Enhancing Text Classification Models with Generative AI-aided Data Augmentation

Huanhuan Zhao
Data Science and Engineering
The University of Tennessee
Knoxville, TN, USA
hzhao31@vols.utk.edu

Haihua Chen
Department of Information Science
The University of North Texas
Denton, TX,USA
haihua.chen@unt.edu

Hong-Jun Yoon
Computational Sciences and Engineering Division
Oak Ridge National Laboratory
Oak Ridge, TN, USA
yoonh@ornl.gov

2023 IEEE International Conference On Artificial Intelligence Testing (AITest) | 979-8-3503-3629-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/AITest58265.2023.00030

***Abstract*—This study investigated the potential of enhancing the performance of text classification by augmenting the training dataset with external knowledge samples generated by a generative AI, specifically ChatGPT. The study conducted experiments on three models - CNN, HiSAN, and BERT - using the Reuters dataset. First, the study evaluated the effectiveness of incorporating ChatGPT-generated samples and then analyzed the impact of various factors such as sample size, sample variability, and sample length on the models' performance by varying the number, variety, and length of the generated samples. The models were assessed using macro and micro-averaged scores, and the results revealed that the macro-averaged scores improved significantly across all three models, with the BERT model showing the greatest improvement (from 49.87% to 65.73% in macro F1 score). The study further found that adding 30 distinct samples produced better results than adding 6 duplicates of 5 samples, and samples with 150 and 256 words had similar performance, while those with 50 words performed slightly worse. These findings suggest that incorporating external knowledge samples generated by a generative AI is an effective approach to enhance text classification models' performance. The study also highlights that the variability of articles generated by ChatGPT positively impacted the models' accuracy, and longer synthesized texts convey more comprehensive information on the subjects, leading to higher classification accuracy scores. Additionally, we conducted a comparison between our results and those obtained from EDA, a widely used data augmentation generator. The findings clearly demonstrate that our method surpasses EDA and offers additional advantages by reducing computational costs and solving zero-shot problem. Our code is available on GitHub.[1]**

***Index Terms*—text classification, data augmentation, ChatGPT, imbalanced data, natural language processing, machine learning, artificial intelligence**

I. Introduction

The classification of natural language texts is a highly researched topic in the fields of artificial intelligence (AI) and machine learning (ML). Since the emergence of deep learning, various applications such as automatic data collection, filtering, and curation have been significantly improved. Recent developments in self-attention mechanism-based language models have made significant strides and have had a profound impact on our daily lives. In essence, ML models rely heavily on the corpus of training data. Thus, their ability to produce accurate inferences is limited to the knowledge included in the training data. When exposed to data samples covering topics not found in the training data (referred to as out-of-distribution samples), these models are unable to generate reliable classifications.

The incorporation of external knowledge into the training of ML models has been extensively researched, with numerous studies and efforts made thus far [1]–[5]. One of the earliest attempts in natural language processing (NLP) models is the application of pre-trained word embeddings [6]. This approach involves training a word embedding matrix using publicly available text data corpus and/or text data of domain knowledge (e.g., PubMed [7] data corpus for comprehending medical texts). One of the key advantages of this approach is that it enables rich vocabulary coverage, which may increase the credibility and effectiveness of NLP models. Additionally, this approach is easily applicable to the conventional deep learning models such as convolutional neural network (CNN) [8] and hierarchical self-attention network (HiSAN) [9] models, making it a promising tool for improving the performance of these models. One limitation of this approach is that the knowledge captured by the word embedding layer may not necessarily propagate to the final decision layer. Although the external knowledge is present in the latent representation of the embedding layer, the identification of keywords and key phrases that drive the inference depends solely on the training data corpus. In addition, during the training process, the word vectors of vocabularies present in the training dataset may undergo alterations. However, the word vectors of vocabularies that are not present in the training dataset remain unchanged, leading to a potential disparity in the latent representation of word embeddings. While it is possible to leave the pre-trained word embedding matrix non-trainable, doing so may result in a degradation of classification accuracy in the models.

A recent study [10] investigates the development of an algorithm that integrates external data augmentation to train a text comprehension model for information extraction from

This manuscript has been authored in part by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

[1]https://github.com/HuanhuanZhao08/AI-data-augmentation

979-8-3503-3629-0/23/$31.00 ©2023 IEEE
DOI 10.1109/AITest58265.2023.00030

cancer pathology reports. The study created an augmented dataset for the training corpus where the input consisted of the definition of the International Classification of Diseases for Oncology, 3rd edition (ICD-O-3) codes [11] from the National Cancer Institute (NCI) thesaurus [12] and the Unified Medical Language System (UMLS) [13], and the target was the ICD-O-3 site and histology codes. The study shows that the approach is effective in addressing the under-represented class labels, which are rare cancers in this context, as demonstrated by a substantial improvement in accuracy scores in macro F1 scores. This study is critical because increasing the sample size of rare cancer cases is physically limited, and the study suggests a cost-effective solution to boost classification performance. However, it should be noted that one caveat of this approach is that it requires manual labor to curate an augmented dataset that includes exploration of external knowledge bases, collecting relevant content from the knowledge bases, and composing them into the form of a training dataset.

This study examines the concept of data augmentation through domain knowledge for natural language processing. Rather than manually curating external knowledge datasets, we utilized generative AI models that were trained on millions of publicly available natural language texts. Our approach involves using generative AI-synthesized sentences as augmented data samples. Generative AI models, such as generative pre-trained transformer (GPT) model [14], are trained on various topics, making them suitable for covering most general topics. To test the feasibility of this concept, we augmented the datasets to enhance the classification accuracy scores for identifying topics of the Reuters dataset. Specifically, we generated texts related to the topics covered by the Reuters dataset and then added them to the training procedure.

To check the effectiveness of our methods, we compared our results with those obtained from easy data augmentation (EDA) data generator, an automated data augmentation technique that generates additional samples by modifying the original text data [15]. EDA utilizes various techniques such as synonym replacement, random insertion, random swap, and random deletion. EDA is widely recognized as an automated data augmentation generator that can effectively enhance datasets.

The primary contribution of this study is the development of an innovative approach that leverages the latest advancements in generative AI models to augment external knowledge for enhancing text classification model development. This approach has the potential to provide an automatic and cost-effective means of boosting task performance scores. To be more specific:

a. We demonstrated the effectiveness of generative AI-aided data augmentation by integrating ChatGPT-generated samples into three deep learning models.
b. We conducted comprehensive experiments to investigate the impact of the length and variability of the generated samples on the performance of the models.
c. We validated the superiority of our approach by comparing it with the existing method of EDA, showcasing the advancements achieved through our proposed methodology.

Section 2 provides a detailed overview of the approach, including data, classification models, and performance measurement. Comparative experimental results are presented in Section 3, and Section 4 discusses the findings and potential for further improvement.

II. Methods

We conducted an experimental study to investigate the efficacy of training ML models with augmented datasets of external knowledge samples generated by ChatGPT [16]. This section provides details about the ML models we tested and the experimental design employed in the study.

A. Dataset

For the study, we utilized the Reuters corpus provided by the Natural Language Tool Kit (NLTK) [17] Python library. The corpus consisted of 10,788 news articles, with a total of 1.3 million words. The corpus contains pre-defined ”training” and ”test” sets with 7,769 and 3,019 cases, respectively. Note that we randomly held out 10% of the training samples for validation of the model training. Each news article belongs to one or more of the 90 pre-defined categories, forming multi-class labels. The corpus provides a multi-labeled dataset for text classification tasks. Each article is labeled with a number from one to fifteen. However, this is a long-tail imbalanced dataset, as the number of samples (articles) for each label (topic) varies greatly, ranging from 1 to 2877. Figure 1 displays the count of samples for each label through a bar plot. The number of words in each article ranges from 2 to 1316, with an average of 130 words per article. The distribution of the articles' lengths is shown in Figure 2.

B. Machine learning models for text comprehension and classification

Given that the dataset is labeled as multi-class, it is imperative for the model to be designed to enable multiple choices of labels. The final decision layer's output nodes should be equipped with a sigmoid activation function that applies binary cross-entropy for optimization in the back-propagation. To implement the proposed approach of augmenting external knowledge for natural language text classification, we utilized the following three widely-adopted ML models. The ML models have been implemented using the PyTorch [18] platform on Python 3.10.

1) Convolutional neural network model: The text classification model based on CNN [8] consists of three parts: a word embedding layer, a one-dimensional convolution layer, and a fully-connected decision layer. The word embedding layer learns a representation of terms by mapping a set of words to numerical vectors. Within the vector space, proximity indicates the similarity of semantic meaning within the context of a text corpus. The convolution layer employs a set of one-dimensional convolution filters to capture the features of the

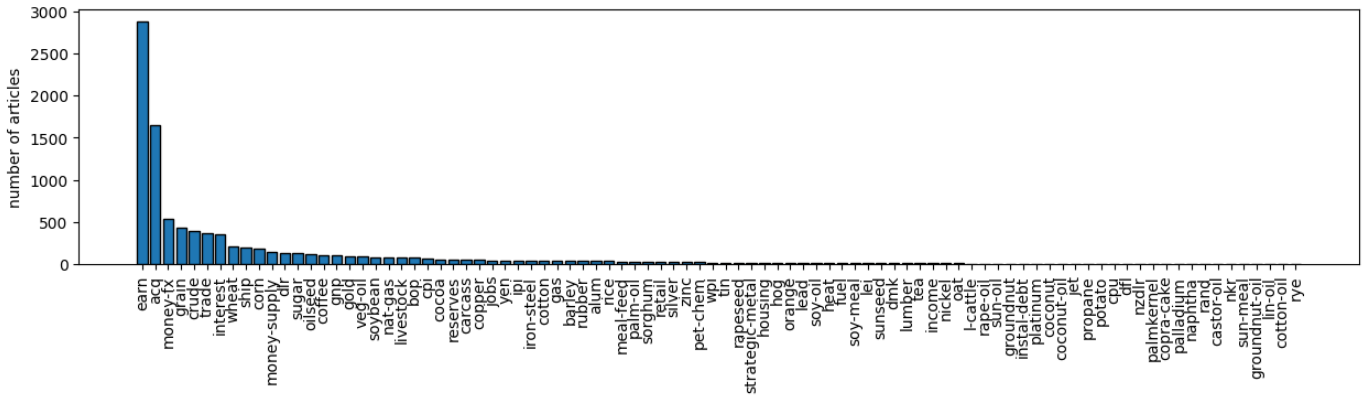

Fig. 1. The number of articles for each topic in the Reuters corpus, illustrates that the dataset is severely imbalanced.

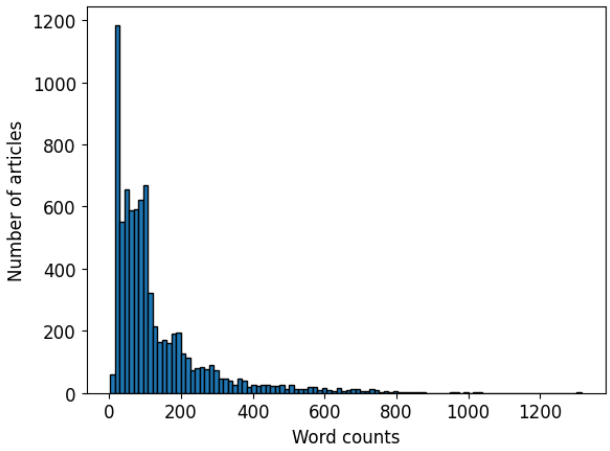

Fig. 2. The article length in the Reuters corpus.

word vectors. The decision layer collects the captured features and makes inferences.

2) Hierarchical self-attention network model: The HiSAN [9] model uses a self-attention mechanism to address the challenge of long-term dependencies between words in natural language texts. This model analyzes document data in a hierarchical manner. At the lower level, words are treated as composing lines, and at the upper level, lines are treated as composing the document. The self-attention mechanism compares a sequence of embeddings with itself to identify relationships between components of the sequence.

3) BERT model: Bidirectional encoder representations from transformers (BERT) [19] is currently the most successful ML model for natural language processing. It has achieved superb classification accuracy scores across many applications. BERT applies multiple layers of self-attention mechanism to identify keywords that characterize documents at scale. We apply a fully-connected layer at the top to make final inferences. For our study, we utilized the pre-trained `bert_base_uncased` model from the HuggingFace [20] library, which is widely recognized for its high performance in natural language processing tasks.

C. Augmenting external data to the text classification

Incorporating external knowledge into ML model training is a crucial aspect of design. Previous studies have been limited by utilizing truth label descriptions as augmented input, which restricts the content of the augmented dataset to the knowledge source. As a result, there is a lack of variability of expression and a limited vocabulary regarding the dataset. To address this issue, this paper proposes a solution that uses newly introduced generative AI products to synthesize the infused dataset. The rationale is that such AI products, e.g., ChatGPT, have been trained on millions of contents and are expected to include relevant knowledge in their training corpus. Thus, ChatGPT can synthesize text with relevant knowledge and background. The proposed approach is expected to significantly increase classification accuracy by incorporating variability of expressions and vocabularies about the target labels.

1) Obtain external knowledge from ChatGPT and EDA: To obtain external knowledge datasets from ChatGPT, we utilized the ChatGPT API (GPT3.5) by submitting queries in the format: ”write an article with N words about LABEL in Reuters news format.” Here, LABEL represents the topic for which we aimed to create data, and N represents a designated word count. We applied three specific word counts (50, 150, and 256) in our experiment.Automated data generation was implemented through Python.

According to the EDA paper, the recommended number of augmented samples depends on the size of the original sample. In our experiment with the Reuters dataset, we generated four additional samples for each original sample, resulting in a total of 31,076 samples (7,769 multiplied by 4). To ensure consistency, we applied the same text data pre-processing pipeline, which included tokenization and vectorization, as used in the training corpus from the Reuters dataset.

2) Integrate the external knowledge to the model: In order to incorporate external knowledge into our model, we added an external knowledge training loop after each batch of the

original data. During a given training update within each epoch, the following steps are taken:

a. The binary cross-entropy loss is calculated from the given minibatch of training samples and backpropagation is performed.
b. A minibatch of augmentation samples is randomly selected, the binary cross-entropy loss is calculated, and backpropagation is performed.

D. Experimental design

Our approach to integrating external knowledge into the training procedure involves augmenting auxiliary data samples generated from that knowledge. We introduce these augmented samples during each training batch to expose the ML models to the external knowledge. The following are four scientific questions related to this strategy.

1) Quantification of performance improvement: The central inquiry of this study is to determine if augmented external knowledge data samples synthesized by ChatGPT models can enhance performance. Specifically, we anticipate an improvement in accuracy scores for under-represented class labels. Our ML models have already acquired sufficient information from the rich training dataset to achieve good performance scores for prevalent classes. However, the lack of training data samples for minor classes impedes the model's ability to acquire adequate knowledge about these class labels. We designed experiments using three different ML models - CNN, HiSAN, and BERT - to validate the effectiveness of our proposed approach, regardless of the model architecture.

2) Sample size vs. variability of expression: As previously discussed, the Reuters dataset is afflicted with severe class imbalance. This imbalance results in a decrease in classification accuracy for the under-represented class labels, as half of the topics have fewer than 20 samples for training. To address this issue, augmenting synthesized data samples with external knowledge has been proposed as a potential solution to boost performance of the minority labels. However, it is important to consider the possibility that the observed improvements may be attributable to a sample size issue rather than the efficacy of the external knowledge. Specifically, one could argue that the repetition of a few synthesized data samples may yield the same improvements as introducing fresh synthesized samples. To better understand this issue, a key question is how much variability can be expected when repeating six sets of five synthesized samples versus when introducing 30 new synthesized samples. The primary question is whether the ChatGPT-generated texts can effectively articulate subjects and topics through various types of expressions and perspectives. If this is the case, then the ML models exposed to such variability may exhibit greater robustness and make more accurate decisions about class labels.

3) Optimal article length: Suggesting longer synthesized documents can potentially enhance specificity and enrich the vocabulary of the subject. However, specifying an extended maximum document length to ChatGPT may result in higher computational time and resource consumption. In the Reuters dataset, articles range from 2 to 1316 words in length. However, the majority of articles (87%) contain less than 250 words, and 73% are less than 150 words in length. The average article length is 130 words. Based on the statistics, we decided to set the article length to 256 words for our experiment. We also conducted tests with maximum lengths of 50 and 150 words to evaluate the potential negative effects.

4) Comparison with EDA data generator: In this study, we have conducted a comparative analysis of augmentation techniques using EDA and ChatGPT with the following two text classification models, namely HiSAN and BERT. We evaluated the performance of these models using samples generated by ChatGPT (90 labels, each with 20 samples of 256 words) and EDA-generated data. The results of the comparison are presented in Tables IV and V.

E. Performance measure

Due to the severe class imbalance and multi-label annotations present in our data corpus, it is necessary to calculate both macro- and micro-averaged F1 metrics using a class-wise multi-label confusion matrix. In this context, macro-averaged F1 scores are equally weighted among the class labels, while micro-averaged F1 scores are equally weighted among individual decisions. To calculate these scores, we use the Scikit-Learn [21] Python library.

For each class label i, we obtain a_i, b_i, c_i, and d_i, where a stands for true positives, b represents true negatives, c represents false negatives, and d represents false positives.
To calculate the macro-averaged precision, recall, and F1 scores, we computed those scores for each label separately, and then took their average over all the labels. In contrast, to compute the micro-averaged precision, recall, and F1 scores, we aggregated the a_i, b_i, c_i, and d_i across all labels and computed the corresponding overall scores. Equations (1) and (2) illustrate the method of calculating macro- and micro-averaged scores.

$$p_{macro} = \frac{\sum_{i=1}^{N} p_i}{N} \quad p_i = \frac{a_i}{a_i + c_i} \quad (i = 1, ..., N) \tag{1}$$

$$p_{micro} = \frac{\sum_{i=1}^{N} a_i}{\sum_{i=1}^{N} a_i + \sum_{i=1}^{N} c_i} \quad (i = 1, ..., N) \tag{2}$$

where N is the total number of the labels. In our case, $N = 90$.

III. Results

We conducted a comparative analysis of three ML models, namely Convolutional Neural Network (CNN), Hierarchical Sampling Network (HiSAN), and Bidirectional Encoder Representations from Transformers (BERT), to classify data. To evaluate the models, we used macro- and micro-averaged metrics and augmented them with external knowledge samples synthesized by ChatGPT. We repeated the training process 10 times for each model and dataset, and then averaged the accuracy scores to mitigate the inherent variability in ML training. Table I summarizes our findings.
Our study has shown that augmenting external knowledge

leads to improved accuracy across both macro- and micro-averaged metrics in ML models. We found a statistically significant improvement in macro-averaged scores, while the improvement in micro-averaged scores was not statistically significant. Our results indicate that BERT models experienced the greatest improvement (with macro F1 scores increasing from 49.87 to 65.73) compared to the other two models. The enhancement was more pronounced in macro-averaged scores than in micro-averaged ones, suggesting that augmenting external knowledge significantly improves the accuracy of minor class labels. These findings validate the use of generative AI models to synthesize texts and augment external knowledge as an effective approach to boost the performance of classification models. Furthermore, our study demonstrates that significant improvements in macro F1 scores can be achieved without sacrificing micro F1 scores, indicating that the proposed method is highly effective and does not compromise overall performance. This suggests that this same concept can be applied to other applications of natural language text classification and information extraction tasks.

In the analysis of three different model architectures, the BERT model exhibited significantly higher accuracy scores than the HiSAN and CNN models. Furthermore, the HiSAN models performed significantly better than the CNN models. These results indicate that self-attention-based text classification models are more effective for natural language text comprehension. It is important to note that the BERT model employed a pre-trained version, whereas the other two models were trained from scratch. The utilization of pre-trained models may have resulted in a higher likelihood of exposure to the topics of the queries in the Reuters dataset, which could have led to the observed increase in accuracy scores.

Table II presents the results of the training of HiSAN models with various augmentations, including the one, three, and six repetitions of five synthesized samples, and thirty newly synthesized samples. The findings indicate that the use of thirty newly synthesized samples resulted in a significant improvement in performance compared to using six repetitions of five synthesized samples. Therefore, the issue at hand is not solely a matter of sample size, which can be resolved by oversampling the training dataset. Furthermore, the accuracy scores for the one, three, and six repetitions of samples did not differ significantly. Additionally, the generative AI synthesized articles possess rich information, which positively influences the target ML models.

Table III presents the classification accuracy scores from the HiSAN models that were trained with 10, 20, and 30 synthesized samples, but limited the article length to 50 words. The results reveal higher accuracy scores than those without external knowledge samples (45.69 for macro F1 and 85.90 for micro F1 from Table I); however, they are lower than the scores with 30 synthesized external knowledge samples with sample length equal to 150 (54.95 for macro F1 from Table III and 56.00 for macro F1 from Table II). This implies that longer synthesized text articles could convey richer information about the subjects, resulting in higher classification accuracy scores. Interestingly, the accuracy scores for 10, 20, and 30 synthesized 50-word-limit samples do not differ significantly.

Table IV and table V present the average performance scores of HiSAN and BERT models trained on two different augmented datasets: EDA and 20-samples ChatGPT dataset. It is evident that the inclusion of EDA data led to enhancements in the models' performance, although these improvements were not as substantial as those achieved with the data generated by ChatGPT. Specifically, for the macro F1 score, the HiSAN model improved by 3.2% with EDA data, while ChatGPT yielded a superior improvement of 9.51%. Similarly, the BERT model improved by 11.53% with EDA data, while the ChatGPT data showcased a more remarkable improvement of 15.86%. Furthermore, it is noteworthy that the BERT model required more epochs to converge when trained on EDA data (30-35 epochs) compared to the ChatGPT data (13-16 epochs). Both EDA and generative AI-aided data augmentation serve as automated data generators used to augment external knowledge to the training process. However, EDA, along with other data augmentation methods [22], [23], heavily depends on the original datasets. This dependence restricts its ability to generate samples for labels that are absent in the original data. In contrast, generative AI solely relies on labels and is not hindered by this limitation. The capability to handle zero-shot labels, combined with the diverse nature of the generated data, enables our approach to surpass the performance of EDA.

IV. Discussion

Generative AI models have gained popularity since the debut of ChatGPT. With their large number of trainable parameters, pre-training with a substantial amount of articles and documents, they achieved noteworthy performance in chatting, question answering, and information retrieval. Early adoption studies have already shown remarkable results, making it clear that these models have great potential for various NLP tasks. However, it is still too early to expect that GPT models can solve complex real-world problems independently. Nonetheless, with proper guidance, we can leverage the vast amounts of information they provide to enhance various NLP tasks.

This paper proposes an application of ChatGPT as an external knowledge data source to enhance the accuracy of natural language text classification, which is a prime topic of NLP research. Our results suggest that emerging generative AI models could be a valuable source of external knowledge. In particular, our method achieved superb macro-averaged scores, surpasses the performance of other common used text data augmentation technologies such as EDA, demonstrating that it is highly effective in improving under-represented class labels. Our approach could be particularly useful in areas that suffer from severe class imbalance issues, such as clinical and health-related document classification and information extraction.

Generative AI-aided data augmentation is easy to implement and provides more flexibility than traditional methods.

Model		Macro (Unit:%)			Micro (Unit:%)		
		Precision	Recall	F1 score	Precision	Recall	F1 score
CNN	without EK	33.39 (30.13, 36.64)	28.96 (27.30, 30.61)	**30.05** (28.15, 31.95)	81.92 (78.64, 85.21)	77.02 (75.99, 78.06)	79.32 (77.71, 80.94)
	with EK	36.98 (35.10, 38.87)	39.6 (38.15, 41.05)	**36.25** (35.15, 37.36)	80.41 (78.79, 82.05)	77.76 (77.10, 78.41)	79.04 (78.32, 79.76)
HiSAN	without EK	56.86 (54.99, 58.72)	41.15 (40.07, 42.23)	**45.69** (44.68, 46.69)	90.49 (89.83, 91.15)	81.76 (81.17, 82.34)	85.90 (85.67, 86.12)
	with EK	67.69 (65.76, 69.61)	50.17 (48.28, 52.05)	**55.20** (53.56, 56.85)	91.22 (90.67,91.78)	81.91 (81.33, 82.49)	86.31 (86.05, 86.57)
BERT	without EK	57.17 (53.82, 60.53)	47.25 (44.02, 50.48)	**49.87** (46.70, 53.03)	91.30 (90.84, 91.75)	87.69 (86.80, 88.58)	89.45 (89.15, 89.75)
	with EK	75.23 (74.00, 76.46)	61.44 (59.65, 63.23)	**65.73** (64.46, 66.99)	92.50 (91.80, 92.62)	87.90 (87.74, 89.06)	90.13 (90.07, 90.45)

TABLE I
MODELS PERFORMANCE WITH AND WITHOUT EXTERNAL KNOWLEDGE AUGMENT

The table above shows the mean scores and 95% confidence intervals of macro precision, recall, f1, micro precision, recall and f1 for each model respectively, with and without external knowledge. The augmentation data contains 20 samples for each label, and each sample contains 256 words. The abbreviations EK represents External Knowledge.

	Macro (Unit:%)			Micro (Unit:%)		
	Precision	Recall	F1 score	Precision	Recall	F1 score
5 samples	65.65 (64.66, 66.63)	47.25 (46.39, 48.10)	52.52 (51.80, 53.25)	90.63 (90.17, 91.10)	82.19 (81.45, 82.93)	86.20 (85.82, 86.58)
5×3 samples	67.12 (66.26, 70.52)	48.10 (48.17, 51.67)	53.80 (54.00, 56.87)	91.32 (90.73, 92.05)	82.00 (81.53, 83.27)	86.40 (86.32, 86.98)
5×6 samples	65.12 (63.32, 66.92)	47.47 (46.02, 48.93)	52.56 (51.32, 53.79)	90.58 (89.70, 91.46)	82.21 (81.56, 82.86)	86.18 (85.79, 86.57)
30 different samples	69.59 (68.50, 70.68)	50.64 (49.23, 52.06)	**56.00** (55.05, 56.96)	91.92 (90.21, 92.17)	81.69 (80.84, 82.55)	86.16 (85.95, 86.38)

TABLE II
HISAN PERFORMANCE WITH REPEATED AND DISTINCT SAMPLES

The table above shows the HiSAN models' performance with repeated and distinct samples. The first row shows the result of adding five samples for each label. The second row shows the result of duplicate the first dataset three times and the third row shows the result of duplicate the dataset six times. Row four shows the result of adding 30 distinct samples. Each sample with 150 words.

Researchers can specify the number of samples required for specific labels, allowing for more targeted augmentation. Moreover, it overcomes the limitations of relying solely on the original text dataset, making it effective in handling zero-shot problems.

However, it is important to note that when interacting with generative AI models, researchers must carefully formulate their query sentences. The models are highly sensitive to input questions, and the quality of generated samples depends on the quality of the queries. For instance, when augmenting the Reuters dataset, a query like "write the definition of [label]" may not provide as rich information as a query like "write an article about [label] in Reuters news format." Therefore, researchers should invest effort in crafting precise and informative queries to obtain desirable augmented samples from the generative AI models.

Despite their success, generative AI models have raised issues regarding their limitations. For example, in some cases, GPT models can provide misleading information on certain topics. This raises questions about how we can ensure that the external knowledge is proper and sound and whether GPT models may negatively affect our underlying tasks. These concerns point to the need for further research.

V. CONCLUSIONS

Our study has demonstrated the substantial performance improvement achievable in text data classification through the application of generative AI-aided augmentation across three deep learning models. The automated and easy implementation of this approach presents a promising avenue for enhancing text classification tasks. Future research should focus on

	Macro (Unit:%)			Micro (Unit:%)		
	Precision	Recall	F1 score	Precision	Recall	F1 score
10 samples	66.56 (65.10, 68.01)	47.75 (46.00, 49.50)	53.11 (51.71, 54.50)	90.94 (90.41, 91.47)	81.70 (80.87, 82.54)	86.07 (85.75, 86.38)
20 samples	68.44 (67.34, 69.54)	50.65 (48.56, 52.74)	55.69 (54.18, 57.20)	90.47 (89.65, 91.30)	82.54 (81.51, 83.57)	86.30 (85.99, 86.63)
30 samples	68.60 (67.33, 69.89)	49.29 (47.56, 51.02)	54.95 (53.60, 56.30)	91.26 (90.46, 92.06)	81.87 (81.13, 82.61)	86.3 (86.06, 86.54)

TABLE III
HISAN PERFORMANCE WITH SAMPLES' LENGTH EQUAL TO 50

The table above shows the mean scores of macro precision, recall, f1 and micro precision, recall, f1 of running the HiSAN model with samples' length equal to 50. The first row shows the result of adding 10 distinct samples for each label. The second row shows the result of 20 distinct samples and the third row shows the result of 30 distinct samples. Each sample with 50 words.

	Macro (Unit:%)			Micro (Unit:%)		
	Precision	Recall	F1 score	Precision	Recall	F1 score
without EK	56.86 (54.99, 58.72)	41.15 (40.07, 42.23)	**45.69** (44.68, 46.69)	90.49 (89.83, 91.15)	81.76 (81.17, 82.34)	85.90 (85.67, 86.12)
EDA	60.29 (58.62, 61.96)	44.24 (43.06, 45.41)	**48.89** (47.69, 50.07)	91.18 (90.58, 91.79)	82.44 (82.06, 82.82)	86.59 (86.34, 86.83)
ChatGPT	67.69 (65.76, 69.61)	50.17 (48.28, 52.05)	**55.20** (53.56, 56.85)	91.22 (90.67,91.78)	81.91 (81.33, 82.49)	86.31 (86.05, 86.57)

TABLE IV
HISAN PERFORMANCE WITH EDA AND CHATGPT DATA

The table above shows the mean scores of macro precision, recall, f1 and micro precision, recall, f1 of running the HiSAN model without external knowledge, with 31076 EDA samples, and with 90*20 ChatGPT samples with 256 words length.

	Macro (Unit:%)			Micro (Unit:%)		
	Precision	Recall	F1 score	Precision	Recall	F1 score
without EK	57.17 (53.82, 60.53)	47.25 (44.02, 50.48)	**49.87** (46.70, 53.03)	91.30 (90.84, 91.75)	87.69 (86.80, 88.58)	89.45 (89.15, 89.75)
EDA	68.04 (66.77, 69.30)	59.28 (58.01, 60.56)	**61.40** (60.35, 62.44)	90.78 (90.29, 91.26)	89.38 (88.97, 89.80)	90.07 (89.87, 90.27)
ChatGPT	75.23 (74.00, 76.46)	61.44 (59.65, 63.23)	**65.73** (64.46, 66.99)	92.50 (91.80, 92.62)	87.90 (87.74, 89.06)	90.13 (90.07, 90.45)

TABLE V
BERT PERFORMANCE WITH EDA AND CHATGPT DATA

The table above shows the mean scores of macro precision, recall, f1 and micro precision, recall, f1 of running the BERT model without external knowledge, with 31076 EDA samples, and with 90*20 ChatGPT samples with 256 words length.

expanding the investigation of our approach to different topics and datasets to assess its effectiveness in diverse domains. Such studies would provide valuable insights into the generalizability and adaptability of our method across various textual data sets. Overall, our findings highlight the potential of generative AI-augmented augmentation as a powerful technique for improving the performance of text data classification, opening doors for further advancements in natural language processing research.

REFERENCES

[1] Wei, J. and Zou, K. (2019) Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196,*.

[2] Akkaradamrongrat, S., Kachamas, P., and Sinthupinyo, S. (2019) Text generation for imbalanced text classification. In *2019 16th International Joint Conference on Computer Science and Software Engineering (JCSSE)* IEEE pp. 181–186.

[3] Hu, Z., Tan, B., Salakhutdinov, R. R., Mitchell, T. M., and Xing, E. P. (2019) Learning data manipulation for augmentation and weighting. *Advances in Neural Information Processing Systems,* **32**.

[4] Xu, B., Qiu, S., Zhang, J., Wang, Y., Shen, X., and de Melo, G. (2020) Data augmentation for multiclass utterance classification–a systematic

study. In *Proceedings of the 28th international conference on computational linguistics* pp. 5494–5506.

[5] Chen, H., Pieptea, L. F., and Ding, J. (2022) Construction and Evaluation of a High-Quality Corpus for Legal Intelligence Using Semiautomated Approaches. *IEEE Transactions on Reliability,* **71**(2), 657–673.

[6] Qi, Y., Sachan, D. S., Felix, M., Padmanabhan, S. J., and Neubig, G. (2018) When and why are pre-trained word embeddings useful for neural machine translation?. *arXiv preprint arXiv:1804.06323,*.

[7] Canese, K. and Weis, S. (2013) PubMed: the bibliographic database. *The NCBI handbook,* **2**(1).

[8] Kim, Y. (2014) Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882,*.

[9] Gao, S., Qiu, J. X., Alawad, M., Hinkle, J. D., Schaefferkoetter, N., Yoon, H.-J., Christian, B., Fearn, P. A., Penberthy, L., Wu, X.-C., et al. (2019) Classifying cancer pathology reports with hierarchical self-attention networks. *Artificial intelligence in medicine,* **101**, 101726.

[10] Blanchard, A. E., Gao, S., Yoon, H.-J., Christian, J. B., Durbin, E. B., Wu, X.-C., Stroup, A., Doherty, J., Schwartz, S. M., Wiggins, C., et al. (2022) A keyword-enhanced approach to handle class imbalance in clinical text classification. *IEEE journal of biomedical and health informatics,* **26**(6), 2796–2803.

[11] Fritz, A., Percy, C., Jack, A., Shanmugarathnam, K., Sobin, L., Parkin, D., and Whelan, S. (2000) International Classification of Diseases for Oncology 3rd edition WHO: Geneva. *Switzerland.[Google Scholar],*.

[12] Sioutos, N., de Coronado, S., Haber, M. W., Hartel, F. W., Shaiu, W.-L., and Wright, L. W. (2007) NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *Journal of biomedical informatics,* **40**(1), 30–43.

[13] Bodenreider, O. (2004) The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research,* **32**(suppl_1), D267–D270.

[14] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018) Improving language understanding by generative pre-training.

[15] Wei, J. and Zou, K. (November, 2019) EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* Hong Kong, China: Association for Computational Linguistics pp. 6382–6388.

[16] ChatGPT: Optimizing language models for dialogue. https://openai.com/blog/chatgpt/ Accessed: 2023-03-29.

[17] Loper, E. and Bird, S. (2002) Nltk: The natural language toolkit. *arXiv preprint cs/0205028,*.

[18] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019) Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems,* **32**.

[19] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805,*.

[20] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019) Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771,*.

[21] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research,* **12**, 2825–2830.

[22] Min, J., McCoy, R. T., Das, D., Pitler, E., and Linzen, T. (July, 2020) Syntactic Data Augmentation Increases Robustness to Inference Heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* Online: Association for Computational Linguistics pp. 2339–2352.

[23] Jin, D., Jin, Z., Zhou, J. T., and Szolovits, P. (4, 2020) Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. *Proceedings of the ... AAAI Conference on Artificial Intelligence,* **34**(05), 8018–8025.