Contemporary Readings in Law and Social Justice
ISSN: 1948-9137, e-ISSN: 2162-2752
Vol 15 (02), 2023
pp. 1037- 1049

Impact of Using Synthetic Data Generated by GPT-4o-mini on the Performance of Supervised Machine Learning Models

Luis Eduardo Muñoz Guerrero [1]

[1] Universidad Tecnológica de Pereira

Abstract

Text classification models rely on large labeled datasets to achieve effective classification. However, collecting this data can be costly and laborious, especially for classification tasks in complex environments with data scarcity. Manual labeling of large data volumes requires significant time and human resources, which can be a barrier for research and various projects by limiting the performance of machine learning models.

The use of synthetic data can improve performance in machine learning models without incurring the costs associated with collection and labeling. This approach is not only more economical but also allows the generation of relevant data that improves model results.

This document presents the results of creating synthetic data with the objective of improving text classification in machine learning models. To carry out this exploration, a subset of Amazon product reviews in Spanish was used, taken from Hugging Face under the MTEB project called Amazon Review Multi.

To conduct this study, 10% of the 200,000 available Spanish reviews were randomly extracted, thus recreating a data scarcity scenario. Our objective is clear: to demonstrate that, even with a limited amount of original data, the generation of synthetic data can notably improve the accuracy and effectiveness of machine learning models.

The choice to work with a reduced subset is not coincidental. We want to emphasize that synthetic data can complement and improve results, especially in contexts where data is limited. Despite having a large dataset available, this scarcity simulation allows us to highlight the true potential and benefit of synthetic data.

With the extracted dataset, fine-tuning was performed on the GPT-4o-mini model, which resulted in the generation of synthetic data that was subsequently used to evaluate and compare the performance in classification of supervised learning models on two datasets: original and combined (original + synthetic).

The results demonstrated significant improvements of up to +5.7% in the combined data group in metrics such as Accuracy, F1 Macro, and Recall. Likewise, improvements of up to +1% were observed in metrics such as ROC-AUC. These findings indicate that the combined use of real and synthetic data through this methodology can considerably improve the performance of learning models, providing a viable alternative in scenarios where data collection is complex and costly.

Keywords: Synthetic data, Machine learning, Natural language processing, Fine-tuning, Text classification

Received: 15 Sep 2023 **Accepted:** 27 Nov 2023 **Published:** 20 Dec 2023

Introduction

Working with synthetic data has become fundamental for enriching Natural Language Processing, and for good reason: we constantly encounter barriers when trying to collect information - whether due to the difficulty of obtaining it or those thin ethical lines. This is precisely where synthetic data generation comes into play. The above, combined with recent advances in research on transformer architecture-based models such as GPT-4o-mini, opens a new window for data generation in complex environments. These models can create highly relevant data with a very high level of quality and the ability to solve multiple problems simultaneously, generating results that reduce costs and protect privacy.

The main objective of this study is to evaluate how synthetic data affects the performance of supervised machine learning models. To do this, the following workflow will be followed globally: 1.) Perform fine-tuning

on the GPT-4o-mini model 2.) Generate a synthetic dataset that faithfully captures the characteristics of the original set. 3.) Evaluate and compare performance in a supervised machine learning model.

Through this document, we aim to provide valuable insights into the effectiveness and viability of using synthetic data, offering an alternative to the challenge of data scarcity and the high costs associated with manual labeling. The results of this study have the potential to influence future strategies for improving the classification of supervised learning models, particularly in contexts where obtaining labeled data is complex or costly.

2. Problem Description

Text classification in natural language processing faces several critical challenges: First, the collection and manual labeling of large volumes of data is a costly and time-consuming process (Frénay & Verleysen, 2014). Additionally, when dealing with sensitive, medical, or ethically delicate data, the situation becomes even more complicated due to restrictions and ethical concerns (Bender & Friedman, 2018). This data scarcity negatively affects the possibility of researching and generating solutions to problems we currently face.

Synthetic data, generated through advanced models such as GPT-4o-mini, offers a promising solution to these problems (Brown et al., 2020). This data can faithfully mimic the characteristics and distributions of original data, reducing costs and overcoming ethical, collection, and privacy limitations. In this way, synthetic data can improve accessibility to large volumes of data necessary to improve the performance of machine learning models.

3. Methodology

3.1. Methodological Introduction

In this document, a set of Amazon reviews taken from Hugging Face under the MTEB project called Amazon Review Multi was used. This set classifies and labels various reviews written about a product, including a rating that reflects the user's perception on a scale of 0 to 4 (where 0 represents a bad experience and 4 the best experience).

This dataset served as the main input for this study, from which 2,000 data points were extracted, consisting of 1,400 training data, 300 test data, and 300 validation data. With the training data, fine-tuning was performed on the GPT-4o-mini model to generate synthetic data. Subsequently, 2 datasets were created: the first of synthetic data and the second called 'combined' which contains synthetic + original data. The objective was to compare the performance of these two datasets on the Multinomial Naive Bayes supervised machine learning model.

3.2. Dataset Structure

For this study, the Amazon Review Multi dataset extracted from Hugging Face under the MTEB project (Massive Text Embedding Benchmark, 2023) is used as a reference. This dataset is composed of Amazon product reviews and contains a total of 1.26 million data points in 6 different languages. Although the dataset includes multiple languages, Spanish was chosen to demonstrate that the synthetic data generation methodology is effective regardless of language. This approach underscores the versatility of the process and allows replicating the results in other languages and contexts, reaffirming that the objective is to improve the accuracy of machine learning models through synthetic data generation, regardless of the language of the original reviews. The Spanish dataset has 210,000 data points: 200,000 training data, 5,000 test data, and 5,000 validation data.

The dataset is structured as follows:

- id: A unique identifier for each review, composed of a prefix indicating the language (in this case, 'es' for Spanish) and a unique number.
- text: The complete text of the product review. The length of this field varies depending on the review content.
- label and label_text: A label representing the review rating, in a range from 0 to 4.

3.3. Data Cleaning

The text contained in the 'text' column, which contains product reviews, goes through a data cleaning process that includes, among others:

- Replacement of problematic characters: Correction of common problems in UTF-8 character encoding, such as 'Ã¡' for 'á', 'â€' for quotation marks, and other similar errors.
- Unicode normalization: Use of the unicodedata.normalize method from Python's standard library to

standardize the representation of special characters.

- Replacement of line breaks: Substitution of carriage returns (\r) and line breaks (\n) with blank spaces to maintain text coherence. This is necessary because line breaks can interrupt text coherence, causing phrases that should be together to appear on separate lines.
- Elimination of redundant spaces: Reduction of multiple spaces to a single space.
- Emojis: Emojis are kept in the text. This decision is based on the findings of Kunneman et al. (2014), who demonstrated that 'paratextual elements in social networks are not mere ornaments, but fundamental carriers of emotional and contextual meaning.'

3.4. Custom Review Mapping and Classification

For this study, we start from a dataset that includes numerical labels in the 'label' column. These labels assign each review a unique value between 0 and 4. We decided to group these ratings as follows:

- 0 and 1: Reviews with a negative perception of the product (bad).
- 2: Reviews with a neutral perception (neutral).
- 3 and 4: Reviews with a positive perception (good).

This scale was specifically adapted for our analysis. As in the work of González-Ibáñez, Muresan and Wacholder (2011) where a study of sarcasm on Twitter was conducted, we chose to exclude neutral reviews and focus only on positive and negative ones. The methodological decision to exclude neutral reviews and proceed with binary classification is also based on technical decisions aligned with the scope of this study. Shanto et al. (2023) demonstrated that machine learning models achieve better performance in binary sentiment classification compared to multiclass classification. This evidence, along with the following considerations, supported the decision to implement a binary model.

First, neutral opinions frequently contain ambivalent information that can introduce ambiguity in the classification model, potentially affecting its accuracy. Binary classification allows for a more direct evaluation of model performance, especially in the context of synthetic data generation and within the project framework.

The relevance of the objective is oriented toward identifying a clear polarity between good and bad opinions. Since the neutral class does not provide conclusive information about such polarity, it was considered dispensable in the study context. Finally, reducing classification to a binary problem (bad and good) allows for a clearer and more effective approach in interpreting results and applying machine learning models.

3.5. Distribution Verification and Class Balancing

In the Amazon Review Multi dataset extracted from Hugging Face under the MTEB project (Massive Text Embedding Benchmark, 2023), after performing sentiment classification, a review of the class distribution of the Spanish dataset was conducted. The following relative frequencies were observed: bad and good each represented 40% of the observations, while the neutral class had a significantly lower frequency, representing only 20%. This distribution is likely due to the dataset structure created by its owners, which may not necessarily reflect the actual proportion of comments in general. For the purposes of this study, it was decided to maintain this distribution for several reasons:

- Consistency in Fine-Tuning: Maintaining the original distribution ensures that fine-tuning of the GPT-4 mini model is performed with representative and balanced data, which is crucial for obtaining good results.
- Synthetic Data Generation: By working with a balanced distribution, the generated synthetic data will reflect this proportion, which is beneficial for avoiding biases in future data simulation. A balanced dataset facilitates the generation of synthetic data that equitably covers the classes of interest (bad and good).
- Avoiding bias in analysis: An equitable distribution between bad and good classes prevents the model from leaning toward a specific class due to data imbalance. This is particularly important when the objective is to demonstrate the effectiveness of synthetic data generation in improving model accuracy.
- Study Replicability: By maintaining the structured distribution of the dataset, it facilitates other researchers being able to replicate the study under similar conditions, adding value to the research by making it more robust and reliable.

With the exclusion of the neutral class, the dataset was balanced equitably between the two remaining classes. This ensures that both classes are equally represented, avoiding the bias that could arise from an unbalanced distribution. The final result is that 50% of the observations belong to the 'bad' class and 50% to the 'good' class. We verified and maintained this distribution for our tests, ensuring that the balance was achieved

through a subsample that reflects the desired proportionality for both groups.

3.6. Dataset Selection for the Study

For the specific development of this study, 1,000 reviews from each class are randomly selected using pandas' sample function with a fixed random state (random_state=42) to ensure process reproducibility.

This generates a balanced data subset with 2,000 total records distributed equitably between the good and bad categories (50% each). This decision responds to methodological and strategic reasons, aligned with the objective of demonstrating that, even with limited data, it is possible to generate synthetic data that improves analytical model results.

Demonstration with reduced data: The approach seeks to validate that, with a small dataset, synthetic data can compensate for information scarcity and optimize model performance, evidencing its applicability in limited resource contexts.

Balance between classes: The balanced sample ensures fair representation of the good and bad categories, improving model learning and eliminating possible biases.

The selection of this reduced sample reflects the study's purpose: to explore how synthetic data generation can resolve original data limitations and open opportunities in environments where data is scarce, reaffirming the relevance of this approach in practical applications.

3.7. Dataset Split for the Study

For this study, the dataset extracted in section 3.6 'Dataset Selection for the Study' is divided into training, validation, and test subsets, following a stratified sampling strategy:

- Split Proportion
- 70%: Training set - 1,400 data points (training).
- 15%: Test set - 300 data points (test).
- 15%: Validation set - 300 data points (validation).
- Stratification ensures that the proportion of classes (good and bad) is maintained in each subset. This is achieved using sklearn's train_test_split function with the stratify parameter.

3.8. Fine-tuning

3.8.1 Model

The model to be used in this study is GPT-4o mini, a proprietary model developed by OpenAI. According to OpenAI's website and as mentioned by Kotzé and Senekal (2023) in their study, GPT-4o mini is OpenAI's most cost-effective small model. GPT-4o mini surpasses GPT-3.5 Turbo and other small models such as Gemini 1.5 Flash or Claude 3 Haiku in academic parameters in both textual intelligence and multimodal reasoning. This model has a context window of 128,000 tokens and supports up to 16,384 output tokens per request. The fine-tuning for this study will be performed through the API that OpenAI has available on its website.

3.8.2 Starting the Fine-tuning Process

Using the training dataset extracted in section 3.7 'Dataset Split', equivalent to 1,400 review data points classified as 'good' and 'bad' according to section 3.4 'Review Mapping and Classification', the GPT-4o-mini model is used to perform the corresponding fine-tuning. This process aims to adapt a general model, previously trained on large amounts of data, to a specific and small dataset, in order to improve its performance on particular tasks, such as generating product reviews with specific sentiments (positive and negative).

The data was structured in a format compatible with the fine-tuning process, with the creation of specific 'prompts' for each sentiment, as shown below:

Table 1. Example of training prompt for positive perception

The entire loading and fine-tuning process was carried out through the OpenAI API.
Hyperparameter Tuning: To ensure adequate learning and avoid overfitting, specific hyperparameters were configured for fine-tuning:

Epochs: A value of 8 epochs was established. Studies by Dodge et al. (2020) demonstrate that language models

generally reach their optimal performance between 5-10 epochs.

Batch Size: A small batch size of 4 was chosen, which allows the model to generalize better and avoid overfitting to the training data. Smith et al. (2017) demonstrated that small batches can function as a form of natural regularization.

Learning Rate Multiplier: A low value of 0.05 was used for the learning rate, which favors a more stable and controlled learning process.

In the results observed when generating synthetic reviews, adequate coherence, consistency, tone, and naturalness of generation are observed. Some examples of this can be seen in Table 4.

3.9. Synthetic Data Generation

Once fine-tuning has been performed on the model, synthetic data is generated through a defined Prompt, using the resources of the OpenAI API. The prompt plays a fundamental role in data generation as it is responsible for obtaining the desired responses, ensuring contextually accurate results and increasing the utility of generative language models (Wu, Terrand, & Cai, 2022).

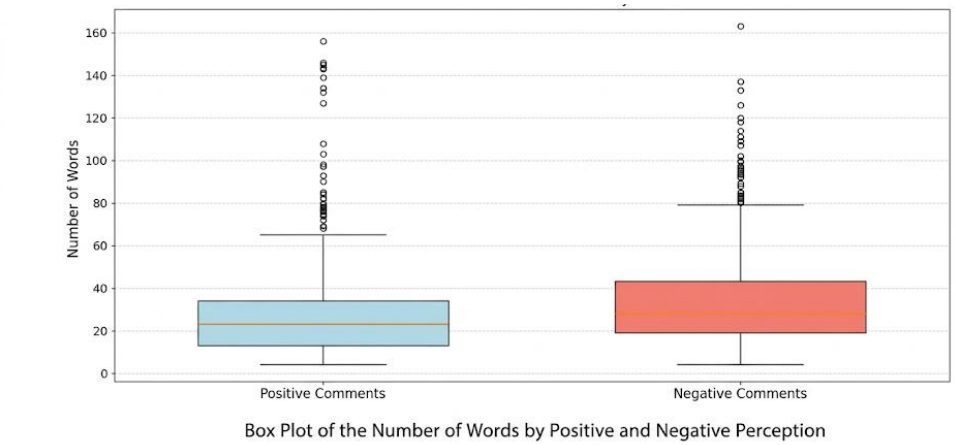
3.9.1. Preliminary Analysis

To begin synthetic data generation and define the Prompt, a preliminary analysis of the training dataset is performed. The objective is to explore data behavior and thereby enrich the instructions that will be given to the model for synthetic data generation. To do this, the following exploratory analyses are conducted.

3.9.1.1 Text Length

The objective is to evaluate the maximum, minimum length and box plot analysis of the reviews in the dataset.

Graph 1. Box plot diagram of the number of words by positive and negative perception

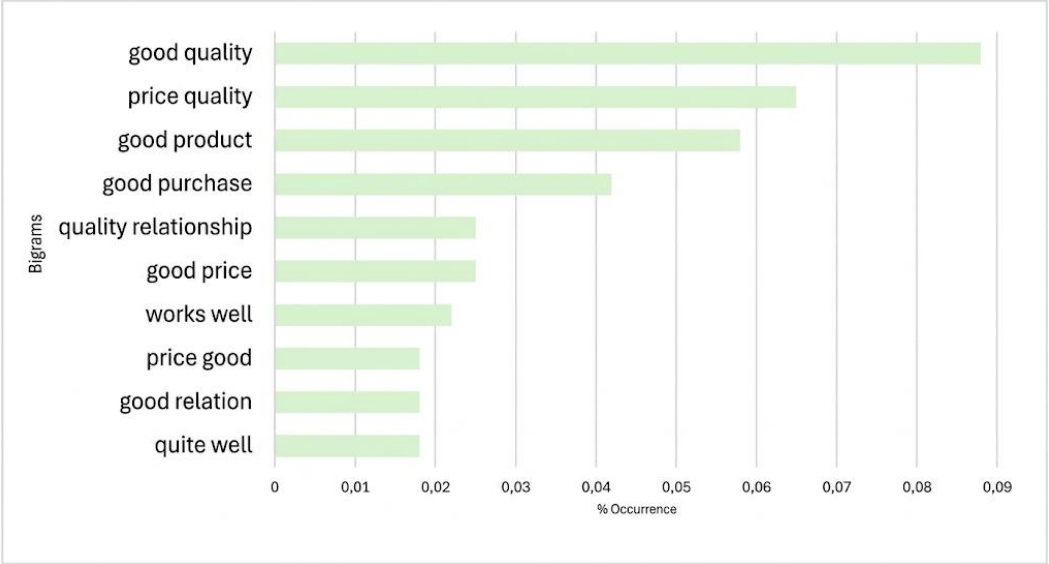


Box Plot of the Number of Words by Positive and Negative Perception

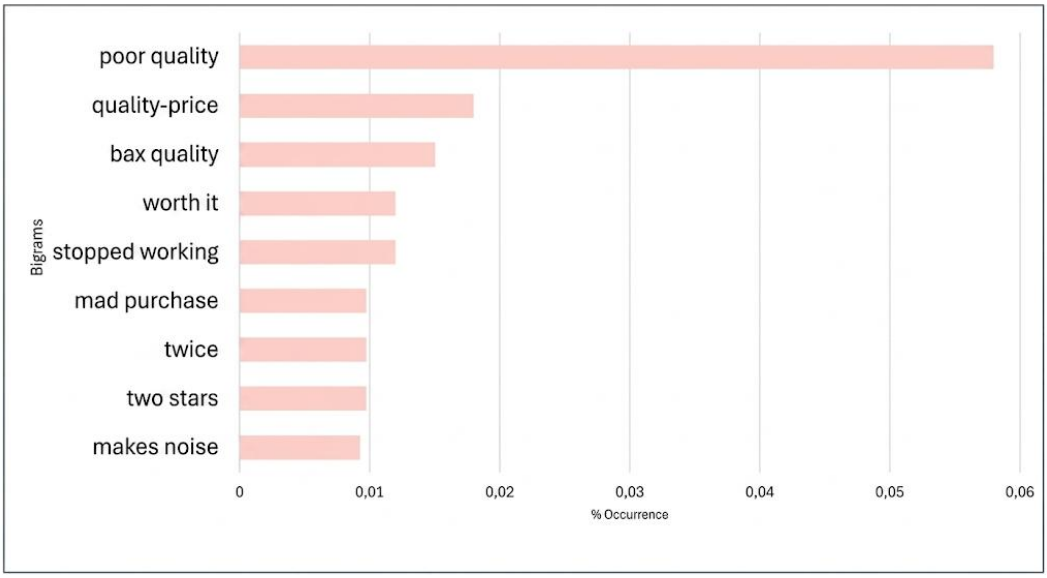
3.9.1.2 Bigram Analysis

The objective is to examine the combination of two consecutive words within reviews, identifying patterns to evaluate their insertion in the model in order to generate more effective classifiers, as has been demonstrated in various studies (Peng & Van Huurmans, 2003).

Graph 2. Bar chart, relevant bigrams with positive perception



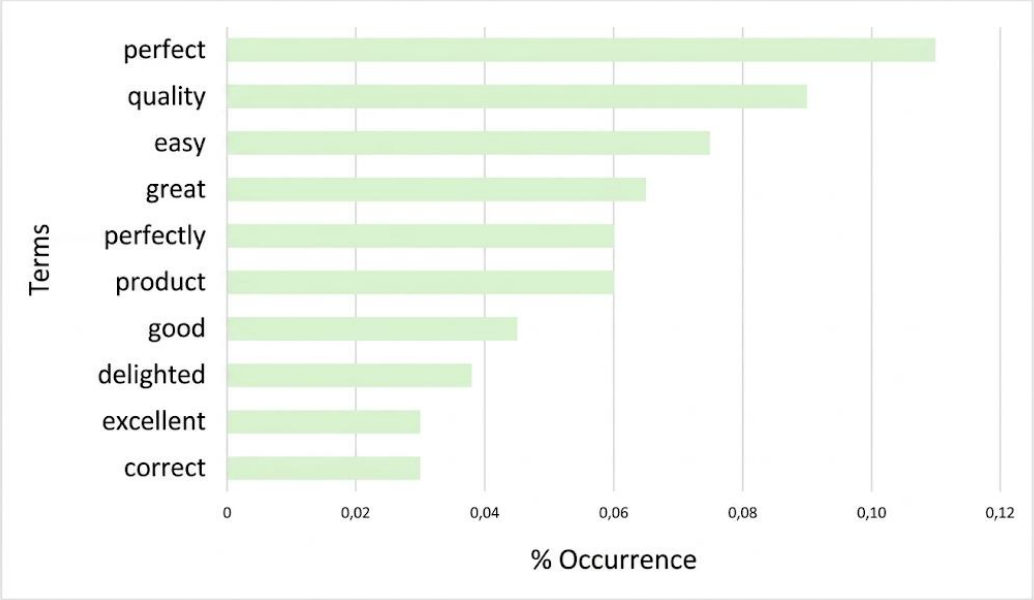
Graph 3. Bar chart, relevant bigrams with negative perception



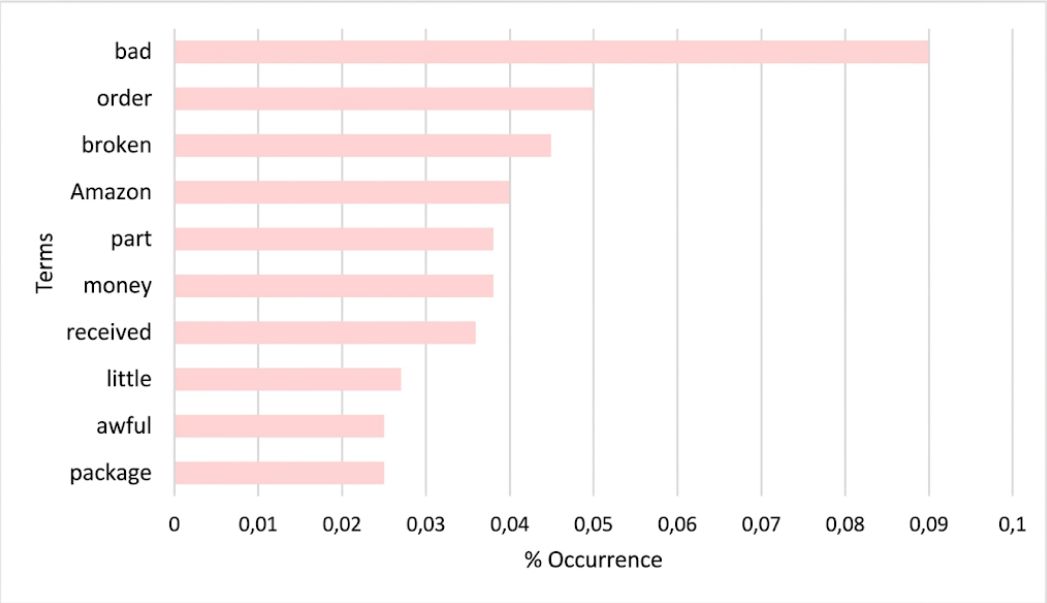
3.9.1.3 Relevant Terms Analysis

The objective is to identify and assess the importance of specific words in reviews that reveal strong patterns and/or keywords in texts through TF-IDF. This TF-IDF method generates keyword lists with a score or weight that indicates how relevant the word is with respect to the selected document and the corpus in general (Cabrera Diego, 2011).

Graph 4. Bar chart of relevant TF-IDF terms with positive perception



Graph 5. Bar chart of relevant TF-IDF terms with negative perception



3.9.1.4 Data Analysis Conclusions

Preliminary analyses have revealed certain patterns. For example, it is observed on average that reviews with negative perception are +21% longer than those with positive perception, which may suggest that depending on whether it is one or the other, they should be created synthetically with different lengths to generate diverse data that is representative of the general average. Therefore, in terms of text length, the following references will be taken, with the objective of controlling diversity and aligning with the original data.

Table 2. Maximum and minimum length by review type

On the other hand, preliminary analyses have revealed certain patterns in text analysis. For example, the maximum frequency observed for a specific bi-gram or term in our analysis was 10%. While this does not suggest a dominant trend, it does reflect a relevant value on which the form of inclusion in the study can be evaluated. Based on these findings, two different strategies are proposed to generate synthetic data through prompt elaboration in the fine-tuned model.

Standard Prompt: Consists of creating a prompt that includes a simple instruction, designed according to the model's fine-tuning. This strategy also incorporates a word count restriction based on preliminary analysis results, with the objective of adjusting content generation to the specific observed context.

Probability-Based Random Prompt: This strategy expands the Standard Prompt by incorporating n-grams and keywords according to their frequency in the original data, with the objective of increasing their relevance in the generated synthetic data. Its operation is simple: for each generation of synthetic data, a random number is calculated for each listed term and/or n-gram. This number is compared with the previously calculated probability of each term, derived from its frequency of occurrence in the original dataset. If the random number is less than or equal to the probability assigned to the term, it is selected for inclusion in the prompt; otherwise, it is omitted. This method seeks to ensure that only the most relevant and frequent elements are selected, maintaining fidelity with the original data. If no term meets the criterion in a generation, the prompt is created without those terms.

3.9.2. Synthetic Data Generation

On the fine-tuned model, 4,000 synthetic data points were generated: 2,000 generated by a standard prompt and 2,000 generated by the probability-based random prompt. The model temperature for synthetic data generation was 0.7, which generates more diverse and creative results in a moderate-high manner (Namitha, Manjula, & Belavagi, 2023) while maintaining text coherence. These datasets were generated through prompt definition as follows:

Table 3. Number of synthetic data generated by prompt type
Some of the synthetic data generated according to positive or negative perception are as follows:
Table 4. Random sample of synthetic data generated according to user perception

3.10 Implementation and Training of Multinomial Naive Bayes Model

Cleaning: It begins with the general cleaning described in section 3.3. Additionally, a process of stopword removal and dimensionality reduction was performed.

Processing and separation: Tokenization and text vectorization are performed for each dataset: training, validation, and test.

Training: The Multinomial Naive Bayes model is used to perform training using the validation set and the test set to evaluate final results.

Evaluation: The models will be evaluated and compared with metrics such as Accuracy, F1-Macro, ROC-AUC, and Recall in different evaluation scenarios.

Results: Results will be generated in tables and graphs including comparison of the main metrics to facilitate data comprehension and clarity in findings.

3.11. Results

To see the results of supervised machine learning model performance, different scenarios will be evaluated on the Multinomial Naive Bayes models. Below, the 2 scenarios to be evaluated are defined:
Table 5. Evaluation scenarios of the presented model

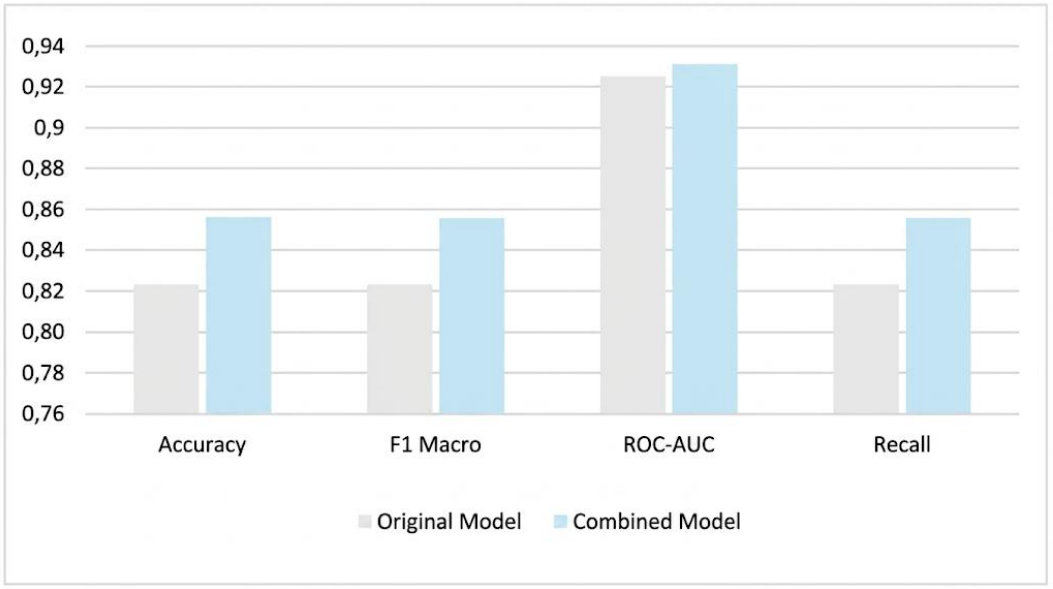
Scenarios 1 and 2 will be validated with the same test and validation dataset that were defined in section 3.7 'Dataset Split for the Study'. The metrics that will be used to evaluate and compare the results will be those defined in section 3.10 'Implementation and Training of Multinomial Naive Bayes Model'.

3.11.1 Training and Testing Results of Multinomial Naive Bayes Model Applying Synthetic Data Generation with Standard Prompt

The synthetic dataset generated through standard prompt contains 2,000 data points distributed as shown in Table 3. The results generated by applying the different scenarios defined in Table 5 on the Multinomial Naive Bayes model are as follows:

Table 6. Training and testing results of Multinomial Naive Bayes model with Standard prompt
Below, we observe a comparative graph with the results from Table 6, which visually shows the key differences in the performance of the Multinomial Naive Bayes model when using original and combined datasets with standard prompt.

Graph 6. Bar chart of original vs combined model results on Multinomial Naive Bayes model with standard prompt



Regarding the metric results, the following results are observed:

Improvement in Accuracy and Recall: The increase in accuracy and recall of approximately 0.0334 reflects a notable improvement in the model's general ability to correctly classify both positive and negative instances. This suggests that synthetic data has effectively contributed to improving the representation of features in the model, which has resulted in better discrimination between classes.

Increase in F1 Macro: A similar increase in F1 Macro indicates that the model's precision and sensitivity have been better balanced with the inclusion of synthetic data.

Slight Improvement in ROC-AUC: Although the increase in ROC-AUC is less pronounced (+0.0061), it remains significant. This shows that the model's ability to correctly classify between positive and negative classes at different thresholds has slightly improved.

The inclusion of synthetic data significantly reduced false positives (from 30 to 20) without increasing false negatives, thus improving the model's specificity without compromising sensitivity. This underscores the value of synthetic data for refining the model's ability to correctly discriminate between negative classes. Likewise, the model with combined data increased the number of true positives (from 120 to 130), highlighting the utility of synthetic data in improving the detection of positive cases. True negatives and false negatives remained constant, indicating that the introduction of synthetic data did not negatively affect the model's ability to identify negative cases, maintaining a solid performance base while improving other aspects.

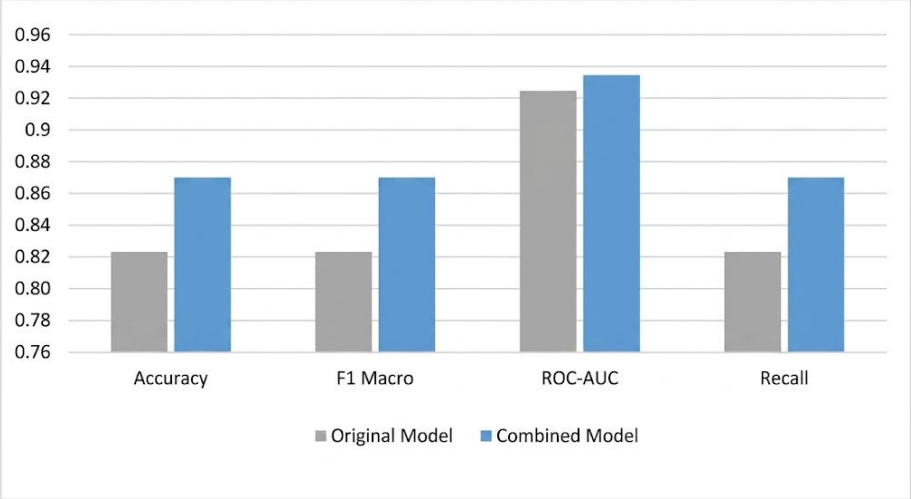
3.11.2 Training and Testing Results of Multinomial Naive Bayes Model Applying Synthetic Data Generation with Probability-Based Random Prompt

The synthetic dataset generated through probability-based random prompt contains 2,000 synthetic data points distributed as shown in Table 3. The results generated by applying the different scenarios defined in Table 5 on the Multinomial Naive Bayes model are as follows:

Table 7. Training and testing results of Multinomial Naive Bayes model with Probability-Based Random Prompt

Below, we observe a comparative graph with the results from Table 7, which visually shows the key differences in the performance of the Multinomial Naive Bayes model when using original and combined datasets with random prompt.

Graph 7. Bar chart of original vs combined model results on Multinomial Naive Bayes model with random prompt



Regarding the metric results, the following results are observed:
Accuracy and Recall: Both Accuracy and Recall showed significant improvement in the combined model compared to the original model, with an increase of approximately 0.0467 in both cases. This indicates that the combined model is more effective at correctly classifying both positive and negative cases.

F1 Macro: Like Accuracy and Recall, F1 Macro increased by 0.0468, reflecting an improvement in the balance between precision and sensitivity of the combined model. This metric is particularly important because it takes into account both false positives and false negatives.

ROC-AUC: The increase in the ROC-AUC metric, although smaller (+0.0099), remains relevant because it indicates a better ability of the combined model to discriminate between classes at different decision thresholds.

On this occasion, the combined model not only reduced false positives and increased true positives, but also improved the true negative rate (from 127 to 131) while reducing false negatives. These improvements indicate an overall optimization in the model's accuracy and reliability. Although the reduction in false negatives was low (from 23 to 19), it is an indication that the combined model retains the ability to correctly classify negative cases without compromising the recognition of positive cases.

4. Overall Results

It is very important to highlight that the results obtained demonstrate the significant value that synthetic data brings to improving the performance of classification models. This contribution is manifested in all key performance metrics, from general accuracy to the model's ability to discriminate between classes. Below, we can compare the results obtained in evaluation metrics using a standard prompt or a probability-based prompt in the combined data model (synthetic + original data).

Table 8. Comparison of results when using combined data (original + synthetic) in a supervised machine learning model with different Prompting techniques

The standard prompt had moderate improvements in all metrics, with the greatest increase observed in Accuracy and F1 Macro (+0.0334 and +0.0328 respectively). On the other hand, the probability-based prompt had even greater increases in all metrics, especially highlighting Accuracy and F1 Macro (+0.0467 in both cases).

Although both methods show improvements in ROC-AUC, the increase is more pronounced with the probability-based random prompt. This indicates that data generated from this strategy can provide better

discrimination between classes at different thresholds, which is crucial for applications that require fine adjustments in model sensitivity and specificity.
For the analysis, it is important to compare and analyze the percentage increases generated by each of the measurement indicators by comparing the results obtained by the combined models vs the original model according to the generation prompt.

Table 9. Percentage increases in evaluation metrics of a combined dataset vs original dataset with different Prompt types

The results obtained demonstrate that both strategies for generating synthetic data show notable improvements in Accuracy, F1 Macro, and Recall. However, synthetic data generation based on Probabilities proves to be more effective, with increases approximately 1.6% higher in these metrics compared to the Standard Prompt.

This suggests that the inclusion of n-grams and specific keywords, selected by their frequency of appearance in the original data, can provide the combined model with an advantage in identifying and classifying both positive and negative instances in a better way.

5. Conclusions

This study has demonstrated that synthetic data generation through the GPT-4o-mini model offers significant improvements in key performance metrics in supervised machine learning models. Specifically, the combined dataset (original + synthetic) presents an increase of up to +5.7% in the main evaluation metrics compared to the original dataset.

For synthetic data generation, the importance of conducting prior analyses on the original dataset is highlighted to be taken into account as a strategy in generation. The inclusion of n-grams, keyword identification, and descriptive statistics analysis can help increase overall performance using this data in generation prompts.

The Probability-based prompt strategy vs the standard prompt shows superior performance of up to +1.6% in the main metrics. This opens the possibility for new strategies in synthetic data generation with customized prompts to be investigated and experimented with.

Experimenting with combinations of various synthetic data generation techniques could help identify synergies that offer additional improvements in model performance.

The implementation of natural language processing models such as GPT-4o-mini highlights the need for more research in natural language processing, encouraging researchers to explore practical applications that maximize their potential.

The results offer a foundation for companies and industries to adopt this type of technique without the need to invest significant resources in acquiring or generating new data. This approach allows for efficient and cost-effective implementation, facilitating continuous improvement and innovation.

References

1. Shanto, S. S., Ahmed, Z., Hossain, N., Roy, A., & Jony, A. I. (2023). Binary vs. multiclass sentiment classification for Bangla e-commerce product reviews: A comparative analysis of machine learning models. International Journal of Information Engineering and Electronic Business (IJIEEB), 15(6), 48-63. https://doi.org/10.5815/ijieeb.2023.06.04
2. Wu, T., Terrand, M., & Cai, C. J. (2022). AI chains: Transparent and controllable human-AI interaction and chaining large language model prompts. In Proceedings of the Conference on Human Factors in Computing Systems (pp. 1-10). https://doi.org/10.1145/3491102.3517582
3. Cabrera Diego, L. A. (2011). TF-IDF para la obtención automática de términos y su validación mediante Wikipedia (Tesis de ingeniería, Facultad de Ingeniería, Universidad Nacional Autónoma de México, México D. F., México). https://repositorio.unam.mx/contenidos/3505934
4. Peng, F., & Van Huurmans, D. S. (2003). Combining Naive Bayes and n-gram language models for text classification. In F. Sebastiani (Ed.), ECIR (Vol. 2633, pp. 335-350). Springer. https://doi.org/10.1007/3-540-36618-0_24
5. Frénay, B., & Verleysen, M. (2014). Classification in the presence of label noise: a survey. IEEE Transactions on Neural Networks and Learning Systems, 25(5), 845-869.

6. https://doi.org/10.1109/TNNLS.2013.2292894
7. Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics, 6, 587-604. https://doi.org/10.1162/tacl_a_00041
8. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901. https://doi.org/10.48550/arXiv.2005.14165
9. Kotzé, E., & Senekal, B. A. (2023). Benchmarking political bias classification with in-context learning: Insights from GPT-3.5, GPT-4, LLaMA-3, and Gemma-2. In A. Gerber, J. Maritz, & A. W. Pillay (Eds.), Artificial intelligence research. SACAIR 2023. Communications in computer and information science (Vol. 2326, pp. xx-xx). Springer. https://doi.org/10.1007/978-3-031-78255-8_10
10. Namitha, M. V., Manjula, G. R., & Belavagi, M. C. (2023). StegAbb: A cover-generating text steganographic tool using GPT-3 language modeling for covert communication across SDRs. IEEE Access, 12, 82057–82067. https://doi.org/10.1109/ACCESS.2023.3411288
11. Kunneman, F., Liebrecht, C., van Mulken, M., & van den Bosch, A. (2014). Signaling Sarcasm: From Hyperbole to Hashtag. Information Processing & Management, 50(5), 726-733. https://doi.org/10.1016/j.ipm.2014.07.006
12. Massive Text Embedding Benchmark. (2023). Amazon Reviews Multilingual Dataset [Dataset]. Hugging Face. https://huggingface.co/datasets/mteb/amazon_reviews_multi
13. González-Ibáñez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in Twitter: A closer look. In ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (pp. 581-586). (ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies; Vol. 2).
14. Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., & Smith, N. (2020). Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. arXiv preprint arXiv:2002.06305.
15. Smith, S. L., Kindermans, P. J., Ying, C., & Le, Q. V. (2017). Don't decay the learning rate, increase the batch size. arXiv preprint arXiv:1711.00489.

Table 1. Percentage increases in evaluation metrics of a combined dataset vs original dataset with different Prompt types

Role	Content
system	You are an expert in sentiment analysis and product review generation. Your task is to understand and generate reviews that clearly express positive sentiment towards products.
user	Generate a positive review for a product, maintaining a natural and convincing tone, similar to how a real customer would do it.
assistant	it's worth it, it's a good product, quality price

	Positive perception reviews	Negative perception reviews
Max Length (Equiv Median)	23	28
Min Length (Equiv 90th Percentile)	53	64

		Standard Prompt (Set 1)	Probability-Based Random Prompt (Set 2)
Synthetic Data	Positive perception	1,000	1,000
Synthetic Data	Negative perception	1,000	1,000
	Total	2,000	2,000

Perception	Generated Review
Positive	good price and works well, very easy to install and for my car model (Opel Astra) it fits great, the only downside I see is that it comes in the box looser than it should and it's likely it could fall
Positive	quite good. I bought it for a 2-year-old child who is in the learning phase and he's loving it. I like that you can do it with your finger and not just with the pencil, so it doesn't break
Negative	it broke on first use, it arrived broken, the plastic base was cracked. I returned it and they sent me another one that got stuck and couldn't be used. I returned that one too and now I have to wait for them to refund my money. it caught me on vacation and I've been more than a week without it
Negative	bad characteristics, the truth is they're not good at all, since lights leak through and you can't see the device screen well. I don't recommend it, although the price is very attractive.

Scenario 1	Scenario 2
Original Data	Synthetic + Original Data

Metric	Original Model	Combined Model	Difference
Accuracy	0.8233	0.8567	0.0334
F1 Macro	0.8232	0.8560	0.0335
ROC-AUC	0.9249	0.9310	0.0061
Recall	0.8233	0.8560	0.0334
Metric	Original Model	Combined Model	Difference
Accuracy	0.8233	0.8700	0.0467
F1 Macro	0.8232	0.8700	0.0468
ROC-AUC	0.9249	0.9348	0.0099
Recall	0.8233	0.8700	0.0467
Metric	Standard Prompt (Combined Model)	Probability Prompt (Combined Model)	Difference
Accuracy	0.8567	0.8700	0.0133
F1 Macro	0.8560	0.8700	0.0140
ROC-AUC	0.9310	0.9348	0.0038
Recall	0.8560	0.8700	0.0140

Metric	Combined Model (Standard prompt) vs Original model	Combined Model (Probability prompt) vs Original model
Accuracy	4.1%	5.7%
F1 Macro	4.0%	5.7%
ROC-AUC	0.7%	1.1%
Recall	4.0%	5.7%