Adaptive Synthetic Data Augmentation for Low-Resource Sentiment Analysis

Huzaifa Rabbani, 21i-2496. Faizan Ali, 21i-0422

30th March, 2025

Project Description

Sentiment analysis, the task of automatically determining the emotional tone behind text, is crucial for various applications, including social media monitoring, customer feedback analysis, and opinion mining [7]. However, the performance of sentiment analysis models heavily relies on the availability of large, labeled datasets. Low-resource languages and specialized domains often suffer from data scarcity, hindering the development of effective sentiment analysis systems [1]. This project addresses this challenge by leveraging the power of generative AI to create synthetic data that augments limited seed datasets and improves the performance of sentiment analysis models in low-resource scenarios.

This project proposes an adaptive synthetic data augmentation framework that utilizes pre-trained language models and reinforcement learning to iteratively generate high-quality, diverse, and domain-relevant synthetic text examples for sentiment analysis. The system will begin with a small seed dataset and employ open-source libraries like Hugging Face Transformers [10] and TRL (Transformer Reinforcement Learning) [8] to implement a three-stage pipeline:

Synthetic Data Generation, utilizing techniques like back-translation, synonym replacement, and conditional generation to create sentiment-preserving variations of seed data. [7]

Automated Feedback and Reinforcement Loop, where a lightweight sentiment classifier, trained on the seed data, acts as a reward model to evaluate the quality of generated examples. [1]

Reinforcement learning algorithms, such as Proximal Policy Optimization (PPO), will then be used to fine-tune the generative model based on this feedback, encouraging the generation of higher-quality and more relevant synthetic data. [10] Integration and Evaluation, which involves combining the synthetic data with the original seed data to train and evaluate a robust sentiment analysis classifier (e.g., BERT-based [2]). The effectiveness of the augmented dataset will be rigorously assessed using standard sentiment analysis metrics, demonstrating the potential performance gains achieved through this adaptive synthetic data augmentation approach.

This project embodies the principles of generative AI by using AI models to create valuable new data, ultimately enhancing the capabilities of sentiment analysis in data-constrained settings.

Main Functions

This project is structured around three core modules, each contributing to the overall goal of adaptive synthetic data augmentation:

1. Synthetic Data Generation Pipeline:

Functionality:

This module is responsible for generating synthetic text data from a limited seed dataset. It leverages pre-trained language models (e.g., GPT-2 [8], T5 [9] available through Hugging Face Transformers) as the foundation for text generation.

Techniques Implemented:

Classic Augmentation Methods: Incorporate traditional data augmentation techniques like synonym replacement (using resources like WordNet [6]), random insertion, deletion, and swapping to introduce minor variations while preserving sentiment.

Back-Translation: Utilize machine translation models to translate seed text to another language and back to the original language, generating paraphrased versions while ideally maintaining sentiment.

Conditional Generation: Explore conditional text generation capabilities of models like T5 to generate synthetic examples specifically targeting positive or negative sentiment, guided by sentiment labels. This can be achieved by prompting the model with sentiment-related keywords or fine-tuning on sentiment-labeled data.

Output: A set of synthetic text examples, initially generated based on the seed dataset and augmentation techniques.

2. Automated Feedback and Reinforcement Loop:

Functionality:

This module implements an automated feedback mechanism to evaluate the quality and relevance of the generated synthetic data and iteratively refine the data generation process. [5] It employs a reinforcement learning loop to guide the generative model towards producing better synthetic examples.

Components:

Reward Model (Sentiment Classifier): A lightweight sentiment classifier (e.g., a simple logistic regression model or a fastText classifier [6]) will be trained on the limited seed dataset. This classifier acts as the reward model, scoring the generated synthetic examples based on their predicted sentiment label and potentially confidence scores.

Reinforcement Learning Agent: Utilizing open-source RL frameworks like TRL, a reinforcement learning agent (e.g., PPO or DPO agent) will be implemented. This agent interacts with the synthetic data generation pipeline.

Feedback Loop: The RL agent receives feedback (rewards) from the sentiment classifier based on the generated synthetic data. [3] This feedback is used to update the parameters of the generative model, guiding it to generate synthetic data that is more likely to be classified correctly (and with high confidence) by the reward model, and thus, hopefully, more aligned with the desired sentiment and domain.

Output: An iteratively refined synthetic data generation model that is optimized to produce higher quality and more sentiment-aligned synthetic examples based on the feedback loop. [4]

3. Integration and Evaluation Framework:

Functionality:

This module focuses on integrating the synthetic data generated by the adaptive pipeline into the sentiment analysis training process and rigorously evaluating the impact of synthetic data augmentation on the performance of a target sentiment analysis classifier.

Components:

Sentiment Analysis Classifier Training: A robust sentiment analysis classifier (e.g., a fine-tuned BERT-based model) will be trained using different datasets:

(a) the original seed dataset only, (b) the seed dataset augmented with synthetic data generated without feedback, and (c) the seed dataset augmented with synthetically data generated using the adaptive feedback loop.

Evaluation Metrics: The performance of the sentiment analysis classifiers will be evaluated using standard sentiment analysis metrics such as accuracy, precision, recall, F1-score, and potentially AUC-ROC.

Output: Performance evaluation results demonstrating the effectiveness of the adaptive synthetic data augmentation approach, a functional Jupyter Notebook demonstration, and a comprehensive project report summarizing the methodology, results, and findings.

References

[1] Naeem Ahmed, Rashid Amin, Hamza Aldabbas, Muhammad Saeed, Muhammad Bilal, and Houbing Song. A novel approach for sentiment analysis of a low resource language using deep learning models. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, September 2024. Just Accepted.

[2] Thomas Body, Xiaohui Tao, Yuefeng Li, Lin Li, and Ning Zhong. Using back-and-forth translation to create artificial augmented textual data for sentiment analysis models. *Expert Systems with Applications*, 178:115033, 2021.

[3] Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. An empirical survey of data augmentation for limited data learning in nlp. *Transactions of the Association for Computational Linguistics*, 11:191–211, 03 2023.

[4] Marzieh Fadaee, Arianna Bisazza, and Christof Monz. Data augmentation for low-resource neural machine translation. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[5] Fei Gao, Jinhua Zhu, Lijun Wu, Yingce Xia, Tao Qin, Xueqi Cheng, Wengang Zhou, and Tie-Yan Liu. Soft contextual data augmentation for neural machine translation. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the*

Association for Computational Linguistics, pages 5539–5544, Florence, Italy, July 2019. Association for Computational Linguistics.

[6] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomás Mikolov. Bag of tricks for efficient text classification. *CoRR*, abs/1607.01759, 2016.

[7] Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 2022.

[8] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[9] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), January 2020.

[10] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.