SentiSynth: Teacher-Verified Synthetic Reviews and Soft-Label Distillation Lift Low-Resource Sentiment to 0.84 F1 with 10x Fewer Human Labels

Param Kapur
paramkapur@reed.edu

May 13, 2025

Abstract

Sentiment analysis systems often struggle when applied to new domains or nuanced language phenomena like sarcasm. Labeled data for each new domain is costly to obtain, as exemplified by the ∼67,000 annotated examples in the popular SST-2 benchmark (with roughly 90% for training and 5% each for validation and testing) [?]. We propose **SentiSynth**, a framework that addresses data scarcity by combining large-scale synthetic text generation with knowledge distillation in a closed-loop teacher–student architecture. A powerful teacher model first verifies and filters synthetic examples produced by a generative model, ensuring quality control via a confidence threshold. The teacher's rich probabilistic outputs are then used as soft labels to train a student sentiment classifier, augmented by a weighting scheme to balance synthetic and real data. On the SST-2 dataset, SentiSynth achieves a 5 percentage point F_1 score improvement while using only 10% of the human-labeled data, demonstrating a practical reduction in annotation requirements [?]. We release our code and synthetic data generation pipeline to facilitate further research.

1 Introduction

1.1 Motivation and Problem Definition

Sentiment analysis is integral to many real-world applications such as brand monitoring, customer feedback analysis, and crisis response on social media [?]. Organizations seek to automatically gauge public sentiment about products or events in real time, which can inform marketing strategies or immediate interventions. However, building high-quality sentiment classifiers for specific domains or emerging topics requires substantial labeled data. For example, the Stanford Sentiment Treebank (SST-2) contains roughly 67,000 annotated text snippets, which were expensive and time-consuming to collect [?]. Typically, about 90% of such data is used for model training, with only 5% reserved for validation and 5% for testing. In specialized domains (e.g., financial news, medical forums) or low-resource languages, obtaining tens of thousands of labeled examples is often impractical.

Pre-trained Transformer models have revolutionized sentiment analysis by enabling fine-tuning on relatively small datasets. Despite their success, these models still face significant limitations under domain shift and when encountering linguistic nuances. A sentiment classifier trained on movie reviews, for instance, may falter when applied to social media posts or product reviews that differ in tone and vocabulary [?]. Moreover, Transformers can misinterpret sarcasm or idiomatic expressions—e.g., the phrase "Well, that's just great" might be labeled positive by a literal model despite its negative sarcastic intent [?]. These challenges highlight the need for methods that can adapt sentiment models to new domains and capture subtle cues without relying on prohibitively large labeled datasets.

1.2 Research Gap

Prior research has explored two general strategies to tackle data scarcity in sentiment analysis: synthetic data generation and knowledge distillation. However, these strategies have largely been pursued independently, leaving a gap in how they might complement each other. On one hand, synthetic text generation techniques (such as back-translation or language model prompting) can produce additional training examples to expand a small dataset [?]. On the other hand, knowledge distillation uses a high-capacity teacher model to guide a student model, often assuming a fully labeled training set or using unlabeled data that the teacher labels [?]. Few works have closely integrated these approaches in a unified framework.

Notably, existing methods that do employ synthetic augmentation often lack a feedback loop between the data generator and the classifier. A language model might generate hundreds of examples, but without any verification step, low-quality or label-noise instances can slip into the training data [?]. In contrast, a teacher model (a strong classifier) could be used to vet these generated samples, yet previous studies seldom incorporate such a mechanism during data generation. This absence of teacher feedback means the potential synergy between generation and model guidance is underutilized.

Another underexplored aspect in past work is the use of quality thresholds for accepting synthetic data. While it is intuitive to filter out generated examples on which the teacher model has low confidence (to avoid training on dubious data), the choice of confidence threshold τ is typically ad-hoc or not explicitly discussed in literature [?]. If τ is set too low, the training set may be polluted with ambiguous or incorrect examples; if set too high, many useful synthetic examples are discarded, limiting the data boost. To date, there is little systematic study on how to balance this trade-off in the context of sentiment analysis. This research gap motivates an approach that tightly couples synthetic data generation with teacher verification and carefully chosen quality controls.

1.3 Contributions

Our work proposes a novel framework, **SentiSynth**, which addresses the above gaps. The main contributions are:

- **Teacher-Verified Generation**: We introduce a closed-loop data generation process where a pre-trained generator (GPT-2) creates candidate text samples and a high-accuracy teacher classifier (DeBERTa) filters them by requiring a predicted sentiment confidence $\geq \tau$ [?]. This ensures only high-quality, label-consistent synthetic examples are added to the training data.

- **Soft-Label Caching**: Instead of using only hard labels, we store the teacher's full probability distribution (soft labels) for each accepted synthetic example, with temperature smoothing to retain informative uncertainty patterns [?]. These soft labels are cached and later used to train the student model, transferring richer knowledge from teacher to student.

- **Weighted Training**: We design a weighted training regimen where synthetic examples are down-weighted by a factor β in the loss function [?]. This balances the influence of synthetic data relative to the original human-labeled data, mitigating potential noise from machine-generated texts while still benefiting from the augmented examples.

- **Empirical Gains**: Through extensive experiments on SST-2, we demonstrate that SentiSynth achieves a substantial improvement of approximately **+5 percentage points** in F_1 score over strong baselines, despite using only **10% of the human-labeled data** for training. This result highlights the framework's practicality in reducing annotation costs.
- **Open-Source Release**: We publicly release our full codebase, synthetic data generation scripts, and the augmented datasets produced by SentiSynth. We hope this will enable reproducibility and inspire future research into combined generation–distillation methods for NLP.

1.4 Paper Roadmap

The remainder of this paper is organized as follows. In Section 2, we review the background and related work, including the evolution of sentiment analysis approaches and existing techniques in data augmentation and knowledge distillation. Section 3 details the SentiSynth methodology, describing the teacher-verified generation process, the caching of soft labels, and the weighted training strategy. Section 4 then presents experimental results, comparing SentiSynth to baselines and analyzing the impact of key hyperparameters (e.g., τ and β). In Section 5, we discuss ethical considerations of synthetic data generation and knowledge distillation, such as potential bias introduction and data fidelity. Finally, Section 6 concludes the paper and outlines future research directions.

2 Background and Related Work

2.1 Sentiment Analysis Evolution

Rule-based → classic ML → deep learning (CNN/RNN) → Transformers

2.2 Synthetic Text Generation

Template methods, back-translation, GANs (SentiGAN, CatGAN), GPT fine-tuning

2.3 Knowledge Distillation

Soft-labels, temperature scaling, student–teacher paradigm

2.4 Quality Control in Synthetic Data

Automatic filtering, human spot-checks, trustworthiness of LLM output

2.5 Gap Revisited

Why prior methods don't jointly tackle QC + distillation at scale

3 Methodology: The SentiSynth Pipeline

3.1 Datasets

SST-2 seed set (size, class balance, preprocessing) Synthetic target sizes (20 k, 25 k) Train/val/test splits