REGULAR ARTICLE

Data In Disguise: Getting More From Less with AI

Paula Joy Martinez*, Christian Romeo Plan, Joshua Simon and Nico Rafael Ting

*Correspondence:
PMartinez.MSDS2024@aim.edu
Aboitiz School of Innovation, Technology, and Entrepreneurship, Paseo de Roxas, Legazpi Village, 1229 Makati, Philippines
Full list of author information is available at the end of the article

Abstract

Data scarcity remains a major barrier to the widespread adoption of artificial intelligence (AI) across industries. Recent studies project that the stock of high-quality language data will be depleted by 2026, while lower-quality language data and image data will remain available until the 2030s-2060s. This potential data shortage threatens to impede the scalability of machine learning models unless data efficiency improves or new data sources emerge. This paper presents a proof-of-concept solution by leveraging large language models (LLMs) and interpretability techniques to generate high-quality synthetic textual data to augment limited datasets. The study explores using LLMs like ChatGPT and Claude, along with the Shapley Additive Explanations (SHAP) algorithm, to create synthetic data for enhancing emotion detection models. A logistic regression model trained on a combination of real and LLM-generated synthetic data achieved comparable accuracies up to 78.5% and 78.8% respectively, demonstrating the viability of using LLMs to effectively mimic real data distributions. Key factors include providing LLMs with data exemplars and using SHAP to identify emotion-associated keywords to guide the generation process. The study highlights the potential of LLMs to overcome data constraints and unlock AI capabilities through informed prompt engineering and interpretability methods.

Keywords: data augmentation; large language models; shapley additive explanations

1 Introduction

The idea of LLMs first emerged in the 1960s with the creation of Eliza, the world's first chatbot [2]. However, it was only after the debut of ChatGPT in 2020 that the broader public began to recognize their significance [2]. Their ease of accessibility, coupled with their remarkable ability to generate rapid responses and engage in conversational interactions, catalyzed a significant shift in perception about their capabilities and potential applications.

As organizations begin to integrate these tools into their day-to-day operations, the importance of data is becoming more and more pronounced as it takes a central role in the development and maintenance of artificial intelligence (AI) systems. However, a recent study has revealed a disconcerting trend, particularly for language data [3]: the stock of high-language data is projected to be depleted in the near future, likely before the year 2026. In contrast, the reserves of low-quality language data and images are anticipated to sustain for a more extended period, lasting between 2030 and 2050 for low-quality language data, and 2030 and 2060 for images. This disparity implies that the scalability of machine learning models might be

impeded if substantial improvements in data efficiency are not achieved or if new sources of data remain undiscovered.

In light of this impeding data scarcity, LLMs present a promising solution to augment and expand existing data resources. Their ability to generate human-like text opens up opportunities for creating synthetic textual data that can be used to train and enhance machine learning models, thereby providing opportunities for organizations with limited data collection capabilities to overcome their data constraints and leverage the power of AI effectively.

This study adds to the exploration of the seemingly limitless capabilities of LLMs. It aims to determine how organizations, especially those with limited resources, can utilize large language models and the Shapley Additive Explanations (SHAP) interpretability algorithm to augment their limited textual data resources. The paper culminates at the creation of a machine learning pipeline that effectively utilizes large language models (LLMs) and Shapley Additive Explanations (SHAP) to increase the small existing dataset of a company. The materialization of this overall goal hinges on the following objectives:

1 To demonstrate and compare the performance of two prominent LLMs (ChatGPT 4 and Claude.AI) with respect to generating synthetic textual data following certain specifications.
2 To explore the usability of SHAP to enhance and guide the outcomes of the selected LLMs.
3 To assess the performance of a chosen model trained on augmented data versus real data to validate the augmented data's usability.

A key point the paper emphasizes is that LLMs still rely on pre-existing data to generate new, synthetic text. This means that the synthetic data produced is a derivative of an organization's current data to reflect its specific attributes and complexities. Consequently, the paper reaffirms the ongoing need for robust data collection practices to ensure the acquisition of high-quality data.

This paper is organized as follows: Section 2 discusses the data collection, preprocessing, and machine learning pipeline implemented. Section 3 presents the research findings and provides an analysis of the results. Section 4 offers concluding thoughts and insights derived from the study. Lastly, Section 6 proposes future directions of the work presented.

2 Data and Methods

2.1 Data Source

The dataset used in this paper comes from an evaluation framework called TweetEval, a repository consisting of seven heterogeneous Twitter-specific classification tasks, one of which is on emotion recognition [1]. A sample from this dataset is presented in Table 1.

While the dataset contains four emotions——anger, joy, sadness, optimism——the study only focused on two polar opposite emotional states, anger and joy, to simplify the dataset into a binary classification task. This allows for more generalizable conclusions while still capturing the broad characteristics of emotional expression on Twitter.

Table 1 A subset of the Emotion Recognition Dataset from TweetEval.

Text	Label	Emotion
My roommate: it's okay that we can't spell because we have autocorrect. #terrible #firstworldprobs	0	anger
No but that's so cute. Atsu was probably shy about photos before but cherry helped her out uwu	1	joy
Rooneys fucking untouchable isn't he? Been fucking dreadful again, depay has looked decent(ish)tonight	0	anger
Tiller and breezy should do a collab album. Rapping and singing prolly be fire	1	joy
@user I would never strategically vote for someone I don't agree with...	0	anger

2.2 Summary of Methodology Pipeline

The overall methodology follows a narrative close to organizations with limited data resources. As such, the steps are as follows:

1 **Data Preprocessing and Partitioning.** The original dataset contained 3,281 rows. This subset of 1,000 rows was then extracted and used as the baseline for training and validation of the model. This subset was also the basis for extracting the influential keywords based on interpretability algorithm (SHAP) used. These keywords were then utilized to guide LLMs in generating new, synthetic textual data. An additional 40 rows were used as exemplars to provide guidance to the LLMs. These 1,040 rows were removed from the original dataset, leaving 2,241 rows. This remaining subset was used later for comparing the performance of models trained on synthetic data versus real data.

2 **Model Training**. Using the initial 1,000 rows, the authors identified the worst-performing machine learning model for the dataset. The rationale was that if a sub-optimal model could be enhanced through the proposed methodology, the positive results would likely be observable in better-performing models as well. Of the 1,000 rows, 80% was used for training, and 20% was used as a test set.

3 **Obtain Significant Words through SHAP**. In addition to providing exemplars to the LLMs, the Shapley Additive Explanations (SHAP) algorithm was used to better inform the generation of synthetic data. Specifically, the top words defining each emotion were fed into the LLMs.

4 **Generate the Synthetic Data.** This study utilized Anthropic's Claude.AI and OpenAI's ChatGPT to generate the synthetic textual data. Both LLMs were given the same set of exemplars (20 tweets per emotion) and were asked to generate 500 happy and 500 angry statements, totaling 1,000 AI-generated data points.

5 **Augment the Original Training Data.** In this experiment, the performance of the chosen machine learning model was compared when the training set was supplemented with AI-generated data versus real data. The former was the output of the previous step, while the latter was sourced from the remaining pool of real data (2,241 rows) from Step 1. Both types of supplementary data (synthetic and real) were added to the existing training set (800

rows), resulting in augmented training sets of 1,800 rows each. The models trained on these augmented sets were then evaluated to assess the usability of the synthetic data against real data.

2.2.1 Data Preprocessing and Partitioning

The original dataset contained 3,281 rows. It had no missing values and had sufficient balance between the two emotional states—anger and joy—across the entire dataset. While the two classes are not equal in distribution (65% anger, 35% joy), resampling was not applied as the models used could generalize well on their own.

This subset of 1,000 rows was then extracted and used as the baseline for training and validation of the model. This subset was also the basis for extracting the influential keywords based on interpretability algorithm (SHAP) used. These keywords were then utilized to guide LLMs in generating new, synthetic textual data. The subset was further divided into 80% for training (800 rows) and 20% for testing (200 rows). An additional 40 rows were set aside from the original dataset and used as exemplars to provide guidance to the LLMs. These 1,040 rows (1,000 + 40) were removed from the original dataset, leaving 2,241 rows. This remaining subset was used later for comparing the performance of models trained on synthetic data versus real data.

The data preprocessing steps included tokenizing each tweet into single words and transforming these words into numerical vectors representing the frequency of each word in the entire corpus of tweets. To highlight words that are distinctive to a particular document, the term frequency-inverse document frequency (TF-IDF) technique was applied to adjust the frequency vectors by how common the words are across all documents.

2.2.2 Model Training

Using the initial 1,000 rows, the authors identified the worst-performing machine learning model for the dataset. The rationale behind this approach was that if a sub-optimal model could be enhanced through the proposed methodology, the positive results would likely be observable in better-performing models as well. The random chance of correctly classifying an instance in this dataset is 54.24%, which translates to an accuracy of 67.80% based on 1.25 times the proportional chance criterion (PCC).

Furthermore, only the default parameters were used to emphasize the practicality of the proposed methodology for organizations that do not have the capacity to use additional computational resources. Table 2 shows the predictive performance of different classical machine learning models.

Despite Support Vector Machines (SVM) having the lowest test accuracy, the authors chose to explore Logistic Regression for the following reasons:

1 **Logistic Regression is highly interpretable.** It assigns weights to each word, reflecting how strongly associated that word is with the two emotions. This interpretability makes it easier to understand the influence of each word on the emotion prediction.
2 **Logistic Regression is more computationally efficient than SVM.** This efficiency emphasizes the practicality of the proposed methodology for organizations with limited computational resources.

Table 2 Model Performance on Training Accuracy

Model/PCC	Train Accuracy
K-Nearest Neighbors	0.790000
Gradient Boosting Method	0.776250
Extreme Gradient Boosting Method	0.772500
Random Forest	0.766250
Logistic Regression	0.715000
Support Vector Machine	0.715000
Proportional Chance Criterion (Random Guess)	0.677951

3 **Logistic Regression is easier to use compared to SVM.** It is straightforward to implement and understand, making it a good choice for a proof-of-concept aimed at organizations that might not have advanced machine learning expertise.

By choosing the worst-performing yet interpretable and computationally efficient Logistic Regression model, the authors aimed to demonstrate the effectiveness of the proposed methodology in a practical setting with limited resources. In this case, the accuracy to beat is 76%.

2.2.3 Obtain Significant Words through SHAP

In addition to providing exemplars to the LLMs, the Shapley Additive Explanations (SHAP) interpretability algorithm was used to better inform the generation of synthetic data. For the purpose of producing a proof of concept, only a random sample of 200 tweets were interpreted using SHAP and only words whose corresponding SHAP value belong to the top 20% were considered.

Specifically, the top 20% words that were most influential in defining each emotion (anger and joy) were identified and fed into the LLMs. Ideally, all the top words should be extracted, but this number was chosen as a proof of concept. The SHAP algorithm identified words that had a significant contribution and words that had no significant contribution to each emotion. For example, for the emotion anger, words like 'fuming', 'hate', and 'love' were identified as having significant contribution, while words like 'hilarious', 'birthday', and 'amazing' were identified as having no significant contribution.

It is worth noting that some words might seem counterintuitive to the emotion being considered. For instance, the word 'love' had a significant contribution to anger, even though it typically connotes the opposite emotion. This is because the calculation of the top words was based on the mean absolute weight of each word, focusing on the magnitude of each word's impact on the model's predictions without considering the direction of the effect. In other words, the algorithm did not differentiate whether a word pushed the model's prediction toward joy or anger; it only considered how strongly that word influenced the prediction.

To further refine the list of top words, an additional filtering step was applied based on the average importance of each word for tweets belonging to each emotional class (anger or joy). This was done by computing the average term frequency-inverse document frequency (TF-IDF) value of each word for the documents associated with each label, as produced in Section 2.2.1. If the mean TF-IDF value of a word

was greater than 0, it was considered to have a significant effect on the model's predictions. Otherwise, it was considered insignificant.

The final list of significant words was then used as keywords that the LLMs should consider when generating the synthetic data. Conversely, the LLMs were instructed not to consider the insignificant words.

2.2.4 Generate the Synthetic Data

This study utilized Anthropic's Claude.AI and OpenAI's ChatGPT to generate the synthetic textual data. Both LLMs were given the same set of exemplars (20 tweets per emotion) and were asked to generate 500 happy and 500 angry statements, totaling 1,000 AI-generated data points. Below is a sample prompt for generating the synthetic data.

> Hi Claude! I need your help. Can you generate 500 angry statements? This is for my experiment. Here are some examples: @user The hatred from the Left ought to concern everyone—-who wants a police state-the left, so than can spy on all of us.
> @user : Whaaaat?!? Oh hell no. I was jealous because you got paid to fuck, but this is a whole new level. #anger #love #conflicted
> If Monday had a face I would punch it #monday #horrible #face #punch #fight #joke #like #firstworldproblems #need #coffee #asap #follow
> ... (and so on)
> Here are the words that you may consider in your angry statements (note: each word is separated by a space): hilarious rage fuming fucking furious bad bully people angry ... (and so on)

2.2.5 Augment the Original Training Data

In this experiment, the performance of the chosen machine learning model (Logistic Regression) was compared when the training set was augmented with two different types of data: AI-generated synthetic data and real. The synthetic data was the output from the previous step, where LLMs generated 1,000 new data points. The real data was sourced from the remaining pool of 2,241 rows from Section 2.2.1 (the original dataset after removing the 1,040 rows used for training, testing, and exemplars).

Both types of supplementary data (synthetic and real) were added to the existing training set of 800 rows at an interval of 5% of the total supplementary data size. For example, at the first interval, 50 synthetic data points (5% of 1,000) were added to the original 800 training rows, creating an augmented training set of 850 rows. Similarly, 50 real data points (5% of 1,000) were added to the original 800 training rows, creating a separate augmented training set of 850 rows. This process was repeated at intervals of 5% until all 1,000 supplementary data points (synthetic or real) were added to the original training set, resulting in augmented training sets of 1,800 rows each. Table 3 illustrates this process.

The final result consisted of two kinds of augmented training sets, each containing a total of 1,800 rows:

1 **Synthetic Data Augmentation**: The original 800 training rows supplemented with 1,000 AI-generated synthetic data points.

Table 3 Improvements in Model Accuracy with Addition of Supplementary Data

Supplementary Data Added (%)	Real Data Accuracy (%)	Synthetic Data Accuracy (%)
0	74.15	74.00
5	72.30	73.80
10	73.05	78.50
15	72.15	75.65
20	73.75	76.30
25	74.50	77.60
30	76.10	77.55
35	76.75	79.65
40	76.25	78.30
45	77.50	76.35
50	77.70	76.75
55	76.45	77.40
60	79.25	77.55
65	78.15	75.55
70	79.00	77.05
75	78.20	78.05
80	79.45	78.10
85	78.70	77.70
90	79.70	77.20
95	79.65	78.25
100	81.20	78.95

2 **Real Data Augmentation**: The original 800 training rows supplemented with 1,000 additional real data points from the remaining pool.

At each interval, the augmented training set was used to train the Logistic Regression model, and the corresponding accuracy on the test set was recorded. The Logistic Regression models trained on these two augmented training sets were then evaluated and compared to assess the usability and effectiveness of the synthetic data against real data for enhancing the model's performance.

3 Results and Discussion

The consistent upward trend in both performance curves indicates that the model's predictive capabilities steadily improved as more data was introduced, regardless of its source. This observation reinforces the well-established principle that increasing the volume and diversity of training data generally leads to improved model performance, as the model is exposed to a broader range of patterns and examples.

One striking observation is that the performance of the model augmented with synthetic data generated by LLMs performs just as well as the performance of the model augmented with real data. The model achieved a maximum accuracy of 78.50% when augmented with the entire set of 1,000 synthetic data points generated by LLMs. This level of performance is nearly indistinguishable from the peak accuracy of 78.80% when the model was augmented with the full set of 1,000 real data points.

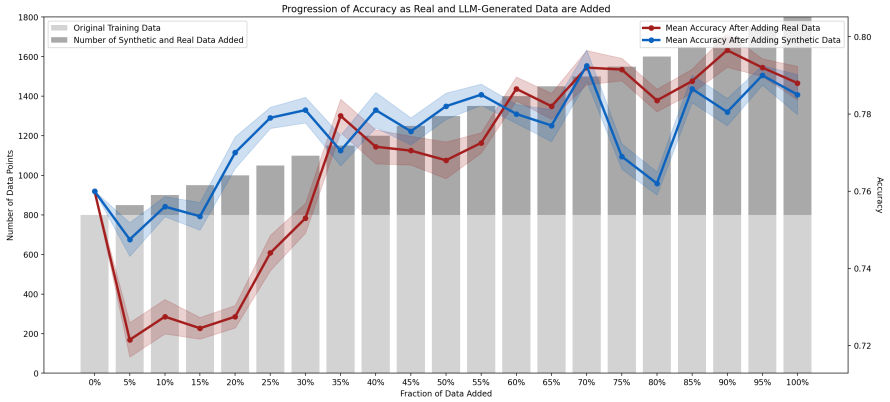

Figure 1 Progression of Accuracy as Real and LLM-Generated Data are Added

The close alignment between the performance curves for synthetic and real data augmentation is particularly noteworthy as it suggests that the synthetic data generated by LLMs captures the essential characteristics and patterns present in the real data with remarkable fidelity, as long as prompt engineering is informed. This finding has remarkable implications as it demonstrates that LLMs can effectively mimic the underlying data distribution, enabling the generation of high-quality synthetic data that can serve as a viable substitute for real data in emotion prediction tasks.

4 Conclusion

Overall, the results of this paper highlight the immense potential of leveraging large language models to generate high-quality synthetic data, thereby helping organizations overcome data scarcity challenges and take advantage of the full capabilities of state-of-the-art AI systems, particularly in the context of emotion detection or sentiment analysis tasks.

However, it is important to note that the key to unlocking this insight lies in the process of prompt engineering, which involves carefully crafting the prompts given to LLMs. In this study, the prompt engineering process was informed by two crucial components:

1 **Exemplars:** A set of 40 real data samples were provided as exemplars to the LLMs, allowing them to grasp the desired structure, tone, and style of the target textual data.

2 **SHAP-derived keywords:** The Shapley Additive Explanations (SHAP) algorithm was employed to identify the most influential words that contribute to the prediction of each emotion class (anger and joy).

These SHAP-derived keywords were then incorporated into the prompts, guiding the LLMs to generate synthetic data that accurately reflects the linguistic patterns associated with each emotion. By combining these two components, the prompt engineering process ensured that the LLMs had access to both representative examples and the most salient features of the real data.

5 Recommendations

The following recommendations are highly suggested to deepen the general community's understanding of the capabilities of large language models:

1. **Apply the Shapley Additive Explanations interpretability algorithm on all tweets.** For this proof of concept, a mere sample of 200 tweets were considered. Allowing for a more comprehensive scope could possibly lead to better keywords for prompt engineering.
2. **Increase the percentage of top words considered for prompt engineering.** The study only focused on the most influential 20% of the words when generating synthetic data with LLMs. Expanding this selection to include a larger percentage of words may enhance the generation process. More words can provide a richer context for LLMs to create more nuanced and comprehensive synthetic data.
3. **Perform a grid search on all the machine learning models.** While the study focused on the Logistic Regression model as a proof of concept, conducting a comprehensive grid search across various machine learning algorithms could identify the algorithms that benefit the most from this approach. Furthermore, exploring the performance of more complex and state-of-the-art models, such as deep neural networks, could provide a more comprehensive understanding of the scalability and generalizability of the proposed approach.
4. **Explore other LLMs as they could possibly provide better results.** The landscape of large language models is rapidly evolving. Exploring the capabilities of other LLMs, such as Gemini and PaLM, could yield interesting comparative insights.
5. **Investigate the impact of domain expertise on prompt engineering.** It would also be interesting to see whether domain expertise could further elevate and refine the quality of the generated data. Incorporating domain expertise into the prompt engineering process could to the generation of synthetic data that more accurately reflects the complexities and idiosyncrasies of real-world data.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
We would like to acknowledge our Machine Learning 2 Mentors Dr. Christoper Monterola, Prof. Leodegario Lorenzo II, and Prof. Kristine Ann Carandang for helping us refine the narrative and methodology of this study.

References

1. Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. Tweeteval: Unified benchmark and comparative evaluation for Tweet Classification. Findings of the Association for Computational Linguistics: EMNLP 2020 (2020). DOI:http://dx.doi.org/10.18653/v1/2020.findings-emnlp.148
2. Keith D. Foote. 2023. A brief history of large language models. (December 2023). Retrieved March 10, 2024 from https://www.dataversity.net/a-brief-history-of-large-language-models
3. Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 2022. Will We Run Out of Data? An Analysis of the Limits of Scaling Datasets in Machine Learning. arXiv:2211.04325. Retrieved from https://arxiv.org/abs/2211.04325