird: Transformers for Longer Sequences. 7 2020.

[16] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to Fine-Tune BERT for Text Classification?5 2019.

[17] Franz A. Heinsen. An Algorithm for Routing Vectors in Sequences. 11 2022.

[18] Zijun Sun, Chun Fan, Qinghong Han, Xiaofei Sun, Yuxian Meng, Fei Wu, and Jiwei Li. Self-ExplainingStructures Improve NLP Models. 12 2020.

[19] Kevin Clark, Minh Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-trainingText Encoders as Discriminators Rather Than Generators. *8th International Conference on Learning Representations, ICLR 2020*, 3 2020.

[20] Jean Y Wu and Yuanyuan Pao. Predicting Sentiment from Rotten Tomatoes Movie Reviews. *Nlp.Stanford.Edu*, 2012

[21] Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y.Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2013.

AUTHORS

L Do received her Bachelor of Science degree in Industrial and Enterprise Systems Engineering from the University of Illinois at Urbana Champaign. She is currently pursuing a Master of Science in Computer Science at Northeastern University, based in Silicon Valley, with an expected graduation in 2025. Her research interests incl ude fine-grained sentiment analysis, responsible AI, and the application of advanced machine learning models in decision-making and user personalization.

T Amjad is an Associate Teaching Professor in the Khoury College of Computer Sciences at Northeastern University, based in Silicon Valley. Before that, she has served International Islamic University, Islamabad, Pakistan, as an Assistant Professor and an Incharge of Faculty of Computing and Information Technology. She has been actively engaged in academic and research activities at the graduate as well as undergraduate levels for the last 18 years. Amjad teaches Information Retrieval, Da ta Mining, Machine Learning, Database systems, Discrete Structures, Digital Logic Design, Data Warehousing, and Research Methods. She has published her research work in leading impact factor journals in the fields of Data Science, Information retrieval, Semantics, Digital libraries, and Scientometrics.

APPENDIX

SST-5 and Yelp Reviews Examples

Table 7: SST-5 and Yelp Reviews Examples with Better Class Distinction

Class	SST-5 Examples	Yelp Reviews Examples
1 (Very Negative)	The film is all over the place, really.	The worst, sugar free will give you stomach cramps unbelievable. I order four for Thanksgivings and all were a disappointment.
2 (Negative)	The overall feel of the film is pretty cheesy, but there's still a real sense that the Star Trek tradition has been honored as best it can, given the embarrassing script and weak direction.	Disappointed. I tried to return a $2 outlet cover (wrong color match to existing)..and was told by the cashier I would need to pay a 25% restocking fee (Yes…I did have to open the package in order to install…which is when I discovered it was the wrong color match) I was purchasing another household product also...so my $2 'return' (plus some) was going right back into their register. Not the best approach for customer loyalty.
3 (Neutral)	The story is so light and sugary that were it a Macy's Thanksgiving Day Parade balloon, extra heavy-duty ropes would be needed to keep it from floating away.	We were staying at Aria for 2 nights last week, and didn't want to go too far for food when I was starving. Since I had a craving for eggs so we decided to have our lunch. I got an egg benedict and my husband got a spa omelet (which has no egg yoke). The food was not bad and satisfactory enough. So that's a good thing enough already. And I agree with other yelpers that the service was not that great. The server didn't attend to us while it wasn't busy at the time we were there. The server wasn't rude, but just not attentive. So I give 3 stars to this restaurant and it's an ok restaurant. It's a good choice when you stay in Aria and don't wanna go too far.
4 (Positive)	Strange and beautiful film.	Above average pub fare at good prices. Daily food/drink specials and great service at this non-chain pub. Live music on weekends. 4 stars.
5 (Very Positive)	The production design, score and choreography are simply intoxicating.	Can't give this enough stars. All you can eat lobster tails, rack of lamb, sushi, caviar, lobster tails, king crab, lobster tails, champagne and bloody Mary's. oh did I mention lobster tails. As a fan of great food I recommend this spot. It isn't cheap at $125 each but I know I drank that amount in champagne.