DSA4213 Final Project Report:
Fine-Tuning Transformer Models for Sentiment Analysis on Steam Reviews

By **GROUP 31**:

Ream Chan Sovath (A0258393R)

Tian Zhuoyu (A0258724W)

Professor: Doudou Zhou

Natural Language Processing for Data Science (DSA4213)

Semester 1, AY2025/2026

1 Abstract

This project examines the effectiveness of transformer fine-tuning strategies for sentiment classification in the domain of **Steam game reviews**, a setting characterized by informal language, sarcasm, and domain-specific slang. The study systematically evaluates six representative approaches covering rule-based, neural, and transformer architectures: (1) VADER, a lexicon-based sentiment analyzer; (2) zero-shot **DistilBERT**, pre-trained on general sentiment data; (3) fully fine-tuned DistilBERT, updating all network parameters; (4) head-only fine-tuning, where only the classifier layer is adapted; (5) **LoRA**, a parameter-efficient variant that trains approximately one percent of model weights; and (6) an **LSTM + Attention** network augmented with review-length features.

A balanced corpus of 50 000 reviews was cleaned and split 80/20 for training and evaluation. Models were assessed using **Average Precision (AP)**, **Accuracy**, and **F1-score**, with consistent precision–recall analysis across experiments. Full DistilBERT fine-tuning achieved the strongest performance (AP = 0.917, F1 = 0.826), while LoRA delivered near-comparable results (AP = 0.905, F1 = 0.814) despite its drastically reduced computational footprint. Traditional baselines, including VADER (F1 = 0.701) and LSTM + Attention (F1 = 0.787), lagged behind, particularly on sarcastic or context-dependent reviews.

The findings demonstrate that domain-specific fine-tuning of pre-trained transformers markedly enhances sentiment classification accuracy on informal, user-generated text. Moreover, LoRA offers a compelling balance between performance and efficiency, enabling high-quality adaptation under CPU-only or resource-limited conditions.

2. Introduction

Steam, the world's largest online gaming platform, hosts millions of player-written reviews. These reviews represent rich, spontaneous sentiment toward games—covering satisfaction, frustration, humor, and irony. Extracting sentiment signals from such informal text is valuable for:

- Developers tracking player satisfaction
- Publishers monitoring early access feedback
- Researchers studying online discourse in gaming communities

However, Steam reviews are linguistically messy: full of slang, sarcasm, and domain-specific abbreviations. Traditional sentiment analysis tools (e.g., VADER, lexicons) cannot capture these subtleties.

The project explores **domain-specific fine-tuning of pre-trained transformer models**, specifically **DistilBERT**, for **sentiment classification** on Steam reviews.
To rigorously examine how different adaptation strategies affect model performance, we conducted a **series of ablation studies**, each varying a specific training configuration while keeping the dataset and evaluation pipeline fixed:

1. Lexicon-based (VADER)
2. Zero-shot transformer (frozen DistilBERT)
3. Fully fine-tuned transformer
4. Head-only fine-tuned transformer
5. Parameter-efficient fine-tuning (LoRA)

6. Classical deep model (LSTM + Attention)

This progression, from static rules to adaptive language models, tests how model adaptation impacts performance on noisy, real-world data.

Our central question:

Can fine-tuning transformer models on domain-specific reviews outperform general-purpose or lexicon-based sentiment models, both in accuracy and robustness?

2. Related Works

Early sentiment analysis methods were dominated by **lexicon-based approaches**, where polarity was derived from predefined word lists. Systems like **VADER** (Hutto & Gilbert, 2014) improved upon earlier rule-based tools by introducing heuristic adjustments for negation, capitalization, and intensifiers. While effective for short, structured text such as tweets, lexicon methods inherently assume that word sentiment is context-invariant. In informal domains like gaming, this assumption breaks down—expressions such as *"grind", "OP"*, or *"broken"* fluctuate in polarity depending on gameplay context, community norms, and sarcasm.

The advent of **deep learning** shifted sentiment analysis from manually defined rules to learned representations. **LSTM** networks (Hochreiter & Schmidhuber, 1997) captured sequential dependencies and improved sentiment prediction on longer text spans. The subsequent integration of **attention mechanisms** allowed models to emphasize emotionally salient tokens. Yet, despite these advances, recurrent architectures remained constrained by their sequential nature and limited capacity to capture global relationships across words, leading to degraded performance on complex or sarcastic reviews.

A breakthrough arrived with **transformer-based models** such as **BERT** (Devlin et al., 2019) and its distilled variant **DistilBERT** (Sanh et al., 2019), which leverage self-attention to encode bidirectional context efficiently. These models learn transferable linguistic and semantic representations through large-scale pretraining, which can be fine-tuned for domain-specific tasks. More recently, **parameter-efficient tuning methods** like **LoRA** (Hu et al., 2022) have enabled adaptation with minimal memory and compute cost. However, while transformers excel in formal or standardized datasets, **their performance on informal, sarcasm-rich domains like Steam reviews remains underexplored**—a gap this study aims to address through systematic evaluation of fine-tuning depth and efficiency.

4. Methods

4.1 Overview

This section describes the architecture, training configurations, and mathematical underpinnings of all six models implemented. Each model represents a distinct approach to sentiment classification—ranging from rule-based heuristics to deep contextual learning. All were evaluated on the same preprocessed Steam dataset to ensure comparability.

Formally, given a review $x = (w_1, w_2, \dots, w_T)$, the goal is to predict a sentiment label $y \in \{0,1\}$(negative or positive). Each model approximates the posterior probability:

$$\hat{y} = \arg\max_{y} P(y \mid x; \theta)$$

where θrepresents the model parameters.

4.2 Rule-Based Model: VADER

Library: `vaderSentiment.vaderSentiment`

VADER computes sentiment using a lexicon of word valences $s(w_i) \in [-4,4]$. For a review with tokens w_i, the overall sentiment score is:

$$S = \sum_i \alpha_i s(w_i)\,, \text{compound} = \frac{S}{\sqrt{S^2 + 15}}$$

where α_i amplifies or diminishes word strength based on surrounding punctuation, capitalization, or negation.
The final prediction is mapped to binary sentiment:

$$\text{label} = \{ \begin{matrix} 1, & \text{if compound} > 0.5 \\ 0, & \text{otherwise} \end{matrix}$$

This method requires no training and serves as a benchmark for learned models.

4.3 Transformer Models

All transformer-based models are built upon **DistilBERT** (Sanh et al., 2019), a 6-layer compressed version of BERT retaining 97 % of its performance at 60 % of its size.
Each input review is tokenized, padded to 128 tokens, and converted to contextual embeddings:

$$\mathbf{X} = [\mathbf{e}_{[CLS]}, \mathbf{e}_{w_1}, \ldots, \mathbf{e}_{w_T}]$$

Within each attention head, relevance between tokens is computed via:

$$\text{Attention}(Q, K, V) = \text{softmax}\,(\frac{QK^\top}{\sqrt{d_k}})V$$

The contextualized [CLS] vector ($\mathbf{h}_{CLS}$) is passed to a classification head:

$$P(y = 1 \mid x) = \text{softmax}(W_c \mathbf{h}_{CLS} + b_c)_1$$

All transformer variants use **Cross-Entropy Loss**:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N} \sum_{c \in \{0,1\}} y_{ic} \log\,(\hat{y}_{ic})$$

A. Zero-Shot DistilBERT (Frozen Baseline)

Weights remain unchanged; only inference is performed using the pre-trained SST-2 model.

B. Full Fine-Tuning

All parameters θ trainable.
Optimizer: AdamW (lr = 2×10^{-5}); Epochs = 3.
All gradients flow through every transformer layer.

C. Head-Only Fine-Tuning

Encoder frozen; only classification head (W_c, b_c) updated.
Optimizer: AdamW (lr = 5×10^{-4}); Epochs = 3.

D. LoRA (Low-Rank Adaptation)

Instead of updating the full weight matrix $W_0 \in \mathbb{R}^{d\times k}$, LoRA injects a trainable low-rank offset:

$$W = W_0 + AB, \qquad A \in \mathbb{R}^{d\times r}, B \in \mathbb{R}^{r\times k}, r \ll d$$

Only A, B are trained (≈ 1 % of weights).
Hyperparameters: r = 8, α = 16, dropout = 0.1, lr = 1×10^{-3}.

4.4 LSTM + Attention + Length Feature

The LSTM baseline learns sequential dependencies.
Input embeddings (100 dimensions, 20 000 vocabulary) are passed through an LSTM followed by self-attention:

$$\alpha_{ij} = \frac{e^{h_i^\top h_j}}{\sum_k e^{h_i^\top h_k}}, \tilde{h}_i = \sum_j \alpha_{ij} h_j$$

A second LSTM encodes context. The final hidden state is concatenated with normalized review length L, then passed to a sigmoid layer:

$$\hat{y} = \sigma(W_o[h_{\text{final}}; L] + b_o)$$

This model introduces an auxiliary feature to test whether text length improves classification.

5. Experiments

5.1 Dataset and Pre-Processing

A balanced dataset of **50 000** reviews (25 k positive, 25 k negative) was extracted from the Steam corpus. Text was lowercased, stripped of non-ASCII symbols, and validated for missing values.

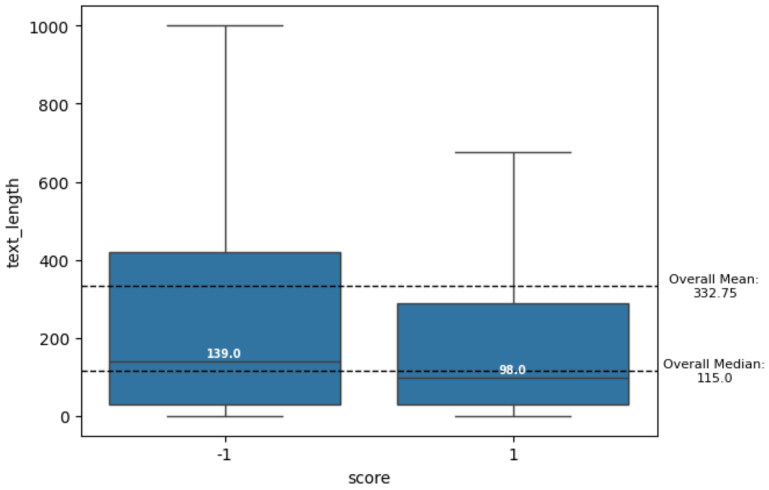

Figure 1: Review Length Distributions

Negative reviews are typically longer (median = 139 tokens) than positive ones (median = 98), indicating that dissatisfied users tend to write more.

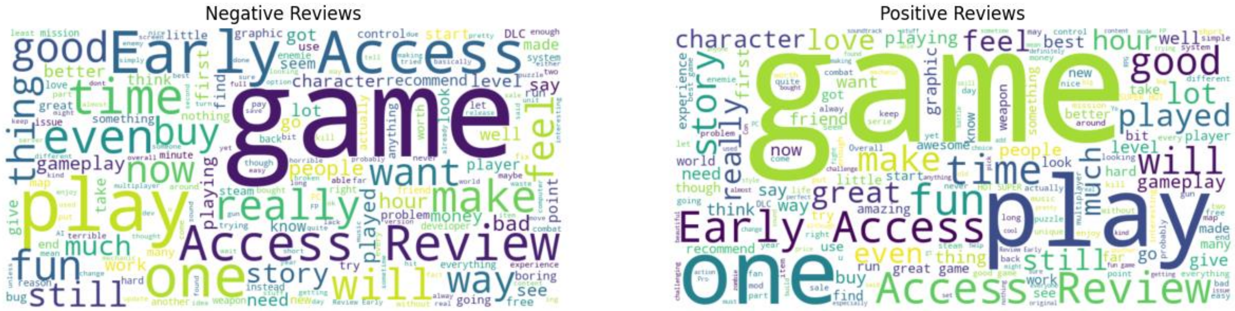

Figure 2: Word Clouds of Positive vs Negative Reviews

We generated word clouds from both positive and negative reviews to visualize dominant vocabulary. Common across both: “game”, “play”, “fun”, “early access”, “time”. Negative reviews contain problem-oriented words (“bug”, “fix”, “bad”, “crash”). Positive reviews highlight experience-oriented terms (“love”, “great”, “friends”, “character”).

5.2 Training Environment

All models were trained on a GPU environment.
Batch size = 16–32; epochs = 3 (Transformers) and 10 (LSTM with early stopping on validation PR-AUC).

5.3 Evaluation Protocol

Performance was assessed on an identical 20 % test split using:

- Average Precision (AP)
- Accuracy
- Maximum F1-score
- Best Threshold (derived from PR curve)

F1 was computed as:

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

6. Results & Analysis

6.1 Quantitative Results

Model	Avg Precision	Accuracy	Max F1	Best Threshold	Trainable %
VADER	0.699	0.620	0.701	0.49	–
Frozen DistilBERT	0.858	0.748	0.782	0.010	0
Full Fine-Tune	**0.917**	**0.804**	**0.826**	0.260	100
Head-Only	0.888	0.771	0.801	0.323	≈ 10
LoRA	0.905	0.790	0.814	0.458	1.09
LSTM + Attention	0.869	0.751	0.787	0.441	–

- The **full fine-tuned DistilBERT** achieved the highest scores across all metrics (F1 = 0.826, AP = 0.917), confirming that end-to-end domain adaptation yields the most robust performance.
- **LoRA**, despite training only ~1 % of weights, achieved 98 % of the full model's performance, demonstrating its effectiveness under computational constraints.
- **Head-only fine-tuning** was competitive (F1 = 0.801) but less adaptable, indicating that domain-specific sentiment cues reside within deeper encoder layers.
- The **LSTM + Attention** model performed strongly relative to VADER but fell short of transformer-based contextual models.
- **VADER**, relying purely on lexical rules, lagged significantly (F1 = 0.701), validating that static sentiment dictionaries struggle with sarcasm and slang.

6.2 Graphical Results (Precision–Recall Curves)

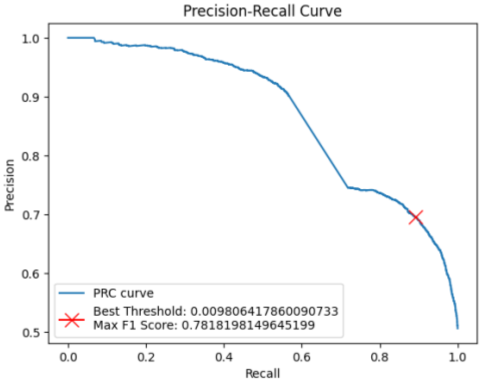

Figure 3: Zero-shot transformer (frozen DistilBERT)

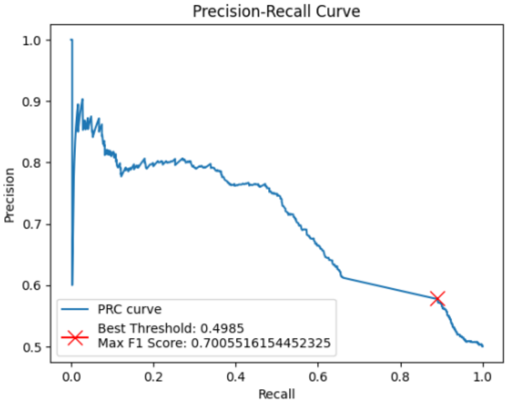

Figure 4: Lexicon-based (VADER)

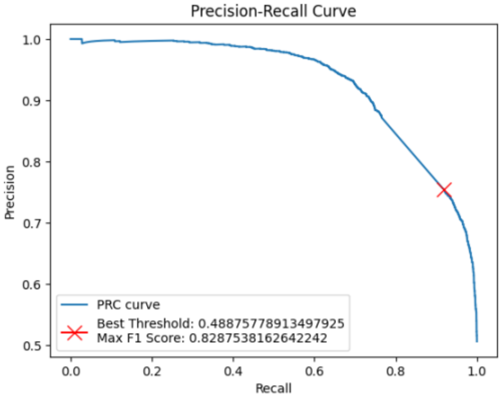

Figure 5: Fully fine-tuned transformer

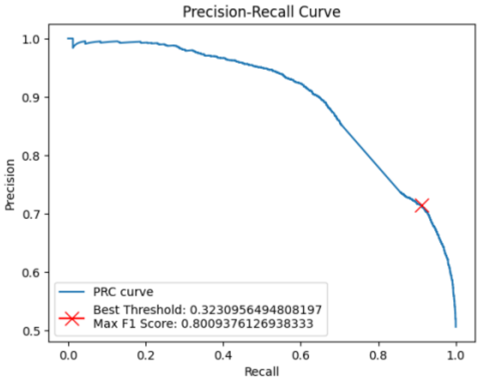

Figure 6: Head-Only fine-tuned transformer

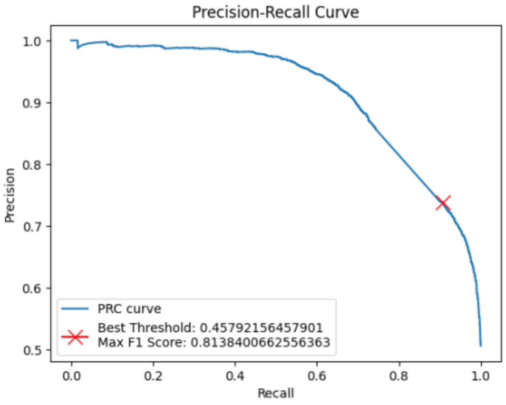

Figure 7: Parameter-efficient fine-tuning (LoRA)

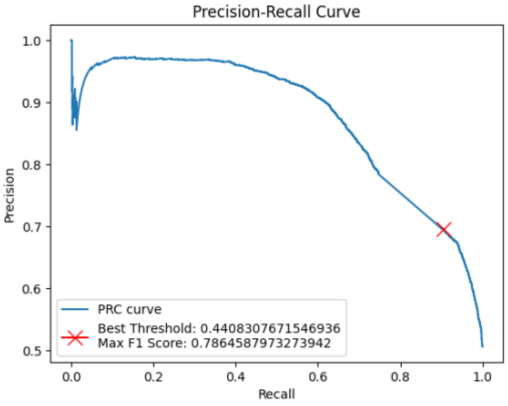

Figure 8: Classical deep model (LSTM + Attention)

6.3 Ablation Analysis

The series of fine-tuning variants functions as a structured ablation study.
Key findings include:

- **Impact of Adaptation Depth:**
  Freezing all encoder layers (Head-Only) reduces F1 by ≈ 2.5 points relative to Full Fine-Tuning, confirming that mid- and lower-layer adaptation contributes meaningfully to domain alignment.
- **Parameter Efficiency:**
  LoRA achieves 98 % of full fine-tuning performance while training only ≈ 1 % of weights. However, the runtime per epoch is almost the same as full fine-tuning, around thirty minutes on CPU. That's because LoRA still runs the same forward computations, the real savings are in memory and storage, not speed.
- **Comparison to Classical Architectures:**
  The LSTM + Attention model outperforms rule-based baselines but underperforms transformer variants, reinforcing that pre-trained contextual embeddings capture semantics far beyond sequential recurrence.

6.4 Qualitative Results and Error Analysis

To complement the quantitative evaluation, a qualitative inspection was performed on the **best-performing model — the fully fine-tuned DistilBERT**.
The goal was to identify residual weaknesses that numerical metrics alone could not capture.

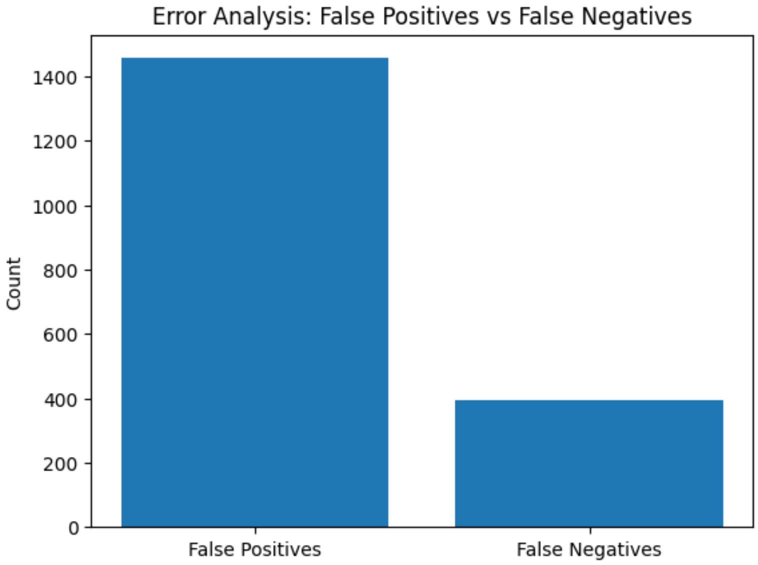

Figure 9: False Positives vs False Negatives

Out of the total test samples, the model produced **1,456 false positives** and **395 false negatives**.
This asymmetry suggests a mild bias toward predicting the positive ("recommended") class, a common tendency in review datasets where enthusiastic language dominates.
To understand the causes behind these misclassifications, we closely examined the highest-confidence **false positives** and **false negatives**.

False Positives

Most false positives displayed overtly positive wording but conveyed **sarcasm or ironic criticism**. Representative examples include:

- “Wow what a fun game you know! So fun! FUN FUN FUN FUN FUN FUN!”
- “Best game ever. I love this game more than my family. Can’t wait for chapter 2.”
- “Very good teenagers in heat simulator. I give it a 9/11.”

Although clearly negative to human readers, the model misclassified them as positive because it heavily weighted emotionally charged tokens such as *“fun”*, *“best”*, and *“love”*.
This pattern shows that even after domain-specific fine-tuning, DistilBERT interprets sentiment primarily at a **lexical level** and lacks sensitivity to **pragmatic cues** like sarcasm or parody that invert literal polarity.

False Negatives

In contrast, false negatives were typically **long, mixed, or contrastive reviews** that began critically but concluded with praise or recommendation.
Examples include:

- “It’s buggy and horribly optimized… but it’s such vibrant, chaotic fun that you’ll ignore the negatives.”
- “The game itself is incredible after fixes and with a controller.”

In these cases, the model was influenced more by the **early negative context** and underweighted later positive segments.
This reveals a **positional bias in attention**, where tokens appearing early in a long review dominate the sentiment representation, leading to errors when polarity reverses mid-text.

Interpretation

Across both categories, the misclassifications demonstrate that the fully fine-tuned DistilBERT excels at capturing **lexical sentiment patterns** but still lacks **pragmatic and discourse-level understanding**.
Fine-tuning successfully adapted the model to gaming-specific vocabulary (e.g., *“OP”*, *“nerfed”*, *“grindy”*), yet the model remains brittle when processing **humor, irony, or evolving sentiment** within multi-sentence reviews.

Overall, these results show that while fine-tuning enhances domain awareness, it does not inherently grant models the capacity to interpret tone, intent, or sarcasm—elements central to human-like sentiment understanding.

7 Discussion and Limitations

7.1 Overview

The preceding results demonstrate that fine-tuning pre-trained transformer models significantly improves sentiment classification accuracy on informal text such as Steam reviews. However, a closer inspection reveals subtler theoretical and practical implications. This section discusses how linguistic complexity, computational constraints, and modeling design interact to shape the outcomes observed in Section 6.

7.2 Interpretation of Findings

Effectiveness of Fine-Tuning.
Full DistilBERT fine-tuning yielded the highest overall performance (F1 = 0.826, AP = 0.917). The improvement over zero-shot inference (F1 = 0.782) confirms that domain adaptation—however limited in data size—enables the model to realign its internal attention toward task-specific lexical and syntactic patterns. In particular, the model learns to reinterpret gaming terms such as *"grindy"* or *"nerfed"*, which would otherwise remain semantically neutral to a general-domain model.

Efficiency of Parameter-Reduced Tuning.
LoRA achieved F1 = 0.814 with only 1 % of trainable parameters, validating that sentiment information resides in a low-rank subspace of the full weight matrix. This result carries practical significance: comparable accuracy can be achieved under CPU-only conditions without expensive GPU infrastructure. In contexts such as low-resource academic labs or real-time industry systems, LoRA provides a sustainable alternative to exhaustive retraining.

Relative Role of Representation Depth.
The gap between head-only fine-tuning (F1 = 0.801) and full fine-tuning highlights that the middle transformer layers contribute most to domain generalization. Freezing them restricts semantic adaptation, while fully updating them allows sentiment polarity to propagate through the entire contextual hierarchy.

Comparative Baselines.
The LSTM + Attention baseline (F1 = 0.787) remains substantially stronger than rule-based VADER but weaker than transformer models, demonstrating that static word embeddings and sequential recurrence alone cannot encode discourse-level contrast or sarcasm. The relative ranking—Transformers > LSTM > Lexicon—echoes established trends in contemporary NLP but within a domain that has not been extensively tested.

7.3 Error Behavior and Linguistic Challenges

The qualitative inspection (Section 6.4) clarified that the full fine-tuned model primarily fails on **sarcasm, contrastive sentiment, and long multi-clause structures**.
These cases expose two deeper limitations:

1. **Literal Sentiment Modeling.**
   The model still equates lexical polarity with emotional intent. Words such as *"great"* or *"fun"* dominate representation even when embedded in ironic contexts.
2. **Contextual Myopia.**
   Attention distributions favor tokens at the beginning of a review, producing misclassifications when sentiment reverses mid-text.
   This bias is exacerbated by the review length disparity observed during EDA, where negative reviews tend to be longer and more elaborate.

Both phenomena suggest that transformer encoders fine-tuned purely on categorical sentiment labels may not capture pragmatic relations that extend beyond sentence boundaries.

7.4 Practical Limitations

1. **Compute Environment.**
   All experiments were conducted on CPU. While adequate for DistilBERT and LoRA, the absence of GPU acceleration limited batch size, epoch tuning, and exploration of larger model variants (e.g., RoBERTa, DeBERTa).
2. **Dataset Scope.**
   The dataset, though balanced at 50 000 samples, represents a narrow slice of gaming discourse and may not generalize to other informal platforms such as Reddit or Discord.
3. **Annotation Noise.**
   The "recommended / not recommended" tags on Steam can reflect not only enjoyment but also value for money or jokes, which makes the labels less clear.

7.5 Reflection and Broader Implications

The progression from VADER → Frozen → Full → LoRA fine-tuning illustrates a continuum between interpretability and adaptability.
While rule-based systems remain transparent but rigid, transformer models learn flexible contextual representations at the expense of explainability.
LoRA provides a middle ground—retaining interpretability of the base model while enabling domain-specific tuning through limited parameters.
From a research standpoint, this study affirms that **model efficiency and linguistic robustness need not be mutually exclusive**; thoughtful architectural modification can yield both.

8 Conclusion and Next Steps

8.1 Summary of Contributions

This study systematically evaluated six sentiment analysis paradigms on a large corpus of Steam game reviews, spanning rule-based, recurrent, and transformer architectures.
Through consistent preprocessing, tokenization, and evaluation protocols, it provides a controlled comparison of how fine-tuning depth influences performance in a highly informal, sarcasm-rich domain.

Key outcomes include:

- **Full Fine-Tuning Superiority:** DistilBERT fine-tuned end-to-end achieved the strongest results (AP = 0.917, F1 = 0.826).
- **LoRA Efficiency:** Parameter-efficient adaptation reached near-equivalent accuracy (AP = 0.905, F1 = 0.814) with ≈ 1 % of trainable weights.
- **Contextual vs Lexical Modeling:** Transformers surpassed both LSTM + Attention and VADER by over 10 F1 points, highlighting the advantage of contextualized embeddings in noisy user-generated text.
- **Persistent Pragmatic Gaps:** Even the best models misinterpret sarcasm, contrastive phrasing, and humor—demonstrating limits of purely text-level sentiment learning.

8.2 Implications

The findings confirm that **domain-specific fine-tuning is not optional but essential** when applying large language models to informal online content.
Moreover, the success of LoRA illustrates that efficient adaptation can democratize transformer research by lowering computational and environmental costs.
From a methodological perspective, combining quantitative metrics with qualitative inspection provided a more comprehensive understanding of model behavior, a practice that should become standard in applied NLP evaluation.

8.3 Future Work

1. **Domain-Adaptive Pre-Training (DAPT):**
   Further pre-train DistilBERT on unlabeled Steam text to enrich its latent vocabulary before supervised fine-tuning.
2. **Sarcasm-Aware Learning:**
   Integrate auxiliary sarcasm detection or contrastive sentiment objectives to handle ironic polarity inversion.
3. **Hierarchical or Document-Level Modeling:**
   Employ hierarchical transformers or memory-based attention to capture sentiment progression across multiple clauses or paragraphs.
4. **Cross-Domain Transfer:**
   Evaluate the fine-tuned and LoRA models on other review datasets (IMDB, Amazon) to assess generalizability beyond gaming.

8.4 Concluding Remark

Overall, this project demonstrates that **contextual language models can meaningfully decode the chaotic sentiment landscape of gaming communities** when properly fine-tuned, even under constrained resources.
Yet the persistent errors on sarcasm and mixed tone reveal that true sentiment understanding remains an open challenge, one that future research must tackle through richer context modeling and multi-task pragmatic learning.

Github Repository: **https://github.com/tzy815/DSA4213-project**

Reference:

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., & Google AI Language. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019*. Association for Computational Linguistics. https://aclanthology.org/N19-1423.pdf

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021, June 17). *LORA: Low-Rank adaptation of Large Language Models*. arXiv.org. https://arxiv.org/abs/2106.09685

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019, October 2). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. arXiv.org. https://arxiv.org/abs/1910.01108

AI Tool Declaration

We used AI tool like Chatgpt 5 to help clean and debug our codes, and to make our report sound more professional. We are responsible for the content and quality of the submitted work.