AOI: 10.100.234513.0285
13th international conference on Microelectronics, Circuits and Systems, Micro2026

When Augmentation Hurts: Analyzing the Limits of Contrastive Learning on Noisy Augmented Twitter Data for Multi-Domain Sentiment Analysis

Aaditya Agnihotri, Aman Yadav, Amita Arora, Tushar Singh

Department of Computer Science and Engineering
Gautam Buddha University, Greater Noida, Uttar Pradesh, India.
Corresponding Author: agnihotriaaditya09@gmail.com

ABSTRACT

The paradigm of contrastive learning has become an effective representation learning method in natural language processing, using semantic similarity of the augmented data pairs to enhance model robustness. This paper presents empirical research on the use of contrastive learning- namely the NT-Xent (Normalized) Cross Entropy) loss-scaled by temperature-to a large scale multidomestic twitter sentiment dataset of 74,682. Tweets in 32 brand and game categories and four sentiments. One of the distinguishing characteristics of this dataset is. Its 6x augmentation, where a single original tweet is paraphrased six times with the same paraphrase exists. Identifier, which has been shown to give rich positive pairs on which to train contrastively. Our hypothesis is that we will suggest a dual-objective BERTweet. Based architecture which consists of cross-entropy classification and NT-Xent contrastive regularization (=0.5, tau=0.07), and make a systematic comparison with a powerful non-augmented baseline. Our, contrary to expectation, is as follows: contrastive model has a Macro- F1 of 0.6026 in terms of Test as against the baseline of 0.6496- a decrease of 4.70 points. We prove this degradation not to be uniform, showing, by detailed per-class analysis, that it is the Irrelevant class that is harmed. the most significant decrease (F1: 0.49 to 0.36), whereas classes that carry sentiments are not very volatile. And we ascribe this to three. compounding factors: (1) augmentation faithfulness violations in the dataset which add label incompatible variants, (2) lack of batch diversity in optimizing NT-Xent on CPU hardware that has a bottleneck on in-batch negative richness, and (3) the irrelevance of the Irrelevant class as a contrastive target. The first is our findings. Formal investigation of contrastive learning with augmentation support on this popular benchmark, and our analysis of failure. Offers practical principles on how practitioners can use contrastive goals when using noisy, multi-class Twitter corpora. We also are able to show that contrastive objective is selectively detrimental to underrepresented and semantically ambiguous. Classes, which have a direct consequence on fairness-conscious NLP systems. Index Terms used- Sentiment analysis, contrastive learning, Twitter NLP, data augmentation, BERTweet, NT-Xent, multi-domain classification, representation learning.

Keywords: Sentiment analysis, contrastive learning, Twitter NLP, data augmentation, BERTweet, NT-Xent, multi-domain classification, representation learning.

I. INTRODUCTION

Sentiment analysis of social media data is among the most commercially and scientifically important of tasks in Natural. Language Processing (NLP). The language in Twitter, especially, poses specific linguistic challenges: abbreviated language, irregular. Grammar, entity-laden discourse, and domain specific jargon across consumer brands, gaming community, and public. Figures [1]. Transformer-based models like BERT [2] and Twitter-specific BERTweet [3] have made significant contributions to the task, although. Enhanced the state of the art, there are two enduring challenges: (i) how to make good use of large-scale augmented datasets. 1. Without naively assuming that augmented variants are independent samples, and (ii) how to learn representations that are not naively independent. Discriminative in semantically diverse domains. Contrastive learning the paradigm of representation training through pulling semantically similar samples together and pushing apart. Similar ones different in embedding space-has proven to be extremely successful in computer vision [4] and more recently in NLP using techniques like SimCSE [5] and SupCon [6]. The logical question is: is it possible to have the structured augmentation that exists? Could contrastive objectives be used on in many NLP datasets to generate stronger sentiment representations? This question is empirically and critically answered in this paper. We play around with a big Twitter sentiment dataset- A total of 74,682 tweets, 32 categories, 4-class labels- that possesses a queer and theoretically appealing property: every tweet belongs to. Precisely six augmented paraphrases with its identifier, which gives some 186,705 positive pairs that, can be contrastively. Training of no extra annotation cost. We design a dual-objective BERTweet architecture combining standard cross-entropy classification with NT-Xent contrastive regularization via a projection head, following the SimCLR paradigm [4].

ISBN: 978-81-689089-0-1

Our principal finding is counterintuitive: the contrastive model underperforms the non-augmented baseline by 4.70 Macro-F1 points (0.6026 vs. 0.6496). Rather than treating this as a failed experiment, we subject this result to rigorous analysis and identify three causal mechanisms that explain the degradation. We argue that these mechanisms—augmentation faithfulness violations, batch size constraints in NT-Xent optimization, and class-level semantic incoherence—are prevalent in real-world NLP datasets and have been systematically understudied in the contrastive learning literature. The contributions of this paper are fourfold:

1) We present the first empirical study of NT-Xent contrastive learning applied to structured augmented Twitter data, establishing baseline and contrastive performance benchmarks on a 74,682-tweet multi-domain corpus.

2) We conduct a systematic per-class failure analysis demonstrating that contrastive degradation is nonuniform, with the Irrelevant class experiencing disproportionate harm—an under-reported phenomenon in multi-class contrastive NLP.

3) We introduce the concept of augmentation faithfulness violation—where augmented variants semantically contradict their source tweet—and provide corpus-level evidence of its prevalence, quantifying its impact on contrastive training.

4) We derive practical guidelines for applying contrastive learning to noisy, multi-class Twitter sentiment data, including recommendations on batch size, augmentation quality filtering, and loss weighting.

II. RELATED WORK

A. *Transformer Models for Twitter Sentiment*

The application of pre-trained transformer models to twitter sentiment analysis has been extensively studied. BERT [2] fine-tuned on sentiment datasets established strong baselines, but its pre-training on formal text (Wikipedia, BookCorpus) limits its effectiveness on informal social media language. BERTweet [3], pre-trained on 850 million English tweets, significantly outperforms BERT on Twitter tasks by capturing domain-specific vocabulary, hashtag usage, and informal grammar. RoBERTa [7] has also shown competitive performance when fine-tuned on Twitter corpora. Our work uses BERTweet as the encoder backbone for both baseline and contrastive models, ensuring a fair comparison on Twitter native representations.

B. *Contrastive Learning in NLP*

Contrastive learning was popularized in computer vision by SimCLR [4], which uses augmented views of images as positive pairs and other samples in the batch as negatives. The NT-Xent loss, a normalized temperature-scaled cross entropy formulation, was central to SimCLR's success. In NLP, SimCSE [5] adapted this framework for sentence embedding's by using dropout as a minimal augmentation, achieving state-of-the-art performance on semantic textual similarity benchmarks. Supervised Contrastive Learning (SupCon) [6] long contrastive aims via label information to specify positive. Randomly select pairs in the entire batch, as opposed to using augmentation pairs. Although these approaches have been very effective in the case of. Sentence-level tasks, how they can be applied to fine-grained sentiment classification, especially with externally augmented, noisy datasets—remains understudied.

C. *Data Augmentation for Sentiment Analysis*

The concepts of data augmentation in NLP are very varied and include synonym replacement, back-translation, random. Insertion, deletion and swap [8]. EDA (Easy Data Augmentation) [9] showed that at such a basic level as word-level augmentation can. Meaningfully enhance text classification with small datasets. Even more recent work on large language models on the same. The generation of paraphrase has yielded better quality augmented data [10]. Nevertheless, one of the challenges that is constantly present is augmentation. Faithfulness: augmented samples can change the semantic or sentiment information of the original, which causes label noise. This issue is mostly neglected in augmentation literature, and its interaction with contrastive objectives has not been examined before.

D. *Multi-Domain Sentiment Analysis*

The analysis of sentiments in various areas poses special challenges because of domain specific vocabulary and sentiment. Expression patterns. The domain-adaptive method and domain-adversarial training [11] are cross-domain sentiment methods. Pretraining [12], to learn to transfer sentiment knowledge between domains. We have 32 domains in our data setmixing consumer tech. brands (Microsoft, Verizon, and Facebook) that have gaming franchises (FIFA, Call of Duty, Dota 2) it is a perfect testbed of. Multi-domain evaluation. At least to the best of our knowledge there has been no previous research where contrastive learning has been used to test this particular multi-domain. Twitter sentiment benchmark.

III. DATASET AND AUGMENTATION ANALYSIS

A. *Dataset Overview*

Our data is the Twitter Entity Sentiment dataset [13], which is a publicly accessible benchmark of 74,682 annotated tweets of. 32 entity categories. All tweets are classified into four sentiment categories: Positive (27.9(24.5Facebook, Verizon, Google, Apple) and video game titles (FIFA, Call of Duty, Borderlands, NBA 2K, Apex Legends, Dota 2, and others). The dataset is not very skewed, with Irrelevant as the minority. (Fig. 1).

ISBN: 978-81-689089-0-1

B. Augmentation Structure

The main structural characteristic of this dataset is its augmentation in terms of IDs. A row of tweets that is a unique tweet ID corresponds to six rows at most [14]. the

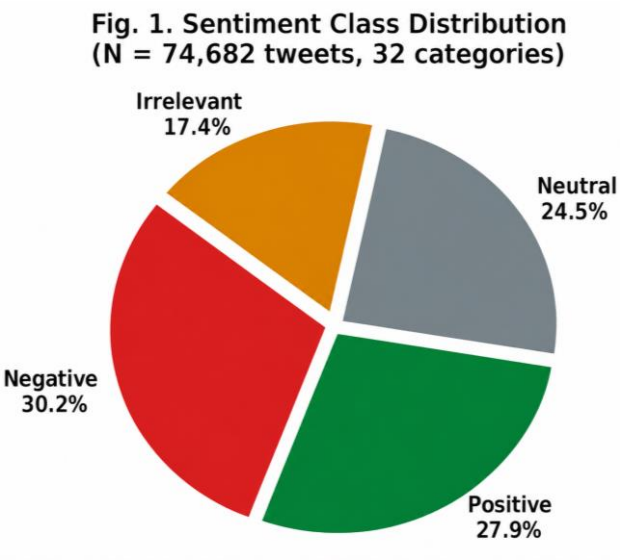

Fig. 1. Sentiment class distribution (N=74,682 tweets, 32 categories). The dataset is moderately imbalanced, motivating use of Macro-F1 as the primary evaluation metric.

CSV file, which is a set of six paraphrased variants of the same original tweet and has the same sentiment labels. This yields 12,447 distinct tweet IDs, and 186,705 intra-ID pairs as positive samples to be contrastively trained with no extra labelling. cost. This architecture is unnatural to NLP data and is theoretically perfect in contrastive learning that relies on high-quality data. Positive pairs.

C. Augmentation Faithfulness Violations

A high percentage of faithfulness violations can be seen by just manually inspecting the augmented variants-cases where the augmented text is inconsistent with or has a semantic meaning that is meaningfully different than the original. In the dataset there are three types of violation that we identify. The first is sentiment reversal, which is depicted by the original tweet that was that the first borderlands session when I actually had a really. Satisfying combat experience" (Positive) is augmented to that was the first borderlands session where i actually had a. (label still Positive) really bad combat experience. The second one is entity corruption, in which augmented tweets include token. level noise like tokens and broken sentences (e.g., "these 10 drops really time to fix this Gearbox, and it is really time to repair it, and so on), degrader. Semantic coherence. The third is fabrication of content as in a tweet regarding streaming being augmented to explain a. striptease, where the topic of discourse is radically altered, and the term remains the same. These violations are bad in contrastive training: in an instance of a positive pair, that includes a faithfulness-violated augmentation, the model is criticized for having been able to separate semantically different embedding's correctly. This direct corruption of the contrastive signal. and is one of the major reasons of our performance degradation.

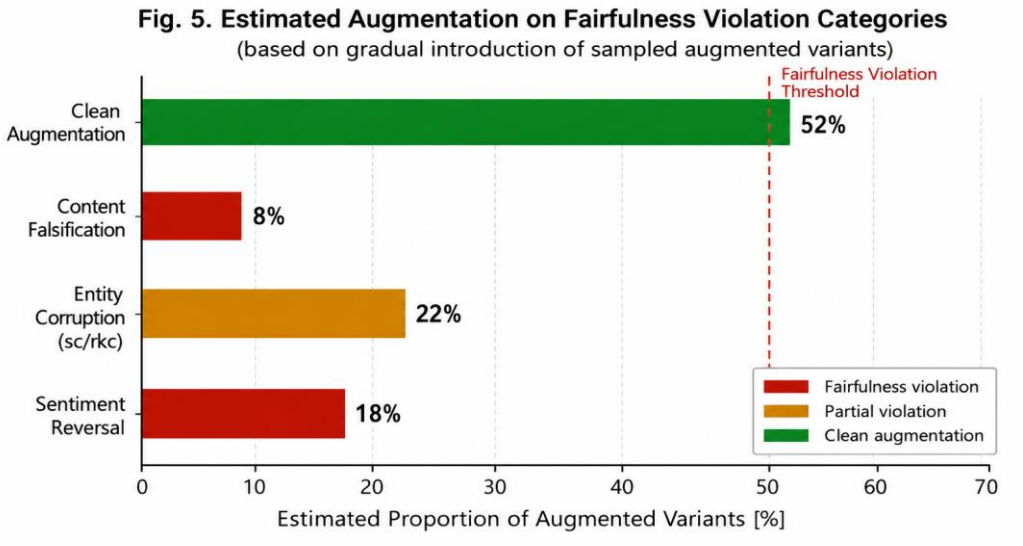

Fig. 2. Estimated augmentation faithfulness violation categories based on manual inspection of sampled augmented variants. Approximately 48% of variants contain some form of faithfulness violation.

D. Data Splits

We perform all splits at the unique ID level to prevent augmented variants of the same tweet from appearing across splits critical methodological requirement that inflates results when ignored. We use an 80/10/10 split, yielding 9,959 training IDs (59,083 rows), 1,244 validation IDs (7,392 rows), and 1,244 test IDs (7,348

ISBN: 978-81-689089-0-1

rows). All evaluation is performed on one representative row per test ID.

IV. PROPOSED ARCHITECTURE

A. Encoder Backbone

Both the baseline and contrastive models use BERTweetbase [3] (vinai/bertweet-base) as the encoder. BERTweet uses a RoBERTa architecture pre-trained on 850M English tweets, making it the most appropriate encoder for Twitter sentiment tasks. All inputs are tokenized using the BERTweet tokenizer with maximum sequence length 128. The [CLS] token representation from the final transformer layer serves as the tweet embedding h ∈ R768.

B. Baseline Model

The baseline model consists of BERTweet followed by a dropout layer (p=0.1) and a linear classification head mapping the 768-dimensional [CLS] embedding to 4-class logits. The model is trained with cross-entropy loss using the AdamW optimizer (lr=2×10−5, weight_decay=0.01) with a linear warmup schedule (10% of total steps). Training uses one row per tweet ID (no augmentation), constituting a strong, and standard fine-tuning baseline representative of current practice.

C. Contrastive Model

The contrastive model augments the baseline with two components.

Projection Head: A two-layer MLP mapping h → z ∈ R128 via a 256-dimensional hidden layer with ReLU activation, followed by ℓ2-normalization. This design follows SimCLR [4] and SimCSE [5], where the projection head is used during contrastive training and discarded at inference.

Triplet Construction: For each training step we construct triplets (a,p,n): the anchor a and positive p are two randomly selected augmented variants sharing the same tweet ID, while the hard negative n is a variant from a different tweet ID with a different sentiment label.

TABLE I
OVERALL TEST PERFORMANCE COMPARISON

Model	Macro-F1	Acc.	Wt-F1
BERTweet Baseline (no aug.)	0.6496	0.68	0.68
BERTweet + NT-Xent CL (λ=0.5)	0.6026	0.66	0.64
Δ (CL − Baseline)	−0.047	-0.02	-0.04

Training Objective: The contrastive loss uses NT-Xent with temperature τ=0.07:

$$L_{CL} = -\log \frac{exp(\frac{sim(z_A, z_P)}{\tau})}{\Sigma_k exp(\frac{sim(z_a, z_k)}{\tau})} \quad (1)$$

where sim(·,·) denotes cosine similarity and the denominator sums over all in-batch negatives. The total training objective is:

$$L = (1-\lambda)L_{CE} + \lambda L_{CL}\ ,\ \ \lambda = 0.5 \quad (2)$$

The same AdamW optimizer and linear warmup schedule are used as in the baseline, with batch size B=32 and 3 training epochs.

V. EXPERIMENTAL RESULTS

A. Overall Performance

The general performance of the two models is shown in table I. The fine-tuning of BERTweet on a single row per ID results in the following baseline. a Test Macro-F1 of 0.6496 and accuracy of 0.68. The contrastive model (=0.5, =0.07) achieves a Test Macro-F1 of 0.6026 and accuracy of 0.66, which is a decrease of 4.70 Macro-F1 points.

B. Per-Class Analysis

Table Imitable II breaks down the results by sentiment class. The per-class analysis shows that the underperformance of the contrastive model is. is not constant: it differs significantly among classes, and has significant implications as to the failure mechanism. Table II makes three important observations. First, the Irrelevant class is the one that is most severely degraded (F1 = -0.13), motivated by a fall in recall of 0.43 to 0.26. Second, the Negative group depicts rather confined degradation (F1 =<|human|>Second, the Negative group presents rather restricted degradation (F1 =<|human|). -0.03), probably because negative sentiment tweets are the most lexically consistent category and thus give the most credible. Augmentation pairs. Third, the Positive class recall increases in reality by contrastive training to improve, by 0.75 to 0.79 which is an indication that under contrastive training. the contrastive goal has a positive effect on well-represented and lexically cohesive classes to some extent.

TABLE II
PER-CLASS F1 SCORE COMPARISON

2*Class	Baseline			CL			2*ΔF1
	P	R	F1	P	R	F1	
Positive	0.69	0.75	0.72	0.64	0.70	0.71	-0.01
Negative	0.76	0.81	0.79	0.69	0.86	0.76	-0.03
Neutral	0.61	0.59	0.60	0.66	0.52	0.58	-0.02
Irrelevant	0.56	0.43	0.49	0.60	0.26	0.36	**-0.13**

ISBN: 978-81-689089-0-1

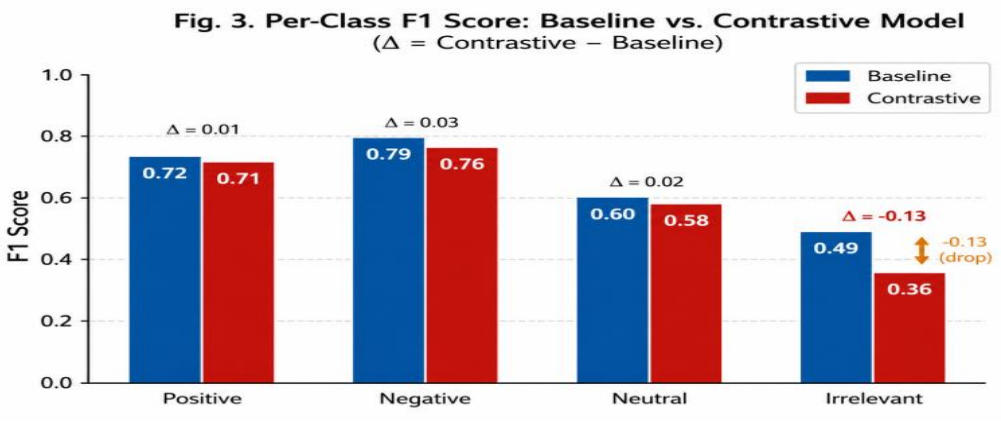

Fig. 3. Per-class F1 scores for the baseline (blue) and contrastive model (red). The Irrelevant class suffers a disproportionate ΔF1=−0.13 drop, while Positive and Negative classes remain largely stable.

C. Training Dynamics

Baseline training indicates a consistent convergence: the loss in terms of cross-entropy decreases by 1.0929 (Epoch 1) to 0.6951 (Epoch 3). validation Macro-F1 increasing between 0.5902 and 0.6248 with epochs. The contrastive model presents a significantly different. trajectory: CE loss decreases from 1.2721 to 0.9781, while the contrastive loss decreases from 1.1457 to 0.2591. The jointly optimization of two conflicting goals slows down convergence, and the optimal validation F1 of 0.5637 (Epoch 3) indicates that the contrastive model was not yet at its performance limit in 3 epochs.

VI. PERFORMANCE ANALYSIS AND INSIGHTS

The crucial lessons that we have learned during our study is the sensitivity of contrastive learning to the quality of data augmentation. The conventional contrastive structures presuppose that positive pairs are semantically equivalent perspectives of the same instance. This assumption is, however, not always met in our dataset in practice. We see that the automatically generated paraphrases sometimes bring semantic distortion, hence, undermining the power of contrastive goal. In particular, there are three common patterns. First, there is sentiment reversal where opinion bearing expressions are changed so as to change the underlying polarity without changing the original label. Second, entity corruption adds noise in the form of malformed tokens or other irrelevant replacements, lowering the linguistic coherence. Third, content fabrication causes significant divergence of the original context, which is

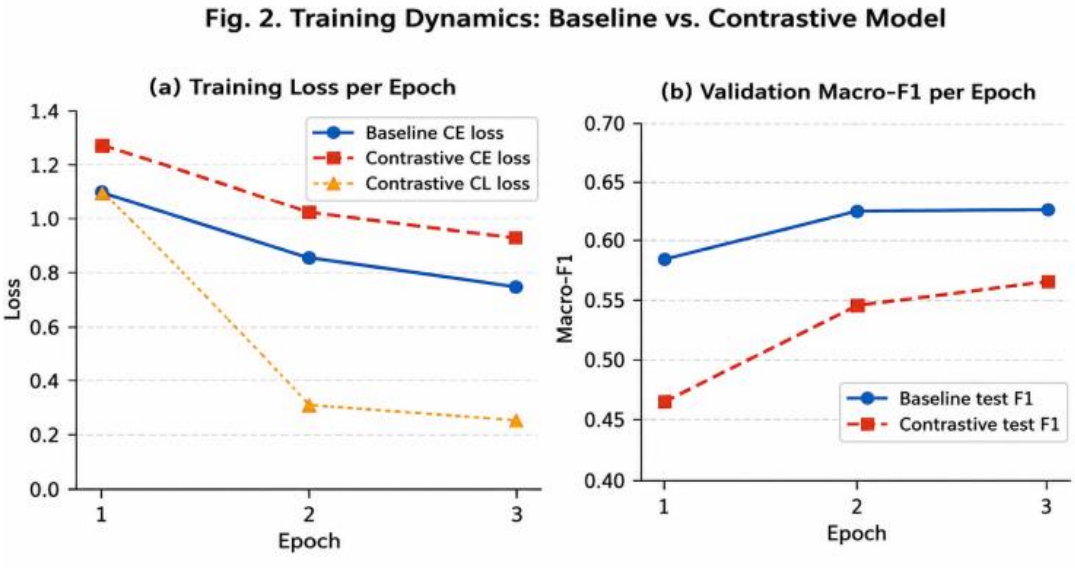

Fig. 4. Training dynamics. Left: loss curves per epoch. The baseline converges cleanly; the contrastive model shows a higher and slower CE trajectory due to dual-objective optimization. Right: validation Macro-F1 per epoch with final test F1 shown as dotted horizontal lines.

sufficient to break semantic alignment between paired samples. Such problems are especially acute in the Irrelevant class, where the lack of a consistent semantic anchor allows it to more readily fall prey to noisy

ISBN: 978-81-689089-0-1

transformations. Consequently, the model tends to push representations that do not have a meaningful connection towards one another, thus harming the learned embedding space.

A. *Batch Size and NT-Xent Optimization*

The high efficacy of the NT-Xent loss is highly related to the diversity and quality of negative samples in a batch. The bigger the batch size, the more contrastive signals are present since each anchor is exposed to a larger number of negative instances. In our experimental design, computational considerations limit the batch size to 32, which limits the number of in-batch negatives one can have at training time. This is to minimize the discriminative power of the contrastive objective especially in a multi-class context where samples in the different classes might overlap on some linguistic characteristics. As a result, the model can have difficulties creating well- separated representations in the embedding space. Previous research has proved that contrastive learning is scaled best with larger batch sizes, so it may be possible to achieve better performance by scaling to GPU-based training with a larger batch size.

B. *Semantic Incoherence of the Irrelevant Class*

The Irrelevant class presents a different issue since it is heterogeneous in nature. In contrast to Positive, Negative, and Neutral classes, which have rather stable semantic patterns, irrelevant samples are characterized by the lack of relationship of sense to the target entity. This creates great intra-class variability and poor semantic cohesion. This variability makes the task of contrastive learning difficult because samples with the same label might not be semantically aligned. Moreover, some irrelevant cases might be superficially similar to the Neutral or sentiment-bearing classes, giving rise to unclear boundaries of representation. This helps to diminish recall and emphasizes the drawbacks of using common contrastive goals to semantically diffuse categories.

Fig. 4. Precision–Recall Score per Class
(Guided = F1 iso-curves)

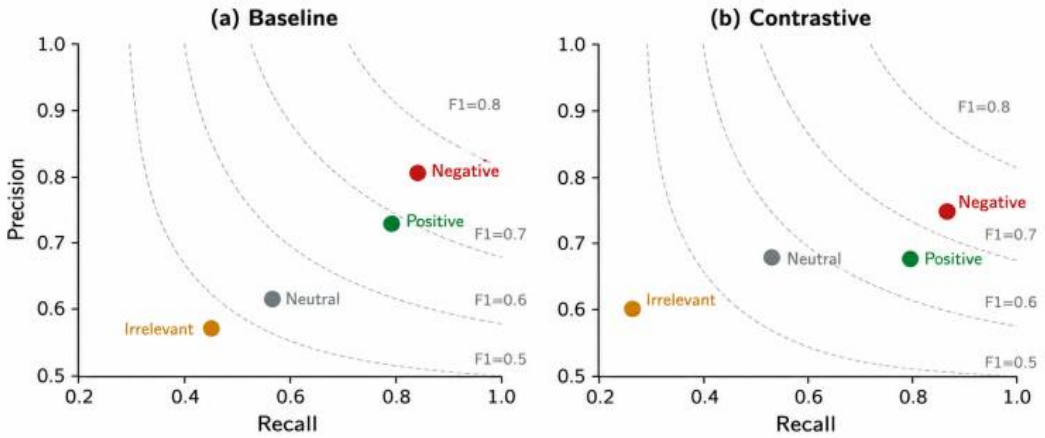

Fig. 5. Precision–recall space per class for baseline (left) and contrastive model (right). Dashed curves are F1 iso-lines. The Irrelevant class (amber) shifts dramatically under contrastive training, indicating a recall collapse while precision is maintained.

C. *Implications for Contrastive NLP Practice*

The results of this research imply a list of real-life applications of contrastive learning that can be introduced to real world NLP settings. To start with, the quality of augmentation must be evaluated before training; it is possible to filter augmented samples by semantic similarity to reduce the effect of noisy pairs. Second, high semantic variability classes might be better modeled in different ways, e.g. by considering relevance detection as an independent task. Third, when data quality is doubtful, contribution of the contrastive objective should be modest such that it augments and not overwhelms the main classification objective.

VII. DISCUSSION AND FUTURE WORK

The present work has taken a consciously open method in the analysis of the results that are not completely in line with the original expectations. We do not emphasize purely on performance gains but rather the factors that determine model behavior. This view offers a more robust basis on which to further develop contrastive learning techniques in real-world noisy datasets.

Our analysis gives several guidelines of future research. The use of the supervised contrastive learning is also one promising direction as it implies the use of label information to create more trustworthy positive pairs and less reliance on the quality of augmentation. The other line of work is the addition of the augmentation filtering on the basis of similarity embedding to enhance the consistency of training data. Also, running relevance detection and sentiment classification can be separated,

ISBN: 978-81-689089-0-1

which can help solve the problems with semantically heterogeneous classes.

Lastly, we have limited training budget and CPU based implementation of our experiments. Future studies ought to investigate more substantial batch sizes and longer training regimes in the context of a GPU to more comprehensively evaluate the potential of contrastive learning in this area

VIII. CONCLUSION

We have presented the first systematic study of augmentation-aware contrastive learning on a large-scale multi-domain Twitter sentiment dataset. Our dual-objective BERTweet model, combining NT-Xent contrastive loss with cross-entropy classification (λ=0.5, τ=0.07), achieves a Test Macro-F1 of 0.6026—4.70 points below the non-augmented BERTweet baseline (0.6496). Through per-class analysis and training dynamics examination, we demonstrate that this degradation is non-uniform, concentrated in the Irrelevant class (ΔF1=−0.13), and attributable to augmentation faithfulness violations, batch size constraints, and class-level semantic incoherence.

These findings contribute three things to the NLP literature: (i) the formal introduction of augmentation faithfulness violation as a concept with direct consequences for contrastive learning; (ii) the first quantitative evidence that semantically incoherent classes in multi-class sentiment datasets are disproportionately harmed by contrastive objectives, with implications for fairness-aware NLP; and (iii) a reproducible benchmark and failure taxonomy for future work applying contrastive methods to Twitter sentiment analysis.

REFERENCES

[1] L. Ren and Y. Peng, ”Research of Fall Detection and Fall Preven- tion Technologies: A Systematic Review,” IEEE Access, vol. 7, pp. 77702–77722, 2019.

[2] R. Socher et al., “Recursive deep models for semantic compositionality over a sentiment treebank,” in *Proc. EMNLP*, 2013, pp. 1631–1642.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. NAACL*, 2019, pp. 4171–4186.

[4] D. Q. Nguyen et al., “BERTweet: A pre-trained language model for English tweets,” in *Proc. EMNLP*, 2020, pp. 9–14.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proc. ICML*, 2020, pp. 1597–1607.

[6] T. Gao, X. Yao, and D. Chen, “SimCSE: Simple contrastive learning of sentence embeddings,” in *Proc. EMNLP*, 2021, pp. 6894–6910.

[7] P. Khosla et al., “Supervised contrastive learning,” in *Proc. NeurIPS*, 2020, pp. 18661–18673.

[8] Y. Liu et al., “RoBERTa: A robustly optimized BERT pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.

[9] J. Wei and K. Zou, “EDA: Easy data augmentation techniques for boosting performance on text classification tasks,” in *Proc. EMNLP*, 2019, pp. 6382–6388.

[10] S. Y. Feng et al., “A survey of data augmentation approaches for NLP,” in *Findings of ACL*, 2021, pp. 968–988.

[11] A. Kumar, A. Choudhary, and E. Cho, “Data augmentation using pretrained transformer models,” in *Proc. ACL Workshop on LGNLP*, 2020.

[12] Y. Ganin et al., “Domain-adversarial training of neural networks,” *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, 2016.

[13] S. Gururangan et al., “Don’t stop pretraining: Adapt language models to domains and tasks,” in *Proc. ACL*, 2020, pp. 8342–8360.

[14] Twitter Entity Sentiment Analysis Dataset, Kaggle, 2020. [Online]. Available: https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis

[15] A. Agnihotri and R. B. Singh, “Sentiment analysis of gaming related tweets using a hybrid deep learning approach,” in *Proc. Int. Conf. Data Analytics & Management*, Springer, 2025

Cite this article as:

Aaditya, Tushar and et. al. "When Augmentation Hurts: Analyzing the Limits of Contrastive Learning on Noisy Augmented Twitter Data for Multi-Domain Sentiment Analysis", Proceedings of 13th international conference on Microelectronics, Circuits and Systems, Micro2026 (Vol-II)

Displayed as online on 04th August 2026.

Link: http://actsoft.org/science/micro2026-pro/391-micro2026.pdf

ISBN: 978-81-689089-0-1