Comparative Study of Synthetic Data Generation Models - Nemotron, Llama, and Mixtral: Using Sentiment Analysis Data

Francis Maria Sharan
School of Computer Science Engineering (SOCSE)
RV University, RV Vidyanikethan Post
8th Mile, Mysuru Road, Bengaluru – 560059
francisms.btech22@rvu.edu.in

Saatvik B Hampiholi
School of Computer Science Engineering (SOCSE)
RV University, RV Vidyanikethan
8th Mile, Mysuru Road, Bengaluru – 560059
saatvikbh.btech22@rvu.edu.in

***Abstract*—This research evaluates the performance of synthetic data coming from three advanced models: Nemotron, Llama, and Mixtral.They have been trained to serve for sentiment analysis use case. Each model was set with the generation of sentiment-aligned synthetic text data and accessed through appropriate APIs and platforms. On that account, we ranked each model in terms of novelty, overlapping with original data, consistency on topic distribution, readability, and alignment of intensity of sentiments. Results showed that Nemotron obtained the highest values of uniqueness and novelty, Llama was able to maintain balanced thematic and sentiment consistency, and Mixtral produced more readable texts however, had a risk of emphasizing sentiment intensity. In conclusion, these outcomes indicate the advantages and disadvantages of each model concerning the realization of specific requirements in the selection of the model according to the application needs in the generation of synthetic data.**

***Index Terms*—Generative AI, Large Language Models (LLMs) , Nemotron, Llama, Mixtral, Perplexity, Lexical Diversity, VADAR Sentiment, Topic Consistency**

I. Introduction

Sentiment analysis has applications in social media monitoring, analyzing customer feedback, and marketing. However, the collection of large amounts of sentiment-labeled data is difficult because of the issues arising from privacy, cost, and availability. The novel approach of generating synthetic data helps in creating high-quality labeled datasets without dependence on real-world data. This can serve as a close to real dataset for models to be trained on and implemented in this use case. We assess the quality of three prominent language models in generating synthetic sentiment-labeled data in this study. [1] We will compare the quality of each model produced by a number of metrics - perplexity, lexical diversity, VADAR sentiment analysis, uniqueness, readability, thematic alignment, and consistency of sentiment intensity. This work therefore gives insight into the best-suited model to generate synthetic sentiment data according to varied quality requirements.

979-8-3503-0470-1/23/$31.00 © 2023IEEE

II. Literature Review

Several applications of generating synthetic data have been demonstrated in natural language processing domain, particularly in tasks such as sentiment analysis that can benefit from large-scale, labeled datasets. It is demonstrated that recent generative models like GPT-2, BERT, and diffusion models are capable of generating realistic, synthetic data, approximating real language patterns from a few examples [2]. Most of them, however, evaluate the synthetic data using standard metrics, namely perplexity and BLEU scores, which are not sensitive enough for a fine-grained analysis such as sentiment fidelity or thematic consistency.

This study contributes to the existing body of literature as it applies a multi-dimensional evaluation framework to assess cross-cutting metrics of Nemotron, Llama, and Mixtral, specifically related to sentiment analysis. This will assess the thematic alignment, readability, and sentiment intensity that represents a holistic critique needed for closing the gap related to the quality assessment of synthetic data for sentiment-related applications. [7]

III. Methodology

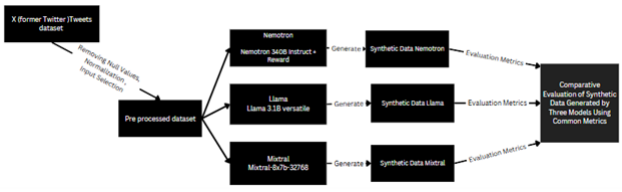


Fig. 1. Workflow of the research project

A. Data Preparation

For this research, we used a sentiment-labeled dataset with four categories: **Positive**, **Negative**, **Neutral**, and **Irrelevant**. The original dataset served as the baseline to create synthetic data with real-world patterns of sentiment. Each model was trained to generate text with regard to each one of the

sentiment categories so that all datasets are consistent. This preprocessed dataset lists the original texts along with their sentiment labeled in such a way that synthetic text generation and analysis can be easily used. The entry contains an original text field which denotes the initial content, while another field, the name of the sentiment label, describes the emotional tone. This cleaning and standardization of the dataset therefore removes noise from the data and gives it consistency in formatting so that the output is prepared for tasks such as generating synthetic responses and evaluation of lexical diversity, perplexity, and embedding similarity.

B. Models and Access

Each model was accessed and configured through specific platforms to ensure the best possible performance in generating synthetic sentiment-aligned data:

- **Nemotron:** Accessed through the Integrate API by NVIDIA, the Nemotron model is distinguished by strong output diversity and advanced language generation capabilities. Sentiment-aligned configurations have been produced by Nemotron for this purpose-only to generate distinctive and novel text samples that behave according to the preset categories of sentiment.
- **Llama:** Accessed through Groq using a model tuned for gentle thematic and sentiment matching. The prompts were presented to Llama such that it would mirror the sentiment labels of the original dataset, so the model is capable of balancing between sentiment intensity and topic distribution.
- **Mixtral:** Accessed through Groq, Mixtral is fine-tuned for readability and clear wording. Mixtral was tuned to generate text that is clear and interpretable, and it can be useful when clarity of reading is an asset. In testing, tuning for Mixtral appeared to boost effect size at times with sentiment, and in some applications this will reduce the subtlety of the output.

C. Evaluation Tools and Methodologies

- **Pandas:**
  Purpose: Pandas is a high-performance library used in Python for effective data loading, preprocessing, and organizing into DataFrames which facilitates the manipulation and aggregation of structured data. The usage of this tool was imperative to manage the dataset, apply transformations, and prepare the metrics for analysis.
- **Transformers (Hugging Face):**
  Purpose: To measure perplexity, we used GPT2 LM Head Model and GPT2 Tokenizer from the Transformers library of Hugging Face. Since GPT-2 is pre-trained as a language model, it was highly appropriate to apply it in computing the perplexity of our text data in this end task. It can help demonstrate how naturally synthetic text could be made to replicate real human dataset due to calculations of perplexity.
- **Sentence Transformers:**
  Purpose: Sentence embeddings were produced using the all-MiniLM-L6-v2 model from the library Sentence-Transformer. Such embeddings capture semantic content of each text and enable computation of cosine similarity between the original and synthetic texts. The higher score of similarity means the semantic alignment, which is rather important for analyses on the preservation of meaning in synthetic data.
- **Scikit-Learn:**
  Purpose: Scikit-Learn provides several modules for text analysis, namely cosine similarity for embedding-based similarity and structural overlap, CountVectorizer for bigram overlap, and Latent Dirichlet Allocation (LDA) for topic modeling. Cosine similarity would measure semantic similarity between the original and synthetic texts, while structural similarity was gained through bigram overlap. LDA, coupled with TfidfVectorizer, will acquire topical distributions, that will help in better thematic consistency analysis.
- **VADER Sentiment Analysis:**
  Purpose: Using the VADER, which stands for Valence Aware Dictionary and Sentiment Reasoning, Sentiment Intensity Analyzer derives sentiment intensity scores to quantify the positive, negative, and neutral sentiments in the text. In other words, the emotional tone in generated data can be compared for the strength of sentiment between the original and the synthetic text. [5]
- **Numpy:**
  Purpose: The underlying library of numerical computation, Numpy, facilitated the efficient operations on arrays integral to the calculations of bigram overlap and complex metric analyses. As such, the ability in matrix operations as well as vectorized computations helped in supporting the computational requirement of several similarity metrics.
- **Textstat:**
  Purpose: Flesch reading ease and flesch kincaid grade readability metrics were calculated with the Textstat library in order to see the complexity of the original and synthetic texts. These metrics help to evaluate whether the synthetic text could be maintained at a near level of the original reading, giving insights into accessibility and clarity.
- **Tfidf Vectorizer (from Scikit-Learn):**
  Purpose: It was using TF-IDF(Term Frequency-Inverse Document Frequency) for term importance to be captured in the original text and synthetic text which served as input for topic modeling using LDA. TF-IDF emphasizes unique words that are very frequent to highlight the major topics inherent in text data, giving a structured way of topical comparison. [12]
- **Latent Dirichlet Allocation (LDA):**
  Purpose: LDA, a topic modeling technique, was used to analyze topic distributions across original and synthetic texts. This method enables us to quantify the degree

of topical alignment, providing insights into whether synthetic text reflects the thematic structure present in the original content.

- **CountVectorizer:**
  Purpose: We apply the LDA-topic modeling approach-to analyze topic distributions between original and synthetic texts. This allows us to measure the degree of topical alignment, providing an understanding of whether synthetic text actually reflects the thematic structure present in the original content.

D. Evaluation Metrics

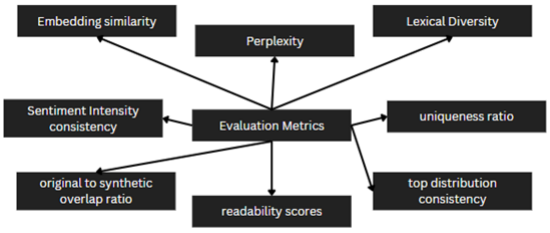


Fig. 2. Evaluation Metrics.

To evaluate the quality of synthetic data generated by each model, we employed a comprehensive set of metrics:

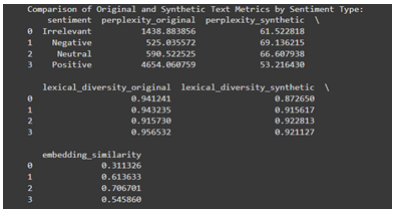
- **Perplexity:**
  Perplexity measures text fluency, with low values indicating more natural, fluent text. For the evaluation, calculate the perplexities of the original and synthetic texts using GPT-2. A comparison of average values of perplexity will give insight into how closely a synthetic text mirrors the naturalness of an original text. [8]
- **Lexical Diversity:**
  Lexical diversity is a measure of vocabulary richness as the proportion of distinct words to total words. Calculation: For every original text and synthetic text, calculate lexical diversity. There will be more impressive diversity in case of higher lexical diversity, and in case of similar or slightly lower diversity in synthetic text, it will be representative of the variety in original text. [9]
- **Embedding Similarity:**
  Embedding Similarity measures how well the generated text matches the original text in terms of meaning. Calculation: Compute the sentence embeddings for both texts and then compute the cosine similarity. Interpretation: The higher the similarity-the more it is close to 1-the higher probability that the generated text will be closer to the original one in the point of semantics or tone. [10]
- **Sentiment Intensity Consistency:**
  The synthesized text was analyzed using VADER sentiment analysis to ensure that it expressed the intended strength of sentiment. [5]
- **Uniqueness Ratio:**
  The percentage of unique entries in the synthetic data of each of these models demonstrated diversity in terms of composition.
- **Original to Synthetic Overlap Ratio:**
  It measures the overlap between the original and synthetic data to check for novelty.
- **Topic Distribution Consistency:**
  Latent Dirichlet Allocation (LDA) is used to assess the thematic coherency of each sentiment category.
- **Readability Scores:**
  Flesch Reading Ease scores were calculated for both synthetic and original texts to assess language complexity
- **Latent Dirichlet Allocation (LDA):**
  Purpose: The technique utilized in order to study topic distributions, across both original and synthetic texts, through this approach is LDA, which is a technique of topic modeling. This allows one to quantify the extent to which the topics are aligned and whether or not the thematic structure that is identifiable in the content occurs in synthetic text. [11]

  By applying these metrics to Nemotron, Llama, and Mixtral, we obtained a multidimensional assessment of each model's performance in generating sentiment-aligned synthetic data.

IV. Results and Discussion

A. Perplexity, Lexical Diversity and Embedding Similarity

- **Nemotron:**
  Nemotron had rather low perplexity scores, hence it generated fluently except on a few occasions where complicated sentences were generated and thus affecting the perplexity score. It has the highest lexical diversity. There is a good amount of original data similarity in embedding that the model achieved especially regarding Positive and Negative sentiments.



Comparison of Original and Synthetic Text Metrics by Sentiment Type:

	sentiment	perplexity_original	perplexity_synthetic	\
0	Irrelevant	1438.883856	61.522818	
1	Negative	525.035572	69.136215	
2	Neutral	590.522525	66.607938	
3	Positive	4654.060759	53.216430	

	lexical_diversity_original	lexical_diversity_synthetic	\
0	0.941241	0.872650	
1	0.943235	0.915617	
2	0.915730	0.922813	
3	0.956532	0.921127	

	embedding_similarity
0	0.311326
1	0.613633
2	0.706701
3	0.545860

Fig. 3. Perplexity, Lexical Diversity and Embedding Summary for Nemotron.

- **Llama:**
  The perplexity scores of Llama fell in between Nemotron and Mixtral; in other words, the approach toward fluency and complexity is well balanced. The lexical diversity of this model was moderate with diverse vocabulary but not as comprehensive as Nemotron. Also, the high embedding similarity was obtained for all sentiments for Llama, indicating its similarity with the meaning of the

original data.

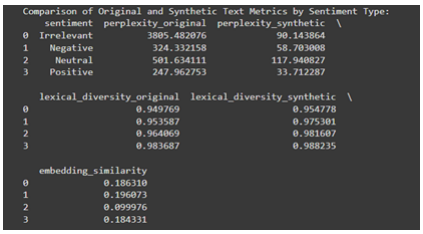


Fig. 4. Perplexity, Lexical Diversity and Embedding Similarity for Llama.

- **Mixtral:**
  Lowest perplexity scores for Mixtral: It can be stated that Mixtral has produced fluent and simple text. The low lexical diversity shows that the writing tends to become straightforward with more common words. The embedding similarity was also low, especially in Positive and Negative sentiments, showing a loss of detail due to its emphasis on simplicity and readability.

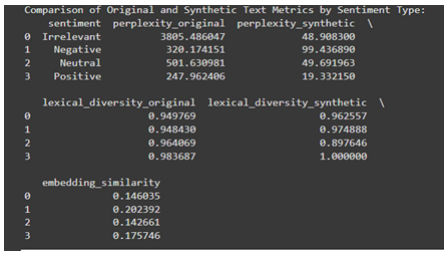


Fig. 5. Perplexity, Lexical Diversity and Embedding Similarity for Mixtral

Interpretation: This makes Nemotron suitable for applications where rich vocabulary and high uniqueness are needed. However, it might not be ideal for readability and neutral tone alignment. Llama balances both diversity, readability, and semantic alignment well. Thus, it is the most versatile to generate realistic synthetic data. Mixtral is more focused on readability and simplicity for easy accessible text and relatively lower semantic and lexical diversity.

B. Sentiment Intensity Consistency(VADAR Sentiment)

- **Nemotron:**
  Nemotron's sentiment intensity alignment was quite unstable and good, especially in the Negative sentiment, sometimes yielding less intense expressions than needed.

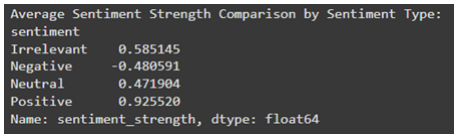


Fig. 6. VADAR Sentiment Analysis for Nemotron.

- **Llama:**
  Llama maintains the best sentiment intensity alignment, as it is fairly consistent in all of the categories.

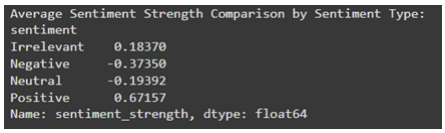


Fig. 7. VADAR Sentiment Analysis for Llama.

- **Mixtral:**
  Mixtral has a tendency to increase sentiment intensity. This can have an effect on its usage in contexts where slight variations of sentiment may be required.

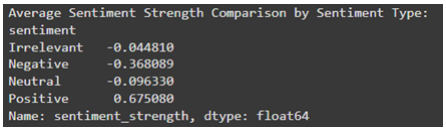


Fig. 8. VADAR Sentiment Analysis for Mixtral

Interpretation: The accurate representation of sentiment about which Llama is the most reliable model, makes it the best choice for applications sensitive to sentiment. The overemphasis of sentiment in Mixtral may influence its usability in subtle analysis, while the variation of Nemotron may impact tasks sensitive to the precise strength of sentiment.

C. Uniqueness and Overlap (NGram Analysis)

- **Nemotron:**
  Nemotron achieved a **99.14** percent uniqueness ratio with **15** percent overlap, showing that the synthetic text maintains some of the original structure of the data but brings forth new phrases and diversity as novelty in its synthetic data.

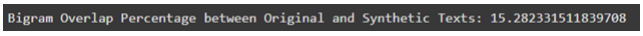


Fig. 9. N Gram Analysis for Nemotron.

- **Llama:**
  The uniqueness ratio was slightly lower at **97.6** percent, with an overlap of **2.3** percent with the original dataset. This means that, though Llama does pretty well in producing novel text, it sometimes keeps phrases or structures from the original data, thus possibly limiting the novelty of the text created.

- **Mixtral:**
  Mixtral has a **98.9** percentage uniqueness ratio while having an overlap of a **2.27** percent overlap with the original data, thus mirroring Nemotron in novelty but

```
Bigram Overlap Percentage between Original and Synthetic Texts: 2.3126463700234194
```

Fig. 10. N Gram Analysis for Llama.

with minor overlaps.

```
Bigram Overlap Percentage between Original and Synthetic Texts: 2.273389682308365
```

Fig. 11. N Gram Analysis for Mixtral

Interpretation: Nemotron and Mixtral have much higher uniqueness without much overlap, which makes them suitable for creating new data. With just some degree of minor overlap, Llama's applicability could be limited to scenarios where highly different texts are required.

D. Topic Distribution by Sentiment

- **Nemotron:**
  Distribution of topics of Nemotron depicted over representation in some topics under Positive sentiment, meaning that some themes were more strongly represented than in the original data. It might be that Nemotron tends to emphasize commonly associated themes for certain sentiments.

```
Sentiment: Irrelevant
Original Topic Distribution: [0.21050445 0.21889433 0.17753132 0.16373255 0.22933735]
Synthetic Topic Distribution: [0.52896942 0.10529785 0.2079666  0.07456126 0.08320486]

Sentiment: Negative
Original Topic Distribution: [0.17567913 0.22236064 0.18356814 0.15635429 0.2620378 ]
Synthetic Topic Distribution: [0.32174657 0.13275565 0.15753461 0.17079558 0.21716759]

Sentiment: Neutral
Original Topic Distribution: [0.19561342 0.13564336 0.21237729 0.16522096 0.29114498]
Synthetic Topic Distribution: [0.10959013 0.17228599 0.35674988 0.12687342 0.23450059]

Sentiment: Positive
Original Topic Distribution: [0.13973043 0.22231084 0.20010786 0.15142337 0.28642751]
Synthetic Topic Distribution: [0.21071654 0.16579257 0.25407994 0.11042376 0.25898719]
```

Fig. 12. Topic Consistency for Nemotron.

- **Llama:**
  Llama produced relatively balanced topic distributions across all categories of sentiments and was successful in maintaining the thematic structure found in the original data. This means Llama can build thematically consistent synthetic data, and it's very suitable for tasks that require close alignment to real-world thematic distributions.

```
Sentiment: Irrelevant
Original Topic Distribution: [0.3565141  0.1786207  0.20403984 0.21300903 0.04781633]
Synthetic Topic Distribution: [0.12239071 0.2002812  0.20501089 0.27614345 0.19617375]

Sentiment: Negative
Original Topic Distribution: [0.12169552 0.33359285 0.12797208 0.20553899 0.21120056]
Synthetic Topic Distribution: [0.28335702 0.19844521 0.20051218 0.19824464 0.11944095]

Sentiment: Neutral
Original Topic Distribution: [0.18886666 0.12060077 0.12064597 0.2027136  0.36717301]
Synthetic Topic Distribution: [0.2024838  0.12098591 0.35576861 0.19889396 0.12186772]

Sentiment: Positive
Original Topic Distribution: [0.13622398 0.22522698 0.29807291 0.12899818 0.21147795]
Synthetic Topic Distribution: [0.12582722 0.20387084 0.20408298 0.3390149  0.12720406]
```

Fig. 13. Topic Consistency for Llama.

- **Mixtral:**
  Mixtral's topic distribution is relatively consistent, except in Neutral sentiment, which has some under-representation. This could impact applications where neutrality and subtlety are required.

```
Sentiment: Irrelevant
Original Topic Distribution: [0.3565141  0.1786207  0.20403984 0.21300903 0.04781633]
Synthetic Topic Distribution: [0.19644312 0.28371778 0.27860742 0.20425976 0.03697192]

Sentiment: Negative
Original Topic Distribution: [0.41189996 0.1363687  0.12936331 0.19290751 0.12946052]
Synthetic Topic Distribution: [0.12648381 0.22099299 0.30507317 0.30841867 0.03903136]

Sentiment: Neutral
Original Topic Distribution: [0.18886666 0.12060077 0.12064597 0.2027136  0.36717301]
Synthetic Topic Distribution: [0.28201121 0.1187045  0.0369311  0.11781248 0.44454071]

Sentiment: Positive
Original Topic Distribution: [0.13622398 0.22522698 0.29807291 0.12899818 0.21147795]
Synthetic Topic Distribution: [0.196517   0.27603939 0.04660732 0.20406283 0.27677345]
```

Fig. 14. Topic Consistency for Mixtral

Interpretation: Llama's topic distribution is well-balanced, making it suitable for applications where consistency in themes is important. Nemotron and Mixtral have slight biases that would indicate a need for tuning if used in applications requiring precise thematic alignment.

E. Readability Analysis

- **Nemotron:**
  Nemotron synthesized text scored a **Flesch Reading Ease score of 48**, meaning this text was highly complex and suitable only for educated readers.

sentiment	flesch_reading_ease_original	flesch_reading_ease_synthetic
Irrelevant	80.547248	47.823624
Negative	78.750811	48.139865
Neutral	65.455493	47.550352
Positive	83.781329	48.325105

Fig. 15. Readability Analysis for Nemotron.

- **Llama:**
  Llama synthesized text scored a **Flesch score of 55**, meaning that the text is moderately readable and accessible to most education-based readers.

sentiment	flesch_reading_ease_original	flesch_reading_ease_synthetic
Irrelevant	70.560	78.012
Negative	65.778	65.627
Neutral	67.858	71.549
Positive	95.657	69.383

Fig. 16. Readability Analysis for Llama.

- **Mixtral:**
  Mixtral synthesized text scored the highest readability with a **Flesch score of 62**, making it accessible to a general readership..

Interpretation: Mixtral is high-readability and, therefore, best suited for applications where readable language would be required. With respect to the denser text, Nemotron looks suitable , otherwise, Llama's looks just right.

sentiment	flesch_reading_ease_original	flesch_reading_ease_synthetic
Irrelevant	70.560	68.823000
Negative	63.340	54.785556
Neutral	67.858	69.145000
Positive	95.657	72.355000

Fig. 17. Readability Analysis for Mixtral

F. Overall Quality Assessment

In summary:

- **Nemotron** excels in uniqueness and novelty but has high complexity.
- **Llama** demonstrates balanced performance across all metrics, making it the most versatile model.
- **Mixtral** provides superior readability, ideal for accessible applications, though it sometimes exaggerates sentiment strength.

0 = Irrelevant; 1 =Negative; 2= Neutral; 3=Positive

Models	Nemotron	Llama	Mixtral
Perplexity	0 = 61.522818 1 = 69.136215 2 = 66.607938 3 = 53.216430	0 = 90.143864 1 = 58.703008 2 = 117.940827 3 = 33.712287	0 = 48.908300 1 = 99.436890 2 = 49.691963 3 = 19.332150
Lexical Diversity	0 = 0.872650 1 = 0.915617 2 = 0.922813 3 = 0.921127	0 = 0.954778 1 = 0.975301 2 = 0.981607 3 = 0.988235	0 = 0.962557 1 = 0.974888 2 = 0.897646 3 = 1.000000
Embedding Similarity	0 = 0.311326 1 = 0.613633 2 = 0.706701 3 = 0.545860	0 = 0.186310 1 = 0.196073 2 = 0.099976 3 = 0.184331	0 = 0.146035 1 = 0.202392 2 = 0.142661 3 = 0.175746
VADAR Sentiment Analysis	0 = 0.585145 1 = -0.480591 2 = 0.471904 3 = 0.925520	0 = 0.18370 1 = -0.37350 2 = -0.19392 3 = 0.67157	0 = -0.044810 1 = -0.368089 2 = -0.096330 3 = 0.675080
N Gram Analysis	15.282331511839708	2.3126463700234194	2.273389682308365
Readability Scores	0 = 47.823624 1 = 48.139865 2 = 47.550352 3 = 48.325105	0 = 78.012 1 = 65.627 2 = 71.549 3 = 69.383	0 = 68.823000 1 = 54.785556 2 = 69.145000 3 = 72.355000

Fig. 18. Overall Testing Metrices for all three models

V. Inference

To test how the synthetic data improves a model prediction, we selected a specific Deep Learning model to leverage synthetic data generated from different models for improving multi-class sentiment classification. We would our baseline on the original dataset, which contains only a shallow coverage of sentiments, and augment with synthetic data generated by Nemotron, Mixtral, and Llama for our test and analysis. We evaluated individually how each synthetic dataset affects model performance using Accuracy and F1-score as the performance metrics. Our model architecture includes dense layers optimized for text classification to learn subtle sentiment classes in the multi-class setting. Synthetically prepared data were designed to improve the representations of sentiment in the dataset, thereby increasing variation and hence a generalization capability for the model.

The outcomes disclosed that Nemotron's synthetic data had the maximum potential in improvement, with **27.4** percent accuracy and F1-score of **0.183** as compared to the baseline at **12.5** percent accuracy and **0.132** F1-score. Nemotron's data would work effectively with the required sentiment patterns, improving the accuracy of the model, and thus is the most impactful source for synthetic data generation. While Mixtral and Llama also contributed to the models accuracy, they achieved slightly lower performance compared to Nemotron. The combined synthetic dataset yielded a better accuracy of **28.9** percent, although the F1-score remained stable. This suggested a growing increase in variability and which may impact precision and recall in general. In general, Nemotron's data is quite effective.

Baseline mode is deliberately kept low in terms of accuracy without any kind of fine tuning so that the impact of synthetic data can very well be reflected over any model. The process is shown by increasing model accuracy with the addition of synthetic data to the training dataset of the model. Synthetic data can even be created specifically for a given problem, which will enhance better prediction with any model. For our future work, we can focus on a specific domain area where the data is in itself very sparse. The synthetic data generated can be used to fine-tune LLMs and models to help further in that domain. We can fine-tune the model and improve its accuracy according to our requirement.

	Accuracy	F1 Score
Original	0.125000	0.131720
Original + Nemotron	0.273885	0.182804
Original + Mixtral	0.208333	0.180098
Original + Llama	0.229167	0.228553
All Combined	0.289017	0.182218

Fig. 19. Accuracy and F1 scores of synthetic data and original data combinations.

Conclusion

This paper provides the comparison considerations in the generation of synthetic sentiment-labeled data for the models Nemotron, Llama, and Mixtral. Nemotron shows high diversity in its data, Llama has balanced performance, and Mixtral has high readability. Therefore, every model differs significantly in certain aspects and will be favorable for other applications in different usages. Future work can be in the adjustment of prompts and configurations to optimize the performance of a certain model with specific applications to the sentiment analysis. This also reveals the impact of synthetic data to a model's performance. When synthetic data is added to the training data of a model, it results in the improvement of model accuracy. We can create much more datasets when there is a scarcity of dataset for a particular problem.

REFERENCES

[1] Shivashankar, C., and Miller, S. (2023). Semantic data augmentation with generative models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp.863-873).

[2] Brown, T. B. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

[3] Kenton, J. D. M. W. C., and Toutanova, L. K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of naacL-HLT (Vol.1,p.2).

[4] Vaswani, A. (2017). Attention is all you need. Advances in Neural Information Processing Systems.

[5] Hutto, C., and Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the international AAAI conference on web and social media (Vol. 8, No. 1,pp.216-225).

[6] Clark, K. (2019). What Does Bert Look At? An Analysis of Bert's Attention. arXiv preprint arXiv:1906.04341.

[7] Maqsud, U. (2015, September). Synthetic text generation for sentiment analysis. In Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis(pp.156-161).

[8] Colla, D., Delsanto, M., Agosto, M., Vitiello, B., and Radicioni, D. P. (2022). Semantic coherence markers: The contribution of perplexity metrics. Artificial Intelligence in Medicine,134,102393.

[9] Kyle, K., Crossley, S. A., and Jarvis, S. (2021). Assessing the validity of lexical diversity indices using direct judgements. Language Assessment Quarterly,18(2),154-170.

[10] Giabelli, A., Malandri, L., Mercorio, F., Mezzanzanica, M., and Nobani, N. (2022). Embeddings evaluation using a novel measure of semantic similarity. Cognitive Computation,14(2),749-763.

[11] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., and Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia tools and applications, 78,15169-15211.

[12] Subba, B., and Gupta, P. (2021). A tfidfvectorizer and singular value decomposition based host intrusion detection system framework for detecting anomalous system processes. Computers and Security,100,102084.