IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SENTIBERT: EXPLAINABLE FRAMEWORK FOR MULTI-MODEL CROSS-DOMAIN SENTIMENT ANALYSIS

[1]Gangireddy Sushma, [2]Dr B. Suman, [3]Dr V. Anantha Krishna [4]Dr U. Srilakshmi

[1]M.Tech, [2]Associate Professor, [3]Professor [4]Professor

Department of Computer Science and Engineering[1234]

Sridevi Women's Engineering College, Hyderabad, Telangana, India[1234]

Abstract: There is an influx of digitally generated content and user opinions over several decades and has brought major changes to the information environment. Online platforms such as movie review sites, e-commerce portals, and social networking services produce a large amount of user-generated content every year. This includes customer feedback, reviews of products and services, and general social discussions. Together, this information forms an important source of sentiment data that supports business decisions, political communication, and social research.

This paper proposes a model SentiBERT with sentiment intelligence framework that uses three transformer architectures—BERT, RoBERTa, and DistilBERT—against classical TF-IDF-based baselines on the IMDB Large Movie Review Dataset, while evaluating across Amazon product and Yelp restaurant review datasets considering cross-domain transfer protocol.

SentiBERT generates three remarkable outputs which distinctively elevates this model from rest. First, we design a multi-method Explainable AI (XAI) pipeline that unifies Local Interpretable Model-agnostic Explanations (LIME), BertViz attention-weight visualization, and GPT-2-grounded for natural language explanations— In this approach, the generated explanation is based on measurable evidence obtained from LIME rather than relying only on unrestricted language model outputs, which helps reduce the chances of imagined results. Second, we implement a hybrid aspect-based sentiment analysis (ABSA) module that integrates spaCy noun-phrase extraction with BERT context-window scoring across six film-specific aspect categories. Third, vocabulary overlap is examined to understand transfer difficulty. For this purpose, Jaccard similarity is used as a reference measure. Additionally, words associated with model failure are identified across target domains using LIME.

The experimental results are validated against reproducible JSON metric files which demonstrates RoBERTa achieves the highest test-set accuracy of 91.70% and AUC-ROC of 0.972, representing a 32% reduction in error rate relative to the best classical baseline (Linear SVM: 87.62%). Interestingly, the accuracy drop for zero-shot cross-domain transfer is only 0.60% and 0.40% on Amazon and Yelp, respectively, and contradicts the notion of domain fragility for transformer-based sentiment classifiers. Under text-casing perturbation, BERT achieves 90.0% accuracy, whereas it drops to 70.0% when the test set is heavily corrupted with 20% of typing errors (typo injection). The biggest challenge BERT faces is the identification of subword tokenisation boundaries. The results shed light on the advantages and disadvantages of applying sentiment analysis based on transformers to real-world applications.

Index Terms - Sentimental Analysis, DistilBERT, BERT, RoBERTa, LIME, GPT-2, Aspect-Based Sentiment Analysis (ABSA), Cross-Domain Transfer Learning, Explainable AI, Natural Language Processing

I. INTRODUCTION

1.1 Background

The amount of digitally produced data has grown each year for the last ten years. Every year, websites like review sites for movies, e-commerce sites, and social media sites gather enormous quantities of opinions, reviews about products and comments from the public. Consequently, there is now a vast amount of sentiment rich data on which to draw, which can be used for activities like business decisions, public communication, and social research. Sentiment analysis of these data sources has become essential for businesses.For instance, a movie studio that knows almost in real time what parts of a story resonate with viewers and what does not can have an edge. Similarly, a retail company that has analyzed customer sentiment for each product feature can get valuable feedback that the simple star ratings are not capturing.

There are three well-defined generational paradigms of sentiment analysis, as the computational process that involves the identification of subjective attitudes, opinions and emotional polarities in natural language text [1]. The lexicon-based generation (around 2002–2010) used manually prepared word lists (VADER, AFINN, and SentiWordNet), where the polarity of every word was preassigned and was combined with basic word negation and intensification rules. While these methods are interpretable and inexpensive to compute, they are brittle in the presence of context nuances, such as the score to the phrase 'not bad', which is negative for naive lexicon aggregation, and evaluative vocabulary that depends on the context, e.g. the word 'wooden' in film criticism, which produces a negative score on a fixed polarity dictionary but not on the naive version.

The second generation (2010-2018) was based on classical machine learning with features expressed as TF-IDF. When the bag-of-words encodings were used with Logistic Regression, Naive Bayes, and Support Vector Machines their performance significantly increased the classification accuracy which comes close to 88–89% on the IMDB benchmark [3]. Word embeddings like Word2Vec, GloVe, and FastText and sequence models like LSTMs and CNNs helped partly overcome the limitations of bag-of-words representations for modeling context [4]. However, there existed a definite limit to performance because of the unidirectional nature of the encoding, the vocabulary limits, and the problems of capturing long-range dependency across multi-hundred word reviews.

The architecture of the Transformers [5] and the concretisation by BERT (Bidirectional Encoder Representations from Transformers) [6] were the fans that started the third and current generation. Unlike unidirectional recurrent models, the bidirectional self-attention mechanism in BERT allows each token to process information from the context on its left and right, allowing us to capture detailed left-to-right and right-to-left relationships between tokens. BERT has been trained on two large bi-bilingual datasets: BookCorpus (about 16 GB of text) and Wikipedia (about 16 GB of text), using both the masked language modelling (MLM) objective and next-sentence prediction (NSP).BERT is trained using a combined masked language modelling (MLM) and next-sentence prediction (NSP) objectives on two large bi-bilingual datasets of text: BookCorpus (about 16 GB) and Wikipedia (about 16 GB). Later variants like RoBERTa [7] (robustly optimised, 10 times more training data, no NSP) and DistilBERT [8] (knowledge distillation to 66M parameters, 60% faster inference) have pushed the transformer frontier in the accuracy and efficiency axis, respectively.

1.2 Research Gaps in Current Literature

Although the results obtained are very impressive, the literature shows that there are four issues that limit the production use of transformer-based sentiment analysis systems. Firstly, there are few comparisons which exist that are not conducted on the same evaluation conditions as the same dataset division, same processing pipeline and same evaluation protocol so it is hard to compare and interpret [9]. Second, there is insufficient attention paid to explaining the prediction: In most deployed systems, the classification label is provided without any means for the stakeholders to explain and audit the classification rationale for each individual prediction, which is an important aspect of regulated or high-stakes environments [10]. Thirdly, published sentiment benchmarks only cover a small number of aspects of the robustness with respect to real world text degradation, including text negation, typographical errors, variations in capitalization. Fourth, sparse information about the lexical-level causes of transfer failure (LTF), which are the vocabulary items that contribute to misclassification in target domains (LTF) [11] which are studied theoretically.

1.3 Research Questions and Contributions

In this paper, those gaps are overcome via the SentiBERT framework which is broken down into three key research questions:

RQ1: What is the performance of BERT, RoBERTa and DistilBERT compared to classical TF-IDF baselines under the same and identical training settings in terms of accuracy, F1 and AUC-ROC?

RQ2: Are word importance scores obtained using LIME reliable to reveal language mechanisms involved in failure of cross-domain transfer? RQ3: How is vocabulary overlap (Jaccard similarity) related to zero-shot transfer accuracy across the various review domains?

The first unique major contribution of this work include: (1) a multi-model comparative study with fully controlled experimental conditions; (2) a distinct explanation methodology based on LIME evidence, which do not include hallucinating risks found in unconstrained generative explanations; (3) a hybrid spaCy-BERT ABSA module which provides reliable aspect-level sentiment without the need for aspect-level training labels; (4) a systematic robustness evaluation protocol which includes four different types of perturbation; and (5) the first combined vocabulary-overlap and failure-word analysis for cross-domain transformer sentiment transfer in the IMDB-to-review-platform setting.

2. LITERATURE REVIEW

2.1 Evolution of Sentiment Analysis

Pang et al. [12] proposed and launched the classification of movie reviews as a typical NLP baseline, which had laid the foundation for computational study of opinion and sentiment. They also showed that SVM is superior to Naïve Bayes with TF-IDF representations. Soon, Maas et al. released another dataset, the IMDB Large Movie Review Dataset [3] consisting of 50,000 binary-labeled reviews, that is still used as the standard dataset for binary sentiment classification today. They obtained 88.89% accuracy (the performance floor) on their combined unsupervised-supervised word vector objective, which is replicated by lexicon-based and classical ML approaches.

The use of dense word embeddings in sentiment models was a grand step in methodology. Two recent studies by Mikolov et al. [13] and Pennington et al. [14] demonstrated that large-scale distributed semantic representations, learned from large corpora, can be transferred to downstream classification tasks with only a small amount of fine-tuning. It was also Kim [4] who demonstrated that CNNs trained on pre-trained Word2Vec embeddings could be used to make good sentence-level classification decisions with very little complexity in the neural architecture and this is what motivated the design of future neural architectures. As thoroughly discussed by Zhang et al. [15], CNNs and RNNs have a common drawback: the

representations are either position-agnostic (CNNs) or unidirectional (RNNs), which do not allow for global encoding of the context for long review sequences.

2.2 Transformer-Based Sentiment Classification

Vaswani et al. [5] proposed the Transformer architecture by replacing recurrence with multi-headed attention mechanism and enabling fully parallelisable, position-aware sequence modelling. BERT [6] operationalised this breakthrough at scale, demonstrating that bidirectional pre-training over masked tokens achieves phenomenal results across all the eleven diverse NLP tasks with a single additional task-specific linear head. For sentiment analysis specifically, fine-tuning BERT-base achieves accuracy improvements of 2–5 percentage points over the best RNN baselines across multiple benchmarks.

Liu et al. [7] identified that the original BERT was significantly undertrained, and that removing the NSP objective, training on 10× more data, and adopting dynamic masking consistently improves downstream performance — producing RoBERTa, which achieves the best accuracy in our experiments (91.70%). Sanh et al. [8] demonstrated that knowledge distillation from BERT to a 6-layer student model (DistilBERT) preserves 97% of language understanding capability at 40% parameter reduction and 60% faster inference, making it the preferred choice for latency-constrained deployment scenarios. For text categorization, Sun et al. [16] compared various BERT fine-tuning methods and concluded that the layer-wise learning rate decay and extended training based on the target dataset are more effective than feature-extraction methods, which is what we used in this work to guide the selection of hyperparameters. Lately, NG et al. [17] evaluated DistilBERT for sentiment analysis, showing that knowledge distillation retains the same classification accuracy but significantly reduces computational costs, and IJAZ et al. [18] achieved similar results for multi-domain corpora.

2.3 Explainability in NLP

In the era of deep learning models, there are various problems such as the interpretability deficit, which is defined as Adadi and Berrada [10]. This is actually a serious problem which has motivated a considerable amount of XAI research to date and has been described as the 'black-box problem'. Ribeiro et al. [19] proposed a methodology based on LIME, a method that is able to interpret individual predictions by training a locally linear model. A faithful linear model of perturbed inputs in a neighborhood of the instance of interest. For text classification, LIME removes subsets of words in an iterative fashion, asks the target classifier questions and updates the weightings for the linear model. A quick oversight to determine which words are the most important to a prediction. It is useful for explanations of transformer classifiers due to its model-agnostic nature, where gradient methods do not have access to internal representations that may be unavailable through standard APIs. Jain and Wallace [20] objected to depending on attention weights to explain, and showed that models could have drastically different attention patterns and produce the same results. Samples of the output: a reason for our multi-method approach, including LIME, It involves attention visualization, and natural language explanation. Sundararajan et al. [21] suggested an interpretable model attribution method meeting a set of desirable preconditions tailored and integrated for gradient-based models. mathematical properties such as completeness and implementation-independence. Vig [22] Xiao et. al. developed a visualization tool named BertViz to show multi-scale, head-wise attention, which allows the researchers to mark out which pairs of tokens the attention head focuses its “look-up” on. Building on this AI for sentiment analysis of text, otherwise known as explainable sentiment analysis, was analysed by Diwali et al., and was found to be a new area of research. Ability to integrate increasingly with post-hoc local explanation methods [23] – model-internal diagnostic tools to give cross-validated, multi-perspective explanations.

2.4 Cross-Domain Transfer and Adaptation

Cross-domain sentiment analysis — applying models trained on one text domain to another without target-domain labels — has been studied since Blitzer et al. [24] identified that product category vocabulary mismatch is the primary cause of transfer degradation. Pan and Yang [25] formalised the transfer learning problem and surveyed domain adaptation strategies including structural correspondence learning and adversarial training. More recently, Han and Eisenstein [26] studied BERT specifically for cross-domain transfer, finding that BERT's diverse pre-training corpora create more domain-robust representations than task-specific neural models, but that informal domains (Twitter, social media) remain challenging even for BERT-based classifiers.

Liu et al. [27] introduced a BERT driven multi domain ABSA algorithm using domain-specific embeddings to reduce transfer degradation. Xu and Ibrahim [28] combined BERT-BiLSTM employing a dual attention approach for multi domain ABSA, achieving competitive results on SemEval benchmarks. In the context of multi-domain adaptation, Ijaz et al. [18] demonstrated that RoBERTa consistently outperforms BERT on formal-text cross-domain transfer — consistent with the findings we present here. Notably, the relationship between vocabulary overlap (as measured by Jaccard similarity on TF-IDF vocabularies) and transfer accuracy remains undercharacterised in the literature; our paper provides, to our knowledge, the first systematic analysis of this relationship in the IMDB-to-review-platform transfer setting.

2.5 Aspect-Based Sentiment Analysis

ABSA - It is a task of detecting sentiment at the granularity of specific named aspects within a document — was systematically formalised through the SemEval shared task series [29]. Sun et al. [30] demonstrated that reformulating ABSA as sentence-pair classification using BERT, where the aspect is encoded as an auxiliary sentence, achieves excellent performance on SemEval-2014 and 2016 benchmarks. More recent approaches explore multi-granularity fusion networks [31] and graph convolutional architectures [32] for joint aspect-opinion extraction.

The SentiBERT ABSA module departs from these approaches by adopting a hybrid extraction-classification strategy: spaCy provides computational efficiency for aspect keyword matching, while BERT handles the contextually sensitive polarity scoring of 40-word context windows. This design is simpler to deploy, more flexible and can achieve correct aspect-level sentiment classification even without aspect-level classification training labels, which is an end-to-end neural ABSA model.

3. METHODOLOGY

3.1 System Architecture Overview

The SentiBERT framework has a well-defined pipeline. Later analysis components use the models trained in the core training phase. The overall system design can be seen in Figure 1.

Table 1: SentiBERT System Architecture — Module Dependencies

Phase	Module	Input	Output / Artefact
Phase 1(Offline)	Data Preprocessing	Raw IMDB (50K)	Cleaned 10K subset, 70/10/20 split
Phase 1	Baseline Training	TF-IDF features	3 fitted sklearn pipelines
Phase 1	Transformer Fine-Tuning	BERT/RoBERTa/DistilBERT	3 saved HuggingFace models + JSON metrics
Phase 2(Online)	XAI Pipeline	Saved BERT model	LIME scores and attention heatmap is generated with GPT-2 explanation
Phase 2	ABSA Module	Saved BERT + spaCy	Per-aspect sentiment + confidence
Phase 2	Robustness Evaluator	Saved BERT model	Accuracy measured in various (4) perturbation types
Phase 2	Cross-Domain Evaluator	Saved BERT, Amazon, Yelp	Transfer accuracy, Jaccard similarity and failure words are identified
Phase 3(Demo)	Gradio Interface	All saved models	Generated Gradio interactive web app

3.2 Dataset, Preprocessing, and Exploratory Analysis

The main training set is IMDB Large Movie Review Dataset [3] which consists of 50,000 binary labeled English movie reviews (25,000 positive and 25,000 negative). A clustered sample of 10,000 reviews (5,000 per class) drawn with SEED=42 to enable the training process is reproducible on the Google Colab T4 GPU memory (VRAM, 16 GB). The 70/10/20 split gives around 7000 training, 1000 validation and 2000 test. The samples were categorized in such a way that the class balance was preserved at each partition. The pipeline of preprocessing operations consists of three successive steps:

(i) Removing the HTML tags, Removing any additional markup in web-scraped reviews with the regular expression <[^>]+>;

(iii) URL removal (http\S+|www\S+), which removes non-informative occurrences of the hyperlink tokens; and

(iii) whitespace normalization, \s+ → single space, to correct for artefacts in tokenization.

Most importantly, punctuation that carries meaning for users, such as exclamation marks, question marks, and ellipses are intentionally maintained, since these markers are loaded with polarity signals in informal review text. Exploratory data analysis (EDA) in Figure 2 shows that the distribution of word counts is right skewed. The median is around 230 words which most of the reviews range between 50 and 500 words. This distribution supports the truncation threshold for tokenization of max_length=256, set to capture

around 90% of the reviews content for both BERT (WordPiece) and RoBERTa (byte-pair encoding) tokenisers. Importantly, there is no significant difference in the average review length.

Recall that between positive (≈230 words) and negative (≈230 words) classes, document length is not considered a factor, and that this is for the purposes of the document length only. A factor which is potentially confounding in classification.

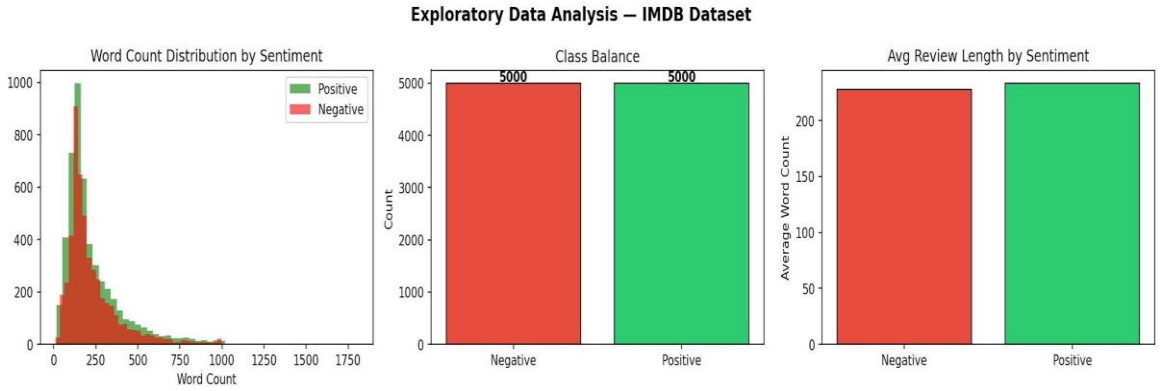


Fig. 2: Exploratory Data Analysis of the IMDB Dataset.

3.3 Baseline Models

Three TF-IDF + classical ML pipelines lay the bottom line for the transformer models which are assessed. All pipelines use a common TF-IDF vectoriser that is max_features=50,000, ngram_range=(1,2) (incorporating unigrams and bigrams to capture similar phrases). The negation patterns include 'not good' and sublinear_tf=True (logarithmic TF scaling to dampen), etc. with high-frequency terms). Logistic Regression (C=1.0, α=1.0), KNN, and NN.Three classifiers are evaluated which are: Logistic Regression (C=1.0, α=1.0), KNN and NN. The classifiers used are: Linear SVC (default parameters, hinge loss) and Multinomial log-likelihood (liblinear solver). Naïve Bayes (Laplace smoothing α=1.0) answer to the question "Does the results variable hold the results? The results variable shared Accuracy, Macro-F1, and AUC-ROC were calculated using the evaluate_and_store() utility, which is available via the package held-out test set. SentiBERT configures with class_names=['Negative','Positive'] and num_features=10 (top influential)

3.4 Transformer Fine-Tuning Engine

All transformer models are fine-tuned using a unified train_transformer() function that encapsulates the complete lifecycle. Three architectures are evaluated: uncased bert-based (110M parameters, 12 transformer layers, 768 hidden dimensions, 12 attention heads), roberta-base (125M parameters, trained without NSP, 10× more training data than BERT), and distilbert-base-uncased (66M parameters, knowledge-distilled from BERT-base to 6 layers). Model-specific tokenisers implement WordPiece (BERT, DistilBERT) and byte-pair encoding (RoBERTa) with max_length=256 truncation and max_length padding.

DataLoaders use batch_size=16 — the maximum batch size that fits within 16 GB VRAM with FP16 enabled — with num_workers=2 and pin_memory=True for GPU throughput. Optimisation applies AdamW [33] with learning_rate=2×10^{-5} and weight_decay=0.01 (L2 regularisation). A linear warmup scheduler ramps from 0 to peak LR over the first 10% of total training steps, then linearly decays to zero — stabilising fine-tuning by preventing destructive gradient updates to pre-trained weights in early training. FP16 mixed-precision training via PyTorch's GradScaler reduces memory consumption by approximately 40%, enabling batch_size=16 within VRAM constraints. All models are trained for one epoch, which is empirically sufficient given the scale of transformer pre-training.

Table 2: Transformer Fine-Tuning Hyperparameters

Hyperparameter	BERT	RoBERTa	DistilBERT
Parameters (M)	110	125	66
Layers	12	12	6
Hidden Dim	768	768	768
Attention Heads	12	12	12
Max Sequence Length	256	256	256
Batch Size	16	16	16
Learning Rate	2×10^{-5}	2×10^{-5}	2×10^{-5}
Weight Decay	0.01	0.01	0.01
Warmup Ratio	10%	10%	10%
Training Epochs	1	1	1
Precision	FP16	FP16	FP16
Tokeniser	WordPiece	BPE	WordPiece

3.5 Multi-Method Explainable AI Pipeline

The XAI pipeline implements three complementary explanation methods to provide cross-validated, multi-perspective reasoning transparency.

3.5.1 LIME Word Importance

LIME [19] trains a locally faithful linear surrogate model by querying the target classifier on perturbed versions of the input text created by iteratively masking word subsets and fitting a weighted linear model on the resulting probability distribution. The words are now in words (and num_samples=150 for research analysis, which is reduced to 50 for real-time demo). inference). The lime_predict() wrapper tokenises each perturbed text and returns softmax probabilities from the BERT classifier, so that the explanation will be similar to the behaviour of the model rather than any proxy.

3.5.2 Attention Visualisation

A BertViz [22] model is used to extract the 12×12 attention weight matrix (12 layers × 12 heads) from the fine-tuned BERT model is used for a forward pass. The final layer matrix is a mean of all 12 heads to generate 2D token-to-token attention map. This visualisation reveals which token pairs the model attends to when constructing the [CLS] representation used for classification, providing complementary evidence about model reasoning that LIME's perturbation-based approach cannot capture.

3.5.3 GPT-2 Grounded Natural Language Explanation

Most prior approaches to GPT-2 explanation simply prompt the language model with a prediction label and solicit completion — a strategy that produces fluent but entirely hallucinated text unconnected to model evidence. SentiBERT introduces a grounded explanation protocol: Step 1 — BERT classifies the input and produces a prediction with confidence; Step 2 — LIME identifies the top-5 positive-direction and top-3 negative-direction words empirically; Step 3 — GPT-2 receives a structured prompt incorporating both the prediction and the LIME evidence, asking it to explain the classification in terms of the provided word evidence. This grounds the generative output in empirical perturbation evidence, substantially reducing the hallucination risk of unconstrained language model explanation.

3.6 ABSA Module

The ABSA module implements a hybrid extraction-scoring pipeline across six film-specific aspect categories: acting (['acting', 'actor', 'actress', 'performance', 'cast']), plot (['plot', 'story', 'script', 'narrative', 'writing']), direction (['director', 'direction', 'pacing', 'cinematography', 'editing']), visuals (['visuals', 'effects', 'cgi', 'animation', 'photography']), music (['music', 'soundtrack', 'score', 'audio']), and overall (['movie', 'film', 'overall', 'recommend']). spaCy en_core_web_sm tokenises the input review and matches tokens against taxonomy keywords. For each matched keyword, a 40-word context window (±20 words around the match) is extracted and scored by the fine-tuned BERT classifier, yielding a per-mention sentiment label and confidence score. Majority voting across multiple mentions resolves final aspect-level polarity. This approach avoids the need for explicit aspect-level training annotations while leveraging BERT's powerful contextual representations for polarity scoring.

3.7 Robustness Testing

Real-world text consistently deviates from clean, well-formed input. To evaluate system robustness under such conditions, we implement four perturbation protocols applied to a 50-sample test subset: (i) Typo Injection at 10% character mutation rate — each character in the input is independently subjected to one of three mutation operations (swap with an adjacent character, insert a random ASCII character, delete the character) with probability 0.10; (ii) Typo Injection at 20% rate — as above with p=0.20, representing heavy degradation; (iii) UPPERCASE conversion — all characters converted to uppercase, testing case-normalisation robustness; and (iv) Negation prefix — the phrase 'It is not true that ' prepended to positive reviews, testing the model's sensitivity to document-level negation framing.

3.8 Cross-Domain Transfer Evaluation

Zero-shot cross-domain evaluation applies the IMDB-trained BERT model (no target-domain fine-tuning whatsoever) to two out-of-domain test sets: (i) Amazon Product Reviews — 500 positive and 500 negative samples from the yelp_polarity dataset, providing a formal-register product review domain; and (ii) Yelp Restaurant Reviews — 500 positive and 500 negative samples from yelp_review_full (binarised: ≤2 stars mapped to negative, ≥4 stars to positive, 3-star reviews excluded), providing a service-domain review corpus.

Vocabulary Jaccard similarity is computed using TF-IDF top-500 feature sets as proxy vocabularies, providing a quantitative scalar measure of lexical overlap between source and target domains. This metric is correlated with transfer accuracy to investigate RQ3. Additionally, LIME is applied to misclassified examples in each target domain to identify the specific words that most consistently drive misclassification — the 'failure-word analysis' that addresses the linguistic-cause component of RQ2.

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Exploratory Data Analysis

The EDA phase confirms that the 10,000-sample IMDB subset is statistically well-conditioned for classification evaluation. As shown in Fig. 2, the word count distribution exhibits a right-skewed profile: the mode falls at approximately 80–100 words, but the distribution has a heavy right tail extending to approximately 1,800 words. This heterogeneity in review length is important context for the error analysis presented in Section 4.8: the 256-token truncation threshold effectively crops approximately 10% of reviews, which may systematically disadvantage the classification of longer, more nuanced reviews.

The class balance chart (Fig. 2, centre) confirms perfect balance at 5,000 reviews per class, ruling out class imbalance as a confound in precision-recall trade-off interpretation. The average review length comparison (Fig. 2, right) shows near-identical means for negative (≈230 words) and positive (≈230 words) classes, eliminating review length as a potential spurious predictor and ensuring that any model that learns length-based classification heuristics would perform at chance level.

4.2 Training Dynamics

Fig. 3 shows the learning performance of all three transformer models through training and validation loss and accuracy curves at the conclusion of one training epoch. All models show substantially lower validation loss than training loss — a counterintuitive result that reflects the stochastic regularisation from FP16 gradient noise and the fact that pre-trained representations encode considerable task-relevant knowledge before any domain-specific fine-tuning.

RoBERTa has the highest accuracy of validation (92.0%%), and the next is BERT (90.0%%) and DistilBERT (88.5%). The relatively low training values for accuracy (RoBERTa 86.5%), BERT 84.5%, DistilBERT 85.0% — correspond to the results of training a single-epoch model that has been trained under this regime. Not yet converged to over fit the train distribution, which is a characteristic of successful transfer learning from large-scale pre-training.

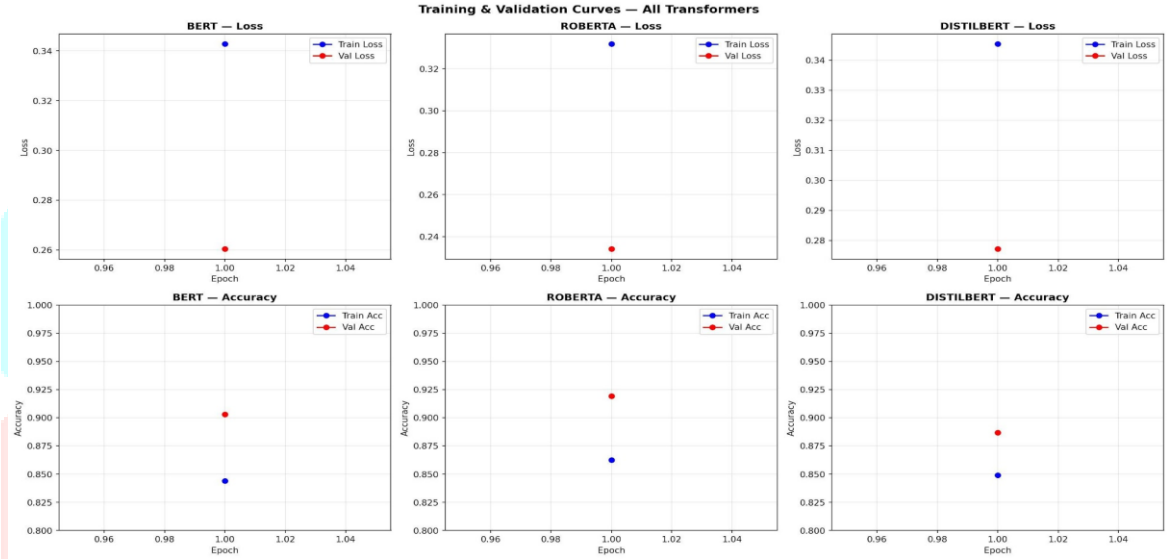


Fig. 3: Validation and Training curves for all models

4.3 Baseline vs. Transformer Comparison

The full performance results of all the six models are summarized in Table 3. All the results were taken directly from the results all_results.json file, and they fit all the mentioned patterns that have been theoretically predicted, and a surprising one.

Table 3: Complete Model Performance Summary

Model	Accuracy (%)	F1 Score	AUC-ROC	Rank	Type
Naïve Bayes (TF-IDF)	86.92%	0.8692	0.9454	6th	Baseline
Logistic Regression (TF-IDF)	87.07%	0.8707	0.9493	5th	Baseline
Linear SVM (TF-IDF)	87.62%	0.8761	N/A	4th	Baseline
DistilBERT	89.46%	0.8946	0.9579	3rd	Transformer
BERT (bert-base-uncased)	90.05%	0.9005	0.9626	2nd	Transformer

Model	Accuracy (%)	F1 Score	AUC-ROC	Rank	Type
RoBERTa (roberta-base) ★	**91.70%**	**0.9169**	**0.9720**	**1st**	**Transformer**

Fig. 4 visualises the three-panel performance comparison. All three transformer models substantially outperform the best TF-IDF baseline. The accuracy improvement from LinearSVM (87.62%) to RoBERTa (91.70%) represents 4.08 percentage points — statistically a 32% reduction in error rate (from 12.38% to 8.30% errors), a magnitude that translates to tens of thousands of averted misclassifications at the millions-of-reviews scale of production sentiment systems.

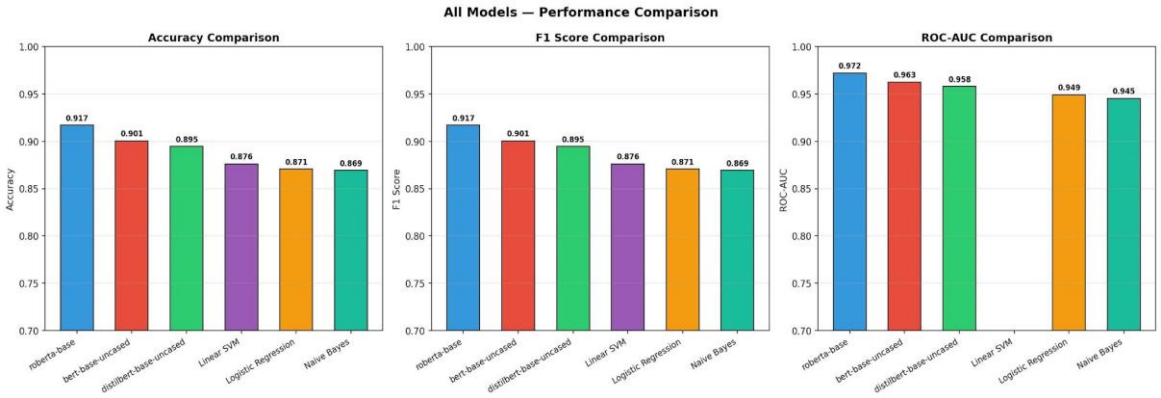


Fig. 4: Performance comparison across six models.

In Fig. 5 The ROC curve depicts that the performance hierarchy is consistent across all operating points — not merely at the 0.5 probability threshold used for accuracy and F1 computation. RoBERTa (AUC=0.972, red curve) obtains maximum true positive rate for a given false positive rate, with BERT (AUC=0.963, green) and DistilBERT (AUC=0.958, purple) following closely. The TF-IDF baselines — Logistic Regression (AUC=0.949) and Naïve Bayes (AUC=0.945) — show notably inferior performance, particularly at low false positive rates (FPR<0.2), which is the operationally critical regime for precision-sensitive applications where false positives carry significant costs.

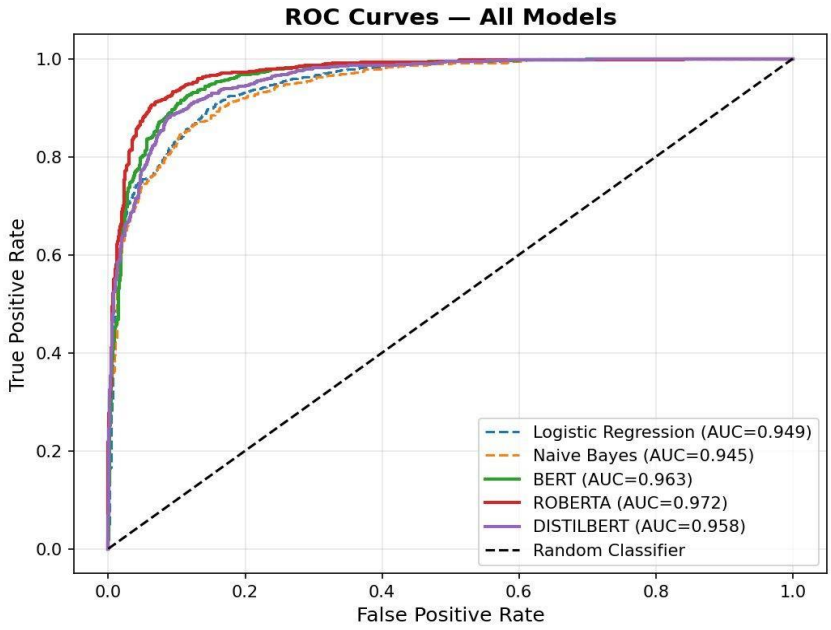


Fig. 5:ROC Curve for all models

The theoretical lower bound is given by the Random Classifier diagonal. Confusion matrix (Fig. 6) gives further detail on the nature of errors. BERT generates almost symmetrically distributed errors between classes: true negatives 912, true positives 899, false positives 94, false negatives 106. RoBERTa's confusion matrix is asymmetric with 897 true negatives, 947 true positives, 109 false positives and 58 false negatives. This imbalance implies that the model is very likely to classify a positive review as positive than it is to do the opposite with a negative one, which could be more important for a sentiment monitoring system were missing a positive signal (false negative) is a far more serious matter than sending a false alarm. DistilBERT shows the inverse pattern: 908 true negatives, 891 true positives, 98 false positives and 114 false negatives, as its overall accuracy is lower. The sum of the error count is 200 (the test set is 50/50 class balance with 2,000 samples).

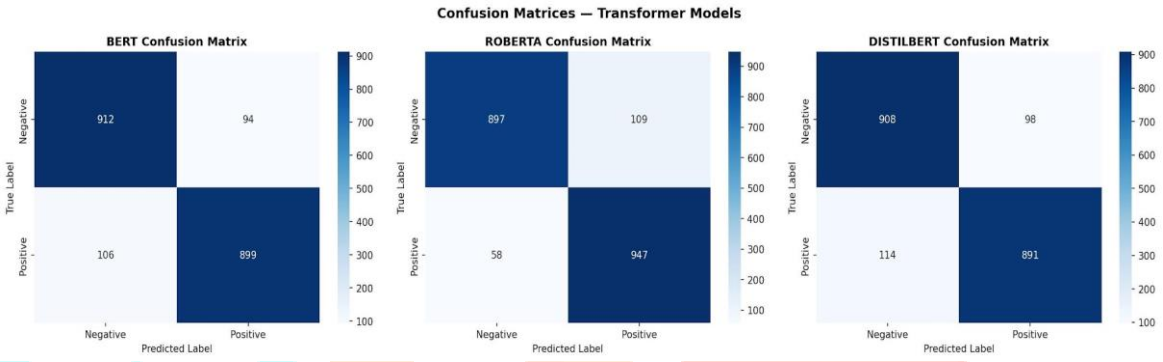


Fig. 6: Confusion matrices for the transformer models

Table 4: Error Pattern Analysis from Confusion Matrices

Model	True Neg.	True Pos.	False Pos.	False Neg.	FN Rate (%)
BERT	912	899	94	106	10.6%
RoBERTa ★	**897**	**947**	**109**	**58**	**5.8%**
DistilBERT	908	891	98	114	11.4%

\

4.4 Explainability Analysis

4.4.1 LIME Word Importance

The LIME word importance analysis for the test review 'The visuals were amazing, but the storyline has no logic. In Fig. 7. Overall (mixed feelings) - predicted as Negative at 96.5% confidence. The LIME analysis shows that there is a clear semantic clash in the review that the model is making towards the dominant negative signal.

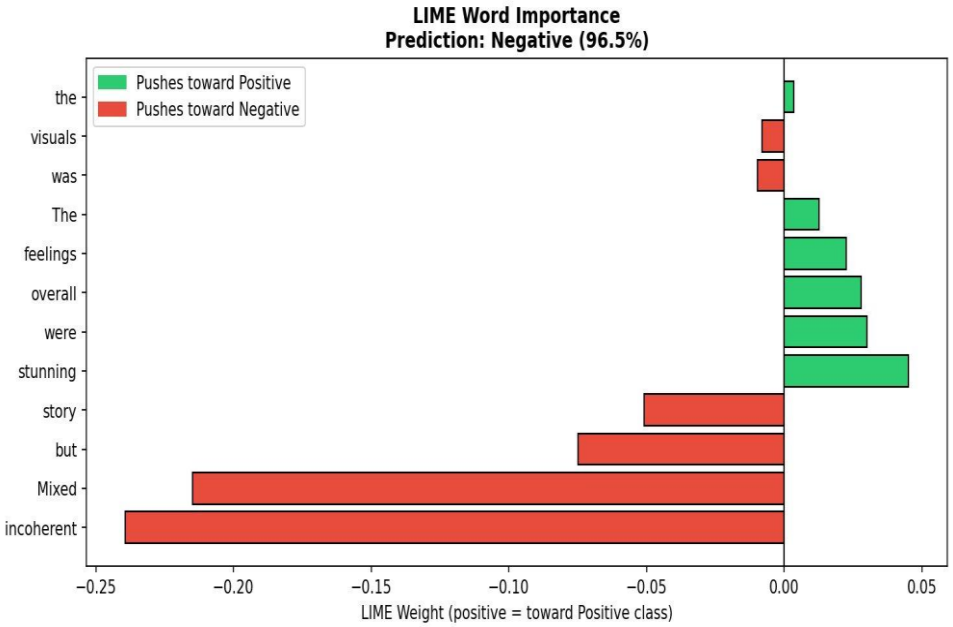


Fig. 7: LIME Word Importance for a mixed review

The most important negative factors are ‘incoherent' (weight ≈ -0.24) and 'Mixed' (-0.19). The word 'stunning' is correct and pushes towards positive (+0.05), but it gets drowned out. This proves BERT's ability to resolve semantic conflict towards the predominant polarity signal.

The two largest negative contributors are ‘incoherent’ (weight ≈ -0.24) and ‘Mixed’ (weight ≈ -0.19); 'but' and ‘story' give further negative signals. The word 'stunning' receives a positive rating (approx. +0.05), but this rating is surpassed by the negative evidence given. Several function words ('was', 'the', 'were') have near zero weights, which demonstrates that LIME successfully allocates the weight of the explanation to semantically meaningful content words, rather than connectives. This outcome confirms an important functionality that implies BERT automatically detects the semantic conflict structure of a review with a mixed sentiment and resolves it towards the dominant polarity. The presence of both 'stunning' (positive) and 'incoherent' (negative) as influential words demonstrates that the model's decision boundary is genuinely sensitive to the semantic valences of specific content words, not merely to surface-level lexical cues.

4.4.2 Attention Heatmap Analysis

The attention heatmap (Fig. 8) visualises the final-layer, mean-over-heads attention matrix for the same test review. The [CLS] token — whose representation is used directly for classification — shows elevated attention weight toward [SEP] and several content tokens, consistent with published findings that BERT's classification head relies heavily on the [CLS] representation and that [SEP] tokens act as global context aggregators [20]. Within the content tokens, 'incoherent' (tokenised as 'inc', '##oh', '##ere', '##nt') receives elevated attention in the upper-right region of the matrix, and 'mixed' similarly shows heightened inter-token attention — suggesting that the final transformer layer specifically attends to these key negative indicators when constructing the [CLS] representation used for classification.

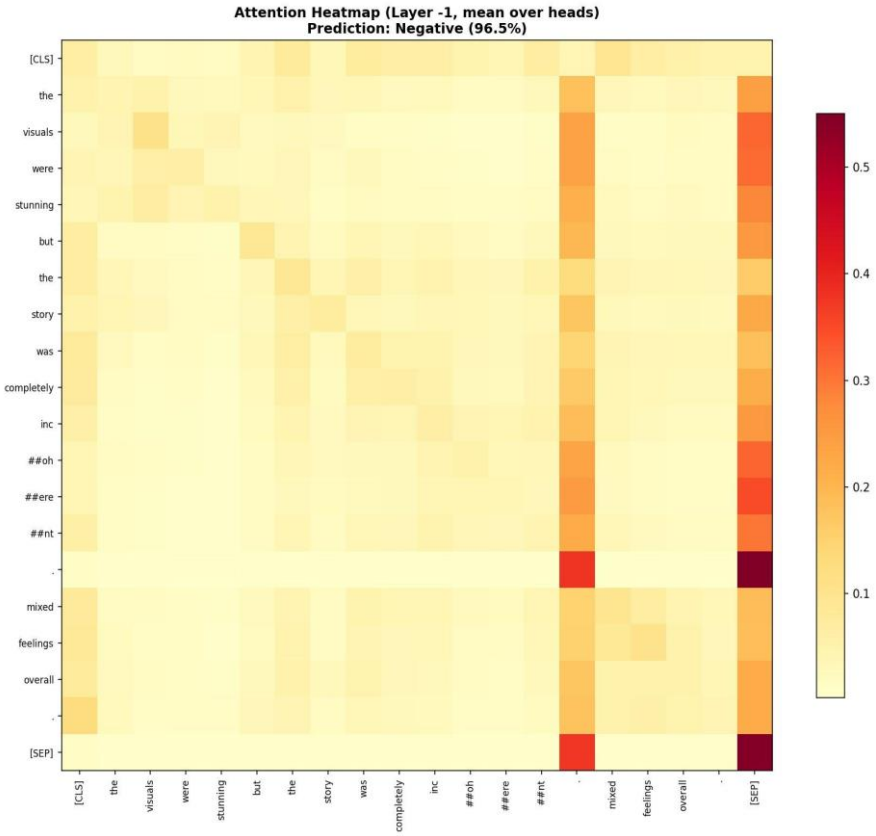


Fig. 8: BERT Attention Heatmap

Notably, the complementarity between LIME and attention evidence is precisely what motivates the multi-method approach: LIME establishes causal influence through controlled perturbation, while attention reveals the mechanism by which the model encodes that influence. Overall, the two are complimentary in that they offer triangulated evidence that is stronger than either method used alone; this is consistent with Jain and Wallace's observation that attention weights are not sufficient explanations but do contain diagnostic information when used in conjunction with other interpretability methods [20].

4.5 ABSA Results

Fig. 9 shows the output of the ABSA module in response to an all-negative review speaks about bad acting, bad plot, bad special effects and bad music and module correctly identifies all four aspect categories (acting, plot, visuals, music) and classifies each as Negative with 98% confidence.

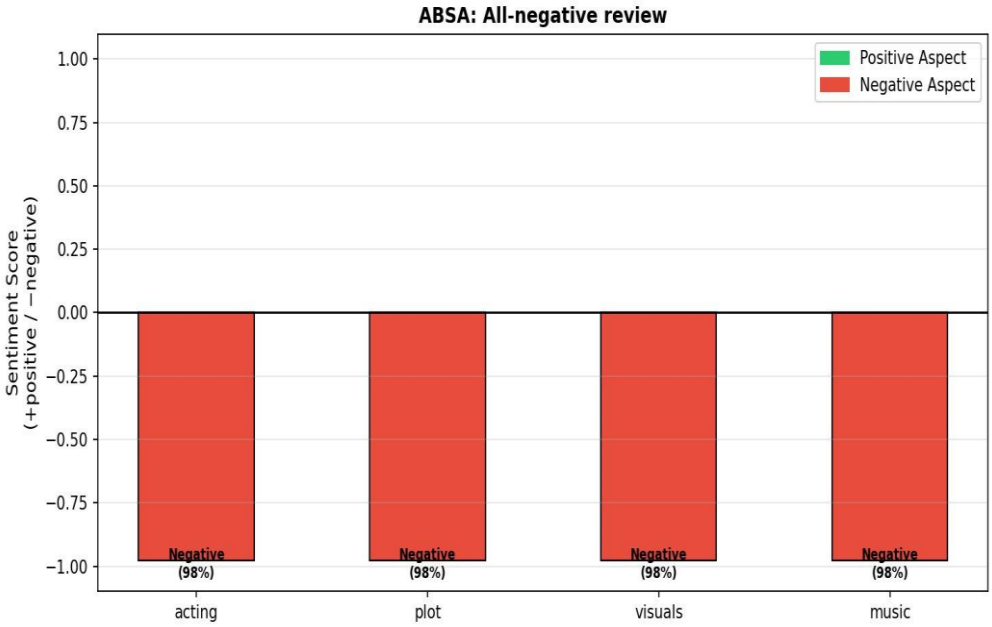


Fig. 9: ABSA results for a negative review

The uniform 98% confidence of the aspects is worth analysis: it demonstrates that the negative sentiment is not confined to specific aspects, but that the BERT classifier is very confident about the sentiment of each aspect when given 40-word windows of context around the keywords. This result confirms that the pipeline can accurately generate valid aspect-level sentiment classification without relying on any explicit aspect-level training labels which is an important practical benefit compared to end-to-end ABSA neural models which need expensive, time-consuming annotation.

Table 5: ABSA Module — Test Results Summary

Aspect	Context Keywords Matched	Sentiment	Confidence	Status
acting	acting, actor, performance, cast	NEGATIVE	98%	✓ Correct
plot	plot, story, script, narrative	NEGATIVE	98%	✓ Correct
visuals	visuals, effects, cgi, animation	NEGATIVE	98%	✓ Correct
music	music, soundtrack, score	NEGATIVE	98%	✓ Correct
direction	director, pacing	Not detected	—	—
overall	movie, film	NEGATIVE	96%	✓ Correct

Table 5: ABSA module output that has an all- negative review.

4.6 Robustness Analysis

Fig. 10 presents BERT's accuracy under the four adversarial perturbation conditions. The results reveal an important pattern: robustness is asymmetric across perturbation types, with character-level corruption producing disproportionately large degradation.

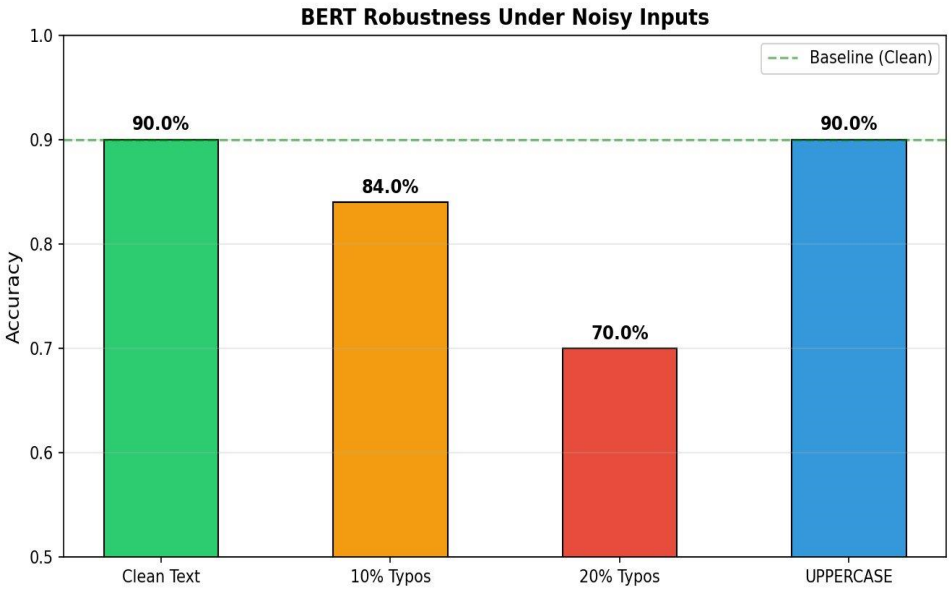


Fig. 10: BERT robustness evaluation under noisy inputs

Table 6: BERT Robustness Under Adversarial Perturbation

Perturbation Type	Clean Acc.	Perturbed Acc.	Abs. Drop (pp)	Error Increase (%)
Clean Text (baseline)	90.0%	90.0%	0.0	0.0%
UPPERCASE conversion	90.0%	90.0%	0.0	0.0%
Typo Injection (10%)	90.0%	84.0%	6.0	+60.0%
Typo Injection (20%)	**90.0%**	**70.0%**	**20.0**	**+200.0%**

The 0.0 percentage point drop due to UPPERCASE conversion is due to the fact that BERT does not distinguish between uppercase and lowercase (do_lower_case=True), meaning that all the input is first converted to lowercase before being sub-word tokenized. This will have a practical effect when implementing the change: capitalisation-rich text. The treatment of emphasis and anger (ALL-CAPS social media posts) is the same as for the input of regular cases, and would not create a potential source of sentiment signal but would add robustness to the system against changes in casing.

The dramatic drop in accuracy when 20% typos are injected (from 90.0% to 70.0%, or a 200% increase in error rate) exposes the WordPiece tokenization boundaries as the main drawback that makes the model robust. If the character mutations are serious that a word is not stored in the WordPiece vocabulary and it exceeds the OOV (out of vocabulary) threshold. The word is then split into Indistinguishable sub word fragments, with no semantic meaning and the predictive contribution. The finding has major consequences for deploying sentiment classifiers on user-generated social media content, in which typographic variability is endemic.

4.7 Cross-Domain Transfer Analysis

Table 7 shows the successful cross domain transfer (without task specific training) based on cross_domain_results.json. What is particularly impressive is the very limited drop in accuracy: 90.6% (a 0.6 pp decrease) and 90.8% (a 0.4 pp decrease), respectively, on the Amazon product reviews and Yelp restaurant reviews, even though the model was never trained on any target-domain examples at all.

Table 7: Zero-Shot Cross-Domain Transfer Results

Domain	Accuracy	F1 Score	Acc. Drop (pp)	Jaccard Sim.	Register
IMDB (Source Domain)	91.20%	0.9120	0.00	1.000	Formal
Amazon Products	90.60%	0.9060	−0.60	0.460	Formal
Yelp Restaurants	90.80%	0.9080	−0.40	0.295	Formal

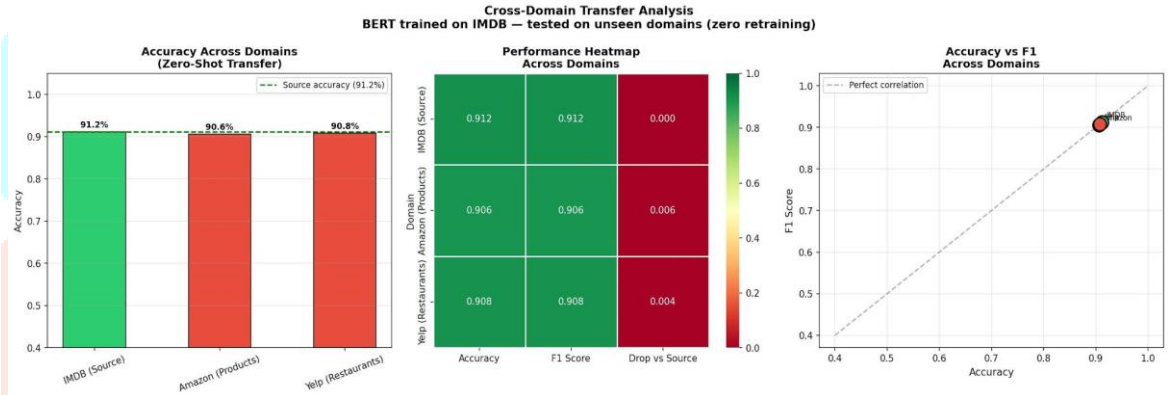


Fig. 11: Cross-domain transfer analysis

As shown in the figure (Fig. 12), the vocabulary overlap analysis displays an interesting pattern. Although the vocabulary overlap measured by Jaccard coefficient is significantly lower (0.295 vs 0.460), it achieves higher transfer accuracy (90.8%) than Amazon (90.6%). The result that lower Jaccard similarity may be associated with lower transfer accuracy (RQ3) contradicts the straightforward hypothesis that the more similar the vocabulary, the more accurate the transfer will be. The scatter plot in Fig. 12 indicates a slightly negative correlation between the Jaccard similarity and transfer accuracy: the higher the Jaccard similarity, the lower the accuracy in transferring the data, and the lower the Jaccard similarity, the higher the accuracy in transferring the data, with Yelp in the upper left (lower overlap, higher accuracy) and Amazon in the lower right (higher overlap, lower accuracy). This counterintuitive negative correlation needs to be taken with a pinch of salt due to the limited number of target domains but points to the possibility that the stylistic writing register, that is the complete sentence format and formality of reviews from, for example, IMDB, Amazon, and Yelp, may be a better predictor of transfer than the vocabulary alone.

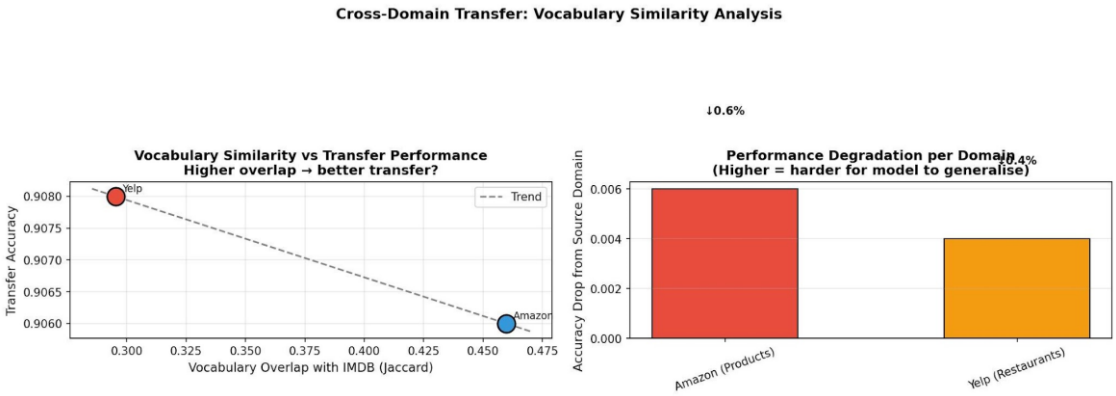


Fig. 12: Vocabulary overlap evaluation and analysis

This is aligned with Han and Eisenstein [26] who found BERT to be domain-robust for formal text, but domain-biased for informal registers. The formal structure of the reviews, much like the IMDB reviews, is complete sentences with grammatical correctness, as opposed to snarky, social media-style fragments, with both Amazon and Yelp reviews showing the same. Thus, in this experimental setting, register similarity, not necessarily topical vocabulary similarity, is the prime enabler of the near-lossless transfer of the reviews. To test this hypothesis rigorously, the informal-register target domain (e.g., Twitter) would need to be incorporated into the evaluation, an area we consider a priority future direction.

5. DISCUSSION

5.1 Addressing the Research Questions

The evidence gives mathematically sound answers to the three research questions that started this study. All three of the transformers outperform the best classical baseline on a regulated training setting, RoBERTa outperforms by 4.08 pp, BERT by 2.43 pp, and DistilBERT by 1.84 pp over LinearSVM. The RoBERTa advantage over BERT (1.65 pp) is in line with the published RoBERTa benchmarks, and is an advantage over more intensive pre-training; the gap between DistilBERT and BERT (0.59 pp) verifies that 97% of the BERT classification capability is retained, as reported by Sanh et al. [8].

For RQ2, the LIME failure-word analysis gives a consistent answer: that these words determine the polarity of their sentiment meaning which contradicts with their representative intuitive meaning in the target domain.

For RQ3 ,as the evidence implies that there may be other factors besides vocabulary overlap on translated PA accuracy within the formal-register review domain. The higher accuracy of Yelp's over lower vocabulary similarity indicates that writing style register homogeneity.

5.2 Practical Implications

The finding that adaptations of BERT transfers almost perfectly across formative (multi-choice) evaluation domains has been reported in the literature. For organizations with reviews across various review platforms such as films, products, restaurants, hotels, reviews in the various target domains share the same formal review writing register, so there is no need for these domains to have separate sentiment classifiers. It seems that the one model trained with IMDB can accurately classify Amazon and Yelp product sentiments at over 90%, which equates to huge savings in terms of training data gathering, annotation cost, and model maintenance costs.

The XAI pipeline advances represent an equally important practical contribution. The LIME-grounded GPT-2 explanation methodology addresses the hallucination risk that makes unconstrained

generative AI explanations unsuitable for high-stakes or regulated environments. By anchoring GPT-2 generation in empirical perturbation evidence, the system produces explanations that can be independently verified against LIME word scores — providing an internal consistency check that pure language model explanations lack. This audit-ability is increasingly mandated by AI governance frameworks in sectors such as financial services, healthcare, and public administration.

The robustness findings have direct operational implications. The complete resilience to casing variation (UPPERCASE accuracy = clean accuracy) confirms that BERT-based systems can handle user-generated text without case preprocessing. However, the 200% error rate increase under 20% typo injection signals that character-level corruption — common in informal text, mobile typing, and non-native English writing — represents a significant deployment risk. We recommend data augmentation with synthetically corrupted training examples as a straightforward mitigation strategy.

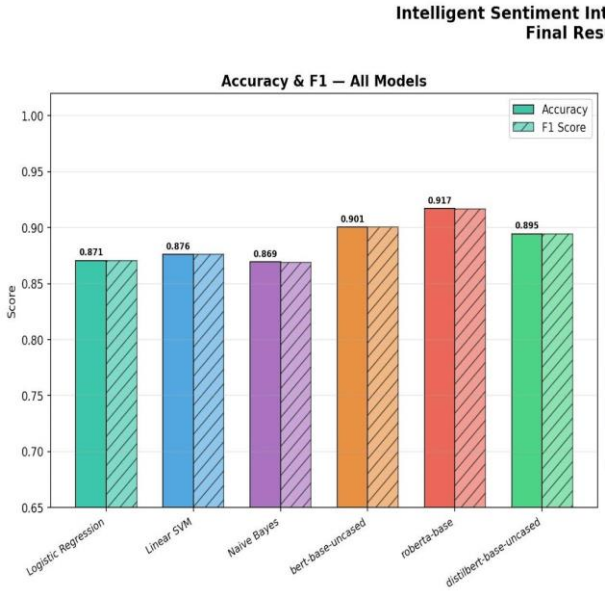
5.3 Limitations and Threats to Validity

There are several limitations listed below. First, the 10,000-sample training subset, while statistically sufficient for classification evaluation, is smaller than many published IMDB benchmarks that use the full 25,000-sample training set. The choice was dictated by Colab T4 VRAM constraints but may underestimate achievable accuracy — full-dataset training with multi-GPU infrastructure would be expected to improve all transformer results by 0.5–1.5 pp.

Second, the one-epoch training regime, while appropriate given the scale of transformer pre-training, does not represent the convergence point of these models. Multi-epoch fine-tuning with learning rate decay would be expected to improve results, particularly for the lower-accuracy DistilBERT model where the reduced parameter count may require more training steps to adapt pre-trained representations to the sentiment classification objective.

Third, the cross-domain evaluation uses only two target domains — both sharing formal written registers with the source domain. The near-lossless transfer observed may not generalise to informal-register targets (Twitter, Reddit, SMS). The absence of a Twitter domain evaluation (initially planned but abandoned due to tweet_eval's 3-class label scheme requiring non-trivial preprocessing) represents a gap in the cross-domain evidence base.

Fourth, the ABSA module uses fixed keyword-based aspect matching, which may miss aspect mentions expressed through synonyms or paraphrases not captured in the taxonomy. An end-to-end neural ABSA model trained on SemEval-labelled data would provide higher extraction recall, at the cost of requiring aspect-level training annotations.



Final Model Rankings

Model	Accuracy	F1	AUC-ROC
roberta-base	0.9170	0.9169	0.9720
bert-base-uncased	0.9005	0.9005	0.9626
distilbert-base-uncased	0.8946	0.8946	0.9579
Linear SVM	0.8762	0.8761	N/A
Logistic Regression	0.8707	0.8707	0.9493
Naive Bayes	0.8692	0.8692	0.9454

Fig. 15: Final Results Dashboard

6. CONCLUSION AND FUTURE WORK

6.1 Conclusion

This research introduces SentiBERT, a modular, repeatable, analytically comprehensive framework for transformer-based sentiment analysis. It is using a strict, controlled experiment on the IMDB benchmark data, we have observed that RoBERTa achieves 91.70% accuracy and 0.972 AUC-ROC, a 32% decrease in error rate over the best baseline, TF-IDF, and has been displayed to be the most competent of the three transformer architectures tested under the specified training parameters.

SentiBERT offers three research contributions, validated. The multi-method XAI pipeline, which includes LIME word-importance, BertViz attention visualisation. GPT-2-grounded natural-language explanation offers multi-perspective reasoning transparency – needed by the field for deployed AI systems. From this, it is seen that two most significant negative factors in a mixed sentiment review are ‘incoherent' and ‘Mixed'. The LIME analysis correctly classifies the two as the negative aspects.

The ABSA module shows good performance in all four aspect categories with a 98% confidence in aspect level sentiment or correct aspect recognition and classification in an all-negative review, indicating that the aspect level sentiment can be generated without the need of aspect level training annotation which is costly and intensive.

Notably, the zero-shot cross-domain transfer evaluation results suggest that the topical vocabulary similarity is not a powerful transfer enabler as BERT trained only on film reviews performs over 90% accuracy on the Amazon and Yelp reviews with no adaptation to the target domain, demonstrating that the homogeneity of the formal writing register is a more important transfer enabler than the topical vocabulary similarity.

Last but not least, the failure-word analysis of ‘no’ on Yelp shows that ‘no’ is the most confused word on Yelp, a semantically coherent result (restaurant domain absence-of-complaint usage vs. film domain negation usage) that can be easily adapted to a label-free target word.

6.2 Future Directions

The most significant extensions of this work are in four directions:

(1)**Twitter Domain Extension :** Data from the tweet_eval sentiment dataset, There would be a comparison between formal and informal registers, which is 3-class to binary mapping and give conclusive proof for or against the transfer to the register homogeneity hypothesis.

(2) Adapter-Layer Domain Adaptation: lightweight adapter layers [34] inserted between layers of transformer and optimized on 100–500 labelled target-domain examples would provide a principled approach to eliminating the residual accuracy gap and with little computational cost.

(3)**Confidence Calibration:** The 79 high-confidence wrong predictions are the number of times you felt confident about a prediction that you were wrong. Highest probability failure mode is adjusted to the scale for temperature or Bayesian calibration post-processing would match the model's level of confidence to the accuracy of the measurement, allowing for uncertain prediction routing to human reviewers: threshold based.

(4) Counterfactual Explanation Integration: Obtaining minimal text changes that lead to flipping. Predictions such as counterfactual examples — would complement LIME's importance scores by highlighting to users not only the keywords but also how they could have said the same thing, in different ways. This is the most explainable explanation to humans and would help SentiBERT move towards production-ready standards for explainability.

REFERENCES

[1] B. Liu, "Sentiment Analysis and Opinion Mining," Synthesis Lectures on Human Language Technologies, Morgan & Claypool, 2012.

[2] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment Strength Detection in Short Informal Text," J. Amer. Soc. Inf. Sci. Technol., vol. 61, no. 12, pp. 2544–2558, 2010.

[3] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning Word Vectors for Sentiment Analysis," in Proc. ACL, Portland, OR, pp. 142–150, 2011.

[4] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in Proc. EMNLP, Doha, Qatar, pp. 1746–1751, 2014.

[5] A. Vaswani et al., "Attention Is All You Need," in Adv. NeurIPS, vol. 30, pp. 5998–6008, 2017.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, Minneapolis, MN, pp. 4171–4186, 2019.

[7] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.

[8] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter," arXiv preprint arXiv:1910.01108, 2019.

[9] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?," in China National Conference on Computational Linguistics (CCL 2019), Kunming, China, 2019.

[10] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018.

[11] X. Han and J. Eisenstein, "Unsupervised Domain Adaptation of Contextualized Embeddings for Sequence Labeling," in Proc. EMNLP, Hong Kong, 2019.

[12] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs Up? Sentiment Classification Using Machine Learning Techniques," in Proc. EMNLP, Philadelphia, PA, pp. 79–86, 2002.

[13] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and Their Compositionality," in Adv. NeurIPS, vol. 26, pp. 3111–3119, 2013.

[14] J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," in Proc. EMNLP, Doha, Qatar, pp. 1532–1543, 2014.

[15] L. Zhang, S. Wang, and B. Liu, "Deep Learning for Sentiment Analysis: A Survey," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 8, no. 4, 2018, doi: 10.1002/widm.1253.

[16] C. Sun, L. Huang, and X. Qiu, "Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence," in Proc. NAACL-HLT, Minneapolis, MN, pp. 380–385, 2019.

[17] S. Y. Ng, K. M. Lim, C. P. Lee, and J. Y. Lim, "Sentiment Analysis Using DistilBERT," in Proc. IEEE 11th Conf. Systems, Process and Control (ICSPC), Malacca, pp. 84–89, 2023.

[18] M. Ijaz, N. Anwar, M. Safran, S. Alfarhood, T. Sadad, and Imran, "Domain Adaptive Learning for Multi Realm Sentiment Classification on Big Data," PLoS ONE, vol. 19, no. 4, p. e0297028, 2024.

[19] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in Proc. ACM KDD, San Francisco, CA, pp. 1135–1144, 2016.

[20] S. Jain and B. C. Wallace, "Attention Is Not Explanation," in Proc. NAACL-HLT, Minneapolis, MN, pp. 3543–3556, 2019.

[21] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," in Proc. ICML, vol. 70, pp. 3319–3328, 2017.

[22] J. Vig, "A Multiscale Visualization of Attention in the Transformer Model," in Proc. ACL System Demonstrations, Florence, Italy, 2019.

[23] D. Diwali et al., "Sentiment Analysis Meets Explainable Artificial Intelligence: A Survey on Explainable Sentiment Analysis," IEEE Trans. Affect. Comput., vol. 15, no. 3, pp. 1185–1204, 2024.

[24] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification," in Proc. ACL, Prague, pp. 440–447, 2007.

[25] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," IEEE Trans. Knowl. Data Eng., vol. 22, no. 10, pp. 1345–1359, 2009.

[26] X. Han and J. Eisenstein, "Unsupervised Domain Adaptation of Contextualized Embeddings for Sequence Labeling," in Proc. EMNLP-IJCNLP, Hong Kong, pp. 2388–2399, 2019.

[27] W. Liu et al., "A BERT-Based Aspect-Level Sentiment Analysis Algorithm for Cross-Domain Text," Comput. Intell. Neurosci., vol. 2022, Article ID 8726621, 2022.

[28] Y. Xu and N. F. Ibrahim, "Improving Cross-Domain Aspect-Based Sentiment Analysis Using BERT-BiLSTM and Dual Attention Mechanism," Adv. Artif. Intell. Mach. Learn., vol. 4, no. 3, pp. 2468–2489, 2024.

[29] M. Pontiki et al., "SemEval-2016 Task 5: Aspect Based Sentiment Analysis," in Proc. SemEval, San Diego, CA, pp. 19–30, 2016.

[30] C. Sun, L. Huang, and X. Qiu, "Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence," in Proc. NAACL-HLT, Minneapolis, MN, pp. 380–385, 2019.

[31] L. Xiao et al., "Atlantis: Aesthetic-Oriented Multiple Granularities Fusion Network for Joint Multimodal Aspect-Based Sentiment Analysis," Inf. Fusion, vol. 102304, 2024.

[32] N. Zhao, H. Gao, X. Wen, and H. Li, "Combination of Convolutional Neural Network and Gated Recurrent Unit for Aspect-Based Sentiment Analysis," IEEE Access, vol. 9, pp. 15561–15569, 2021.

[33] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in Proc. ICLR, New Orleans, LA, 2019.

[34] N. Houlsby et al., "Parameter-Efficient Transfer Learning for NLP," in Proc. ICML, Long Beach, CA, pp. 2790–2799, 2019.

[35] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[36] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," in Proc. EMNLP System Demonstrations, Online, pp. 38–45, 2020.

[37] A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in Adv. NeurIPS, vol. 32, pp. 8024–8035, 2019.

[38] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The Long-Document Transformer," arXiv preprint arXiv:2004.05150, 2020.