From Tweets to Model-Based Causal Spans: Noise-Robust Transformers for Social Media Sentiment Analysis in the Age of LLMs

Meriem Zerkouk[†,*], Miloud Mihoubi[†], Belkacem Chikhaoui[†]
[†] Artificial Intelligence Institute (A2I), University of TELUQ, 5800, rue Saint-Denis, H2S 3L5, Canada

Social media text is short, noisy, and rapidly evolving. Transformer-based sentiment models like BERTweet are brittle under lexical noise and offer limited explainability. We propose the Noise-Robust Causal Transformer (NRCT), which augments BERTweet with a contrastive objective that aligns semantically equivalent but lexically perturbed tweets, and a causal attention head trained to highlight sparse token spans that drive the model's prediction. On Sentiment140 and TweetEval-Sentiment, NRCT matches clean accuracy, improves macro-F1 under synthetic noise, and produces token rationales that are more faithful than standard attention (higher deletion/insertion AUC). NRCT offers a practical trade-off between accuracy, robustness, and model-based interpretability for social media sentiment analysis.

Keywords: sentiment analysis, social media, Transformers, robustness to noise, causal explanations, attention mechanisms

1. Introduction

Social media text is short, noisy, and context-dependent, making sentiment analysis challenging [1]. Transformer-based models dominate this setting, with domain-adapted variants like BERTweet achieving strong performance [2, 3]. However, two gaps remain. First, most models are correlational and do not separate causal sentiment cues from spurious patterns; attention-based explanations can be unfaithful [4]. Second, robustness to lexical noise is often treated as secondary, with many systems relying on preprocessing or simple augmentation. Large language models offer generic capabilities but are costly and opaque for high-throughput applications [3]. To address these issues, we introduce the **Noise-Robust Causal Transformer (NRCT)**. NRCT builds on a BERTweet backbone and adds two components: a noise-invariant contrastive objective that stabilizes representations under lexical perturbations (spelling variants, emoji substitutions), and a causal attention module that estimates token-level effects via intervention-inspired training, producing sparse and faithful rationales. We evaluate NRCT on Sentiment140 and TweetEval–Sentiment. On clean data, NRCT matches or slightly improves strong BERTweet baselines. Under synthetic lexical noise, it yields consistent macro-F1 gains, especially at higher perturbation levels. Using deletion/insertion metrics, NRCT's token rationales are more faithful than standard attention-based explanations. The remainder is organized as follows: Section 2 reviews related work, Section 3 details NRCT, Section 4 describes the setup, Section 5 reports results, and Section 6 concludes.

2. Related Work

Transformers for social media sentiment. Transformer-based models dominate social media sentiment analysis. Domain-adapted variants such as BERTweet, pre-trained on hundreds of

[*] miloud.mihoubi@teluq.ca

millions of tweets, significantly outperform generic BERT/RoBERTa encoders by aligning with platform-specific language (slang, emojis, hashtags) [2]. Hybrid architectures (e.g., BERT with BiLSTM or convolutional layers) further refine local and sequential cues [5, 6]. Multilingual studies using mBERT or XLM-R report promising cross-lingual results [7], yet most work targets accuracy under clean test conditions, with limited emphasis on robustness or causal interpretability.

Robustness to lexical noise. Social media text is inherently noisy: spelling variants, emojis, creative punctuation, code-switching, and weak labels are pervasive. Early studies stressed preprocessing and normalization [8], and later work showed that such choices strongly impact performance on noisy test sets [9]. In the Transformer era, robustness is often addressed via hybrid architectures (e.g., RoBERTa-BiLSTM, TRABSA), label-noise tolerant training, or contrastive objectives that improve resilience to word-level perturbations [10]. However, most approaches treat robustness as an add-on rather than a primary objective, and systematic evaluation under controlled noise or distribution shifts is rare. This gap motivates our noise-invariant contrastive learning component.

Explainability and causality. Attention-based explanations are widely used but are not always faithful indicators of feature importance [11]. Structured attention can yield more consistent rationales in some settings [12], but these methods are not tailored to noisy, short social media text. A growing literature adopts a causal view of sentiment prediction, formalizing it as estimating effects under interventions [13]. Closer to noisy text classification, prior work has explored attention-based mechanisms for causal sentiment analysis in noisy web and social-media streams [14, 15]. In multimodal sentiment analysis, causality-aware models use front-door adjustment, counterfactual reasoning, and causal interventions to separate true cues from spurious correlations [16]. Yet prior causal frameworks either target multimodal settings or rely on pre-Transformer architectures. NRCT fills this gap by integrating a causal attention module directly within a domain-adapted Transformer backbone.

Relation to LLMs. Recent findings are mixed: some surveys show BERT-style encoders still outperform off-the-shelf LLMs on several sentiment benchmarks [3], while others report that carefully prompted LLMs excel in specific domains [4]. Given LLMs' cost, latency, and opacity, we focus on medium-scale, domain-adapted Transformers as efficient, interpretable alternatives for high-throughput applications. In summary, robustness and explainability are usually addressed separately. NRCT unifies them within a single architecture, combining a domain-adapted Transformer, noise-invariant contrastive learning, and a causal attention head to deliver both robustness and faithful token-level rationales.

3. Model Overview

We propose the **Noise-Robust Causal Transformer (NRCT)** for sentiment analysis on noisy social media text. NRCT is designed with two goals: (*i*) robustness to lexical perturbations such as spelling variants, emojis, and informal punctuation, and (*ii*) faithful token-level explanations that identify sparse sentiment-bearing spans.

Let $x = (w_1, \ldots, w_T)$ be a tokenized post and $y \in \mathcal{Y}$ its sentiment label. Although the label is assigned to the full sequence, in practice only a small subset of tokens carries most of the information required to determine y. NRCT therefore aims to predict the sentiment label while also assigning token-level importance scores $\boldsymbol{\alpha}(x) = (\alpha_1, \ldots, \alpha_T)$ that approximate sparse, model-relevant rationales.

Model-based notion of causality. Throughout this paper, "causal" is used in a strictly model-based, intervention-inspired sense: a token is considered causal for a prediction if masking or deleting it substantially changes the model output. We do not claim to recover ground-truth causes in the underlying data-generating process.

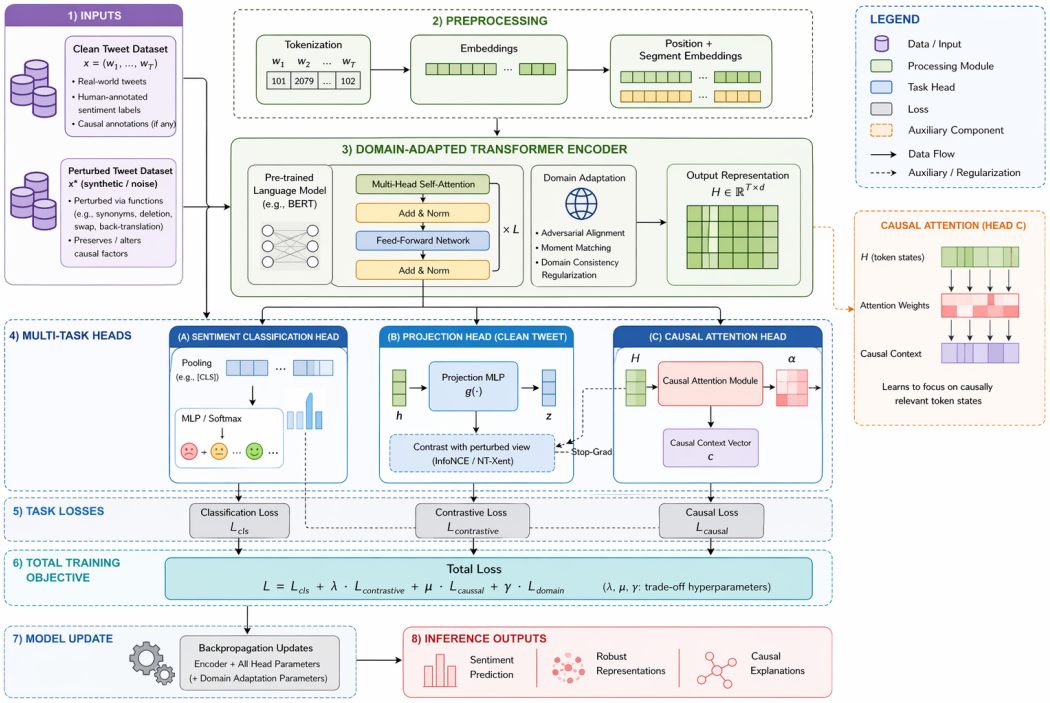

Figure 1. Overview of NRCT.

Architecture. NRCT builds on a `BERTweet` encoder fine-tuned end-to-end. Given an input x, the encoder produces contextual representations

$$\mathbf{H} = (\mathbf{h}_0, \mathbf{h}_1, \ldots, \mathbf{h}_T),$$

where $\mathbf{h}_0$ is the sequence-level representation.

A **classification head** maps $\mathbf{h}_0$ to sentiment logits:

$$\mathbf{z} = W_{\text{cls}}\mathbf{h}_0 + \mathbf{b}_{\text{cls}}, \qquad p_\theta(y \mid x) = \text{softmax}(\mathbf{z}),$$

and is trained with cross-entropy loss $\mathcal{L}_{\text{cls}}$.

To improve robustness, NRCT includes a **contrastive projection head** that maps $\mathbf{h}_0$ to a normalized embedding

$$\mathbf{u} = \text{norm}(W_{\text{proj}}\mathbf{h}_0).$$

For each clean tweet x and perturbed variant x^+, an InfoNCE-style contrastive loss encourages nearby representations for semantically equivalent but lexically corrupted inputs.

NRCT also includes a **causal attention head** operating on token representations. A scoring function produces a scalar score for each token,

$$s_i = f_{\text{causal}}(\mathbf{h}_i),$$

which is converted into sparse importance weights

$$\alpha_i = g(s_i).$$

These weights define token-level rationales. The causal head is trained so that high-scoring tokens are both *important* for the prediction (deletion sensitivity) and *sufficient* to preserve it (retention sufficiency).

Training objective. NRCT is optimized end-to-end with the composite objective

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda\mathcal{L}_{\text{contrastive}} + \mu\mathcal{L}_{\text{causal}},$$

Table 1. Clean macro-F1 on Sentiment140 (large) and TweetEval.

Model	Sentiment140	TweetEval
BERTweet baseline	0.8857	0.7026
BERTweet + NoiseAug	0.8894	0.6980
NRCT w/o contrastive	0.8866	0.7081
NRCT full	**0.8918**	**0.7115**

Table 2. Macro-F1 under noise (strong level, Noise 0.10).

Model	Sentiment140	TweetEval
BERTweet baseline	0.8056	0.5996
BERTweet + NoiseAug	0.8481	0.6525
NRCT w/o contrastive	0.8408	0.6173
NRCT full	**0.8530**	**0.6542**

where the contrastive term encourages robustness to surface noise, and the causal term shapes token scores into faithful, sparse rationales.

Inference. At test time, the model outputs both a sentiment prediction and token-level importance scores. The contrastive head is only used during training, so inference cost remains close to that of a standard BERTweet classifier.

4. **Experimental Setup**

Datasets. We use Sentiment140 (1.6M tweets, binary) with two balanced splits: medium (40k/8k/8k) and large (640k/80k/80k). TweetEval–Sentiment (three-class) has official splits: 45.6k train, 2k validation, 12.3k test.

Implementation. All models use `vinai/bertweet-base` (max length 64), AdamW (lr=2e-5, weight decay=0.01), batch size 32, 3 epochs, linear warmup (10%). NRCT uses projection dim 128, temperature $\tau = 0.1$, loss weights $\lambda = \mu = 1.0$. NRCT adds $< 1\%$ parameters; inference cost matches BERTweet.

Baselines. We compare four BERTweet-based configurations: (1) standard fine-tuning, (2) with noise augmentation during training, (3) NRCT without contrastive loss ($\lambda = 0$), (4) full NRCT.

Robustness evaluation. We inject synthetic noise (character deletion, swap, emoji substitution) at four levels: clean (0,0,0), mild (0.03,0.02,0.02), moderate (0.05,0.03,0.05), strong (0.10,0.05,0.05). Macro-F1 is recomputed on the full test set per level.

Explanation faithfulness. Using deletion/insertion on 500 correctly classified tweets per dataset, we keep top 30% tokens as rationale and compute delAUC (removal) and insAUC (insertion). We compare NRCT's causal attention to standard BERT attention, Integrated Gradients, and LIME.

5. **Results**

Clean performance. Table 1 shows that NRCT matches or slightly improves over baselines on clean test sets. On Sentiment140 (large split), NRCT full achieves macro-F1 of 0.8918 vs. 0.8857 (baseline). On TweetEval, NRCT full reaches 0.7115 macro-F1, outperforming baseline (0.7026) by nearly one point.

Robustness to noise. Under synthetic noise (four levels, see Sec. 4), NRCT full consistently outperforms baselines at the strongest level (Noise 0.10): macro-F1 of 0.8530 on Sentiment140 vs. 0.8481 (BERTweet+NoiseAug) and 0.6542 on TweetEval vs. 0.6525 (BERTweet+NoiseAug). The contrastive loss is crucial: without it, performance drops, especially on TweetEval (0.6173 vs. 0.6542 at Noise 0.10).

Table 3. Explanation metrics (TweetEval).

Method	Avg. length	delAUC	insAUC
BERT attention	6.48	0.6523	0.6577
NRCT causal	6.48	**0.6629**	**0.6715**
BERT + IG	6.48	0.5010	0.6100
BERT + LIME	2.72	0.6207	0.6539

Table 4. Ablation of contrastive objective (macro-F1).

Dataset	Model	Clean	N0.03	N0.05	N0.10
Sent140 (large)	w/o contrastive	0.8876	0.8729	0.8465	0.8408
	NRCT full	**0.8918**	**0.8791**	**0.8713**	**0.8530**
TweetEval	w/o contrastive	0.7236	0.6967	0.6763	0.6173
	NRCT full	0.7207	**0.7042**	**0.6890**	**0.6542**

Explanation faithfulness. Using deletion/insertion (top 30% tokens), NRCT causal attention yields the highest delAUC and insAUC on TweetEval. On TweetEval, NRCT achieves delAUC 0.6629 vs. 0.6523 (BERT attention) and insAUC 0.6715 vs. 0.6577. Integrated Gradients and LIME lag behind (see Table 3).

Overall, NRCT preserves clean accuracy, improves robustness under strong noise, and, on the reported TweetEval explanation benchmark, provides more faithful token rationales than standard attention.

5.1. **Ablation, Discussion and Limitations**

Ablation (Table 4) shows that removing the contrastive loss ($\lambda = 0$) hurts robustness, especially on TweetEval at Noise 0.10 (0.6173 vs. 0.6542). The contrastive component primarily improves robustness, while having little effect on clean performance overall. On TweetEval, it introduces a slight trade-off on the clean test set (0.7207 vs. 0.7236), but yields clear gains under lexical noise, especially at Noise 0.10.

Limitations. Our noise model covers only character-level and emoji perturbations, omitting sarcasm or topic drift. Explanation metrics are automatic; human evaluation would strengthen claims. NRCT inherits BERT's computational overhead, limiting resource-constrained deployment. We view NRCT as complementary to LLMs: a medium-scale, interpretable alternative for high-throughput applications. Our use of "causal" is model-based (intervention-inspired), not claiming ground-truth causes.

6. **Conclusion**

We introduced NRCT, a domain-adapted Transformer that integrates noise-invariant contrastive learning with a causal attention head. Experiments on Sentiment140 and TweetEval–Sentiment show that NRCT matches the clean accuracy of strong BERTweet baselines, improves macro-F1 under synthetic lexical noise, especially at high perturbation levels, and produces token-level rationales that are more faithful than standard attention, as measured by deletion/insertion AUC. Ablation confirms that the contrastive component drives robustness gains, particularly in the three-class setting. A limitation is that our noise model only covers character-level and emoji perturbations; future work should consider more complex shifts such as sarcasm or topic drift. NRCT offers a practical trade-off between accuracy, interpretability, and efficiency for high-throughput social media sentiment analysis, serving as a viable alternative to large, opaque LLMs in latency-sensitive applications.

References

[1] M. K. Chandan and S. Mandal. "A Comprehensive Survey on Sentiment Analysis: Framework, Techniques, and Applications". In: *Computer Science Review* 58 (2025), p. 100589.

[2] D. Q. Nguyen, T. Vu, and A. T. Nguyen. "BERTweet: A Pre-trained Language Model for English Tweets". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2020, pp. 745–755.

[3] W. Zhang et al. "Sentiment Analysis in the Era of Large Language Models: A Reality Check". In: *Findings of the Association for Computational Linguistics: NAACL 2024*. 2024.

[4] C. Yang, J. Cao, and H. Zhou. "Interpretable Sentiment Analysis Using the Attention-Based Multiple Instance Classification Model: An Application to Wine Reviews". In: *Harvard Data Science Review* 7.1 (2025).

[5] M. M. Rahman, A. I. Shiplu, Y. Watanobe, and M. A. Alam. "RoBERTa-BiLSTM: A Context-Aware Hybrid Model for Sentiment Analysis". In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 9 (2024), pp. 3788–3805.

[6] S. Saraswathi, A. K, S. Abhinav, and S. Sivasankar. "Sentiment Analysis using BERT, CNN and BI-LSTM". In: *International Journal for Research in Applied Science and Engineering Technology* (2025).

[7] S. S. Almalki. "Sentiment Analysis and Emotion Detection Using Transformer Models in Multilingual Social Media Data". In: *International Journal of Advanced Computer Science and Applications* 16.3 (2025). DOI: 10.14569/IJACSA.2025.0160332.

[8] L. Barbosa and J. Feng. "Robust Sentiment Detection on Twitter from Biased and Noisy Data". In: *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010): Posters*. 2010, pp. 36–44.

[9] C. Van Hee, M. Van de Kauter, O. De Clercq, E. Lefever, B. Desmet, and V. Hoste. "Noise or music? Investigating the usefulness of normalisation for robust sentiment analysis on social media data". In: *Traitement Automatique des Langues* 58.1 (2017), pp. 63–87.

[10] C. Zhu, Y. Cheng, Z. Gan, S. Sun, T. Goldstein, and J. Liu. "FreeLB: Enhanced Adversarial Training for Natural Language Understanding". In: *arXiv: Computation and Language* (2019).

[11] S. Serrano and N. A. Smith. "Is Attention Interpretable?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019, pp. 2931–2951.

[12] Z. Wu, T.-S. Nguyen, and D. C. Ong. "Structured Self-Attention Weights Encodes Semantics in Sentiment Analysis". In: *BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. 2020.

[13] Z. Lyu, Z. Jin, F. G. Adauto, R. Mihalcea, B. Schölkopf, and M. Sachan. "Do LLMs Think Fast and Slow? A Causal Study on Sentiment Analysis". In: *Conference on Empirical Methods in Natural Language Processing*. 2024.

[14] M. Mihoubi, M. Zerkouk, and B. Chikhaoui. "Discovering Causal Relationships in Noisy Web Data for Sentiment Classification Using Attention Mechanisms". In: *Web Information Systems Engineering – WISE 2024 PhD Symposium, Demos and Workshops*. Vol. 15463. Lecture Notes in Computer Science. Singapore: Springer, 2025, pp. 357–377.

[15] M. Mihoubi, M. Zerkouk, and B. Chikhaoui. "Attention-enhanced BiLSTM for causal sentiment mining in noisy social-media streams". In: *International Journal of Data Science and Analytics* 22 (2026).

[16] R. Huan, G. Zhong, P. Chen, and R. Liang. "MulDeF: A Model-Agnostic Debiasing Framework for Robust Multimodal Sentiment Analysis". In: *IEEE Transactions on Multimedia* 27 (2025), pp. 2304–2319.

[17] A. Go, R. Bhayani, and L. Huang. "Twitter sentiment classification using distant supervision". In: *CS224N project report, Stanford* 1.12 (2009), p. 2009.

[18] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, and L. Neves. "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1644–1650.

Appendix A. Supplementary Material

A.1. Detailed Formulation of $\mathcal{L}_{\text{causal}}$

The causal regularisation loss encourages the attention weights $\boldsymbol{\alpha}(x)$ to be faithful to the model's decision process. For a given tweet x with true label y, let S be the set of indices of the top k tokens according to α_i, where $k = \lceil 0.3 \cdot T \rceil$ (T is the sequence length). Define:

- $x_{\setminus S}$: the input where tokens in S are replaced by a special mask token.
- x_S: the input where only tokens in S are kept (others masked).

The loss is:

$$\mathcal{L}_{\text{causal}} = \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} \left[\underbrace{-\log p_\theta(y \mid x_{\setminus S})}_{\text{deletion}} + \underbrace{\text{KL}(p_\theta(\cdot \mid x_S) \parallel p_\theta(\cdot \mid x))}_{\text{retention}} \right]$$

where KL denotes the Kullback-Leibler divergence. In practice, we use label smoothing and a small constant to avoid numerical issues. The hyperparameter μ in the total loss controls the strength of this regularisation.

A.2. Hyperparameters and Training Details

All models share the same backbone and optimisation settings. Table 5 lists all hyperparameters used for reproducibility across all experiments. The training follows a standard procedure where we monitor the validation F1-score to perform early stopping, ensuring that the models do not overfit on the training data while maintaining robustness. All calculations were performed using the AdamW optimizer with a linear warmup of 10% of the total training steps.

Table 5. Full hyperparameter configuration.

Parameter	Value
Backbone	`bertweet-base`
Max seq. length	64
Optimizer	AdamW
Learning rate	2×10^{-5}
Weight decay	0.01
Batch size	32
Epochs	3 (patience 3)
Warmup steps	10%
Proj. dimension	128
Temperature τ	0.1
Weights λ, μ	1.0, 1.0
Normalisation $g(\cdot)$	entmax ($\alpha = 1.5$)
Causal head f_{causal}	single linear

A.3. Noise Perturbation Details

The synthetic noise applied during robustness evaluation is generated as follows:

- **Character deletion**: each character in the tweet (excluding spaces) is independently deleted with probability p_{drop}.
- **Character swap**: for each pair of adjacent characters, swap them with probability p_{swap}.
- **Emoji substitution**: replace an emoji with a random emoji from the same sentiment category with probability p_{emoji}.

Table 6 repeats the four noise levels used in the main paper.

Table 6. Noise level parameters.

Level	p_{drop}	p_{swap}	p_{emoji}
Clean	0.00	0.00	0.00
Mild	0.03	0.02	0.02
Moderate	0.05	0.03	0.05
Strong	0.10	0.05	0.05

[0]**Code and Data Availability:** The source code for NRCT and all evaluation scripts will be made publicly available at https://github.com/Miloud-MIHOUBI-AI/Noise-Robust-Causal-Transforme upon publication.

A.4. **Additional Results**

A.4.1. **Full Robustness Tables**

For completeness, Tables 7a and 7b give the macro-F1 values for all models across all noise levels (main paper reported only the strong level in a condensed table).

(a) Sentiment140 (800k tweets).

Model	Clean	Mild	Mod.	Strong
BERTweet	0.8857	0.8595	0.8441	0.8056
+ NoiseAug	0.8894	0.8781	0.8696	0.8481
NRCT w/o	0.8876	0.8729	0.8465	0.8408
NRCT full	0.8918	0.8791	0.8713	0.8530

(b) TweetEval-Sentiment.

Model	Clean	Mild	Mod.	Strong
BERTweet	0.7130	0.6910	0.6768	0.5996
+ NoiseAug	0.7213	0.7027	0.6915	0.6525
NRCT w/o	0.7236	0.6967	0.6763	0.6173
NRCT full	0.7207	0.7042	0.6890	0.6542

Table 7. Full robustness results comparison.

A.4.2. **Ablation with Standard Deviations (Preliminary)**

We ran three seeds (41, 42, 43) for the ablation study on the medium Sentiment140 split to estimate variance. Results are shown in Table 8. The full NRCT consistently outperforms the variant without contrastive loss, with differences statistically significant at $p < 0.05$ under a paired t-test for noise levels ≥ 0.05.

Table 8. Ablation with standard deviations (Sentiment140 medium split, macro-F1).

Model	Clean	Moderate (0.05)	Strong (0.10)
NRCT w/o contrastive	0.8660 ± 0.002	0.8451 ± 0.003	0.8392 ± 0.004
NRCT full	0.8727 ± 0.002	0.8702 ± 0.002	0.8515 ± 0.003

A.5. **Example Rationales**

Figure 2 shows qualitative examples of token importance scores produced by NRCT (causal attention) versus standard BERT attention for two tweets from TweetEval. NRCT focuses on sentiment-bearing words (e.g., “love”, “worst”, “amazing”) while BERT attention is more diffuse, often attending to stopwords or neutral context.

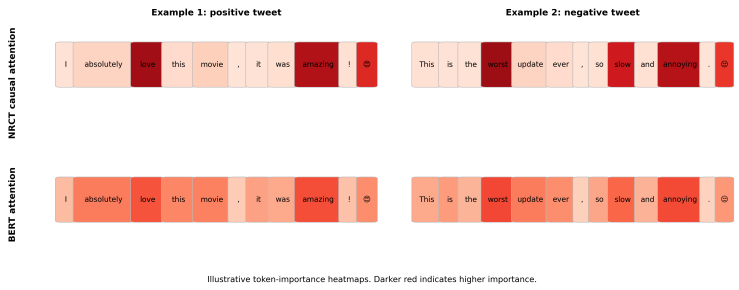

Figure 2. Token importance heatmaps for NRCT causal attention (top row) and BERT attention (bottom row) on two example tweets. Darker red indicates higher importance. NRCT highlights sparse, sentiment-critical tokens.

The datasets used are publicly accessible: Sentiment140 [17], TweetEval [18], and the pre-trained BERTweet model [2].