From Binary to Nuanced: A Deep Learning Study of Three-Class Sentiment Analysis in Movie Reviews

Samuel Veliveli
University of Virginia
School of Data Science
Charlottesville, Virginia
rrs4bw@virginia.edu

Anthony J. Ventura
University of Virginia
School of Data Science
Charlottesville, Virginia
kyg8cd@virginia.edu

Patrick Maloon
University of Virginia
School of Data Science
Charlottesville, Virginia
kvq3jb@virginia.edu

Abraham Tedla
University of Virginia
School of Data Science
Charlottesville, Virginia
wqp7qy@virginia.edu

***Abstract*—Binary sentiment classification oversimplifies the nuanced opinions prevalent in real-world text. This paper reframes IMDb movie review sentiment analysis as a three-class problem (bad, neutral, and good) using rating annotations as weak supervision labels. The full dataset contains 108,133 reviews (train: 75,693 / val: 16,220 / test: 16,220), split under a wide-neutral label scheme (1–3 bad, 4–7 neutral, 8–10 good) yielding a class distribution of 36.7% bad, 19.6% neutral, and 43.7% good. We evaluate five model families—Logistic Regression, TextCNN, Bidirectional LSTM, DistilBERT, and BERT-base—with and without inverse-frequency class weighting, across three random seeds (mean ± std). BERT-base with class weighting or last-2-layer fine-tuning achieves the best macro-F1 of 0.784 ± 0.002 (accuracy $81.7 \pm 0.5\%$), followed by DistilBERT (0.779 ± 0.003), Logistic Regression (0.750), BiLSTM (0.723 ± 0.003), and TextCNN (0.708 ± 0.004). Logistic Regression outperforms both deep neural baselines—a meaningful negative result confirmed by a GloVe ablation showing no improvement over random embeddings for TextCNN. Focal loss ($\gamma = 2$) achieves the highest neutral-class F1 (0.630 ± 0.008). Sequence length ablations confirm that longer contexts monotonically improve neutral recall ($+4$ pp, 128→512 tokens). A binary reference experiment quantifies the cost of adding the neutral class at ∼11 accuracy points and ∼14 macro-F1 points. Qualitative error analysis identifies two dominant failure modes at the neutral boundary: polar-language reviews with moderate ratings, and hedged-praise reviews misclassified as neutral. Code is available at https://github.com/abrish2049/nuanced-sentiment-classification.**

I. Introduction

Sentiment analysis is a foundational task in natural language processing, with applications in recommendation systems, content moderation, and market intelligence. The dominant benchmark paradigm on IMDb frames sentiment as a binary decision: positive or negative. While effective in controlled settings, this formulation discards the nuanced middle ground that characterises a significant fraction of real user-generated text.

In practice, reviewers frequently express mixed or moderate opinions, giving a film 6 out of 10—neither enthusiastic nor dismissive. Binary classifiers must assign such reviews to an extreme label, losing information relevant to downstream recommendation quality. Moreover, the standard IMDb-50K benchmark [4] excludes ratings 5–6 by design, artificially inflating binary accuracy by removing the hardest cases.

This work reframes the problem as three-class classification (bad, neutral, good) using a large-scale IMDb dataset with numerical rating annotations. Our contributions are: (1) a reproducible three-class IMDb evaluation framework with stratified splits and multi-seed variance reporting; (2) a systematic comparison of five model families under weighted and unweighted loss; (3) ablation studies on label thresholds, focal loss, fine-tuning strategy, sequence length, and GloVe embeddings; (4) confusion matrix analysis and qualitative error inspection revealing two failure modes at the neutral boundary; (5) TF-IDF feature analysis identifying the linguistic signature of the neutral class; and (6) quantification of the performance cost of adding the neutral class relative to the binary baseline.

II. Literature Review

Hutto and Gilbert (2014) present VADER, a rule-based sentiment model combining a validated lexicon with punctuation, capitalisation, and negation rules [1]. Its interpretability and zero training-data requirement make it a practical baseline, but dependence on predefined lexical knowledge limits generalisation to the hedged, mixed-opinion language prevalent in our neutral class.

Maas et al. (2011) established the standard IMDb-50K dataset, which excludes ratings 5–6 by design to enforce a clean binary split [4]. Our dataset retains these middle-rated reviews, making three-class classification both possible and non-trivial.

Kim (2014) demonstrated that a simple CNN on pretrained word embeddings achieves strong sentence-level sentiment performance using parallel convolutional filters and max-over-time pooling [3]. Convolutional models excel at local n-gram patterns but struggle with long-range dependencies in document-length reviews.

Socher et al. (2013) introduced SST-5 and a recursive neural tensor network that computes sentiment over syntactic parse trees [6]. By labelling constituent phrases, their work establishes that compositional structure drives sentiment. Our

neutral class shares SST-5's challenge of distinguishing near-neutral text from mildly polar text.

Qaisar (2020) applied LSTMs to full IMDb review classification, showing that gated recurrent mechanisms better capture document-level context than CNNs [5]. Devlin et al. (2019) introduced BERT, whose bidirectional Transformer pretraining transfers effectively to classification with minimal task-specific architecture [2], making it well-suited to long reviews where sentiment cues span many sentences. Together these works trace the progression from lexical baselines through CNNs, RNNs, and Transformers that we evaluate empirically.

III. Dataset and Label Construction

A. Dataset Verification

We use the IMDB Reviews dataset available on Hugging Face,[1] containing reviews paired with integer ratings 1–10. Unlike the standard Maas et al. IMDb-50K benchmark, this dataset retains ratings 5–6, verified prior to training. Figure 1 shows the full rating distribution, confirming IMDb's well-known bimodal structure: ratings 1–2 and 8–10 dominate, with 5–6 ratings forming a small minority. This motivates the wide-neutral scheme, which groups ratings 4–7 to assemble a sufficiently large neutral class for reliable learning.

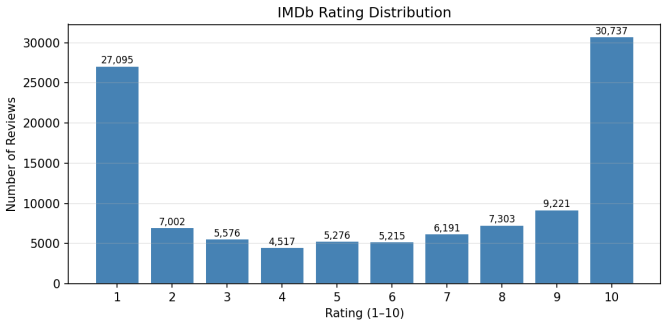

Fig. 1. IMDb rating distribution across all 108,133 reviews. The bimodal structure motivates the wide-neutral label scheme (4–7 neutral) to assemble a learnable neutral class.

B. Label Mapping and Class Distribution

The full dataset contains 108,133 reviews. We apply a wide-neutral label scheme: Bad (1–3), Neutral (4–7), Good (8–10), yielding the class distribution in Table I. The dataset is split via stratified sampling into training (75,693), validation (16,220), and test (16,220) sets.

TABLE I
Class distribution under wide-neutral scheme.

Class	Count	%
Bad	39,673	36.7
Neutral	21,199	19.6
Good	47,261	43.7
Total	108,133	100.0

[1]https://huggingface.co/datasets/Daksh0505/IMDB-Reviews

C. Class Weighting

Inverse-frequency class weights are computed as $w_c = N/(K \cdot n_c)$, where N is total training samples, $K = 3$, and n_c is the count for class c. The resulting weights are bad: 0.909, neutral: 1.700, good: 0.763. The neutral weight of $\approx 2\times$ either polar class confirms meaningful imbalance, motivating both class-weighting and focal loss experiments.

IV. Method

A. Model Architectures

We compare five model families of increasing complexity.

Logistic Regression (LR): TF-IDF features (50K features, unigrams and bigrams, sublinear TF scaling) with L2-regularised logistic regression (solver: lbfgs, $C = 1.0$). Serves as a non-neural performance baseline.

TextCNN: Kim (2014)-style CNN with 300-dimensional embeddings, parallel convolutional filters of widths $\{3, 4, 5\}$ (100 filters each), max-over-time pooling, dropout(0.5), and a 3-way output layer. Trained with Adam (lr $= 10^{-3}$), 15 epochs, batch size 64.

Bidirectional LSTM (BiLSTM): 300-dimensional embeddings into a 2-layer BiLSTM (hidden size 256 per direction). Final hidden states concatenated, passed through dropout(0.5), and fed to a 3-way output layer. Trained with Adam (lr $= 10^{-3}$), 15 epochs, batch size 64.

DistilBERT-base-uncased: 6-layer distilled Transformer (66M parameters) with a linear classification head on the `[CLS]` token. Fine-tuned with AdamW (lr $= 2\times10^{-5}$), linear warmup (10% of steps), batch size 16, 5 epochs.

BERT-base-uncased: 12-layer Transformer encoder (110M parameters) with the same classification head and training protocol as DistilBERT, fine-tuned for 5 epochs.

B. Loss Functions

Each model is evaluated under: (a) unweighted cross-entropy and (b) inverse-frequency class weighting. BERT is additionally evaluated under (c) focal loss with class weights:

$$\mathcal{L}_{\text{FL}} = -\sum_{c} w_c (1 - p_c)^{\gamma} \log p_c, \quad \gamma = 2. \tag{1}$$

Focal loss down-weights easy examples and concentrates gradient signal on hard ones—particularly relevant for the neutral class, which overlaps linguistically with both polar classes.

C. Fine-tuning Strategy Ablation

For BERT we compare three strategies: (a) *full fine-tuning*—all 12 encoder layers updated; (b) *last-2-layers*—only encoder layers 10–11 and the classification head updated; and (c) *head-only*—all encoder layers and embeddings frozen. This tests how deeply the model must adapt its representations to handle the neutral class.

D. Training Protocol

All models are trained on an NVIDIA A100 GPU. For each model, the checkpoint with the highest validation macro-F1 is selected for test evaluation. Transformer models use gradient clipping (max norm = 1.0). All neural models are run with three random seeds (42, 43, 44); results are reported as mean ± std. Logistic Regression is deterministic.

V. Results

A. Main Results

Table II summarises test-set performance across all model–condition pairs at max_len = 512. Figure 2 shows the grouped comparison. The binary reference (LR, neutral reviews dropped) yields accuracy 92.4% and macro-F1 0.923.

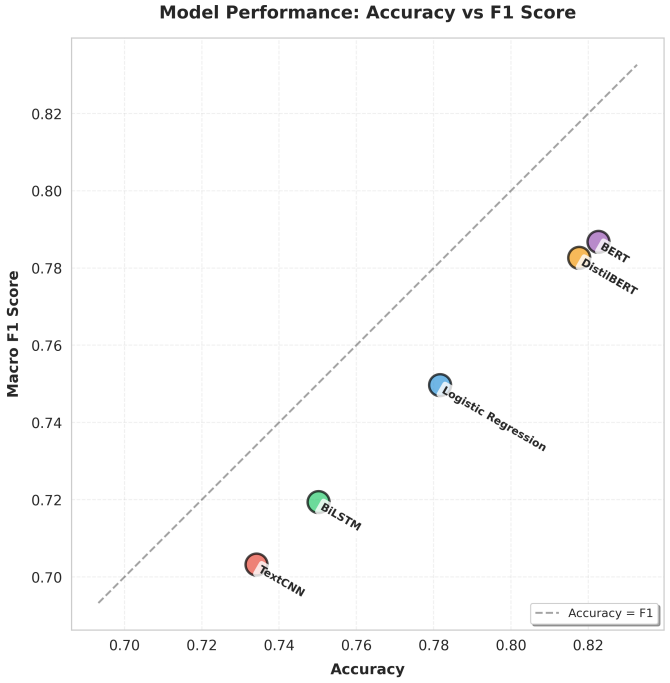

Fig. 2. Accuracy vs. Macro-F1 scatter for all five models (class-weighted, len=512). The dashed line marks Accuracy = F1. BERT and DistilBERT cluster in the top-right; LR sits above both CNN and LSTM baselines despite its simplicity.

B. Key Observations

BERT and DistilBERT lead all models. BERT achieves macro-F1 0.781–0.784 and DistilBERT 0.779, both well ahead of BiLSTM (0.723, +6 pp). Pretraining provides the largest single gain, concentrated in neutral F1 (0.612–0.630 vs. 0.513–0.545 for BiLSTM). DistilBERT nearly matches BERT within 0.5 pp macro-F1 at 66M vs. 110M parameters.

LR outperforms TextCNN and BiLSTM. LR macro-F1 0.741–0.750 beats TextCNN (0.708) and BiLSTM (0.723). The GloVe ablation (Section VI-E) rules out embedding quality; the explanation lies in TF-IDF bigram sufficiency, document-level bag-of-words advantage, and CNN/LSTM overfitting at 15 epochs.

Focal loss best for neutral. BERT with focal loss achieves the highest neutral F1 across all conditions (0.630 ± 0.008), confirming that down-weighting easy polar examples directly benefits the minority class at a small accuracy cost (81.1 vs. 81.7%).

Last-2-layers ≈ full fine-tuning. Last-2-layer BERT (0.784 ± 0.001) is statistically indistinguishable from full fine-tuning (0.784 ± 0.002) and exhibits lower variance. Head-only loses ∼11 pp macro-F1, confirming encoder adaptation is necessary.

Neutral is the consistent bottleneck. Neutral F1 is 15–25 pp below bad/good in every model–condition pair, reflecting linguistic overlap between mildly polar and genuinely neutral reviews.

Class weighting consistently improves neutral F1. Gains range from +0.046 (LR) to +0.013 (DistilBERT) at a modest 1–2 pp accuracy cost. Larger models show diminishing returns, as stronger representations partially compensate for imbalance.

Cost of nuance is large. The binary reference yields 92.4% accuracy and macro-F1 0.923 vs. best three-class 81.7% and 0.784—a ∼11 accuracy-point and ∼14 macro-F1-point gap.

Compute cost contextualises accuracy gains. Training time per seed on a single A100 GPU: LR (<1 min), TextCNN (∼66 min), BiLSTM (∼18 min), DistilBERT (∼21 min), BERT (∼41 min). DistilBERT matches BERT within 0.5 pp macro-F1 at roughly half the training time and 60% of the parameters, making it the preferred efficiency–accuracy trade-off. LR delivers macro-F1 0.750 in seconds, outperforming TextCNN and BiLSTM at orders-of-magnitude less compute—the added cost of CNN and LSTM training is not justified on this task.

C. Confusion Matrix Analysis

Figure 3 shows confusion matrices for LR, BiLSTM, BERT, and DistilBERT (class-weighted, len=512). All matrices share the same structure: bad and good are classified reliably, while the neutral column receives substantial spillover from both polar classes. Neutral reviews are more often misclassified as bad than as good, consistent with the error analysis finding that many neutral reviews use predominantly negative language.

Figure 4 shows the BERT fine-tuning strategy ablation. Head-only shows the most neutral confusion; focal loss visibly improves neutral recall over the unweighted baseline.

D. Training Curves

Figure 5 shows training curves for BERT and BiLSTM (class-weighted, len=512). BiLSTM exhibits classic overfitting: validation macro-F1 peaks at epochs 3–5 then degrades while training loss continues to fall. BERT shows stable convergence through epoch 5, consistent with the regularising effect of large-scale pretraining.

VI. Ablation Studies

A. Label Threshold Ablation

Table III compares three label threshold schemes using LR. Neutral F1 is inversely related to neutral class size. The narrow scheme (4.8% neutral, rating 6 only) yields the highest

TABLE II

TEST-SET RESULTS (WIDE-NEUTRAL SCHEME, MEAN ± STD, 3 SEEDS, MAX_LEN = 512). BINARY REFERENCE (LR, BAD/GOOD ONLY): ACC = 0.924, MACRO-F1 = 0.923. BOLD = BEST PER METRIC AMONG ALL MODELS.

Model	Variant	Acc	mF1	$F1_{bad}$	$F1_{neu}$	$F1_{good}$
LR	No weighting	.793	.741	.831	.536	.857
	Class weights	.782	.750	.820	.582	.848
TextCNN	No weighting	.749±.001	.708±.002	.789±.006	.508±.005	.827±.005
	Class weights	.742±.009	.708±.004	.783±.009	.520±.005	.820±.008
BiLSTM	No weighting	.773±.002	.723±.002	.813±.003	.513±.004	.844±.001
	Class weights	.755±.003	.723±.003	.789±.003	.545±.006	.835±.006
DistilBERT	No weighting	.818±.003	.779±.008	.856±.003	.605±.021	.876±.002
	Class weights	.815±.003	.779±.003	.850±.005	.612±.005	.876±.002
BERT	No weighting	.818±.002	.781±.003	.850±.006	.612±.016	**.881±.001**
	Class weights	.817±.005	**.784±.002**	.853±.005	.625±.008	.876±.004
	Focal (γ=2)	.811±.003	.783±.003	.848±.004	**.630±.008**	.871±.002
	Head-only	.705±.002	.672±.001	.731±.001	.496±.001	.790±.002
	Last-2-layers	**.817±.002**	**.784±.001**	**.852±.002**	.624±.002	.876±.001

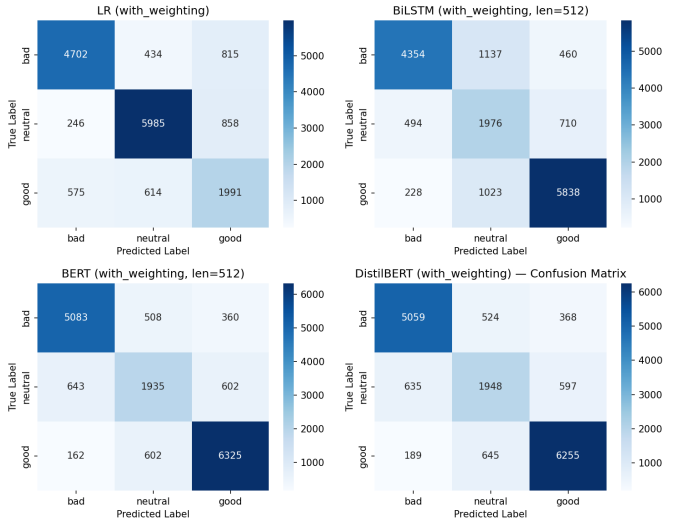

Fig. 3. Confusion matrices (class-weighted, len=512): LR (top-left), BiLSTM (top-right), BERT (bottom-left), DistilBERT (bottom-right). Neutral is the most-confused class; bad→neutral and neutral→bad are the dominant off-diagonal errors.

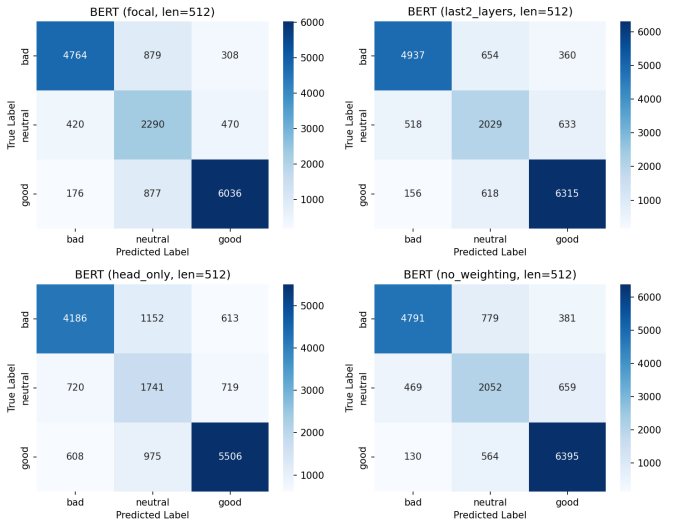

Fig. 4. BERT ablation confusion matrices (len=512, seed 42): Focal loss (top-left), Last-2-layers (top-right), Head-only (bottom-left), No weighting (bottom-right).

accuracy (85.2%) but $F1_{neu}$ collapses to 0.027 unweighted and only 0.256 weighted. Wide neutral (19.6%, ratings 4–7) reduces accuracy by ∼3 pp but raises $F1_{neu}$ to 0.582—more than twenty times the narrow unweighted baseline. The label threshold is the single most impactful design decision for neutral-class performance.

B. Sequence Length Ablation

Table IV reports macro-F1 and neutral F1 for BiLSTM, DistilBERT, and BERT across max_len ∈ $\{128, 256, 512\}$. Neutral F1 gains from 128→512: BiLSTM +3.3 pp, DistilBERT +3.8 pp, BERT +3.9 pp. The consistent monotonic improvement confirms the hypothesis that mixed-sentiment reviews distribute cues throughout the document; truncation

TABLE III

LABEL THRESHOLD ABLATION (LR). ROW PAIRS: UNWEIGHTED / WEIGHTED.

Scheme	Neu.%	Acc	mF1	$F1_{neu}$
Narrow (6 only)	4.8%	.852 .811	.591 .654	.027 .256
Default (5–6)	9.7%	.828 .795	.651 .701	.221 .406
Wide neutral (4–7)	19.6%	.793 .782	.741 .750	.536 .582

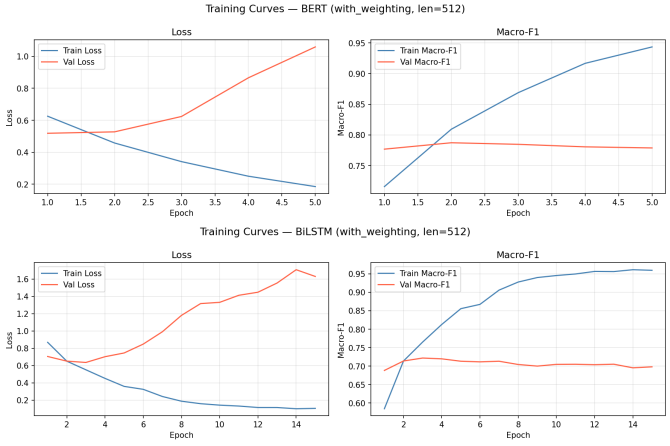

Fig. 5. Training curves (class-weighted, len=512): BERT (top) and BiLSTM (bottom). BiLSTM overfits after epoch 5; BERT converges stably.

at 128 tokens disproportionately hurts neutral recall. Figure 6 illustrates the effect on BERT training dynamics.

TABLE IV
SEQUENCE LENGTH ABLATION (CLASS-WEIGHTED, MEAN ± STD, 3 SEEDS).

Model	len	mF1	$F1_{neu}$
BiLSTM	128	.702±.002	.512±.007
	256	.718±.002	.532±.007
	512	.723±.003	.545±.006
DistilBERT	128	.749±.003	.574±.004
	256	.772±.002	.601±.008
	512	.779±.003	.612±.005
BERT	128	.760±.001	.586±.002
	256	.775±.003	.605±.011
	512	.784±.002	.625±.008

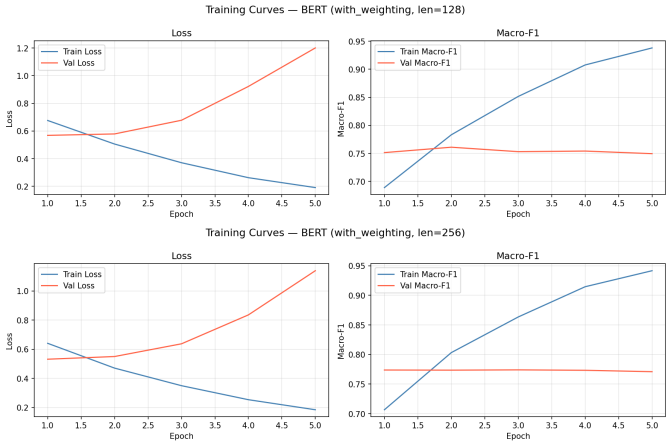

Fig. 6. BERT training curves at len=128 (top) and len=256 (bottom), class-weighted. Longer sequences yield higher peak validation F1.

C. Fine-tuning Strategy Ablation

Table V compares BERT fine-tuning strategies at len=512. Last-2-layer fine-tuning (0.784 ± 0.001) is statistically indistinguishable from full fine-tuning (0.784 ± 0.002) and achieves lower variance, making it the preferred configuration for efficiency. Head-only fine-tuning loses ∼11 pp macro-F1, confirming that frozen BERT representations are insufficient for the three-class task—at least the top encoder layers must adapt to distinguish neutral from polar sentiment.

TABLE V
BERT FINE-TUNING STRATEGY ABLATION (LEN=512, MEAN ± STD, 3 SEEDS, CLASS-WEIGHTED).

Strategy	Acc	mF1	$F1_{neu}$
Full fine-tuning	.817±.005	.784±.002	.625±.008
Last-2-layers	.817±.002	.784±.001	.624±.002
Head-only	.705±.002	.672±.001	.496±.001

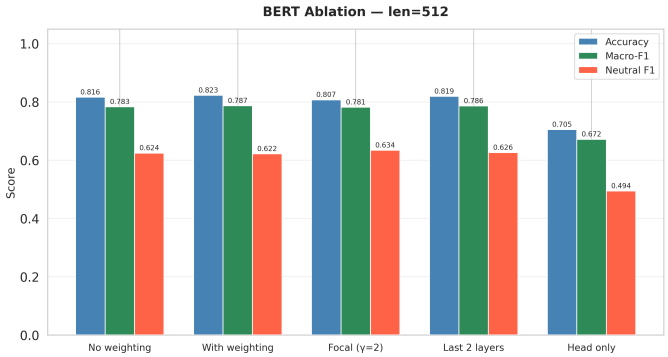

Fig. 7. BERT ablation across all five variants (len=512, seed 42): Accuracy, Macro-F1, and Neutral F1. Head-only shows the largest drop; focal loss achieves the highest neutral F1 (0.634).

D. Focal Loss Ablation

Focal loss ($\gamma = 2$, with class weights) achieves the highest neutral F1 of any configuration (0.630 ± 0.008) while maintaining macro-F1 0.783 ± 0.003—within noise of the class-weighted baseline (0.784 ± 0.002). The slight accuracy cost (81.1 vs. 81.7%) reflects a deliberate redistribution of gradient signal toward the neutral class. For applications where neutral F1 is the primary objective, focal loss is the recommended loss function.

E. TextCNN GloVe Ablation

Table VI compares TextCNN with random versus frozen GloVe-300d embeddings. GloVe provides no improvement in macro-F1 (0.699 vs. 0.703) or neutral F1 (0.483 vs. 0.488), while reducing training time by ∼30%. Figure 8 confirms the confusion matrices are nearly identical.

This result directly addresses why LR outperforms TextCNN: since GloVe does not help, embedding quality is not the explanation. Three factors combine: (1) TF-IDF bigrams on 108K reviews capture sentiment-bearing phrases (e.g., "not good," "highly recommend") without positional assumptions—inspection of the top-weighted features confirms that neutral-class coefficients are dominated by hedging bigrams and modal qualifiers, not merely the

TABLE VI
TEXTCNN GLOVE ABLATION (SINGLE SEED, LEN=512).

Embedding	Variant	Acc	mF1	$F1_{neu}$
Random	No weighting	.752	.703	.488
	Class weights	.741	.708	.530
GloVe-300d	No weighting	.751	.699	.483
	Class weights	.746	.706	.521

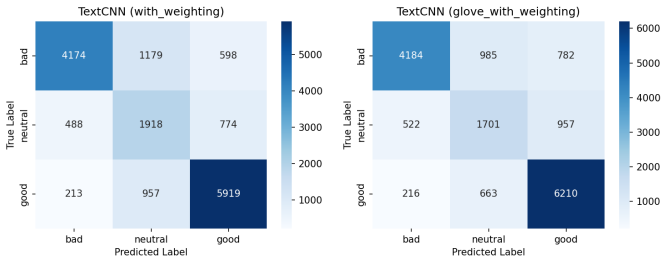

Fig. 8. TextCNN confusion matrices: random embeddings (left) vs. GloVe-300d frozen (right), both class-weighted. Nearly identical, confirming embedding quality is not the bottleneck.

absence of sentiment words; (2) document-length reviews distribute sentiment cues throughout the text, favouring bag-of-words over position-sensitive models; and (3) L2-regularised LR generalises more stably than CNN and LSTM at 15 epochs, as the training curves show clear overfitting for both neural baselines.

VII. QUALITATIVE ERROR ANALYSIS

We manually inspected 30 misclassified reviews from BERT (class-weighted, len=512, seed 42)—10 per true class—and identified two dominant failure modes at the neutral boundary.

Failure Mode 1 — Polar language with moderate rating (rating inconsistency). The most common error pattern involves reviews where the textual sentiment and the numerical rating are misaligned. For the bad class, reviewers frequently use positive or nostalgic language despite giving low scores:

> *"When i was 7 i thought this film was great! I loved every minute of it!!"* [Rating: 2, True: bad, Pred: neutral]

The model correctly responds to positive surface language but cannot recover the sentiment from the rating. Similarly, some neutral reviews use extreme negative language despite a moderate rating:

> *"YUK. Mindless drivel…I regretted WASTING my time with this mess!"* [Rating: 4, True: neutral, Pred: bad]

Here the model is arguably more faithful to the text than to the reviewer's own numerical rating—a fundamental ambiguity in rating-supervised weak labels.

Failure Mode 2 — Hedged praise misclassified as neutral. Good-class reviews that qualify their praise are systematically pulled toward neutral. The model correctly identifies the hedging language but cannot infer that the film was ultimately positive for this reviewer:

> *"I don't think Shawshank is perfect…it would not make my own top 10, despite all the compliments I could make about it."* [Rating: 9, True: good, Pred: neutral]

> *"Really well crafted modern action film that shot adrenaline through the veins but the music choice is horrific."* [Rating: 9, True: good, Pred: neutral]

These examples reveal a ceiling on rating-supervised sentiment learning: the model cannot distinguish genuine ambivalence from rhetorical hedging without understanding reviewer intent. Both failure modes explain why neutral F1 plateaus near 0.63 even for BERT at len=512—the errors are structural ambiguities in the weak supervision signal itself, not modelling failures. Resolving them would require additional supervision beyond rating labels, such as phrase-level annotations or explicit hedging detection.

VIII. FUTURE WORK

The results and error analysis surface several concrete directions for extending this work.

Phrase-level and hedging supervision. Both identified failure modes—rating-inconsistent reviews and hedged-praise reviews—arise from ambiguities inherent in weak rating-based labels. A natural next step is to augment training with phrase-level sentiment annotations (e.g., from SST-5 or aspect-based sentiment datasets) or to train an explicit hedging detector as an auxiliary task. Incorporating such signals could push neutral F1 beyond the ≈ 0.63 ceiling observed across all BERT configurations.

Larger pretrained models. This study evaluated BERT-base (110M) and DistilBERT (66M). Scaling to RoBERTa-large, DeBERTa-v3, or domain-adapted models (e.g., CineBERT) may yield additional gains, particularly for neutral-class disambiguation. Given that last-2-layer fine-tuning matched full fine-tuning in our setting, parameter-efficient methods such as LoRA or prefix tuning could make large-model experiments tractable.

Soft label training. Because IMDb ratings encode continuous preference intensity, treating labels as hard categorical targets discards ordinal information. Replacing one-hot targets with soft labels derived from the rating (e.g., a rating of 5 receives partial probability mass on both neutral and bad) may reduce overconfidence on boundary cases and improve calibration.

Multi-task and cross-domain generalization. The current framework is trained and evaluated entirely within the IMDb domain. Future work should test zero-shot and few-shot transfer to other review corpora (Yelp, Amazon, Rotten Tomatoes) under the same three-class scheme, and explore multi-task objectives that jointly train sentiment classification with related tasks such as rating prediction or review helpfulness scoring.

Interpretability and feature analysis. TF-IDF inspection confirms that neutral-class coefficients are dominated by hedging bigrams and modal qualifiers. A systematic attention-head analysis of BERT—identifying which layers encode sentiment polarity vs. hedging vs. intensity—could both explain current errors and inform targeted architectural modifications. Integrated gradients or SHAP attribution across the full model family would extend the qualitative error analysis to a rigorous quantitative account.

Improved class-imbalance strategies. While inverse-frequency weighting and focal loss both improve neutral F1, the gains are modest relative to the gap between neutral and polar classes. Future work could explore curriculum learning (presenting easy polar examples first, then hard neutral ones), oversampling with back-translation to generate synthetic neutral reviews, or contrastive learning objectives that explicitly separate neutral from polar representations in embedding space.

IX. Member Contributions

Samuel Veliveli: BiLSTM implementation and training; class-weighting experiments; training curve logging; sequence length ablation for BiLSTM.

Anthony J. Ventura: DistilBERT fine-tuning; AdamW optimiser configuration; compute resource management; sequence length ablation for DistilBERT.

Patrick Maloon: BERT-base fine-tuning; fine-tuning strategy ablation (head-only, last-2-layers, focal loss); sequence length ablation for BERT.

Abraham Tedla: Logistic Regression and TextCNN baselines; GloVe embedding ablation; dataset verification and preprocessing pipeline; label threshold ablation; unified training pipeline development; multi-seed aggregation; qualitative error analysis.

Shared: class-weighting ablation; result analysis and write-up.

References

[1] C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," *Proceedings of ICWSM*, 2014.

[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of NAACL*, 2019.

[3] Y. Kim, "Convolutional neural networks for sentence classification," *Proceedings of EMNLP*, 2014.

[4] A. L. Maas et al., "Learning word vectors for sentiment analysis," *Proceedings of ACL*, 2011.

[5] S. M. Qaisar, "Sentiment analysis of IMDb movie reviews using long short-term memory," *IEEE Conference Publication*, 2020.

[6] R. Socher et al., "Recursive deep models for semantic compositionality over a sentiment treebank," *Proceedings of EMNLP*, 2013.