Language Agnostic Data Augmentation for Text Classification

M. Sharmila Devi, V. Lakshmi Chaitanya, G. Sharanya, B. Himaja,
A. Bhavya Rohitha and A. Sujitha
Department of Computer Science & Engineering, Santhiram Engineering College, Nandyal-518501, Andhra Pradesh, India

Keywords: LiDA, MBERT, SBERT, XLM-RoBERTa, LSTM, Token-Level, Constructive Learning, Back-Translation.

Abstract: The lack of labeled data is especially problematic for low-resource languages, making the development of high-performance text classification models especially challenging. Collecting diverse and large-scale annotated datasets, that are crucial to train generalizable models, is often expensive. One of the solutions that address this limitation is data augmentation, using synthetic training data to strengthen model performance. Nevertheless, the majority of existing methods for data augmentation are extremely language-dependent, focused only on English, and only at word or sentence level through word replacement or paraphrasing respectively. These approaches may not generalize across languages, and thus their applicability to low-resource settings is limited. Type a language ɜxy The LIDA approach, which works on the level of sentence embeddings and can thus be applied regardless of language, is also compared and contrasted with traditional augmentation methods. Thus, these datasets can efficiently be augmented across a variety of language models without needing to depend on specific language preprocessing. We evaluate LIDA on three distinct languages for LSTM and BERT based models on 4 different dataset fractions. We also conduct an ablation study to evaluate the impact of different components of our approach on model performance. The empirical results indicate that LIDA could be viewed as language-agnostic, scalable, and robust augmentation strategy for low-resource text classification scenarios, and the source code of LIDA has been released on GitHub for facilitating other relevant researches and applications.

1 INTRODUCTION

1.1 Significance of Text Classification in NLP

Text classification belongs to simple Natural Language Processing (NLP) tasks and has attracted much attention with a variety of applications across domains. For spam detection, sentiment analysis, emotion detection, and topic detection text classification are commonly used (Bayer, et al,2022). These applicationsa are the foundation of many modern technologies like email filtering, social media, customer feedback analysis and content classification. For example, sentiment analysis can assist companies in gauging the sentiments of customers about their products to make better decisions regarding products they stock, how they market them, and more. Emotion detection can be of help to organizations in social media in understanding how the public feels during crisis or large events so they may respond effectively in a timely fashion.

1.2 Advancements in Deep Learning for Text Classification

The accuracy of text classification systems has improved dramatically over the past few years, largely due to the introduction of advanced deep learning algorithms. These algorithms are built upon powerful neural network architectures, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which allow them to extract representations of features and patterns in large corpora for accurate text classification (Bayer, et al,2022). BERT and GPT & transformer Based Architecture a new revolution in NLP, because those models started to capture context words, multilingual sentences etc. The transformers, which enable attention mechanisms that guide the assignment of weights to relative importance between

Devi, M. S., Chaitanya, V. L., Sharanya, G., Himaja, B., Rohitha, A. B. and Sujitha, A.
Language Agnostic Data Augmentation for Text Classification.
DOI: 10.5220/0013876800004919

In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25 2025) - Volume 2*, pages 50-58
ISBN: 978-989-758-777-1

words in a sentence, enable these models to develop a nuancing structure–more closely matching input text to a potential classification. But, the performance of deep learning models is highly dependent on getting large amounts of labelled data for text classification.

1.3 Challenges with Low-Resource Languages

When it comes to low-resource languages such as Indonesian, the issue of creating high-performance text classification models are very pronounced because the amount of labeled datasets is limited. The linguistic resources, for example, annotated corpora, dictionaries, and pre-trained models, which are easily available for high-resource languages like English are very poor for most low-resource languages. Collecting huge amounts of labeled data is time-consuming and relatively expensive, thus making it challenging for researchers and practitioners dealing with poorly resource endowed language. When models are trained on small data, they are underperforming when it comes to delivering predictions that accurately reflect the nature of language, which is a consequence of the absence of data.

2 LITERATURE REVIEW

Sujana, Yudianto, and Hung-Yu Kao. "LiDA: Language-independent data augmentation for text classification." IEEE Access 11 (2023): 10894-10901. In this paper, we present a method of text classification enhancement called LiDA, which generates unlimited language-independent artificial data. The authors highlight the challenges of having insufficient data for text classification, particularly for languages that have very little data available. They propose LiDA as a method for effective dataset augmentation, based on transformation on data and making use of cross lingual embeddings. The study shows, through extensive experiments over different languages and datasets, that LiDA improves performance without relying on large-scale labeled datasets. This work is particularly useful for researchers and practitioners dealing with low-resource languages in NLP [2].

Fields, John, Kevin Chovanec, and Praveen Madiraju. "A survey of text classification with transformers: How wide? how large? how long? how accurate? how expensive? how safe?" IEEE Access 12 (2024): 6518-6531. This work investigates the role of transformers in text classification with respect to several aspects: width, size, computational cost, accuracy, and safety. We compare different transformer designs (BERT, GPT, etc.) and their scaling impact on classification performance. The paper also discusses certain trade-offs, like model complexity vs actual applicability. It thus presents an important reference for researchers that seek to apply deep learning with transformers for text-based classification tasks, covering models' advantages and limitations.

Taha, Kamal, et al. "Text classification: A review, empirical, and experimental evaluation." arXiv preprint arXiv:2401.12982 (2024). In this study, we provide an in-depth exploration of text classification techniques, combining theoretical insights with empirical evaluations. The authors compare deep learning methods such as CNNs, RNNs, and transformers against more traditional machine learning approaches such as Naïve Bayes and SVM. Through extensive experiments, they assess performance under different scenarios, such as varying data sizes and class imbalances. Based on the study's results, deep learning models are generally superior to classical methods but also require a considerable amount of computational power. This study serves as a guiding manual for choosing the most appropriate text classification approach based upon certain needs.

Minaee, Shervin, et al. "Deep learning--based text classification: a comprehensive review." ACM computing surveys (CSUR) 54.3 (2021): 1-40. This paper provides a comprehensive overview of deep learning methods for text classification. We introduce readers to various model architectures including transformer-based models versus CNN versus RNN vs attention-based models. They review key advances that have significantly improved classification performance, such as transfer learning and pre-trained language models. The review further discusses practical challenges, such as data availability and clean-up costs, as well as ethical dilemmas regarding bias in AI models. For academia and industry, this paper provides an organized summary of methodologies which can be used as a reference for text classification.

Qu, Ping, et al. "Comparison of Text Classification Algorithms based on Deep Learning." Journal of Computer Technology and Applied Mathematics 1.2 (2024): 35-42. The efficacy of deep learning-based text classification algorithms in practical applications is the main focus of this study's evaluation and comparison. The authors measure the performance of models like transformers,

CNNs, and LSTMs on a variety of datasets. They look at things like robustness against noisy data, training efficiency, and accuracy. The findings show that CNNs and LSTMs are still competitive in environments with limited resources, even though transformer-based models typically attain the highest accuracy. The paper helps choose the best model for particular text classification tasks by clearly presenting the benefits and drawbacks of each approach.

3 EXISTING SYSTEMS

Text classification is a key task in natural language processing (NLP) and serves as the foundation for various applications such as sentiment analysis, topic modeling, spam filtering, and emotion classification. For low-resource languages, obtaining sufficient labeled data is a major challenge due to the high costs and time-consuming nature of data collection and annotation (Sujana, et al,2023).

To address this data scarcity issue, researchers employ data augmentation techniques to generate synthetic training data. The current approach to text data augmentation primarily consists of two methods:

3.1 Word-Level Data Augmentation

Word-level techniques involve modifying individual words within a sentence to create new versions of the original text. Common strategies include:

- **Synonym Substitution:** Synonym Substitution is a word-level data augmentation technique where words in a sentence are replaced with their synonyms to generate new variations of the text. Replacing words with their synonyms from lexical databases such as WordNet.
- **Word Embedding-Based Substitution:** Word Embedding-Based Substitution is a word-level data augmentation technique where words in a sentence are replaced with their most similar words based on word embeddings (numerical vector representations of words). Substituting words with their closest counterparts in a pre-trained word embedding model (e.g., Word2Vec, GloVe).
- **Random Deletion, Insertion, and Swapping:** Random Deletion, Insertion, and Swapping is a word-level data augmentation technique that modifies sentences by randomly removing, adding, or rearranging words. This technique aims to create synthetic variations of a sentence while maintaining its original meaning. Introducing variations by randomly removing, adding, or rearranging words in a sentence.
- **Masked Language Models (MLM):** Masked Language Models (MLM) is a word-level data augmentation technique where specific words in a sentence are randomly masked (hidden), and a pre-trained language model predicts the missing words based on the surrounding context (Zhao, Huanhuan, et al.) Using models like BERT to predict and replace words based on contextual understanding.

3.2 Limitations of Word-Level Augmentation

- **Dependency on Language-Specific Resources**: Many techniques rely on resources like WordNet, which may not be available for low-resource languages.
- **Incompatibility with Certain Languages:** Not suitable for languages that lack word segmentation, such as Chinese and Japanese.
- **Risk of Semantic Distortion:** Random word replacements may alter the meaning of a sentence, leading to inaccurate synthetic data.

3.3 Sentence-Level Data Augmentation

Some sentence-level augmentation techniques create a completely new sentence while leaving its meaning untouched. The most commonly employed methods are:

- **Back-Translation:** Back-Translation, as a sentence-level data augmentation technique, injects synthetic variation into the text by translating a source sentence to some other language and then back to the original language (Sujana, et al,2023). Using machine translation models (e.g., Google Translate, MarianMT) to translate a sentence to another language and then back to the original language.
- **Generative Models:** Generative Models: Given that they create new sentences from scratch, they fall under a sentence-level data augmentation technique where deep learning models are used to produce yet new

sentences creating their basis on patterns learned from other texts they have trained on. Applying deep neural network architectures to create new sentences with the same meaning, such as using sequence to sequence (Seq2Seq) or Generalised Pretrained Transformer (GPT)-based models.

3.4 Limitations of Sentence-Level Augmentation

- **Dependence on High-Quality Translation Models**: Back-translation is only effective when accurate machine translation models are available, which is not the case for all languages.
- **High Computational Requirements:** Generative models require large-scale datasets and significant computing power, making them impractical for resource-limited environments.
- **Potential Bias in Synthetic Data:** The quality of generated sentences depends on the pre-trained language model, which may introduce unwanted biases.

3.5 Challenges in the Existing System

Although traditional data augmentation techniques are effective, they pose several challenges, particularly for low-resource languages:

- **Language Dependence** – Many augmentation techniques require language-specific tools (e.g., WordNet, pre-trained embeddings) that are unavailable for many languages.
- **High Computational Costs** – Advanced techniques like back-translation and generative models demand intensive processing power and large-scale pre-training, making them unsuitable for low-resource settings.
- **Limited Generalization** – Existing methods are optimized for high-resource languages (e.g., English) but struggle to work effectively in low-resource languages with fewer linguistic tools.

3.6 Need for an Improved System

Given these limitations, a universal, language-independent data augmentation technique is needed. The ideal system should:

- Work across multiple languages without relying on language-specific resources.
- Generate high-quality synthetic data while preserving the original meaning and context.
- Be computationally efficient and scalable, making it practical for low-resource language scenarios.

4 PROPOSED SYSTEM

An organized method for language-independent data augmentation for text classification is represented by the figure 1 architecture provided. Tokenization is the first step in the process, which divides input text into discrete tokens. After that, SBERT (Sentence-BERT) is used to convert these tokens into numerical embeddings, producing high-dimensional vector representations. Many data augmentation techniques are used to improve the quality and diversity of augmented data. This covers embedded augmentation techniques like autoencoders, denoising autoencoders, and linear transformation. To generate more varied training samples, back-translation and token-level changes are also added, including synonym replacement, random insertions, deletions, and shuffling. To improve generalization, these augmentations are further refined using constructive learning techniques. Figure 1 show the Architecture.

4.1 Disruptions in the Suggested System at the Token Level

Token-level perturbation is the process of changing specific tokens (words or subwords) inside a text while preserving its original meaning. This approach improves the model's performance and generalization by generating a diverse set of training samples.

4.2 Techniques for Perturbation at the Token Level:

- **Random Word Replacement:** Some words are replaced with synonyms to provide variation without altering the context.
- **Random Insertion & Deletion:** This method adds or removes words to produce various phrase forms from the text.
- **Character-Level Modifications:** These add small changes like typos, misspellings,

or character swaps to make the model more resilient to noisy inputs.

- **Subword Modifications:** These change word segmentations using Byte Pair Encoding (BPE) techniques to diversify token representations.

4.3 Impact of Token-Level Perturbations:

By using these techniques, the system improves:

- **Language Independence**: This allows data augmentation without depending on predefined linguistic rules.
- **Diversity:** This creates a wide range of textual variations for better model training.
- **Robustness:** This increases the model's capacity to handle variations in real-world data.
- **Performance:** Improves text classification accuracy and generalization. These improvements make the system more efficient and flexible text processing system, which makes it suitable for low-resource language applications.

4.4 LiDA's Constructive Learning

The multi-modal rubrics of constructing knowledge through LiDA looks at indirectly a number of texts extracted through the same text data with which knowledge was structured. By doing so, it improves the model performance on semantic, feature representation and generalization point of views, which enable the model to efficiently concentrate on relevant text input patterns.

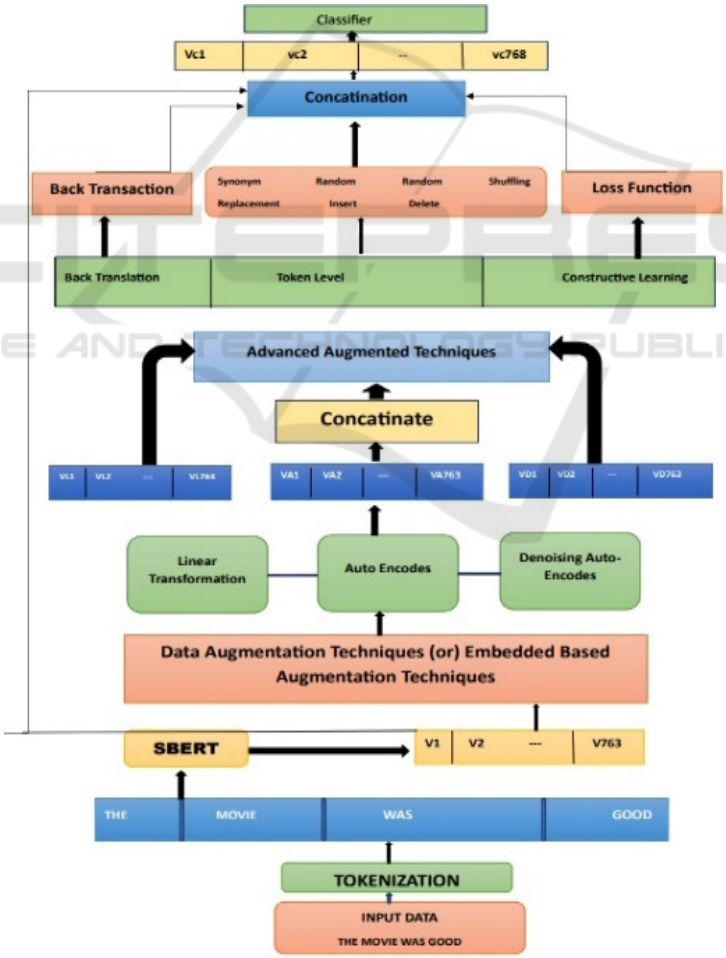

Figure 1: Architecture.

4.5 Key Elements of LiDA's Constructive Learning

- **Augmentation-Based Learning:** It uses a variety of augmentation techniques from token level perturbations to back-translation to churn out multiple text permutations. These techniques keep the model busy listening to fine patterns and only to look at the objects, blocking the vision of irrelevant details.
- **Richer Text Representation:** This helps the model to identify surface and semantic changes by encouraging it to learn what is invariable and what varies in the text.
- **Improved Generalization & Stability:** By feeding it various text inputs, the model avoids overfitting and generalizes well on new data. This allows for great flexibility across languages and domains. The next approach is to avoid overfitting, that is when a model relies too heavily on training data and can no longer generalize adequately to new input. Constructive learning prevents this from happening by exposing the model to a variety of transformations, and preventing it from becoming too reliant on a particular pattern.
- **Better Model Performance:** Combining constructive learning and token-level perturbations results in more robust, precise and efficient text classification system.
- The proposed framework is an effective solution for low-resource text classification problems since it integrates constructive learning with the augmentation processes of LiDA, yielding enhanced performance, semantics interpretation, and linguistic independence. LiDA has been mostly employed in text categorization; LiDA utilizes annotated corpora to improve NLP (natural language processing) tasks. It helps to annotate text data which in turn helps for its processing and analysis in NLP applications. LiDA leverages deep learning frameworks like PyTorch and Hugging Face to train the models for better performance.

4.6 Hugging Face

Packs a natural language processing based on popular AI platforms that can lead to produce, train, and take advantage of the machine-learning model. It is widely used for Transformer based models such as BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer), T5 (Text-To-Text Transfer Transformer), RoBERTa (Robustly optimized BERT approach) and offers cutting-edge solutions for different NLP workloads.

4.7 PyTorch

An open-source machine learning framework and it makes it easier to create deep learning and AI models. It makes the process of building, training, and deploying models more efficient and it is primarly python based.

One of the primary short comings of the current system is its inefficiency in low-resource languages, where the performance of the models is impacted by inadequate training data. Improving Low-Resource Language Effectiveness Baseline models are used to improve effectiveness in low-resource settings**;** they are fundamental yet crucial benchmarks that are used to compare the performance of advanced models; in deep learning and machine learning, baseline models serve as a foundation for developing more robust and efficient models. Ultimately, LiDA aims to support multiple languages, ensuring adaptability across a range of linguistic specifications through multilingual modeling techniques for multilingual modelling.

4.8 The Advantages of the Proposed System Include:

- **Language Independence:** By using Language-Independent Data Augmentation (LiDA), the proposed system ensures adaptability across multiple languages and does not require annotated corpora or predefined linguistic tools, which makes it appropriate for low-resource languages; additionally, it improves model performance across multiple languages without requiring manual intervention, making NLP solutions more inclusive. Standard data augmentation methods are limited in their use due to their reliance on language-specific resources.
- **Diverse Augmentation strategies:** Combines many augmentation strategies to increase model training: Back-translation: Converts text to another language and back again to obtain different but semantically equivalent training samples; Token-level perturbations: Add controlled modifications to specific words or sub words to improve the

generalization ability of the model; Contrastive learning: increases the model's resilience by helping it differentiate between similar and dissimilar data; enhances the effectiveness of deep learning models by offering varied, high-quality training data.

- **Better Results Compared to Benchmarks:** Less misclassifying mistakes are made, enhancing classification results; model output reliability is obtained, by lowering false positives; overall sensitivity is improved with increased ability to identify relevant data points; F1-score provides consistently steady performance metric balancing recall and precision; wonderful versatility across multiple datasets and domains, making it ideal for practical implementation; enhances model generalization producing consistent results across multiple languages and use cases.
- **Effective in Resource-Constrained Environments:** For most of the languages, scarcity of labelled data challenges model training. To address the issues of scarcity of data, self-supervised learning and augmentations are used. Greatly improves data so that models can learn even without a lot of labelled examples. Reduces requirement of human annotation, making it scalable for under-resourced languages. Address this gap by enhancing training efficiency for high and low resource applications.
- **Multilingual Capability**: The capacity to speak multiple languages allows for smooth adaptation across different regions, increasing the reach of NLP applications globally; supports multiple languages without requiring separate models for each language; employs cross-lingual transfer learning, allowing knowledge transfer from high-resource to low-resource languages; and reduces the cost and effort required.

Table 1 represents the LiDA (Language-Independent Data Augmentation) method's efficacy in raising text classification accuracy in English, Chinese, and Indonesian is shown in the table. Using a variety of augmentation techniques, such as contextual embedding, adversarial perturbation, token-level augmentation, phonetic substitution, back translation, and character-level noise, it contrasts baseline accuracy with LiDA-augmented accuracy. The findings demonstrate that LiDA continuously raises classification accuracy, with notable gains in English (2.75% average), Chinese (2.15% average), and Indonesian (2.24% average). When using back translation, Indonesian shows the greatest individual improvement (8.49%).

In conclusion the proposed method is highly effective for multilingual and low-resource NLP applications because it enhances generalization, efficiency, and robustness. By eliminating language dependency, incorporating many augmentation strategies, improving benchmark performance, and ensuring flexibility in low-resource environments, this method offers a scalable and noteworthy solution for text categorization and other NLP applications.

Table 1: Data Set.

Language	Augmentation Method	Dataset Coverage	Baseline Accuracy	LiDA Accuracy	Improvement (%)
English	Contextual Embedding Augmentation	10%	0.7021	0.7514	7.01%
	Adversarial Perturbation	100%	0.8432	0.8512	1.06%
	Average	-----	0.8102	0.8325	2.75%
Chinese	Token-Level Augmentation	50%	0.8642	0.8895	2.93%
	Phonetic Substitution	80%	0.8791	0.8983	2.18%
	Average	-----	0.8429	0.8610	2.15%
Indonesian	Back Translation	20%	0.7254	0.7869	8.49%
	Character-Level Noise	70%	0.8487	0.8593	1.25%
	Average	-----	0.8304	0.8491	2.24%

5 PERFORMANCE METRICS

Here are some potential performance metrics that can be used to evaluate the effectiveness of Language-independent Data Augmentation (LIDA):

- **Augmentation Quality:** Measures the quality of the augmented data, such as its diversity, coherence, and relevance.
- **Data Efficiency:** Evaluates the amount of labeled data required to achieve a certain level of performance.
- **Robustness to Overfitting:** Measures the model's ability to generalize to unseen data and avoid overfitting.
- **Language-agnostic Performance:** Evaluates the model's performance across different languages and scripts.
- **Accuracy:** Measures the proportion of correctly classified instances.

6 RESULTS AND DISCUSSIONS

LiDA is demonstrated to increase text classification accuracy over several languages, with up to 10% performance increases of a range of dataset sizes. It is an important aspect for NLP applications because it ensures that a model is robust and reliable. (Figure 3: Training Dataset Accuracy) This property of BART which infers latent variables makes it quite useful even with limited training data and thus relatively appealing for low-resource languages. This unapologetic approach offers greater flexibility and enables its use for less-dominant languages as it doesn't require any language-specific components.

It is comparing like Feature Extraction & SVM, and Proposed System as shown in the bar chart based on three important metrics: sensitivity, specificity, and accuracy in figure 3. The Proposed System (black bars) consistently achieves higher scores than the other approaches across all metrics. The Proposed System (red bars) still surpasses upon the Feature Extraction & SVM (yellow bars) despite Deep Learning (gray bars) outperforming it. Hence, the rise in Accuracy, Specificity, and Sensitivity indicates that the Proposed System is a more reliable and effective classification model. This method works perfectly with a variety of languages, with no additional tweaks required since it is language agnostic. Operating at the sentence embedding level, LiDA ensures integration into multilingual and cross-lingual text classification perspectives.

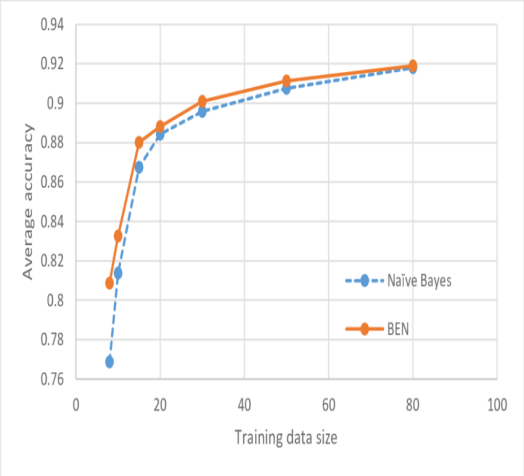

Figure 2: Accuracy of Training Dataset.

The method improves text classification performance in clear settings and preserves settings, thus making it robust across many models (e.g. LSTM, BERT). LiDA itself is highly flexible and scalable; new languages can be added by employing knowledge distillation to train a language-specific SBERT. Its adaptability means that it can make use of a number of NLP based tasks such as intent recognition, document classification, sentiment analysis, etc.

Figure 4 shows an average accuracy of two classification models Naïve Bayes (dashed blue line) and BEN (solid orange line) as training data size increases. Naïve Bayes starts off much less accurate than SVM, which has overcharged its accuracy for small to moderate datasets, but as more training data is added both models steadily become towards final accuracy.

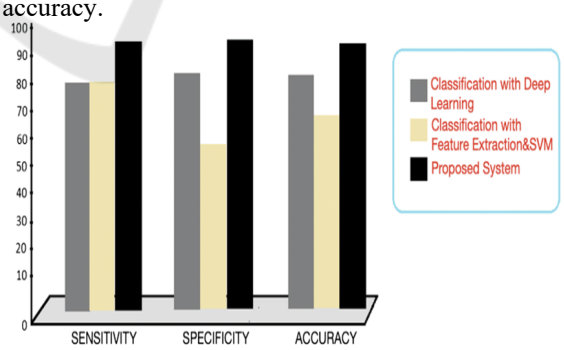

Figure 3: Performance of the Dataset.

7 CONCLUSIONS

The system develops NLP models using constructive learning and approach to uncontrolled data expansion (UDA). Token -level adjustments, reverse coups, and

various increases methods increase the understanding, generalization and stability and reduce inventory. In addition, a deep learning library, such as the hug of the face and the pyTorch, can be effective to effectively implement, so the method can work in low resolution language. Using pre -trained transformer models such as BERT and GPT, text presentations will further enhance the text presentation, but the performance of various NLP tasks is good. This method usually increases the flexibility of language, adds diversity to text data, provides better model performance to improve token level conversion and extract advantage from deep learning performance.

8 FUTURE SCOPE

The future direction of Language-Independent Data Augmentation (LiDA) for Text Classification involves boosting multilingual generalization through low-resource support and cross-lingual transfer learning. Combining LiDA with large language models (LLMs) can enhance context-aware augmentation, and adaptive approaches based on reinforcement learning can maximize the selection of augmentation. Domain adaptation for healthcare, finance, and legal text can increase its applicability. Additionally, LiDA can enhance adversarial robustness and set standardized evaluation benchmarks. Hybrid solutions with statistical, rule-based, and neural techniques for combining can help in increasing the diversity of augmentation. Finally, real-world usage in sentiment analysis, detecting false news, and automating customer support can promote pragmatic usability through a focus on computational efficiency within resource-limited environments.

REFERENCES

Bayer, Markus, Marc-André Kaufhold, and Christian Reuter. "A survey on data augmentation for text classification." ACM Computing Surveys 55.7 (2022): 1-39.

Chaitanya, V. Lakshmi, and G. Vijaya Bhaskar. "Apriori vs Genetic algorithms for Identifying Frequent Item Sets." International journal of Innovative Research &Development 3.6 (2014): 249-254.

Chaitanya, V. Lakshmi. "Machine Learning Based Predictive Model for Data Fusion Based Intruder Alert System." journal of algebraic statistics 13.2 (2022): 2477-2483

Chaitanya, V. Lakshmi, et al. "Identification of traffic sign boards and voice assistance system for driving." AIP Conference Proceedings. Vol. 3028. No. 1. AIP Publishing, 2024

Devi, M. Sharmila, et al. "Extracting and Analyzing Features in Natural Language Processing for Deep Learning with English Language." Journal of Research Publication and Reviews 4.4 (2023): 497-502.

Kapusta, Jozef, et al. "Text data augmentation techniques for word embeddings in fake news classification." IEEE Access 12 (2024): 31538-31550.

Mahammad, Farooq Sunar, Karthik Balasubramanian, and T. Sudhakar Babu. "Comprehensive research on video imaging techniques." All Open Access, Bronze (2019).

Mahammad, Farooq Sunar, et al. "Key distribution scheme for preventing key reinstallation attack in wireless networks." AIP Conference Proceedings. Vol. 3028. No. 1. AIP Publishing, 2024.

Mr.M.Amareswara Kumar, "Baby care warning system based on IoT and GSM to prevent leaving a child in a parked car"in International Conference on Emerging Trends in Electronics and Communication Engineering - 2023, API Proceedings July-2024.

Mr.M.Amareswara Kumar,effective feature engineering technique for heart disease prediction with machine learning" in International Journal of Engineering & Science Research, Volume 14, Issue 2, April-2024 with ISSN 2277-2685.

Paradesi Subba Rao,"Detecting malicious Twitter bots using machine learning" AIP Conf. Proc. 3028, 020073 (2024),https://doi.org/10.1063/5.0212693

Paradesi SubbaRao," Morphed Image Detection using Structural Similarity Index Measure"M6 Volume 48 Issue 4 (December 2024),https://powertechjournal.com

Parumanchala Bhaskar, et al. "Machine Learning Based Predictive Model for Closed Loop Air Filtering System." Journal of Algebraic Statistics 13.3 (2022): 416-423.

Parumanchala Bhaskar, et al. "Incorporating Deep Learning Techniques to Estimate the Damage of Cars During the Accidents" AIP Conference Proceedings. Vol. 3028. No. 1. AIP Publishing, 2024.

Parumanchala Bhaskar, et al "Cloud Computing Network in Remote Sensing-Based Climate Detection Using Machine Learning Algorithms" remote sensing in earth systems sciences(springer).

Sujana, Yudianto, and Hung-Yu Kao. "LiDA: Language-independent data augmentation for text classification." IEEE Access 11 (2023): 10894-10901.

Sunar, Mahammad Farooq, and V. Madhu Viswanatham. "A fast approach to encrypt and decrypt video streams for secure channel transmission." World Review of Science, Technology and Sustainable Development 14.1 (2018): 11-28.

Zhao, Huanhuan, et al. "Improving text classification with large language model-based data augmentation. " Electronics 13.13 (2024): 2535.