Enhancing Low-Resource Text Classification through Selective Text Augmentation (STA) and Large Language Models

Amir Ghavam
AmirGhavam72@gmail.com

1 Problem statement

The goal of this project is to enhance text classification in low-resource environments by utilizing Selective Text Augmentation (STA) and Large Language Models (LLMs). The motivation stems from the challenge of data scarcity in text classification, particularly in settings where conventional data enrichment methods are inadequate. By synergizing STA with LLMs, the project aims to improve the accuracy and generalizability of text classification models in these challenging contexts. This approach seeks to demonstrate the effective use of word role categorization in STA and the potential of LLMs in generating realistic, contextually relevant text augmentations.

2 What you proposed vs. what you accomplished

In the project proposal, we outlined a series of objectives to enhance text classification in low-resource environments. Here's a summary of what was proposed versus what was actually accomplished:

- **Read the STA paper and other related literature:** Successfully completed. We thoroughly reviewed the Selective Text Augmentation (STA) paper and other relevant materials to build a strong foundation for the project.
- **Load and preprocess datasets:** Accomplished. We loaded and preprocessed four datasets - IMDb, SST2, Yahoo, and 20 Newsgroups (20NG) - for sentiment and category classification.
- **Create smaller datasets for experimentation:** Successfully created 16 smaller datasets from the original ones, each containing 50, 100, 500, and 1000 samples, named accordingly (e.g., IMDB_50, IMDB_100).
- **Implement STA on a sample dataset:** Completed. We successfully ran STA on one sample, laying the groundwork for broader implementation.
- **Apply STA to all datasets and create augmented versions:** Accomplished. STA was applied to all 16 datasets, creating augmented versions for each.
- **Classify each dataset with DistillBERT:** Completed. Each original and augmented dataset was classified using DistillBERT, demonstrating the impact of STA on classification accuracy.
- **Prepare Gemini Pro and LLAMA2 for data augmentation:** Partially completed. We made initial preparations for using Gemini Pro and LLAMA2 for data augmentation.
- **Augment data using Gemini Pro and LLAMA2, and classify the augmented files:** Not accomplished. Due to time constraints, we were unable to use Gemini Pro and LLAMA2 for data augmentation and subsequent classification
- **Calculate and plot accuracy for each approach across different dataset sizes:** This was partially successful. We managed to calculate the accuracies for each dataset and charted these metrics to demonstrate performance variations across different sizes. This analysis was specifically focused on the outcomes related to the implementation of Selective Text Augmentation (STA).

In summary, while we achieved most of the objectives, time constraints prevented the implementation of Gemini Pro and LLAMA2 for data augmentation. However, the successful application of STA and analysis using DistillBERT provided valuable insights into text classification in low-resource settings.

3 Related work

In the realm of text classification, the landscape is defined by a variety of innovative approaches and methodologies. The foundational survey by (Bayer et al., 2021), titled ”A Survey on Data Augmentation for Text Classification,” sets the stage by exploring a range of strategies, from simple text editing to complex language model-based augmentations. Complementing this, (Wei and Zou, 2019)’s study, ”EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks,” demonstrates the effectiveness of basic techniques like synonym replacement, while, (Guo et al., 2022), in ”Selective Text Augmentation with Word Roles for Low-Resource Text Classification,” introduce the nuanced STA focusing on word roles.

Advancing into the territory of large language models, (Kaddour and Liu, 2023)’s work, ”Text Data Augmentation in Low-Resource Settings via Fine-Tuning of Large Language Models,” exemplifies the potential of fine-tuning these models for text augmentation in low-resource settings. Their approach, alongside (Yoo et al., 2021)’s exploration of GPT-3 in ”GPT3Mix: Leveraging Large-scale Language Models for Text Augmentation,” underscores the power and challenges of leveraging advanced language models. (Meng et al., 2022), further delve into this by investigating the generation of training data with unidirectional PLMs for zero-shot language understanding in their paper ”Generating Training Data with Language Models: Towards Zero-Shot Language Understanding.

Complementing these studies ”Text Classification via Large Language Models”(Sun and Guo, 2023) scrutinizes the limitations of GPT-3 in text classification, while ”Augmenting Low-Resource Text Classification with Graph-Grounded Pre-training and Prompting”(Wen and Fang, 2023) and ”ChatGraph: Interpretable Text Classification by Converting ChatGPT Knowledge to Graphs”(Shi et al., 2023) propose novel methodologies for enhancing classification in low-resource environments. ”Synthetic Data Generation with Large Language Models for Text Classification”(Li et al., 2023) evaluates the efficacy of synthetic data in training models, and ”WC-SBERT: Zero-Shot Text Classification via SBERT with Self-Training for Wikipedia Categories”(Chi et al., 2023) introduces a zero-shot classification method using SBERT. Lastly, ”Text Classification in the Wild: A Large-Scale Long-Tailed Name Normalization Dataset”(of the paper, 2023) provides insights into handling long-tailed data distributions in text classification.

Together, these studies form a rich tapestry of approaches, each contributing uniquely to the evolving field of text classification and augmentation, reflecting both the diversity and the complexity of challenges in this domain.

4 Your dataset

The project utilizes a selection of diverse datasets, each with unique characteristics and challenges:

- **IMDB:** Sentiment classification of movie reviews. Challenge: Large volume and varied expressions in reviews. Statistics: Thousands of reviews. Example: Input: ”Amazing cinematography.” Output: Positive.

- **Yahoo Answers:** Topic classification from user questions and answers. Challenge: Diverse topics and informal language. Statistics: Multiple topics. Example: Input: ”How to fix a leaky faucet?” Output: Home Improvement.

- **20NG (Twenty Newsgroups):** Document classification across 20 newsgroups. Challenge: Distinct topics, varied writing styles. Statistics: 20,000 documents. Example: Input: ”NASA’s latest mission success.” Output: Science.

- **SST2:** Sentiment classification of sentences from movie reviews. Challenge: Fine-grained sentiment analysis. Statistics: 11,855 sentences. Example: Input: ”A boring plot.” Output: Negative.

Each dataset presents unique challenges in text classification, from sentiment analysis to topic categorization, reflecting the complexity and variety of natural language.

4.1 Data preprocessing

The data preprocessing steps employed in this project are tailored to optimize text for analysis and model training. Initially, all text is converted to lowercase to maintain consistency. URLs are removed to focus on meaningful text content, and any HTML tags are stripped using BeautifulSoup to extract clean text. The text is further processed to expand contractions for clarity and uniformity. Numbers within the text are converted to words, enhancing the model's ability to interpret numerical data. Finally, all punctuation is removed, leaving only textual data. These preprocessing steps culminate in tokenization, which segments the text into tokens, preparing it for input into the model. This methodical preprocessing pipeline ensures that the data is in the most suitable format for the subsequent stages of model training and evaluation.

5 Baselines

In our project, we directly compared the efficacy of STA with the performance of classifications using raw, non-augmented datasets. Given that the EDA function is already implemented within our code, we have the capability to augment datasets using EDA and easily evaluate its impact in future studies.

Our training/validation/test split was carefully balanced to prevent overfitting, ensuring a rigorous assessment of our STA approach compared to basic text classifications.

For our project, the configuration of STA was critical. We set specific hyperparameters: methods including deletion, replacement, insertion, swapping, and selection; a probability of augmentation (p=0.1); and a keyword extraction bar (”Q2”) with one augmentation per method (n_aug=1).

In training the DistilBert model, we adjusted the training epochs based on the classification type of each dataset. For binary classification tasks like IMDb and SST2, we trained the model for 3 epochs. For multi-class classification tasks such as Yahoo and 20 Newsgroups, the training was extended to 10 epochs. Additionally, we used a batch size of 16 per device, a warmup of 500 steps, a weight decay of 0.01, and ensured the best model was loaded at the end of training. These configurations were key in optimizing the model's performance for different types of datasets.

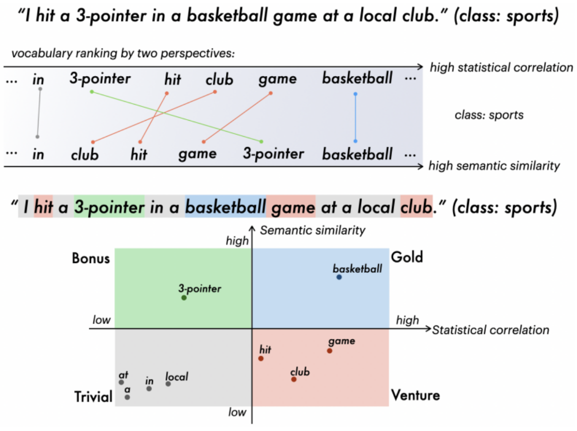
"I hit a 3-pointer in a basketball game at a local club." (class: sports)
vocabulary ranking by two perspectives:
... in 3-pointer hit club game basketball ...
high statistical correlation
class: sports
... in club hit game 3-pointer basketball ...
high semantic similarity
" I hit a 3-pointer in a basketball game at a local club." (class: sports)
Semantic similarity
high
low
Bonus
Gold
Trivial
Venture
3-pointer
basketball
low
high
Statistical correlation
hit
game
club
at
a
in
local

Figure 1: Illustration of the four word roles

6 Your Approach

The foundational approach for this project is Selective Text Augmentation (STA), enhancing text classification in low-resource settings by categorizing words based on their roles (Gold, Venture, Bonus, Trivial) and selectively applying text-editing operations, as illustrated in Figure 1.

- **Gold words:** High statistical correlation and semantic similarity with the category, crucial for class indication.
- **Venture words:** Frequent co-occurrence, low semantic similarity, providing extra information but potentially misleading.
- **Bonus words:** Low statistical correlation, high semantic similarity, rare but beneficial for generalization.
- **Trivial words:** Low in both correlations and semantic similarity, less important for prediction.

Word roles are recognized using metrics like weighted log-likelihood ratio (WLLR) for statistical correlation and cosine similarity for semantic similarity between word vectors. This approach maintains core semantics and creates clean, diverse samples, demonstrating superior performance to non-selective methods. Challenges include complex word role identification and reliance on word vector quality. Figure 2 provides an overview of the STA process.

Integration with Large Language Models (LLMs) leverages the precision of STA in managing word roles, complementing LLMs' generative

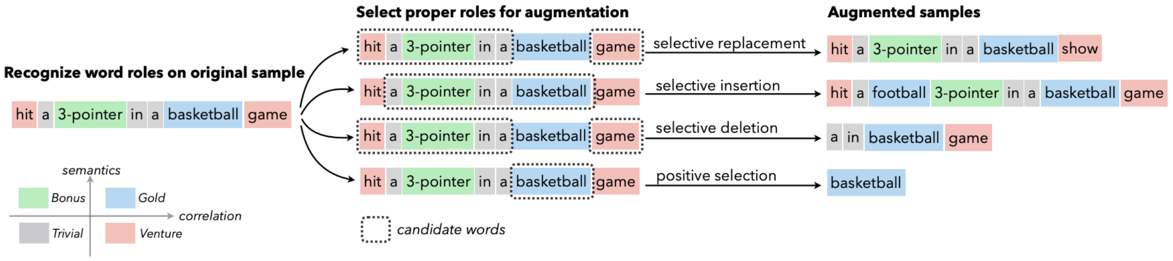
Recognize word roles on original sample
hit a 3-pointer in a basketball game
semantics
Bonus
Gold
correlation
Trivial
Venture
Select proper roles for augmentation
hit a 3-pointer in a basketball game
selective replacement
selective insertion
selective deletion
positive selection
candidate words
Augmented samples
hit a 3-pointer in a basketball show
hit a football 3-pointer in a basketball game
a in basketball game
basketball

Figure 2: An overview of the process of selective text augmentation (STA) method. Different text-editing operations are selectively applied to words with specific roles.

capabilities. LLMs can serve as standalone classifiers or specific task performers within STA. This synergy aims to enhance realistic text sample generation for training classifiers in data-limited scenarios.

Experiments were conducted using a combination of Kaggle, Google Colab, and my personal computer, depending on the specific requirements and limitations of each task. While Google Colab was initially used, its low RAM posed challenges for running larger models like LLama2, prompting a switch to Kaggle for certain tasks. Some experiments were also carried out on my personal computer to overcome these limitations.

Upon concluding our experiments, we observed distinct patterns in the efficacy of STA across datasets of varying sizes, as detailed in Table 1. STA demonstrated considerable effectiveness in smaller datasets, significantly enhancing model accuracy. However, this benefit appears to diminish with larger datasets. These results are visually represented in Figure 3, where the impact of STA on classification accuracy is clearly depicted across all datasets. This visual aid reinforces our findings, highlighting the nuanced performance of STA in different data environments.

For the complete code and implementation details of this project, along with the results, cleaned datasets, and augmented datasets, please refer to the GitHub repository.[1]

7 Error analysis

A critical factor impacting the performance of our models across all datasets was the inherent quality of the data. Many texts in the datasets were complex and occasionally ambiguous, posing interpretation challenges even for human readers. This complexity is likely to have contributed to the models' misclassification errors. Other influencing factors include:

- **IMDB (Sentiment Classification):** For the IMDB dataset, the baseline and STA-enhanced models may have struggled with reviews containing mixed sentiments or subtle irony. For example, the models might have misclassified reviews where positive sentiments were expressed using negative words sarcastically.

- **SST2 (Sentiment Classification):** In the SST2 dataset, errors could have occurred in cases of short texts with ambiguous sentiments or context-dependent meanings. The model might have failed to capture the nuances of sentiment in these succinct expressions.

- **Yahoo (Custom Category Classification):** The Yahoo dataset's challenge was twofold. Firstly, its 10 categories and diverse topics likely caused difficulties in accurately classifying nuanced or overlapping categories, especially with the limited training sizes (50 to 1000 examples). Secondly, the high number of class labels relative to the training size may have led to underrepresentation of some categories, increasing misclassification rates.

- **20 Newsgroups (20NG, Custom Category Classification):** The 20NG dataset faced similar issues. With 20 different class labels and a small training set, the model likely struggled to differentiate between closely related topics, particularly in discussions with overlapping themes or technical language. The sheer number of classes compared to the limited training examples would have sig-

[1] https://github.com/amirghavam93/FMNLP_project_STA_LLM.git

Table 1: Model Accuracies for Non-Augmented vs. Augmented Datasets Across Different Sizes

Size	SST2		IMDb		Yahoo		Newsgroups	
	Non-Aug	Aug	Non-Aug	Aug	Non-Aug	Aug	Non-Aug	Aug
50	50.46%	50.92%	50.00%	50.60%	9.47%	27.78%	5.80%	34.80%
100	49.08%	69.84%	50.00%	67.34%	8.85%	44.44%	12.20%	50.40%
500	74.77%	82.68%	81.46%	86.18%	53.91%	52.88%	65.60%	67.40%
1000	81.88%	81.65%	88.08%	87.56%	58.23%	53.91%	69.60%	72.00%

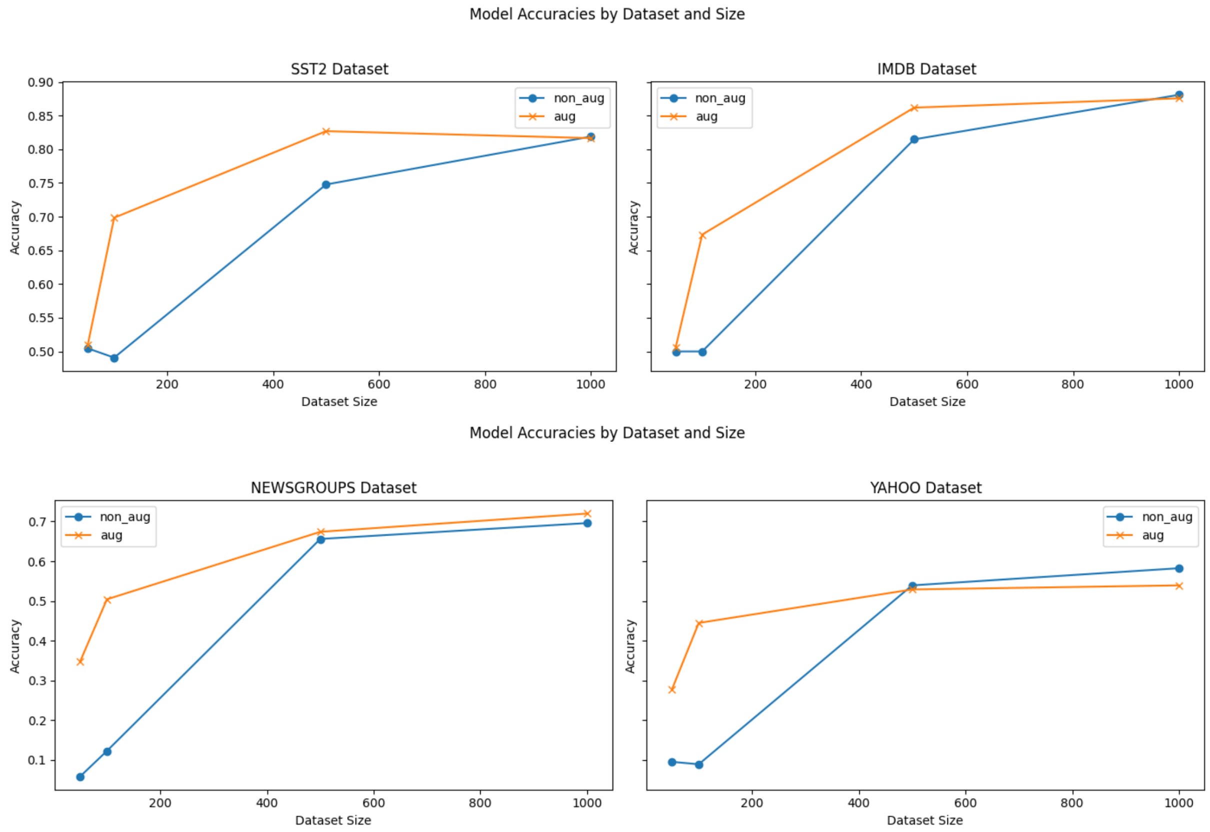

Figure 3: Comparative Model Accuracies for SST2, IMDb, Newsgroups, and Yahoo Datasets. This figure illustrates the impact of Selective Text Augmentation (STA) on model performance across varying dataset sizes, demonstrating the effectiveness of augmentation strategies in different classification contexts.

nificantly impacted the model's learning and generalization capabilities.

In each case, the size of the dataset could have influenced the model's performance. Smaller datasets might have led to overfitting, particularly in complex or nuanced classification tasks. Meanwhile, larger datasets likely provided more comprehensive training but also introduced the challenge of managing diverse and sometimes ambiguous examples. This analysis, although hypothetical, is in line with common challenges faced in text classification tasks using such datasets.

8 Contributions of Group Members

- **Amir Ghavam:** Proudly handled (almost!) the entire project solo, with a little magic from my trusty sidekick, the ever-so-brilliant AI, **ChatGPT 4!**

9 Conclusion

Throughout this project, I've gained substantial knowledge about data augmentation, especially STA, which will be invaluable for my master's thesis on group recommender systems where data scarcity is a common issue.

Challenges Encountered: The project presented several unanticipated challenges. A notable example was the experience with Gemini Pro. Although it initially worked well, it encountered difficulties with large-scale prompts. These issues ranged from blocking texts due to content restrictions to encountering unexplained errors for which solutions were not readily available in existing documentation or online resources. Additionally, the challenges with implementing the STA author's code, which was not only poorly structured and incomplete but also contained inconsistencies in language, added to the complexity of the project.

Surprises in Results: Two aspects were particularly surprising. Firstly, STA proved remarkably effective for very small datasets, exceeding my expectations. Secondly, the accuracy paradoxically decreased with larger datasets and excessive augmentation, contrary to my initial assumption of consistent or improved performance.

Future Directions: I aim to complete the unfinished aspects of the project, particularly involving Large Language Models. An intriguing question has emerged: Would applying STA to the test set, by augmenting and then appending these rows to their original counterparts, impact accuracy? This inquiry opens a path to a broader exploration of the efficacy and limitations of STA, as well as a comparison with other augmentation methods, enhancing our understanding of their impact in various NLP applications.

Throughout this project, I've gained substantial knowledge about data augmentation, especially Selective Text Augmentation (STA), which will be invaluable for my master's thesis on group recommender systems where data scarcity is a common issue.

References

Bayer, M., Kaufhold, M.-A., and Reuter, C. (2021). A survey on data augmentation for text classification. *ACM Computing Surveys*, 55:1 – 39.

Chi, T.-Y., Tang, Y.-M., Lu, C.-W., Zhang, Q.-X., and Jang, J.-S. R. (2023). Wc-sbert: Zero-shot text classification via sbert with self-training for wikipedia categories. *arXiv preprint arXiv:2307.15293*.

Guo, B., Han, S., and Huang, H. (2022). Selective text augmentation with word roles for low-resource text classification.

Kaddour, J. and Liu, Q. (2023). Text data augmentation in low-resource settings via fine-tuning of large language models.

Li, Z. et al. (2023). Synthetic data generation with large language models for text classification: Potential and limitations. *arXiv preprint arXiv:2310.07849*.

Meng, Y., Huang, J., Zhang, Y., and Han, J. (2022). Generating training data with language models: Towards zero-shot language understanding. In *Advances in Neural Information Processing Systems*. NeurIPS.

of the paper, A. (2023). Text classification in the wild: A large-scale long-tailed name normalization dataset. *arXiv preprint arXiv:2302.09509*.

Shi, Y., Ma, H., Zhong, W., Mai, G., Li, X., and Liu, T. (2023). Chatgraph: Interpretable text classification by converting chatgpt knowledge to graphs. *arXiv preprint arXiv:2305.03513*.

Sun, X. and Guo, S. (2023). Text classification via large language models. *arXiv preprint arXiv:2305.08377*.

Wei, J. and Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Wen, Z. and Fang, Y. (2023). Augmenting low-resource text classification with graph-grounded pre-training and prompting. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

Yoo, K. M., Park, D., Kang, J., Lee, S.-W., and Park, W. (2021). GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic. Association for Computational Linguistics.