Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 235 (2024) 102–111

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Machine Learning and Data Engineering (ICMLDE 2023)

Evaluating the Impact of Text Data Augmentation on Text Classification Tasks using DistilBERT

Aarathi Rajagopalan Nair[a], Rimjhim Padam Singh[a,*] , Deepa Gupta[a], Priyanka Kumar[b]

[a]*Department of Computer Science and Engineering, Amrita School of Computing, Bengaluru, Amrita Vishwa Vidyapeetham, India*
[b]*Department of Computer Science, University of Texas at San Antonio, Texas, USA.*

Abstract

Data augmentation entails artificially expanding the dataset's size by applying various transformations to the existing raw data. Enhancing the quality and quantity of the datasets with varying sizes by employing varieddata augmentation techniques has immense importance in the field on Natural Language Processing. Several notable applications for instance text classification, sentiment analysis, text summarization, etc. have proven to be benefitted immensely with the employment of text augmentation techniques. Hence, the paper focuses on efficient text classification using varied datasets of different sizes; small- 500 instances, medium-5564 instances and large-43934 instances.The work considers the standard DistilBERT model, a popular transformer-based language model and presents the impact on the performance of the modelafter employing different text augmentation techniques. The study specifically focuses on three augmentation methods: (a) Synonym augmentation:that involves replacing words with their synonyms to enhance vocabulary diversity and generalization, (b) Contextual word embeddings that enriches semantic understanding by leveraging pre-trained language models, and (c) Black translation that entails translating the text into a another different language and then translating it back, introducing variations in the data and capturing different linguistic patterns.Additionally,the work also discusses the combined effect of employing all three augmentation techniques simultaneously. Moreover, the study also aims compares the relation between the dataset sizes and the performance of the augmentation techniques. The study considers three standard datasets for the analysis and presents a comprehensive analysis using accuracy and F1 score as evaluation metrics. The results highlight the efficacy of each technique across small, medium, and large datasets, enabling a nuanced understanding of their benefits in different data scenarios. The findings indicate the varying degrees of improvement achieved through each augmentation technique.The enhancement achieved by applying text augmentation varied from around 2% on large datasets to 20% on smaller datasets.

*Keywords:*Natural Language Processing; DistilBERT; Text Data Augmentation; Synonym Augmentation; Contextual Word Embeddings; Black Translation

10.1016/j.procs.2024.04.013

1. Introduction

Text classification, a pivotal task in the field of Natural Language Processing (NLP), holds significant importance in structuring and understanding extensive amounts of text data [1]. As the amount of online content continues to grow exponentially, there is an increasing need to extract valuable insights from it, making text classification an active area of research and development [2]. This task involves automatically categorizing and labeling text documents into predefined classes or categories, enabling various applications such as analysis of sentiments, classification of text, detection of spam, and intent recognition [3]. There are numerous advantages of text classification in the research field but it still faces a lot of challenges due to lack of availability of the data that is labeled and rich in the particular domain thereby, leading to poor performance of the models. To solve this issue the text augmentation techniques, play a very significant role in improvising the size and the quality of the data used for training [4].

Data Augmentation is a very innovative technique that aids in increasing the size of the dataset by automatically creating new versions of the existing data for training rather than collecting the new unique data [5]. This helps in reducing the scarcity of the data as well as the absence of the data. Data Augmentation significantly plays a very important role in computer vision and in the present scenarios has been actively adapted in the text domain as well. It is comparatively easy to perform augmentation in the domain of computer vision but applying augmentations in NLP domain tedious to handle due to the intricacies embedded in the language. With the help of data augmentation, when the size of the training data increases the model performance also increases. The distribution of the generated augmented data should strike a balance between being neither excessively similar nor overly dissimilar to the original data. Striking a balance is crucial in effective data augmentation approaches as it helps to prevent issues like over-fitting and poor performance. Hence, the proposed work attempts to analyze the impact of different data augmentation techniques on the datasets of varying sizes. The main contributions of the workpresented in this paper are:

- To study the impact of various of data augmentation techniques on the performance of DistilBERT model for text classification.
- To study the variation in performance of DistilBERT after applying augmenting the data belonging to datasets of varying sizes (small, medium and large sized datasets). The work provides insights into the effectiveness of data augmentation techniques on model's performance with respect to the dataset size.
- To study the relationship between dataset size and the model performance in terms of accuracy.
- The findings contribute to a broader understanding of techniques for improving machine learning models and optimizing the utilization of datasets.

The rest of this document is structured as: Section 2 provides an overview of the current research on text augmentation in the field of text classification. Section 3 explains the approach used, including details on datasets, preprocessing methods, and augmentation techniques. Section 4 presents the experimental outcomes and analysis. Lastly, Section 5 summarizes the conclusions derived from this study and suggests potential avenues for future research.

2. Related Work

Classification of text is the indispensable method of Natural language processing. NLP tasks such as sentimental analysis, topic classification, etc. have significantly achieved higher accuracy. This higher performance generally depends on the size and quality of the data used for training. Scott and Matwin [6] introduced the method of text classification using rule-based methods. Classification of text is basically a method where they automatically use pre-defined labels such that a new unseen data can be grouped. Dalal et al. [7] introduced the semi-supervised machine learningmethod which assigns the label automatically into different classes. The paper illustrated by Kowsari et al. [8] showed how text classification has dominated the NLP domain using different methods. These methods of text classification include categorization of document, analysis of the sentiments and opinion identification [9][10].

Perez et al. [11] explained that the method of data augmentation produces higher accuracy in the case of classification of images. Then, Mosolova et al. [12] introduced augmentation techniques into text data. They implemented and showcased how augmentation techniques on text data can significantly increase size of the training data. Zou et al [13] used basic techniques such as synonym replacement, random insertion, random swap and random deletion to enhance the data and demonstrated its effectiveness on standard datasets. Using synonym replacement, other than the stopwords, it will choose 'n' words at random. Then these selected words will be replaced with its synonym chosen randomly. Similarly random words will be inserted at random position multiple times. Considering random swap, in a sentence, it will select two words and their positions will be interchanged. Lastly, a word will be randomly deleted in a sentence with some probability. This work showed that simple augmentation of data can boost the performance of the classification of the text.

Fadaee et al. [14] explained the method of back translation. From a particular language, the model will translate it into one language and then it will be translated back to the original language. The model demonstrated that back translation is a straightforward yet highly effective technique for data augmentation and is capable of significantly enhancing performance. Silfverberg et al. [15] showed us that, in the majority of instances, word forms can be divided into three components: a prefix that indicates inflection, a central stem, and a suffix that denotes inflection. In this context, the model must initially grasp the core of the word before substituting it with another term thereby, generating fresh training data. But in this paper, the author suggested a problem of over-fitting as the generated strings are too close to the training points present in the datasets. Kobayashi and Sosuke [16] later introduced another way of augmentation technique which is augmentation of the training data by understanding the context of the sentences. Here a bi-directional language model will be used such that it will try to understand the meaning of the sentence and a word will be chosen randomly which will be replaced and a new data point will be created for training. Shorten et al. [17] surveyed and compared the requirement of different data augmentation techniques in multiple tasks in NLP and mentioned that text data augmentation is not really appreciated when compare to computer vision but it is a strategy which is very promising.

Sanhet al. [18] introduced a technique for pre-training a more compact, versatile language model known as DistilBERT. This model can subsequently be fine-tuned to achieve impressive performance on various tasks, similar to its larger counterparts. Barbon and Akabane [19] displayed the significance of DistilBERT method that showed significant progress in text classification. The paper also showcased that DistilBERT takes comparatively less time for training. Inspired by the literature, the proposed work presents a comprehensive analysis of the classification performance of the DistilBERT model when subjected to various data augmentation techniques under the scenarios where datasets have few hundreds of records, thousands of records and tens of thousands of records. The paper presents the performance analysis on three standard datasets namely: Stanford Sentiment Treebank dataset, Twitter Climate Change Sentiment dataset andSpamAssassin dataset using accuracy metric for text classification.

3. Experimental Design

3.1 Bench Mark Dataset

As the proposed work aims at analyzing the impact of different data augmentation methods on the data with varying sizes, the experiments have been conducted using three datasets. First, the proposed work has considered the Stanford Sentiment Treebank (SST) dataset. The Stanford Sentiment Treebank is a corpus that contains parse trees with comprehensive sentiment labels, enabling a thorough analysis of how sentiment is influenced by language composition[13]. This corpus allows for a detailed exploration of the relationship between sentence structure and sentiment thereby, providing valuable insights into how sentiment is expressed at different levels of language hierarchy. Here the class label '0' denotes a negative review and '1' denotes a positive review.

Secondly, the SpamAssassin(SA) dataset [20], a commonly used training dataset for spam detection offering a distinct advantage in its classification of spam and ham into additional classes based on their difficulty levels has been considered. This dataset provides a more detailed breakdown of spam and non-spam samples allowing for a comprehensive assessment of the performance of spam detection models across various complexities.

Finally, Twitter Climate Change Sentiment (CCS) Dataset has been considered. The acquisition of this data has been financially supported by a Canada Foundation for Innovation JELF Grant awarded to Chris Bauch at the

University of Waterloo[21]. The dataset consists of a compilation of tweets related to climate change that were collected from April 27, 2015, to February 21, 2018. A total of 43,943 tweets were annotated in this dataset. Each tweet was independently labeled by three reviewers. Only tweets that received unanimous agreement from all three reviewers are included in this dataset, while those that did not achieve consensus were excluded.Every tweet in this dataset is assigned a label from one of the following classes:

- Label 2 (News): The post includes a link to factual news or information related to climate change.
- Label 1 (Pro): The post demonstrates alignment with the consensus on human-induced climate change.
- Label 0 (Neutral): The post remains impartial regarding the acceptance or rejection of the concept of human-contributed climate change.
- Label -1 (Anti): The post expresses disbelief or skepticism towards man-made climate change.

The goal is to classify tweets into these four categories to better understand the sentiment and information being shared about climate change on social media.

3.2 System architecture

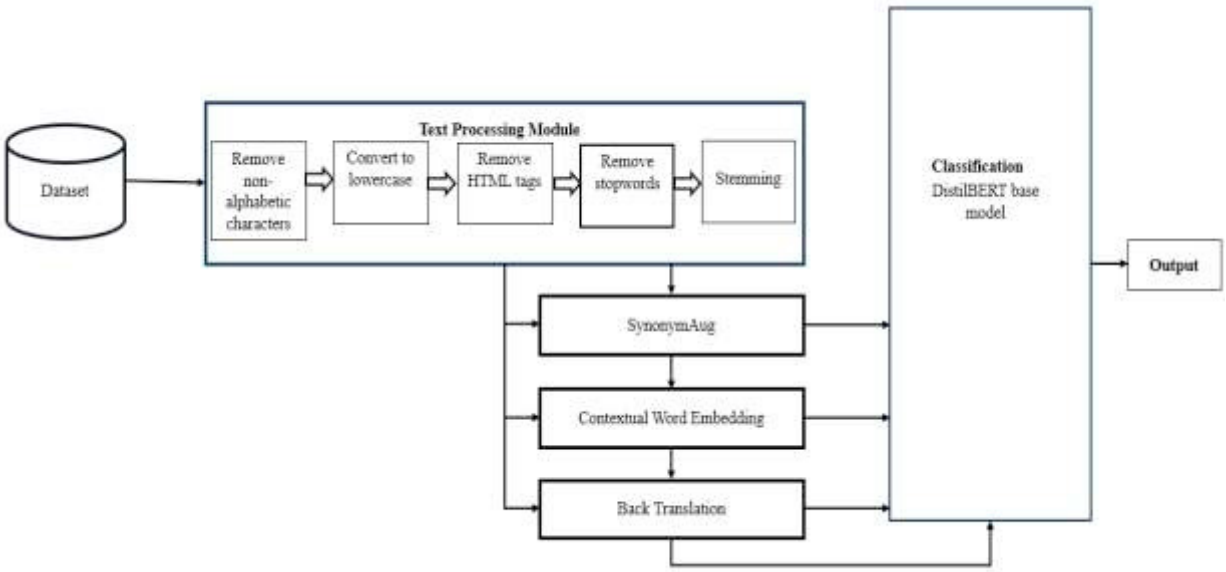

Fig 1: Proposed Framework of Text Data Augmentation

Fig. 1 illustrates the architecture of the proposed model for text data augmentation.The input to the model consists of the unchanged textual data.The datasets are initially pre-processed to make them free from the unwanted information, irrelevant words and noise. Initially, we evaluate the accuracy using a distilled BERT model. Following this initial assessment, we apply augmentation techniques individually to the dataset. These techniques include synonym replacement, contextual word embedding, and back translation, and we calculate the accuracy for each. Once we have the accuracy scores for these three augmentation methods, we combine them and compute an overall accuracy score. This comprehensive evaluation approach allows us to gauge the performance and effectiveness of our model under various augmentation scenarios.In order to evaluate the effectiveness of the proposed model, the experiments on three distinct datasets sourced from various domains have been conducted and analysed. These datasets vary in size with 500(SST dataset), 5564(SA dataset), and 43943(CCS dataset) instances, respectively. For each dataset, we split the data into training and testing sets and performed a comparative analysis of model accuracy with and without augmentation.

3.3 Text Processing

The text pre-processing involves converting unstructured and noisy natural language data into a standardized and coherent format, enabling its utilization in models for analysis and learning purposes [22-27]. Text pre-processing methods can be either generic and suitable for a wide range of applications or tailored to specific tasks. For the proposed work the datasets have been pre-processed using five main techniques namely, (a) removal of HTML tags, (b) removal of non-alphabetical characters, (c) lower casing the sentences, (d) removal of stop words and (e) Stemming

3.4 Data Augmentation Techniques:

The proposed work mainly studied three significant data augmentation techniques when applied to the raw data parallel and sequentially. These augmentation techniques are as under:

i. Synonym Augmentation: SynonymAug, a class that belongs to the "naw" package, which stands for NLP Augmentation for WordNet [28], performs data augmentation in NLP tasks by leveraging WordNet to replace words with their synonyms. WordNet itself is a lexical database that organizes words into groups called ‘synsets’, grouping synonyms together and providing relationships between them. By utilizing SynonymAug, one can identify synonyms for a specific word and replace it with one of the alternatives, while ensuring that the overall meaning of the text remains unchanged. This augmentation technique proves to be beneficial in generating new instances for training, ultimately enhancing the model's capacity to handle diverse word choices and semantic variations. This procedure broadens the dataset and enhances the model's performance.

ii. Contextual word embeddings: These are a type of word representations utilized in natural language processing that capture the meaning of a word by considering its context within a sentence or document [29]. Unlike static word embeddings, which assign a fixed representation to each word, contextual word embeddings take into account the surrounding words, resulting in a more nuanced and dynamic representation. In the case of DistilBERT contextual word embeddings are generated to capture the contextual meaning of words based on their neighbouring words in a sentence or text. In summary, contextual word embeddings, such as those used in DistilBERT, enhance word representations in NLP by considering the context in which words appear, enabling a more comprehensive understanding of their meanings.

iii. Back translation: It is one of the most frequently used data augmentation technique in natural language processing (NLP) that involves the translation of text from one language to a target language and then translating it back to the original language [30-31]. This process helps in automatically generating additional data for training to enhance the models of NLP. The back translation augmentation technique typically follows these steps:
   - Initial dataset: Begin with a dataset consisting of sentences in the original language to be augmented.
   - Translation: Translate the sentences from the original language to a target language using a machine translation system. Different pre-existing translation services can be employed for this purpose.
   - Reverse translation: Convert the translated sentences back to the original language using the same machine translation system. This step generates new sentences that may exhibit slight differences in phrasing or word usage compared to the original ones.
   - Augmented dataset: Merge the original sentences with the back-translated sentences, resulting in an augmented dataset. This expanded dataset can then be utilized for training NLP models, allowing them to learn from a more diverse set of examples.

iv. Combined augmentations: In this module we combined all the three-methods synonym replacement, contextual word embeddings and back translation. By combining these three techniques, data augmentation modules achieve superior performance while comparing it with individual tasks.

3.5 Classification Model: DistilBERT:

DistilBERT is a streamlined version of the original BERT model created to be more compact and quicker while maintaining a significant portion of its effectiveness. When efficiency and compact model sizes are of utmost importance, DistilBERT serves as a significant choice. DistilBERT, a smaller and a faster transformer model,

introduces knowledge distillation into the standard Bidirectional Encoder Representations from Transformers (BERT) model [23]. During its pre-training process for knowledge distillation, it utilizes the corpus which is similar in self-supervised manner with BERT type model acting as its teacher. The detailed architecture has been discussed in [24]. DistilBERT takes a tweet represented by input 'X' as its input, which is a sequence of words from the dataset. These input sequences are transformed into a group of embedding vectors, with distinct vector corresponding to a word in the sequence. The transformer encoder in DistilBERT is then employed to study the information which is contextual for each term. This process involves using a mechanism based on self-attention to create contextual embeddings. The embeddings for each word within the tweet are combined into a unified vector, creating a representation that encapsulates the semantic meaning present in the original message. This concatenated vector is then passed through a fully connected layer which outputs a vector of size 'd', where 'd' represents the number of neurons. Subsequently, a layer of classification is added to the feature extractor model's end, fine-tuning the pre-trained DistilBERT for the event detection task. Using this configuration, the model utilizes the concatenated embeddings to make predictions about the appropriate event class for each given input sequence (tweet). The pre-training involves using raw texts without human labelling and by making use of publicly available data. To achieve this, inputs and labels are automatically generated from the texts leveraging the BERT base model [25]. The pre-training focuses on three primary objectives:

- Distillation loss: During the training process, the model is optimized to generate probabilities that match those of the BERT base model. This means that the model aims to produce similar probability distributions for the predicted event classes as the original BERT base model.
- Masked language modelling (MLM): In the BERT base model's initial training process, the training loss includes a method known as Masked Language Modelling (MLM). MLM comprises the random masking of 15% of the words within a sentence. The entire sentence with the masked words is then fed through the model, which is tasked with predicting the masked words. Unlike conventional sequential models such as recurrent neural networks (RNNs) or autoregressive models like GPT, which mask future tokens within their internal processes, MLM permits DistilBERT to acquire a sentence representation that is bi-directional in nature.
- Cosine embedding loss: The model is trained to generate states which are hidden and closely resemble to those of the BERT base model, aiming to achieve similar representations.

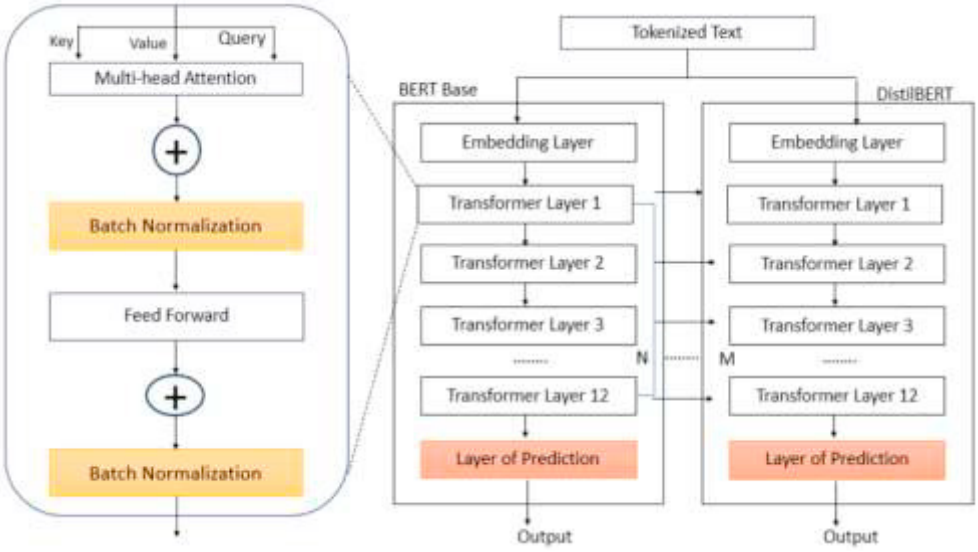

Fig 2: Architecture and components of DistilBERT

DistilBERT, trained by distilling BERT base, has a vocabulary size of 3,052,245. It is more efficient as it uses 40% lesser parameters when comparing with BERT base models. It also supports various activation functions including "gelu", "relu", and "silu". In the development of our model, we have carefully selected a set of

hyperparameters to ensure optimal performance. The batch size was set to 8, allowing for a balance between computational efficiency and gradient estimation accuracy during training iterations. We trained the model over 30 epochs, providing ample opportunity for the model to learn patterns in the data across multiple passes. A learning rate of 0.00002 was chosen to control the rate of optimization, ensuring a steady convergence towards the minimum of the loss function without overshooting. Finally, we introduced a regularization measure in the form of a weight decay factor of 0.01 to mitigate overfitting, consequently improving the model's capability to generalize from the training data to unseen data.

4. Results

In this study, the model's performance is assessed on three distinct datasets, each varying in size. The model's accuracy without text augmentation has also been measured for each dataset for providing valuable insights into the model's capabilities and potential areas for enhancement. Fig. 3 illustrates the graphical representation of the model's performance metrics, accuracy, and F1 score against the number of epochs for the text classification model across three different instances: (a) 500 instances that is the SST dataset, (b) 5564 instances that is the SA dataset, and (c) 43943 instances that is the CCS dataset. The graph presents two lines for each instance: one representing the model's performance without augmentation and the other representing the model's performance after combining all three augmentation methods. The accuracy and F-Score curve against epochs clearly depict the saturation in performance of the model's training process for both raw data and augmented data within merely 30 epochs. Also, it must be noted that even though there was not much performance increase for large sized datasets, the model trained very well within fewer epochs, as presented in Fig. 3(c). This supports the proper training of the model under the two scenarios for text classification.

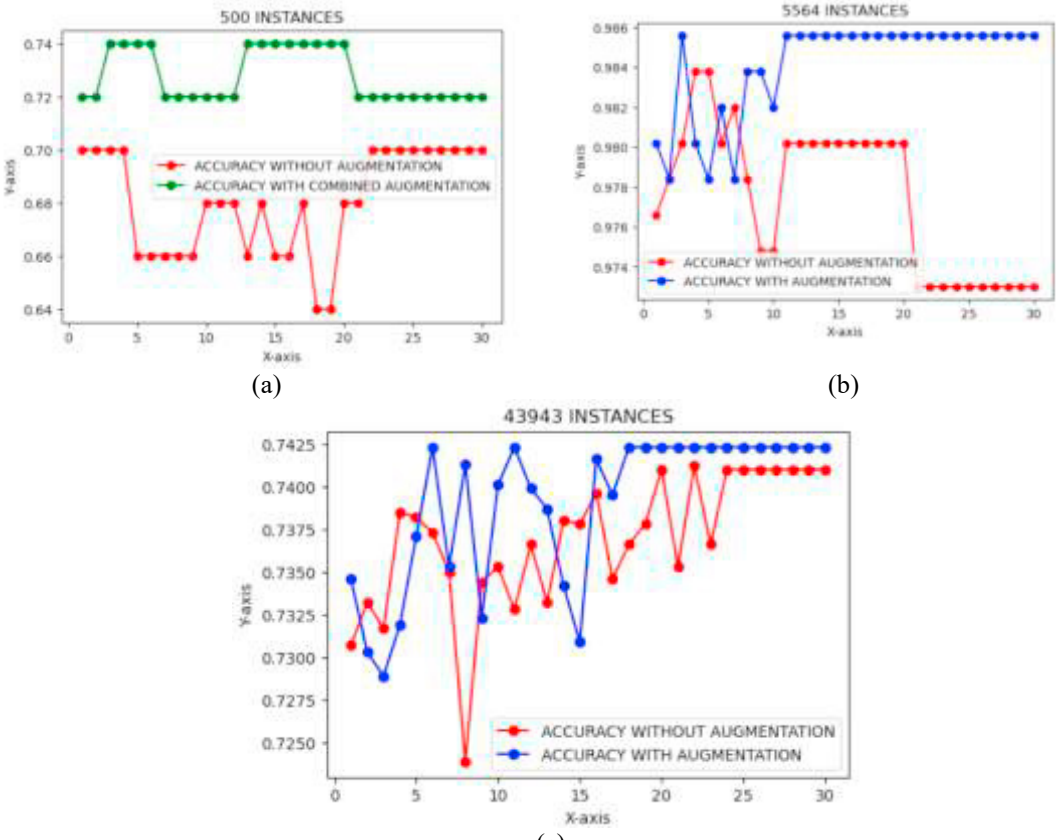

Fig 3: Performance on text classification without and with Augmentation on different datasets (a) SST dataset, (b) SA dataset and (c) CCS dataset.

The model achieved an accuracy of 74%on the SST dataset with 500 instances, indicating relatively lower performance. This suggests that the smaller dataset size might have limited the model's ability to effectively learn underlying language patterns. Consequently, there is ample room for improvement in this instance, and the proposed approach of employing data augmentation techniques could be beneficial in boosting accuracy.Impressively, the model in its base form obtained an accuracy of 97.84% on the SA dataset comprising 5564 instances. This substantially higher accuracy indicates the model's strong performance to learn the language's complex constructs and patterns when trained on a larger dataset. It demonstrates the model's capability to effectively grasp complex patterns, leading to highly accurate predictions with extensive training data.The CCS dataset containing 43943 instances achieved an accuracy of 73.07% without text augmentation. Although this accuracy is lower than that of the larger dataset (5564 instances), it still surpasses the accuracy obtained on the smallest dataset (500 instances). This suggests the model is performing reasonably well on this instance, but there is room for further improvement. The lower accuracy compared to the second instance might be attributed to factors such as dataset noise, class imbalance and the complexity of the data distribution in hued datasets.

Table 1: Accuracy and F1-score obtained by DistilBERT model on raw datasets and augmented datasets of varying sizes; small (500 instance), medium (5564 instances) and large (43943 instances)

	SST dataset	SA dataset	CCS dataset
Without Text Augmentation	0.74	0.978	0.731
Synonym Replacement	0.80	0.992	0.747
Contextual Word Embedding	0.84	0.995	0.740
Back Translation	0.86	0.992	0.745
Combined	0.90	0.993	0.746
	F1-score		
Without Text Augmentation	0.73	0.978	0.724
Synonym Replacement	0.79	0.993	0.742
Contextual Word Embedding	0.83	0.995	0.730
Back Translation	0.84	0.992	0.731
Combined	0.90	0.993	0.741

Table 1 summarizes the outcomes of our experiments before and after data augmentation both in view of F1-Score and accuracy. The synonym replacement technique when applied individually improves the performance on all three datasets compared to the baseline. It boosts both the accuracy and f1 score by approximately 6% on the small instance, 2% on the medium-sized instance, and 1.6% on the larger instance. This indicates that synonym replacement helps the model learn alternative representations of words and improves its ability to understand and generate more accurate responses.

Similar to synonym replacement, contextual word embedding also enhances the performance on all datasets. It alleviated the metrics by approximately 10% on the small instance, 2% on the medium-sized instance, and by approximately 1% on the larger instance. This technique leverages the context in which words appear to provide a richer representation, thereby, enabling the model to capture more nuanced relationships between words and enhance its predictive capabilities. Back translation consistently improves accuracy on all instances as well. It achieves an increase in performance by approximately 12% on the small instance, 2% on the medium-sized instance, and approximately 1.5% on the larger instance. This technique helps the model to learn new sentence structures and linguistic variations, enhancing its understanding and generation of responses.

Finally for the combination of synonym replacement, contextual word embeddings and back translation applied together on the raw datasets, the model achieves a remarkable increment in performance by16%, 2%and 1.5% respectively for SST dataset, SA dataset and CCS dataset. The results indicate a noticeable improvement in both accuracy and F1 score after applying data augmentation. The augmentation process effectively enriched the training dataset, leading to enhanced generalization and better performance on the test datasets.

Overall, the results show that text augmentation techniques, including synonym replacement, contextual word embedding, and back translation, generally improve the accuracy of the model as compared to the model trained on raw data without augmentation. However, the performance of each technique can vary depending on the size of the datasets as the increment in performance for smaller datasets is found to be higher (12% approx.) as

compared to the larger datasets (2% approx.). It is important to note that the results presented here are specific to the given instances and the evaluation metrics used.

The work also provides insights into the impact of dataset size and augmentation on model performance and training time. For larger datasets, we observed that while the accuracy may not be significantly high, the training time is considerably longer. This is likely due to the increased complexity and diversity within larger datasets. Interestingly we observed distinct learning patterns when comparing modelsthat underwent data augmentation during training with those that did not.. The model trained with data augmentation demonstrated an earlier onset of learning with a steady accuracy after merely 17 epochs. This implies that the utilization of data augmentation could potentially improve the model's capacity to generalize from the training data, thereby expediting the learning process.Conversely, the model trained without data augmentation exhibited a delayed learning onset, beginning at epoch 25. This delay may be attributed to the lack of diversity in the training data in the absence of augmentation techniques necessitating additional epochs for the model to effectively learn.These findings underscore the potential benefits of data augmentation in improving learning efficiency and expediting model convergence. They also highlight the importance of considering data diversity and representativeness when designing machine learning experiments.

5. Conclusion And Future Work

The results obtained from applying text augmentation techniques on the given instances provide valuable insights into their impact on accuracy. Text augmentation techniques, such as synonym replacement, contextual word embedding, and back translation, generally improve the accuracy of the model on augmented datasets as compared to that of the raw datasets. The augmentation techniques applied enhanced the model's ability to understand and generate more accurate responses by capturing nuanced relationships between words, sentence structures and variations. The performance of each technique may vary depending on the size of the data thereby highlighting the importance of text augmentation in enhancing the performance of language models. Incorporating techniques such as synonym replacement, contextual word embedding, and back translation can lead to more accurate predictions and better language understanding and generation. However, it's important to note that the results are specific to the provided instances and accuracy as the evaluation metric. Data augmentation accelerates the learning process, seen in earlier accuracy improvements, emphasizing its role in enhancing model generalization and expediting convergence. Further experimentation and analysis are recommended to validate the effectiveness of these techniques in different scenarios by exploring generalization across domains, investigating transfer learning and fine-tuning approaches to validate the practical applicability of text augmentation techniques. By focusing on these aspects, researchers have the opportunity to propel the discipline forward and make meaningful contributions to the enhancement of natural language processing models in terms of their precision, resilience, and efficiency.

REFERNECES

[1] Nair, A. J., Veena, G., & Vinayak, A. (2021). Comparative study of Twitter sentiment on COVID-19 tweets. In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1773-1778). IEEE.

[2] Gontumukkala, S. S. T., Godavarthi, Y. S. V., Gonugunta, B. R. R. T., Gupta, D., & Palaniswamy, S. (2022). Quora Question Pairs Identification and Insincere Questions Classification. In 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.

[3] Mirończuk, M. M., &Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. Expert Systems with Applications, 106, 36-54.

[4] Unnithan, N. A., Shalini, K., BG, H. B., & Soman, K. P. (2018). Amrita_student at SemEval-2018 Task 1: distributed representation of social media text for affects in tweets. In Proceedings of The 12th International Workshop on Semantic Evaluation (pp. 319-323).

[5] Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. Journal of Time Series Analysis, 15(2), 183-202.

[6] Scott, S., & Matwin, S. (1998). Text classification using WordNet hypernyms. In Usage of WordNet in natural language processing systems.

[7] Dalal, M. K., & Zaveri, M. A. (2011). Automatic text classification: a technical review. International Journal of Computer Applications, 28(2), 37-40.

[8] Kowsari, K., et al. (2019). Text classification algorithms: A survey. Information, 10(4), 150.

[9] Khuntia, M., & Gupta, D. (2023). Indian news headlines classification using word embedding techniques and LSTM model. Procedia Computer Science, 218, 899-907. Elsevier.

[10] Sayeed, M. A., & Gupta, D. (2022). Automate Descriptive Answer Grading using Reference-based Models. In 2022 OITS International

Conference on Information Technology (OCIT) (pp. 262-267). IEEE.
[11] Wang, J., & Perez, L. (2017). The effectiveness of data augmentation in image classification using deep learning. Convolutional Neural Networks Vis. Recognit, 11.2017, 1-8.
[12] Mosolova, A., Fomin, V., & Bondarenko, I. (2018). Text Augmentation for Neural Networks. AIST (Supplement), 2268, 104-109.
[13] Wei, J. W., & Zou, K. (2019). EDA: easy data augmentation techniques for boosting performance on text classification tasks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (pp. 6381–6387). Association for Computational Linguistics, Hong Kong, China.
[14] Fadaee, M., Bisazza, A., & Monz, C. (2017). Data Augmentation for Low-Resource Neural Machine Translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 567–573.
[15] Silfverberg, M., et al. (2017). Data augmentation for morphological reinflection. Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection.
[16] Kobayashi, S. (2018). Contextual augmentation: data augmentation by words with paradigmatic relations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 452-457.
[17] Shorten, C., Khoshgoftaar, T. M., &Furht, B. (2021). Text data augmentation for deep learning. Journal of Big Data, 8, 1-34.
[18] Sanh, V., et al. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper, and lighter. arXiv preprint arXiv:1910.01108.
[19] Silva Barbon, R., & Akabane, A. T. (2022). Towards Transfer Learning Techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for Automatic Text Classification from Different Languages: A Case Study. Sensors, 22(21), 8184.
[20] Naem, A. A., Ghali, N. I., & Saleh, A. A. (2018). Antlion optimization and boosting classifier for spam email detection. Future Computing and Informatics Journal, 3(2), 436-442.
[21] Effrosynidis, D., et al. (2022). The climate change Twitter dataset. Expert Systems with Applications, 204, 117541.
[22] Naveenkumar, K. S., Vinayakumar, R., & Soman, K. P. (2019): Twitter dataset for sentimental analysis and application of classical machine learning and deep learning. In 2019 International Conference on Intelligent Computing and Control Systems (ICCS) (pp. 1522-1527). IEEE.
[23] Akash, G., Kumar, H., & Bharathi, D. (2021). Toxic comment classification using transformers. Proceedings of the 11th Annual International Conference on Industrial Engineering and Operations Management, Singapore.
[24] Hari, A., & Kumar, P. (2023). WSD Based Ontology Learning from Unstructured Text Using Transformer. Procedia Computer Science, 218, 367-374.
[25] Hadeer, A., et al. (2022). Improving crisis events detection using distilbert with hunger games search algorithm. Mathematics, 10(3), 447.
[26] Paul, P., & Singh, R. P. (2023). A weighted hybrid recommendation approach for user's contentment using natural language processing. In AIP Conference Proceedings, Vol. 2705, No. 1. AIP Publishing.
[27] Paul, P., & Singh, R. P.(2022). Sentiment Rating Prediction using Neural Collaborative Filtering. In 2022 IEEE 7th International Conference on Recent Advances and Innovations in Engineering (ICRAIE), Vol.7, pp 148-153.
[28] Fellbaum, C. (2010). WordNet. In: Poli, R., Healy, M., Kameas, A. (eds) Theory and Applications of Ontology: Computer Applications. Springer, Dordrecht. DOI: 10.1007/978-90-481-8847-5_10
[29] Miaschi, A., &Dell'Orletta, F. (2020). Contextual and Non-Contextual Word Embeddings: an in-depth Linguistic Investigation. In Proceedings of the 5th Workshop on Representation Learning for NLP, pages 110–119.
[30] Chen, H. Y., & Boore, J. R. (2010). Translation and back-translation in qualitative nursing research: methodological review. Journal of clinical nursing, 19(1-2), 234-239.
[31] Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., &Edunov, S. (2019). Facebook FAIR's WMT19 News Translation Task Submission. In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pages 314–319, Florence, Italy. Association for Computational Linguistics.