computation

MDPI

Article

Does Bigger Mean Better? A Domain-Aware Comparison of Lexicon-Based, TF-IDF, and Transformer Models for Sentiment Analysis

Muhammad Abbas [1,†], Shujat Hussain [2,†], Ahmad Sami Al-Shamayleh [3], Ghulam Mustafa [2,*,†], Azam Mehmood [4,†], Isabel De la Torre Díez [5], Mónica Gracia Villar [6,7,8,9] and Helena Garay [6,10,11]

1 Department of Artificial Intelligence, Shaheed Zulfikar Ali Bhutto Institute of Science and Technology, Islamabad 44000, Pakistan; ma9514610@gmail.com
2 Department of Computer Science, Shifa Tameer-e-Millat University, Islamabad 44000, Pakistan; shujat.ssc@stmu.edu.pk
3 Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Al-Ahliyya Amman University, Amman 19111, Jordan; a.alshamayleh@ammanu.edu.jo
4 Department of Information Technology, University of Gujrat, Gujrat 50700, Pakistan; azam90216@gmail.com
5 eHealth and Telemedicine Group, University of Valladolid, 47011 Valladolid, Spain; isator@uva.es
6 Department of Extension, Universidade Internacional do Cuanza, Cuito EN250, Bié, Angola; monica.gracia@uneatlantico.es (M.G.V.); helena.garay@uneatlantico.es (H.G.)
7 Faculty of Social Science and Humanities, Universidad Europea del Atlántico, Isabel Torres 21, 39011 Santander, Spain
8 Department of Project Management, Universidad Internacional Iberoamericana, Arecibo, PR 00613, USA
9 Department of Projects, Universidad Internacional Iberoamericana, Campeche 24560, Mexico
10 Department of Projects, Fundación Universitaria Internacional de Colombia, Bogotá 111311, Colombia
11 Department of Projects, Universidad de La Romana, La Romana 22000, Dominican Republic
* Correspondence: ghulam.mustafa.ssc@stmu.edu.pk or ghulam.mustafa.ssc@gmail.com
† These authors contributed equally to this work.

Abstract

Sentiment analysis systems range from training-free lexicons to compact supervised models and fine-tuned transformer encoders, but fair comparison requires identical test examples and explicit computational accounting. (1) Background: We reassess three sentiment-classification paradigms under matched binary-label evaluation on IMDb movie reviews and Sentiment140 tweets. (2) Methods: After duplicate and label audits, VADER, TF–IDF with validation-tuned Logistic Regression, and three-epoch GPU-fine-tuned DistilBERT were evaluated on the same 5065 IMDb and 10,000 Sentiment140 held-out records within each corpus. Accuracy, macro-F1, confusion matrices, paired McNemar tests, bootstrap intervals, training time, and inference throughput were computed from saved predictions and execution logs. (3) Results: DistilBERT reached 92.00% accuracy on IMDb and 83.41% on Sentiment140, compared with 91.06% and 78.77% for Logistic Regression and 69.83% and 66.54% for VADER. DistilBERT exceeded Logistic Regression by 0.95 percentage points on IMDb (paired 95% CI 0.08–1.76; $p = 0.0351$) and by 4.64 points on Sentiment140 (CI 3.85–5.41; $p < 0.001$), but required GPU fine-tuning and longer measured training. (4) Conclusions: The relative advantage of a transformer depends on the dataset and available computation. Logistic Regression remains a compact, accurate alternative; results are restricted to these two within-domain configurations and do not demonstrate cross-domain transfer or domain causation.

Keywords: sentiment analysis; IMDb; Sentiment140; VADER; TF–IDF; logistic regression; DistilBERT; reproducibility; matched evaluation; computational efficiency

Check for updates

Received: 16 October 2026
Revised: 30 October 2026
Accepted: 2 November 2026
Published: 5 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Sentiment classification is widely used to monitor user feedback and summarize opinions in reviews and social media. Practical systems can employ lexicon-based polarity rules, sparse statistical features, or contextual language representations, with markedly different requirements for labeled data and computing resources [1–3]. VADER combines word valences with heuristic adjustments for emphasis and negation [2]; TF–IDF features with Logistic Regression (LR) provide a compact supervised baseline; and transformer encoders such as DistilBERT can be fine-tuned end-to-end [4–6]. Model choice is consequently a decision about predictive effectiveness and the computational expense of fitting and applying a classifier.

Existing evidence makes an unconditional ranking of these approaches difficult to defend. In a multi-dataset analysis, Hartmann et al. [7] investigated how the accuracy and practical applicability of sentiment methods vary across settings. Research on transformer fine-tuning also emphasizes the importance of optimization and model-selection procedures [8,9]. These findings motivate a controlled comparison, rather than an assumption that either complex or lightweight models are invariably preferable. Comparisons on separate corpora are especially sensitive to differences in labeling procedures, sample selection, and training size; such differences should not be interpreted as isolated effects of linguistic domain [10].

This study examines two distinct within-domain binary tasks: positive versus negative IMDb movie reviews [11], and positive versus negative posts from the distant-supervision Sentiment140 training corpus [12]. We compare VADER, TF–IDF + LR, and fine-tuned DistilBERT on identical held-out record identifiers within each corpus. We explicitly audit labels and duplicate texts, separate training and model selection from testing, regenerate metrics from saved predictions, and report measured training and inference times. A descriptive subgroup analysis investigates how accuracy varies with text length and selected surface indicators, without treating those associations as evidence of causation.

We organize the investigation around three research questions: **RQ1**: How do the three methods compare in held-out predictive performance under identical binary test sets within each corpus? **RQ2**: What accuracy–efficiency trade-offs arise from their measured training time, inference throughput, and recorded accelerator use? **RQ3**: How do differences in observed performance align with document length and selected noisy-text indicators within each test set? The contribution is a transparent comparative benchmark with reproducible splits, prediction-level evidence, paired uncertainty estimates, and an explicit examination of deployment trade-offs. It is not a new sentiment architecture or a claim of general cross-domain transfer.

The remainder of the paper reviews relevant comparisons (Section 2), specifies the experimental protocol (Section 3), presents the classification and efficiency results (Section 4), and discusses their limitations and implications (Sections 5 and 6).

2. Related Work

2.1. Lexicon and Supervised Sentiment Classification

Lexicon-based sentiment scoring offers an immediately usable baseline when no task-specific labeled training data are available. VADER estimates a normalized compound polarity score and includes rules for capitalization, punctuation, and modifiers [2]. This makes its actual text-input preprocessing important: deleting punctuation or changing case before scoring can discard cues the method was designed to use. Lexical scoring also faces limitations when positive and negative expressions depend on longer context or specialized vocabulary [13].

Sparse supervised methods provide an alternative that learns a decision boundary from domain-specific labeled examples [14]. IMDb has been an established sentiment benchmark since the release of the Large Movie Review Dataset [11]; TF–IDF representations combined with regularized LR are an appropriate strong baseline for such tasks. Their predictive performance depends on tokenization, vocabulary size, feature scaling, regularization, and the training distribution, all of which must be disclosed for a reproducible comparison. These models should therefore be evaluated as tuned supervised baselines, rather than unspecified defaults.

2.2. Transformer Encoders and Comparative Evidence

DistilBERT uses knowledge distillation to obtain a smaller BERT-family encoder suitable for downstream fine-tuning [6]; related transformer designs include the RoBERTa family [15]; however, performance depends on sequence length, the pretrained checkpoint, optimization settings, and validation-based checkpoint selection [8,9]. TweetEval and earlier shared-task research document the distinct character of social-media classification tasks [16,17], while Hartmann et al. [7] provide direct multi-dataset evidence that method selection should reflect the intended application. A 2025 study compared RoBERTa with traditional and deep-learning methods on Amazon book reviews [18], and resource-sensitive DistilBERT adaptation was explored in a parameter-efficient fine-tuning comparison [19]; these studies motivate careful design-specific reporting rather than an absolute architecture ranking. This work focuses on a prediction-matched three-method comparison on two specified binary corpora, alongside runtime measurements, checks of metric consistency, and paired uncertainty analysis.

Ilan and Vilenchik [20] show, in the different task of hate-speech detection, that selection and augmentation of training examples can change in-domain and cross-domain evaluation outcomes. We cite this as a methodological caution about data construction, not as direct evidence of sentiment-model superiority. Similarly, the survey of Pan and Yang [10] distinguishes transfer between domains from separate evaluations of models trained within each domain. The present experiments are of the latter type. Table 1 summarizes the positioning of the present study relative to representative prior work and highlights the methodological distinctions that motivate the proposed comparative evaluation.

Table 1. Positioning of the present evaluation relative to representative prior studies.

Study	Relevant Evidence and Scope
Hutto and Gilbert [2]	Lexicon and heuristic scoring designed for social text; foundational VADER baseline.
Maas et al. [11]	IMDb data and supervised sentiment modeling; an established review-text reference.
Sanh et al. [6]	Compact transformer pretraining; motivates the DistilBERT comparison.
Mosbach et al. [9]	Fine-tuning stability; supports transparent validation and checkpoint selection.
Hartmann et al. [7]	Direct sentiment-method evidence across datasets and applied conditions; contextualizes the novelty claim.
Ilan and Vilenchik [20]	Data-selection effects in hate-speech detection, not a sentiment benchmark.
This study	Matched binary test identifiers for three methods on IMDb and Sentiment140; paired testing, resource accounting, and exploratory subgroups.

3. Materials and Methods

Figure 1 summarizes the pipeline. The same frozen ground-truth labels and record identifiers are supplied to every classifier within each corpus. All hyperparameter and checkpoint decisions precede the held-out evaluation.

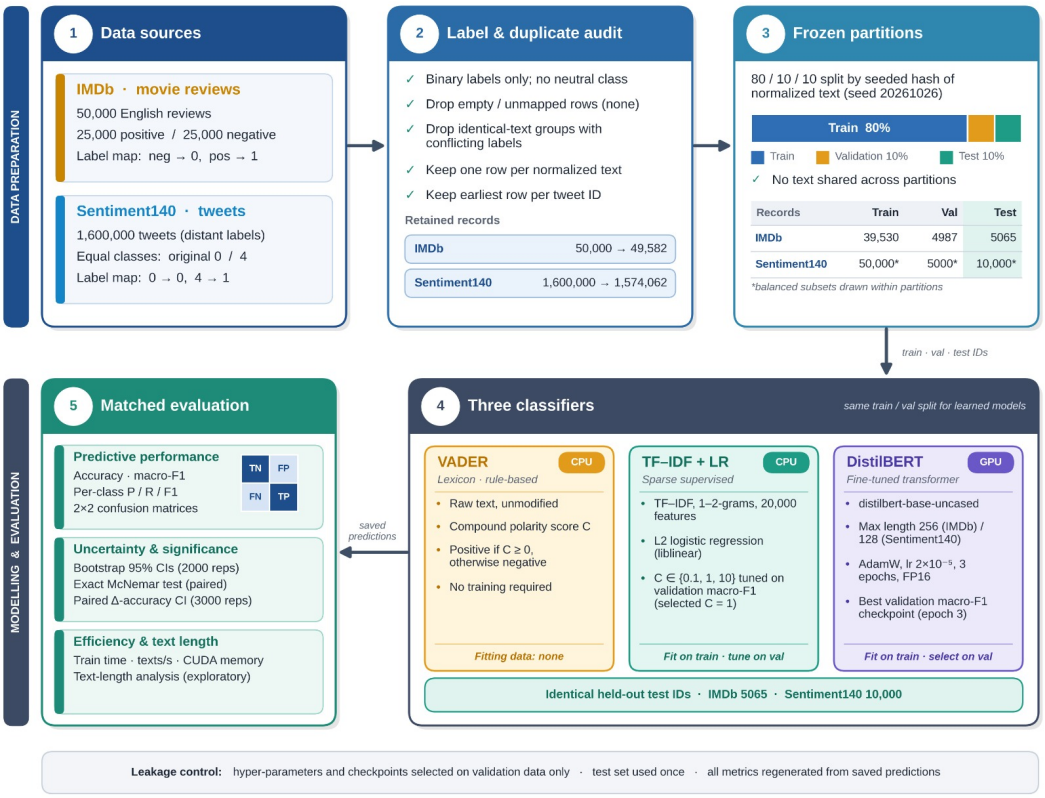


Figure 1. Overview of the proposed reproducible sentiment classification and evaluation framework. The workflow comprises five stages: (1) acquisition of IMDb and Sentiment140 datasets; (2) binary label verification, duplicate removal, and data-quality auditing; (3) deterministic training, validation, and testing partitions with leakage prevention; (4) implementation of three sentiment classifiers—VADER, TF-IDF with Logistic Regression, and DistilBERT—using model-specific preprocessing and validation-based model selection where applicable; and (5) evaluation on identical held-out test records using classification metrics, confusion matrices, bootstrap confidence intervals, paired statistical tests, computational efficiency measurements, and exploratory text-length analysis.

3.1. Data Sources, Label Audit, and Split Construction

The IMDb CSV used in this study contains 50,000 English movie reviews (25,000 positive and 25,000 negative), while the Sentiment140 training CSV contains 1,600,000 tweets, equally divided between original labels 0 and 4 [11,12]. The Sentiment140 training file contains no neutral-labeled observations. We mapped IMDb negative to 0 and positive to 1, and Sentiment140 label 0 to negative (0) and label 4 to positive (1). Both reference labels and classifier outputs were restricted to the resulting binary task.

The dataset audit excluded empty or unmapped records (none in either supplied file), removed entire groups of identical normalized text-bearing contradictory labels, retained one observation per remaining normalized text, and kept the earliest remaining instance of each repeated Twitter identifier. Only the duplicate-detection key is case-folded, whitespace-collapsed, and stripped; the exported review or tweet text itself is not rewritten. This procedure removed 418 IMDb rows and 25,938 Sentiment140 rows, leaving 49,582 and 1,574,062 records, respectively. It addresses identical-text leakage but does not establish the absence of semantic duplicates, mislabeling, or paraphrased copies.

An 80/10/10 split was assigned using a deterministic seeded hash of normalized text (seed 20261026), after the duplicate audit. Thus, repeated normalized texts cannot be divided across train, validation, and test partitions. The IMDb corpus contributed all 39,530 training, 4987 validation, and 5065 test records to learned-model experiments. For Sentiment140, to control the computational burden, balanced subsets of 50,000 training, 5000 validation, and 10,000 test tweets were drawn within the preassigned partitions with fixed seeds. Both learned models used the same train/validation partition within each corpus; VADER required no training. Every model was evaluated on the exact same 5065 IMDb or 10,000 Sentiment140 test IDs. These partitions were derived from the corresponding CSV corpora rather than the official IMDb benchmark split or the separate manually annotated Sentiment140 test set. The detailed statistics of the dataset is presented in Table 2.

Table 2. Source records, cleaning results, and final evaluation subsets. Training and validation counts refer to each learned model; VADER is training-free.

Item	IMDb	Sentiment140
Original supplied CSV rows	50,000	1,600,000
Retained after audit	49,582	1,574,062
Original labels	Positive/negative	0/4 (negative/positive)
Final training records	39,530	50,000
Final validation records	4987	5000
Final test records	5065	10,000
Test class support (negative/positive)	2521/2544	5000/5000

3.2. Model-Specific Inputs and Classifiers

3.2.1. VADER

The VADER baseline was applied directly to the original preserved text without lowercase conversion, stripping of punctuation, or deletion of emotive cues. It produces a compound score $C \in [-1, 1]$ [2]. The fixed binary decision rule was $\hat{y} = 1$ when $C \geq 0$ and $\hat{y} = 0$ when $C < 0$. Thus, zero scores are explicitly positive; no neutral prediction category exists in our evaluation. This rule was specified before test scoring and was not optimized on the held-out labels. Because the conventional VADER three-way recommendation uses an abstention/neutral interval, our binary threshold should not be interpreted as a calibrated probability or equivalent to its native three-class behavior.

3.2.2. TF–IDF with Logistic Regression

We used the scikit-learn pipeline API with a TF–IDF extractor (`TfidfVectorizer`) followed by a logistic classifier (`LogisticRegression`). The vectorizer lowercases and strips Unicode accents, uses the default word-token pattern, unigram and bigram features, a maximum of 20,000 features, `min_df=2`, `max_df=0.99`, sublinear term frequency, smoothed inverse document frequency, and ℓ_2 vector normalization. With N training documents and $\mathrm{df}(t)$ documents containing term t, the smoothed IDF factor is

$$\mathrm{idf}(t) = \log\left(\frac{1+N}{1+\mathrm{df}(t)}\right) + 1. \quad (1)$$

Logistic Regression used an ℓ_2 penalty, `liblinear` solver, `max_iter=500`, default class weighting (none), and a single fixed random seed. Regularization $C \in \{0.1, 1, 10\}$ was selected by validation macro-F1 independently for each dataset; both selected $C = 1$. The vectorizer was fitted only to training texts, and the frozen test set was used once after model selection. The exact parameter dictionary, validation trial scores, and script are included in the link provided in the Data Availability Statement.

3.2.3. Fine-Tuned DistilBERT

We fine-tuned `distilbert-base-uncased` [6] with its matching tokenizer and a two-logit classification head. The model received original text through checkpoint tokenization with dynamic batch padding and maximum sequence lengths of 256 WordPiece tokens for IMDb and 128 for Sentiment140. Truncation at these limits is a constraint of the evaluated configuration, not a property of the underlying task. The model was optimized using AdamW (learning rate 2×10^{-5}, weight decay 0.01), linear warm-up over 10% of training steps, a per-device batch size of 16, up to three epochs, and half-precision CUDA execution. Evaluation and checkpoint saving occurred every epoch; selection was based on validation macro-F1, with early-stopping patience of two validation checks. The best validation checkpoint occurred at epoch 3 for both datasets. A single seed (20261026) was used, and the test set did not guide checkpoint selection. Table 3 presented the detailed setting environment information.

Table 3. Key settings used in the expanded three-method evaluation.

Component	Exact Settings/Selection
Shared protocol	Binary labels 0/1; seed 20261026; disjoint train/validation/test IDs; matched held-out IDs by dataset.
VADER	Raw preserved text; `vaderSentiment`; negative if compound < 0, otherwise positive; no training.
TF–IDF	Default word token pattern; 1–2 g; 20,000 features; `min_df=2`, `max_df=0.99`; `sublinear_tf=True`; `norm=l2`.
LR	`liblinear`; ℓ_2 penalty; `max_iter=500`; `class_weight=None`; $C \in \{0.1, 1, 10\}$ on validation macro-F1.
DistilBERT	`distilbert-base-uncased`; max length 256/128 (IMDb/Twitter); batch size 16/device; AdamW; lr $2 \cdot 10^{-5}$; warm-up ratio 0.1; weight decay 0.01; 3 epochs, validation macro-F1 checkpoint selection.

3.3. Metrics, Uncertainty, and Resource Accounting

For each model–dataset pair we reconstructed a 2×2 confusion matrix with rows representing actual negative/positive labels and columns representing predicted negative/positive labels. Accuracy is $(TP + TN)/(TP + TN + FP + FN)$. Positive-class precision is $TP/(TP + FP)$, positive-class recall is $TP/(TP + FN)$, and its F1 is the harmonic mean of these values. The analogous metrics are also computed for the negative class. Macro-F1 is the unweighted average of class-specific F1 values; macro-precision and macro-recall average the corresponding per-class statistics. We report both class supports and macro results, never conflating micro-averaged recall with positive-class recall. A symmetric confusion matrix does not establish probability calibration, which was not assessed.

Uncertainty of each accuracy estimate was measured by a 2000-replicate nonparametric bootstrap over held-out records. Comparisons between classifiers used exact two-sided McNemar tests on paired prediction correctness and 3000-replicate paired bootstrap intervals for accuracy differences. These intervals quantify test-observation uncertainty conditional on the fitted models, not variability across training seeds or independent model retrainings. We treat analyses of word-count quartiles, URL/mention/hashtag indicators, and punctuation markers as post hoc descriptive comparisons.

All experiments used Kaggle Linux with four logical CPU cores (the CPU processor model was not logged), Python 3.13.15, `transformers` 4.52.4, and PyTorch 2.11.0 with CUDA 12.8; two NVIDIA Tesla T4 devices were detected. LR and VADER executed on CPU, while DistilBERT used CUDA. Training time includes LR's validation search or DistilBERT fine-tuning, respectively; VADER has zero fitting time. Inference time is the measured elapsed time for the full held-out set, from which texts/second and average ms/text are computed. GPU allocated-peak memory was logged for DistilBERT;

however, measured CPU peak RAM, pinned transformer checkpoint commit hash, total serialized transformer model size, and per-device GPU utilization were not available in the supplied logs. Consequently, inference times are not a hardware-matched cost comparison across methods.

4. Results

4.1. Predictive Performance on Matched Test Sets

Table 4 reports accuracy and macro-F1 calculated from the saved test predictions. All three models use the same held-out records within each dataset. On IMDb, DistilBERT achieved 92.00% accuracy, followed by TF–IDF + LR at 91.06% and VADER at 69.83%. On Sentiment140, the corresponding accuracies were 83.41%, 78.77%, and 66.54%. These within-corpus rankings reflect the common evaluation protocol, with identical test IDs for each of the three methods.

Table 4. Classification performance on the common held-out test sets. Confidence intervals are 95% test-record bootstrap intervals (2000 replicates).

Dataset	Model	Test *N*	Accuracy (%)	Macro-F1 (%)	Accuracy 95% CI
IMDb	VADER	5065	69.83	68.94	68.61–71.08
IMDb	TF–IDF + LR	5065	91.06	91.05	90.25–91.87
IMDb	DistilBERT	5065	**92.00**	**92.00**	91.25–92.75
Sentiment140	VADER	10,000	66.54	64.62	65.62–67.47
Sentiment140	TF–IDF + LR	10,000	78.77	78.77	77.98–79.53
Sentiment140	DistilBERT	10,000	**83.41**	**83.41**	82.69–84.10

Figure 2 provides a direct visual comparison of the classification accuracy achieved by VADER, TF–IDF + LR, and DistilBERT on the fixed IMDb and Sentiment140 test sets.

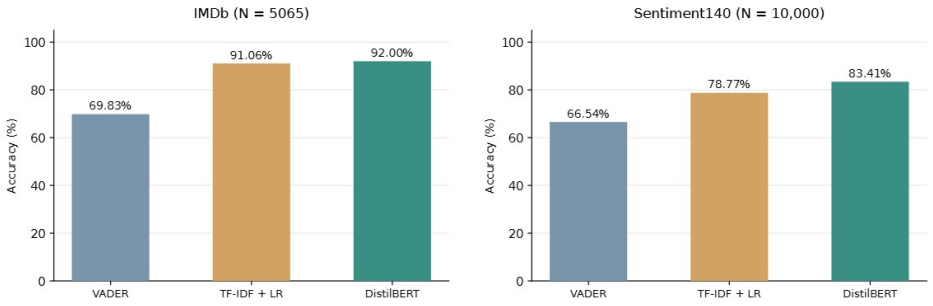


Figure 2. Accuracy of VADER , TF–IDF + LR, and DistilBERT on each fixed test set. Each bar is derived from prediction-level evidence and corresponds to the same test IDs within a dataset.

4.2. Class-Level Results and Confusion Matrices

Tables 5 and 6 disaggregates precision, recall, and F1 by true sentiment class. Class-specific recall is distinct from overall accuracy and from micro-averaged recall on a single-label classification task. For instance, on Sentiment140, VADER's positive-class recall was 89.82% but its negative-class recall only 43.26%, reflecting substantially asymmetric predictions under the specified compound-score threshold. DistilBERT achieved more balanced class-specific recall (84.06% negative; 82.76% positive). Calibration is not inferred from these error counts.

Table 5. Per-class precision, recall, and F1 calculated from the saved binary predictions. N denotes class support.

Dataset	Model	Class	*N*	Precision	Recall	F1
IMDb	VADER	Negative	2521	0.795	0.531	0.637
		Positive	2544	0.650	0.864	0.742
IMDb	TF–IDF + LR	Negative	2521	0.919	0.899	0.909
		Positive	2544	0.902	0.922	0.912
IMDb	DistilBERT	Negative	2521	0.917	0.923	0.920
		Positive	2544	0.923	0.917	0.920
Sentiment140	VADER	Negative	5000	0.810	0.433	0.564
		Positive	5000	0.613	0.898	0.729
Sentiment140	TF–IDF + LR	Negative	5000	0.789	0.786	0.787
		Positive	5000	0.786	0.790	0.788
Sentiment140	DistilBERT	Negative	5000	0.830	0.841	0.835
		Positive	5000	0.839	0.828	0.833

Table 6. Binary confusion-matrix counts (true negatives TN, false positives FP, false negatives FN, and true positives TP). Each row sums to the relevant test size in Table 4.

Dataset	Model	TN	FP	FN	TP
IMDb	VADER	1339	1182	346	2198
IMDb	TF–IDF + LR	2267	254	199	2345
IMDb	DistilBERT	2326	195	210	2334
Sentiment140	VADER	2163	2837	509	4491
Sentiment140	TF–IDF + LR	3928	1072	1051	3949
Sentiment140	DistilBERT	4203	797	862	4138

The matrix counts in Figures 3 and 4 reproduce the reported accuracies exactly. On IMDb, for example, the DistilBERT diagonal is $2326 + 2334 = 4660$ correct classifications among 5065 reviews (92.00%). For Sentiment140, the corresponding count is $4203 + 4138 = 8341$ out of 10,000 (83.41%). All displayed matrices are binary, and their diagonal counts agree with the reported classification accuracies.

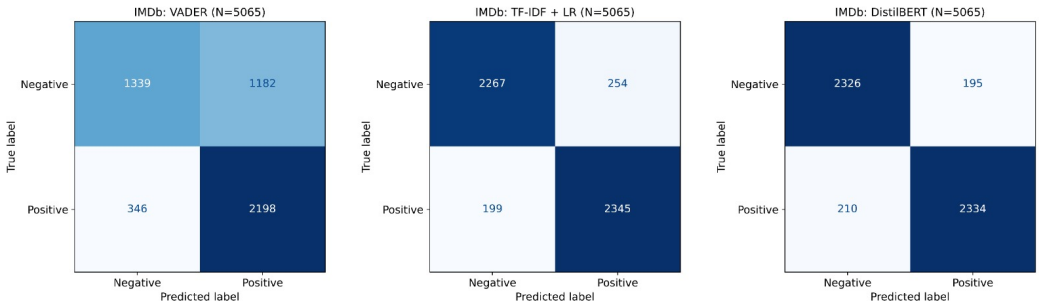


Figure 3. IMDb binary confusion matrices, left to right: VADER, TF–IDF + LR, and DistilBERT. Rows are true classes (0, 1); columns are predictions (0, 1).

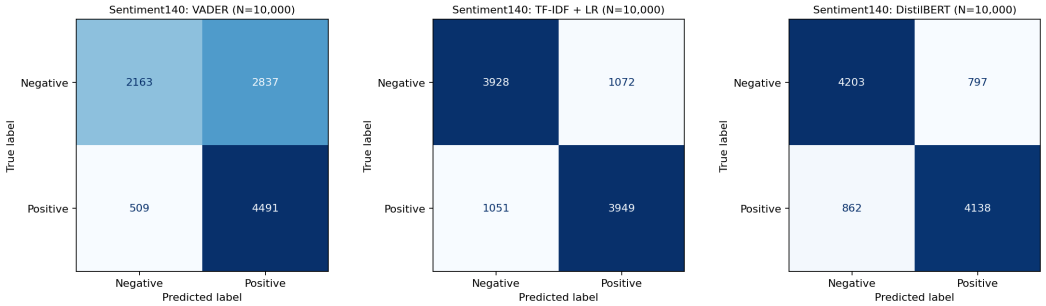


Figure 4. Sentiment140 binary confusion matrices on the common 10,000-record test set, left to right: VADER, TF–IDF + LR, and DistilBERT. No neutral class is present.

4.3. Paired Model Comparisons

Paired tests preserve the fact that all methods labeled the same observations. As shown in Table 7, the DistilBERT improvement over LR was 0.95 percentage points on IMDb (95% paired bootstrap interval 0.08–1.76; McNemar exact $p = 0.0351$), and 4.64 points on Sentiment140 (interval 3.85–5.41; $p < 0.001$). Both differences reach the unadjusted 5% threshold in these fixed-test comparisons, but the IMDb effect is small and should not be described as a decisive universal architecture advantage. The statistical tests do not establish that the difference would recur under other training seeds, datasets, or training budgets.

Table 7. DistilBERT versus TF–IDF + LR on identical test records: paired accuracy gain and exact two-sided McNemar comparison.

Dataset	Gain (Percentage Points)	95% Paired CI (pp)	Exact p
IMDb	+0.95	+0.08 to +1.76	0.0351
Sentiment140	+4.64	+3.85 to +5.41	1.63×10^{-30}

4.4. Training Cost and Inference Throughput

Table 8 places predictive results beside measured computational cost. VADER requires no task-specific training. LR took 88.04 s for IMDb and 8.23 s for Sentiment140, including its validation search over three regularization strengths, whereas DistilBERT fine-tuning took 1700.49 and 646.30 s. On IMDb, the difference between LR and DistilBERT in accuracy was below one point despite a substantially larger fine-tuning investment; on Sentiment140, the improvement was more substantial. DistilBERT peak allocated CUDA tensor memory reached approximately 2.05 GiB and 1.42 GiB, respectively. These are not peak system RAM values or full GPU utilization figures.

Table 8. Accuracy–efficiency under the logged Kaggle configuration. Inference is end-to-end for each dataset's common held-out test set, with LR/VADER on CPU and DistilBERT on CUDA. Device-dependent timings are not hardware-matched.

Dataset	Model	Train (s)	Test Infer. (s)	Texts/s	ms/text	CUDA Peak (GiB)
IMDb	VADER	0	14.04	360.9	2.77	—
IMDb	TF–IDF + LR	88.04	1.72	2947.9	0.34	—
IMDb	DistilBERT	1700.49	24.30	208.4	4.80	2.05
Sentiment140	VADER	0	0.56	17,791.5	0.06	—
Sentiment140	TF–IDF + LR	8.23	0.20	48,916.1	0.02	—
Sentiment140	DistilBERT	646.30	14.31	699.0	1.43	1.42

For further size context, the exported LR pipelines occupied 924,736 bytes (IMDb) and 916,176 bytes (Sentiment140) in the experimental execution environment. The result ZIP does not archive the serialized pipelines or transformer weights, so full checkpoint file sizes

cannot be verified from the retained output files. Timing also includes implementation-specific batching and tokenizer/vectorizer overhead; it should not be extrapolated into normalized energy use or CPU-only transformer performance.

4.5. Exploratory Text Characteristics

To examine within-corpus heterogeneity without inferring causal domain effects, test texts were partitioned into word-count quartiles. Table 9 shows the shortest and longest quarters for each dataset. On IMDb, DistilBERT exceeded LR in the shortest quarter (93.69% versus 89.90%) but underperformed it in the longest quarter (87.28% versus 92.42%). Sequence truncation at 256 tokens is a plausible explanation for the latter pattern, but cannot be established without a controlled length-ablation experiment. On Sentiment140, DistilBERT exceeded LR in both displayed quartiles, although both models performed less accurately in the longest than in the shortest quarter.

Table 9. Exploratory performance by word-count quartile, defined separately within each test set. These are observational subgroups, not causal or cross-domain transfer tests.

Dataset	Quartile	*N*	Median Words	LR Acc. (%)	DistilBERT Acc. (%)
IMDb	Q1 shortest	1267	101.0	89.90	93.69
IMDb	Q4 longest	1266	406.5	92.42	87.28
Sentiment140	Q1 shortest	2500	5.0	82.56	85.48
Sentiment140	Q4 longest	2500	23.0	74.76	80.60

Additional descriptive Twitter subsets based on URLs, mentions, hashtags, and exclamation marks are reported in the accompanying analysis files. These groups overlap, have different supports, and were not selected through randomized interventions. Importantly, VADER assigned exactly zero compound score to 2554 of the 10,000 test tweets (25.54%); all such observations were mapped to positive by the predetermined binary tie rule. Its performance and class imbalance should therefore be interpreted as rule-dependent. A validation-tuned threshold sensitivity analysis remains an appropriate future extension.

5. Discussion

5.1. What Does "Bigger" Buy in These Configurations?

Under the evaluated configuration, DistilBERT achieved the highest classification accuracy on both corpora, but its practical benefit differed in magnitude. On IMDb, the gain over LR was only 0.95 percentage points; on Sentiment140 it was 4.64 points. Both were statistically distinguishable on the paired test sets, while the transformer required additional GPU fine-tuning time and a different inference device. Consequently, the answer to the title question is conditional: a larger model achieved greater accuracy here, but the magnitude of the gain must be considered alongside the training and deployment budget. The measured CPU/GPU timing asymmetry makes a universal performance-per-dollar or latency claim inappropriate.

The observed ordering applies to the measured GPU-assisted DistilBERT configuration, with three fine-tuning epochs and validation-based checkpoint selection; it does not establish how a CPU-only or single-epoch transformer configuration would perform. Fine-tuning stability studies provide a reason to disclose these details [8,9], but cannot identify architecture as the sole cause of the observed rankings.

5.2. Domain Characterization, Label Sources, and External Validity

IMDb reviews and Sentiment140 tweets differ in text length, collection, labeling, training size, and possibly many unmeasured attributes. Comparing separate within-domain

models on these two corpora cannot isolate which attribute caused the observed accuracy levels. The word-count and surface-indicator analyses reveal potentially informative associations within the held-out sets, particularly for long IMDb reviews, but remain exploratory. We have not trained on IMDb and tested on Twitter (or vice versa), so we do not claim a cross-domain transfer experiment or generalized domain-adaptation performance [10].

External validity is limited to the two English-language corpus versions and to the sampled evaluation subsets derived from them. In particular, Sentiment140 contains distant-supervision labels rather than an independently adjudicated human test set, and the present test subset is drawn from the corpus's training source. The official 498-example Sentiment140 evaluation split and other domains (e.g., financial or product sentiment) were not evaluated. Further generalization requires an independently annotated external test set, additional domains, and multiple fine-tuning seeds.

5.3. Reproducibility and Remaining Measurement Gaps

The accompanying code implements binary label mappings, duplicate auditing, fixed split identifiers, model-specific preprocessing, Logistic Regression settings, DistilBERT validation logs, saved predictions, classification reports, and paired statistical tests. Dataset-processing scripts, environment information, and record identifiers are included with the Data Availability Statement. Nevertheless, the run used one random seed, and the available experimental artifacts do not include saved Logistic Regression model files, a pinned transformer checkpoint commit identifier, fine-tuned model weights, measured peak CPU RAM, or a full per-device GPU utilization trace. Prediction files alone are insufficient to reconstruct these measurements or trained artifacts. The reported estimates therefore describe a single-run comparison rather than a multi-seed training study.

6. Conclusions

This study compares VADER, TF–IDF + Logistic Regression, and fine-tuned DistilBERT on cleaned, matched binary test sets from IMDb (5065 reviews) and Sentiment140 (10,000 tweets). All confusion matrices are consistent with their corresponding saved predictions and test sample counts, with explicit label mappings, recorded model configurations, and measured computational times. DistilBERT achieved 92.00% IMDb and 83.41% Sentiment140 accuracy, compared with 91.06% and 78.77% for LR and 69.83% and 66.54% for VADER. Paired comparisons showed a small 0.95-point DistilBERT advantage on IMDb and a larger 4.64-point advantage on Sentiment140, with nonzero paired confidence intervals on both held-out subsets.

These results support a qualified practical conclusion: LR remains an accurate, compact supervised alternative when the additional gains of a transformer do not justify GPU fine-tuning and related deployment demands. VADER remains a training-free baseline but exhibits strong sensitivity to the specified binary tie rule, particularly for zero-scored tweets. Neither two within-domain datasets nor one training seed establishes universal model ranking, domain causation, probability calibration, or cross-domain transfer. Future work should examine additional domains and independently annotated test sets, repeat fine-tuning across seeds, conduct token-length and VADER-threshold ablations, and benchmark models under matched CPU/GPU configurations with measured peak RAM and retained model artifacts.

Author Contributions: Conceptualization, M.A. and G.M.; methodology, S.H.; software, S.H.; validation, M.A., S.H. and A.M.; formal analysis, M.A.; investigation, S.H.; resources, G.M.; data curation, S.H.; writing—original draft preparation, S.H.; writing—review and editing, M.A., S.H., A.S.A.-S., G.M., A.M., I.D.l.T.D., M.G.V. and H.G.; visualization, S.H.; supervision, G.M.; project administra-

tion, G.M.; funding acquisition, A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by FUNDACIÓN UNIVERSIDAD EUROPEA DEL ATLÁNTICO, Santander, Spain (CIF: ESG39764972).

Data Availability Statement: The IMDb Large Movie Review Dataset and Sentiment140 are publicly accessible at https://ai.stanford.edu/~amaas/data/sentiment/ (accessed on 25 May 2026) and http://help.sentiment140.com/for-students (accessed on 5 June 2026), respectively. The experiments use deterministic partitions derived from the corresponding CSV corpora rather than the official benchmark test splits. Dataset-cleaning scripts, fixed test identifiers, experiment configurations, saved predictions, environment information, statistical reports, and the experimental implementation are provided in the following link: https://www.kaggle.com/code/shujathussain1412/notebook96a3280038 (accessed on 25 October 2026). Fine-tuned model checkpoints are not part of the available research artifacts.

Acknowledgments: The authors acknowledge the Department of Artificial Intelligence and the Department of Computing, Shifa Tameer-e-Millat University, and the Department of Information Technology, University of Gujrat, for institutional support. The authors used ChatGPT (5.6), Claude (Sonet 5.5), and Grammarly (1.2.91) to assist in improving the English grammar, clarity, and readability of the manuscript. The authors reviewed the final text and take full responsibility for the scientific content, analyses, interpretations, and conclusions.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NLP	Natural Language Processing
VADER	Valence Aware Dictionary and sEntiment Reasoner
TF-IDF	Term Frequency-Inverse Document Frequency
LR	Logistic Regression
BERT	Bidirectional Encoder Representations from Transformers
CLS	Classification token
CPU	Central Processing Unit
GPU	Graphics Processing Unit
CI	Confidence Interval
CUDA	Compute Unified Device Architecture
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
PEFT	Parameter-Efficient Fine-Tuning
LoRA	Low-Rank Adaptation
IMDb	Internet Movie Database

References

1. Cambria, E.; Schuller, B.; Xia, Y.; Havasi, C. New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intell. Syst.* **2013**, *28*, 15–21. https://doi.org/10.1109/MIS.2013.30.
2. Hutto, C.J.; Gilbert, E. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM), Ann Arbor, MI, USA, 1–4 June 2014; pp. 216–225. https://doi.org/10.1609/icwsm.v8i1.14550.
3. Zhang, L.; Wang, S.; Liu, B. Deep Learning for Sentiment Analysis: A Survey. *WIREs Data Min. Knowl. Discov.* **2018**, *8*, e1253. https://doi.org/10.1002/widm.1253.

4. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
5. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.
6. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *arXiv* **2019**, arXiv:1910.01108. https://doi.org/10.48550/arXiv.1910.01108.
7. Hartmann, J.; Heitmann, M.; Siebert, C.; Schamp, C. More than a Feeling: Accuracy and Application of Sentiment Analysis. *Int. J. Res. Mark.* **2023**, *40*, 75–87. https://doi.org/10.1016/j.ijresmar.2022.05.005.
8. Dodge, J.; Ilharco, G.; Schwartz, R.; Farhadi, A.; Hajishirzi, H.; Smith, N.A. Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping. *arXiv* **2020**, arXiv:2002.06305. https://doi.org/10.48550/arXiv.2002.06305.
9. Mosbach, M.; Andriushchenko, M.; Klakow, D. On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines. *arXiv* **2020**, arXiv:2006.04884. https://doi.org/10.48550/arXiv.2006.04884.
10. Pan, S.J.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. https://doi.org/10.1109/TKDE.2009.191.
11. Maas, A.L.; Daly, R.E.; Pham, P.T.; Huang, D.; Ng, A.Y.; Potts, C. Learning Word Vectors for Sentiment Analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; pp. 142–150.
12. Go, A.; Bhayani, R.; Huang, L. *Twitter Sentiment Classification Using Distant Supervision*; Technical Report; Stanford University: Stanford, CA, USA, 2009.
13. Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K.; Stede, M. Lexicon-Based Methods for Sentiment Analysis. *Comput. Linguist.* **2011**, *37*, 267–307. https://doi.org/10.1162/COLI_a_00049.
14. Pang, B.; Lee, L.; Vaithyanathan, S. Thumbs Up? Sentiment Classification Using Machine Learning Techniques. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), Philadelphia, PA, USA, 6–7 July 2002; pp. 79–86. https://doi.org/10.3115/1118693.1118704.
15. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692.
16. Barbieri, F.; Camacho-Collados, J.; Espinosa Anke, L.; Neves, L. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 16–20 November 2020; pp. 1644–1650. https://doi.org/10.18653/v1/2020.findings-emnlp.148.
17. Rosenthal, S.; Farra, N.; Nakov, P. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 502–518. https://doi.org/10.18653/v1/S17-2088.
18. Pookduang, P.; Klangbunrueang, R.; Chansanam, W.; Lunrasri, T. Advancing Sentiment Analysis: Evaluating RoBERTa against Traditional and Deep Learning Models. *Eng. Technol. Appl. Sci. Res.* **2025**, *15*, 20167–20174. https://doi.org/10.48084/etasr.9703.
19. Nwaiwu, S. Parameter-Efficient Fine-Tuning for Low-Resource Text Classification: A Comparative Study of LoRA, IA3, and ReFT. *Front. Big Data* **2025**, *8*, 1677331. https://doi.org/10.3389/fdata.2025.1677331.
20. Ilan, T.; Vilenchik, D. HARALD: Augmenting Hate Speech Data Sets with Real Data. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 2241–2248. https://doi.org/10.18653/v1/2022.findings-emnlp.165.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.