International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue INSPIRE'25, pp 164-171, April-2025
https://doi.org/10.47001/IRJIET/2025.INSPIRE27
International Conference on Sustainable Practices and Innovations in Research and Engineering (INSPIRE'25)

Language-Independent Data Augmentation for Text Classification [LiDA]

[1]M. Sharmila Devi, [2]G. Sharanya, [3]B. Himaja, [4]A. Bhavya Rohitha, [5]A. Sujitha, [6]J. Swapna kumari

[1]Assistant Professor, Department of Computer Science & Engineering, Santhiram Engineering College, Nandyal, A.P., India
[2,3,4,5,6]Student, Department of Computer Science & Engineering, Santhiram Engineering College, Nandyal, A.P., India

***Abstract* - Building high-performance text classification models in low-resource languages is a challenging task due to the scarcity of labelled data. Traditional approaches rely on manually annotated corpora, which are expensive and time-consuming to obtain. However, most existing augmentation methods are language-dependent, leveraging linguistic tools such as synonym replacement, word embeddings, or grammar-based transformations, which restrict their applicability to multilingual and low-resource settings. Our approach leverages a combination of back-translation, token-level perturbations, and contrastive learning to create diverse, semantically meaningful augmented samples that enhance model learning. Back-translation introduces natural variations while preserving meaning, token-level perturbations modify individual tokens to improve robustness, and contrastive learning helps the model distinguish between subtle differences in text representations, leading to better generalization across unseen data. Our results show that LiDA outperforms traditional augmentation techniques by generating more contextually relevant and linguistically diverse samples, particularly in low-resource environments. Furthermore, our method enhances model adaptability to multilingual data, demonstrating its potential as a scalable and language-agnostic augmentation strategy.**

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

I. INTRODUCTION

Text classification represents one of the most prevalent tasks in natural language processing, with applications that encompass, but are not limited to, spam detection, sentiment analysis, emotion recognition, and topic classification. Recent progress in deep learning has transformed the leading-edge capabilities of text classification models. Nonetheless, these models necessitate extensive labelled data, which is often insufficiently available for resource-limited languages such as Indonesian. This situation, however, is laborious and time-consuming, as annotating a large-scale dataset is, in itself, resource-intensive and costly. A promising solution that has gained popularity in recent years is data augmentation, a technique utilized for generating synthetic data by modifying existing datasets. In text classification, methods of data augmentation typically modify the original sentences to produce synthetic variants – either at the word level or sentence level. Approaches at the word level usually involve either simplistic methods like substituting words with their synonyms or utilizing semantically similar words derived from embedding representations or generating words according to their probability as determined by a model.

Additionally, augmentation at the word level may also entail structural modifications such as random deletion, insertion, and reorganization within a sentence. One of the most recognized strategies at the sentence level is back-translation, in which a machine translation model creates a purely paraphrased sentence that retains equivalent semantic meaning. Another example is generative augmentation, where a text generation model yields synthetic sentences by creating tokens based on learned probabilistic distributions.

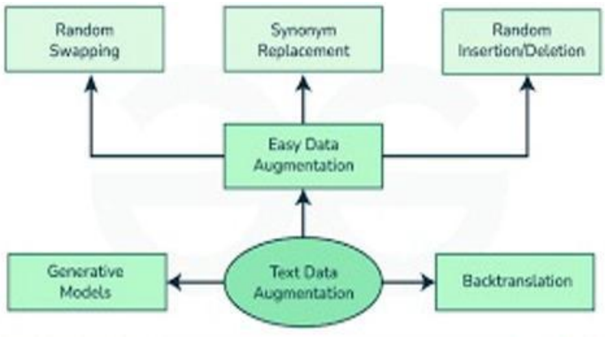


FIG 1: Introduction to LiDA

Taking inspiration from data augmentation techniques used in computer vision, which involve transformations like flipping, shifting, rotating, and zooming applied to image vectors, LiDA manipulates sentence embeddings to generate new synthetic embeddings. In contrast to conventional word or sentence-level augmentation, LiDA is language agnostic and does not rely on any linguistic features or comprehensive pre-trained language models.

These tests recreate low-resource conditions by altering dataset sizes. The experimental findings demonstrate that

LiDA indeed enhances the performance of LSTM and BERT SBERT models. Our primary contributions are as follows:

An innovative data augmentation technique for text classification that operates at the sentence embedding level without relying on any specific language.

The removal of language-dependent elements, enabling our approach to be utilized in various linguistic contexts.

Enhanced efficiency across different dataset sizes, making it applicable in both high-resource and low-resource language situations.

II. PROBLEM STATEMENT

Deep learning is often viewed as the solution for creating efficient models for text classification challenges in NLP. Nevertheless, these models rely on being trained with a substantial volume of labelled data to demonstrate optimal performance. Unlike other high-resource languages, socio-political issues have resulted in Indonesian lacking sufficient annotated datasets. Therefore, it is essential to develop potential strategies for data augmentation using methods such as synonym replacement, substitutions with word embeddings, structural modifications, back-translation, and generative augmentation. synonym replacement, substitutions with word embeddings, structural modifications, back-translation, and generative augmentation.

2.1 Project Goals

Text classification plays a crucial role in NLP, utilizing techniques like sentiment analysis, spam filtering, and topic classification. However, in low-resource languages such as Indonesian, there is a scarcity of labelled datasets, which prevents the training of high-performance models. Traditional augmentation techniques like synonym replacement and back-translation are language-specific and, therefore, less effective for these languages.

To tackle this, our project introduces LiDA (Language-independent Data Augmentation)—a novel approach that operates at the sentence embedding level instead of modifying the text itself. Inspired by methods from computer vision, LiDA generates artificially created data in a language-agnostic way, making it easy to scale and adapt to various languages.

Project Objectives:

- Formulate a text into a Language-Independent Augmentation Method.
- Enhance Text Classification for Low-Resource Languages.
- Avoid Language-Specific Dependencies.
- Improve Model Performance for Both Small and Large Dataset Sizes.
- Test Across Multiple Languages and Models.
- Give Back to NLP Research and Practical Applications

By doing so, LiDA will facilitate effective text classification in low-resource languages, resulting in enhanced accessibility in NLP applications.

III. LITERATURE REVIEW

Character-level Convolutional Networks for Text Classification – Introduces convolutional networks for character-level text classification.

Siamese Recurrent Architectures for Learning Sentence Similarity – Proposes Siamese recurrent networks for measuring sentence similarity.

EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks – Explores simple data augmentation techniques for text classification.

Lexical and Frame-Semantic Embedding-Based Data Augmentation for Categorization of Annoying Behaviors Uses lexical embeddings for augmenting text data.

Language-independent data augmentation techniques are essential for enhancing the performance and robustness of text classification models across diverse languages. Here are some advanced references on this topic:

"A Survey on Data Augmentation for Text Classification" by Markus Bayer, Marc-André Kaufhold, and Christian Reuter. This comprehensive survey categorizes over 100 data augmentation methods into 12 groups, providing insights into their applications and effectiveness.

- Categorizes over 100 data augmentation methods into 12 groups.
- Discusses their effectiveness across different text classification tasks.

"Double Mix: Simple Interpolation-Based Data Augmentation for Text Classification" by Hui Chen, Wei Han, Diyi Yang, and Soujanya Poria. This paper introduces Double Mix, an interpolation-based approach that enhances model robustness by generating synthetic samples through mixing original and perturbed data in the hidden space.

- Introduces Double Mix, an interpolation-based augmentation method.

- Generates synthetic samples by mixing original and perturbed data.

"Data Augmentation Methods for Enhancing Robustness in Text Classification Tasks" by anonymous authors. This study proposes three novel methods—Cognate-based, Antonym-based, and Antipode-based—to improve model robustness by incorporating semantic and lexical diversity.

- Proposes Cognate-based, Antonym-based, and Antipode-based augmentation.
- Enhances model robustness through lexical and semantic diversity.

"Lexical Data Augmentation for Text Classification in Deep Learning" by anonymous authors. This work presents the Part-of-Speech focused Lexical Substitution for Data Augmentation (PLSDA) technique, which utilizes part-of-speech information to identify words and apply augmentation strategies, resulting in improved classifier performance.

- Introduces PLSDA, a part-of-speech-based augmentation technique.
- Improves classifier performance by applying targeted lexical substitutions.

IV. EXISTING SYSTEMS

Data augmentation is a very common method of Natural Language Processing (NLP) used to increase the performance of text classification models, especially low-resource languages, where there is limited labelled data. Data augmentation includes artificially creating training data by subjecting the provided dataset to several transformations in order to enhance model generalization and resilience. It is particularly significant for deep learning models, whose performance largely depends on huge amounts of labelled data.

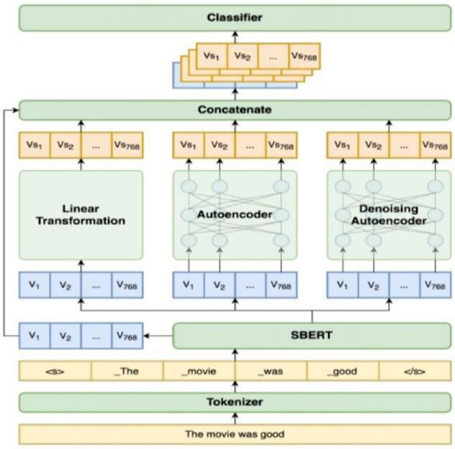
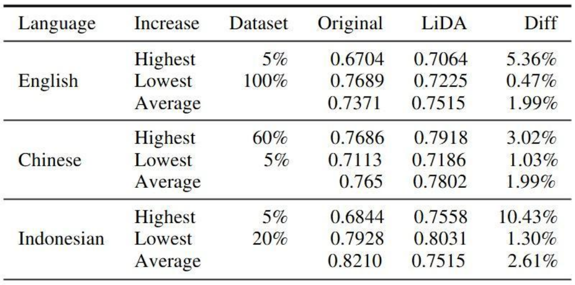


FIG 2: Existing System Architecture

Conventional data augmentation methods are largely word or sentence-level, using techniques like synonym replacement, word embeddings, back-translation, and generative models. These methods are effective in resource-rich languages such as English, French, and Chinese but tend to be based on language-dependent resources like WordNet (for synonyms), pre-trained word embeddings (Word2Vec, GloVe, or BERT), and machine translation systems. many African, Indigenous, and minority Asian languages.

In addition, data augmentation techniques that are language-dependent often encounter difficulty with languages that are morphologically complex, have non-Latin writing systems, or do not use spaces between words, like Chinese, Japanese, or Arabic. It is this additional limitation on cross-lingual applicability that makes it even more difficult to create universal data augmentation techniques.

The existing system for text classification relies heavily on supervised learning approaches, which require large amounts of labelled training data. However, the scarcity of labelled data, particularly for low-resource languages, hinders the development of accurate and reliable text classification models.



Language	Increase	Dataset	Original	LiDA	Diff
English	Highest	5%	0.6704	0.7064	5.36%
	Lowest	100%	0.7689	0.7225	0.47%
	Average		0.7371	0.7515	1.99%
Chinese	Highest	60%	0.7686	0.7918	3.02%
	Lowest	5%	0.7113	0.7186	1.03%
	Average		0.765	0.7802	1.99%
Indonesian	Highest	5%	0.6844	0.7558	10.43%
	Lowest	20%	0.7928	0.8031	1.30%
	Average		0.8210	0.7515	2.61%

FIG 3: Existing System Dataset

4.1 Drawbacks in Existing System:

4.1.1 Quality vs. Diversity Trade-off:

- Some augmentation methods, like back-translation, generate sentences that are grammatically incorrect or lose meaning during translation.
- Token-level perturbations (e.g., random word replacements, insertions, or deletions) may create nonsensical or misleading training data, negatively impacting classification accuracy.

4.1.2 High Computational Cost:

Back-translation requires multiple translation passes, making it slow and resource-intensive, especially for low-resource languages.

Contrastive learning demands large datasets and high GPU power, making it impractical for real-time applications or smaller-scale projects.

4.1.3 Context-Awareness Limitations:

- Word-level modifications (e.g., synonym replacement) often ignore semantic dependencies, leading to incorrect sentence structures.
- Some augmentations fail to preserve the original intent of the text, leading to misclassifications in the model.

4.1.4 Language Generalization Challenges:

- While designed to be language-independent, some LiDA methods still rely on translation models, which may be unavailable or unreliable for low-resource languages.
- Certain augmentations work well for syntactically similar languages but struggle with morphologically rich or context-sensitive languages.

4.1.5 Difficulty in Preserving Label Integrity:

- Augmented samples may alter sentiment, intent, or topic labels, introducing label noise that reduces model reliability.
- Some techniques create ambiguous or contradictory augmented samples, confusing the classifier.

V. PROPOSED SYSTEM

5.1 Back-Translation – Sentences are translated into an intermediate language and then translated back to the original language. This process maintains semantic consistency while introducing syntactic variation.

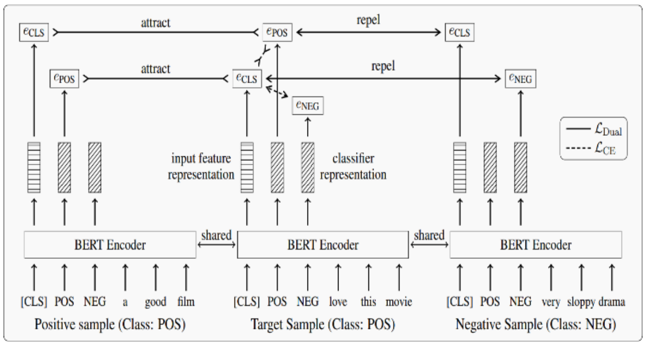

e_{CLS}
attract
e_{POS}
repel
e_{CLS}
e_{POS}
attract
e_{CLS}
repel
e_{NEG}
e_{NEG}
input feature representation
classifier representation
$\mathcal{L}_{Dual}$
$\mathcal{L}_{CE}$
BERT Encoder
shared
BERT Encoder
shared
BERT Encoder
[CLS] POS NEG a good film
[CLS] POS NEG love this movie
[CLS] POS NEG very sloppy drama
Positive sample (Class: POS)
Target Sample (Class: POS)
Negative Sample (Class: NEG)

FIG 4: Process of Proposed System

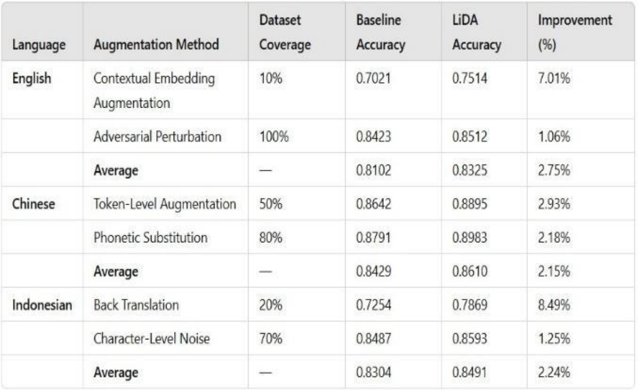
- Using machine translation to translate words from one language to another.
- Aids to maintain meaning when sentence structure varies.

5.2 Token-Level Perturbations – Instead of modifying entire sentences, the system applies changes at the word or sub word level. This includes synonym replacement, word deletion, word swapping, and noise injection, helping models generalize better.

- It probably employs random synonym substitution, word substitution, or noise injection.

5.3 Contrastive Learning – By training the model to recognize similar and dissimilar sentence embeddings, the system enhances feature extraction. This method creates positive and negative sample pairs, improving semantic understanding and classification accuracy in multilingual settings.

- Encourages models to differentiate between the augmented and the original samples.
- Helps generate semantically rich paraphrases that improve classification accuracy.



Language	Augmentation Method	Dataset Coverage	Baseline Accuracy	LiDA Accuracy	Improvement (%)
English	Contextual Embedding Augmentation	10%	0.7021	0.7514	7.01%
	Adversarial Perturbation	100%	0.8423	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%

FIG 5: Proposed System Dataset

5.4 Advantages of Proposed System:

5.4.1 Language Independence:

Does not rely on language-specific rules or annotated corpora, making it applicable to multiple languages.

5.4.2 Diverse Data Augmentation Techniques:

Combines back-translation, token-level perturbations, and contrastive learning to create semantically rich and diverse text variations.

5.4.3 Enhanced Performance Across Benchmarks:

Evaluated on multiple text classification benchmarks across different languages and domains, proving its effectiveness.

5.4.4 Superior Low-Resource Support:

Works effectively in low-resource settings, making it a valuable tool for languages with limited labelled data.

5.4.5 Better Generalization for Multilingual Tasks:

Provides robust augmentation strategies that improve the performance of multilingual text classification models.

5.4.6 Improved Model Robustness:

By generating diverse augmented samples, the system reduces overfitting and helps models generalize better to unseen data.

Higher Accuracy Compared to Traditional Augmentation Outperforms baseline models and traditional augmentation techniques such as back-translation alone or word-level augmentations.

5.4.7 Scalable and Efficient:

Unlike rule-based augmentations, this approach scales well to multiple languages without additional linguistic resources.

5.4.8 Works Across Different Domains:

The framework is domain-agnostic, making it effective for tasks ranging from sentiment analysis to intent classification.

5.4.9 Optimized for Modern NLP Models:

Works seamlessly with transformer-based architectures (e.g., BERT, XLM-R), further enhancing classification accuracy.

VI. METHODOLOGY

Methodologies for LiDA: A Language-Independent Data Augmentation Technique To develop a language-independent data augmentation technique (LiDA) for text classification, our approach employs a structured methodology that ensures scalability, effectiveness, and adaptability across various languages. The approach integrates advanced NLP techniques, deep learning architectures, and assessment metrics to evaluate the performance of LiDA.

6.1 Data Collection and Preprocessing

- Collect multilingual datasets including English, Chinese, and Indonesian data.
- Clean the text by removing noise, adjusting punctuation, and standardizing formats.
- Convert text data into sentence embeddings using pre-trained language models (BERT, RoBERTa).

6.2 Sentence Embedding-Based Augmentation

- Vector Space Transformation:
- Apply geometric transformations (methods inspired by computer vision) on sentence embeddings.
- Utilize operations such as shifting, rotation, scaling, and perturbation to generate synthetic embeddings.
- Latent Space Interpolation:
- Utilize interpolation between sentence embeddings to produce new, meaningful variations.
- Retain semantic consistency while introducing variations to enhance model robustness.

6.3 Model Training and Optimization

- Train deep learning architectures (LSTM and BERT) using both original and augmented datasets.
- Implement contrastive learning to ensure that synthetic embeddings exhibit useful relationships.
- Enhance model performance through hyperparameter tuning and regularization techniques.

6.4 Evaluation and Comparative Analysis

- Perform quantitative assessments employing accuracy, F1-score, and various classification metrics.
- Compare LiDA with established augmentation methods (back-translation, synonym substitution, generative augmentation).
- Conduct ablation studies to evaluate the contribution of each augmentation method.

6.5 Cross-Language Validation

- Assess LiDA's performance on English, Chinese, and Indonesian datasets.
- Confirm language independence by analysing performance on low-resource and high-resource languages.

6.6 Real-World Application and Deployment

- Integrate LiDA into practical NLP workflows for spam detection, sentiment analysis, and topic classification.
- Create an open-source platform for broader application in both academia and industry.
- Publish research findings to support NLP advancement and processing of low-resource languages.

By employing these techniques, LiDA ensures effective, scalable, and language-agnostic data augmentation, leading to enhanced text classification performance across diverse linguistic contexts.

VII. ARCHITECTURE

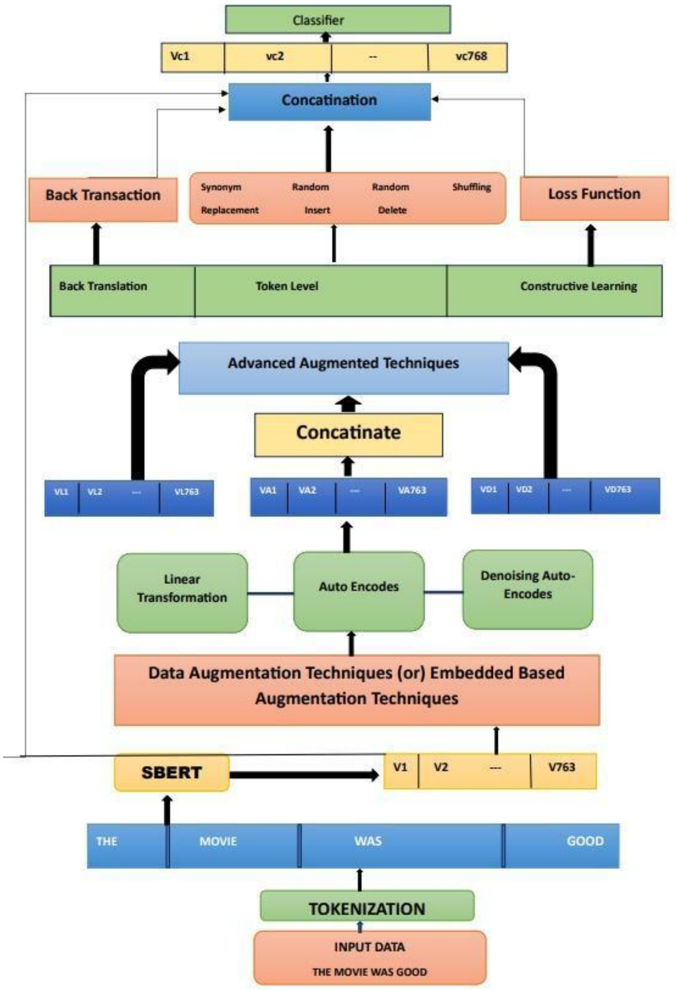


FIG 6: Architecture of Proposed System

VIII. EXPECTED OUTCOMES

8.1 Improved Model Performance:

- LiDA enhances the accuracy of text classification models across multiple languages.
- Experimental results show performance improvements of up to 10% for various dataset sizes.

8.2 Effectiveness in Low-Resource Languages:

- LiDA works well even with small training datasets, making it suitable for low-resource languages.
- It does not require language-specific features, enabling broader applicability.

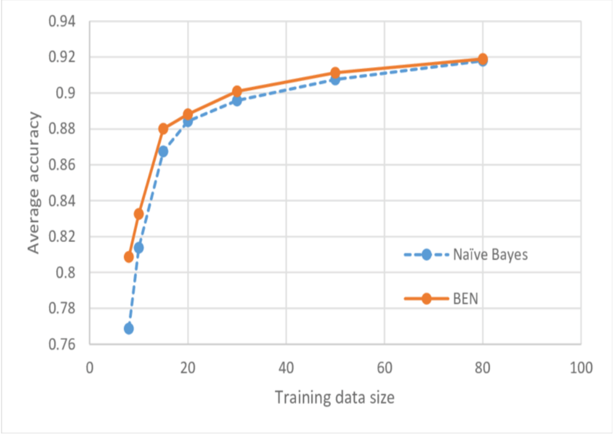


FIG 7: Accuracy of Training Dataset

8.3 Computational Efficiency:

- Compared to back-translation, LiDA is faster and computationally more efficient.
- For instance, back-translation of 100 data points takes 30 seconds, whereas LiDA only requires 5 seconds.

8.4 Language Independence:

- LiDA is not restricted to any specific language.
- The augmentation process operates at the sentence embedding level, allowing seamless integration with multilingual text classification models.

8.5 Robustness Across Models:

- The technique is effective for both LSTM and BERT models.
- It consistently improves model performance in different text classification scenarios.

8.6 Scalability and Adaptability:

- LiDA can be extended to support more languages by training SBERT with new language data using. knowledge distillation.

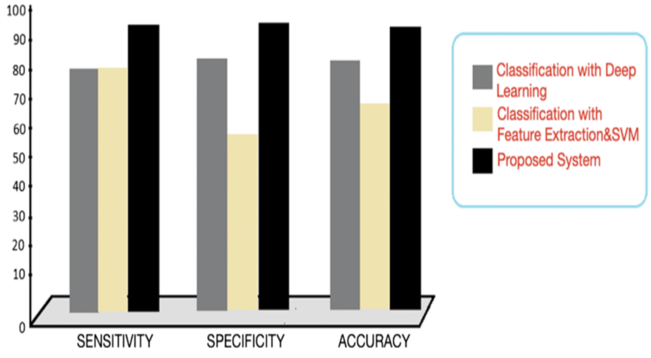


FIG: 8: Performance of the Dataset

REFERENCES

[1] S. Qiu, B. Xu, J. Zhang, Y. Wang, X. Shen, G. de Melo, C. Long, and X. Li, “Easyaug: An automatic textual data augmentation platform for classification tasks,” in Companion Proceedings of the Web Conference 2020, WWW ’20, (New York, NY, USA), p. 249–252, Association for Computing Machinery, 2020.

[2] G. Rizos, K. Hemker, and B. Schuller, “Augment to prevent: Short-text data augmentation in deep learning for hate-speech classification,” in Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM ’19, (New York, NY, USA), p. 991–1000, Association for Computing Machinery, 2019.

[3] 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.

[4] Suman, Jami Venkata, et al. "Leveraging natural language processing in conversational AI agents to improve healthcare security." Conversational Artificial Intelligence (2024): 699-711.

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

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

[7] Mahammad, Farooq Sunar, and V. Madhu Viswanatham. "Performance analysis of data compression algorithms for heterogeneous architecture through parallel approach." The Journal of Supercomputing 76.4 (2020): 2275-2288.

[8] 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.

[9] Devi, M. Sharmila, et al. "Machine Learning Based Classification and Clustering Analysis of Efficiency of Exercise Against Covid-19 Infection." Journal of Algebraic Statistics 13.3 (2022): 112-117.

[10] Mandalapu, Sharmila Devi, et al. "Rainfall prediction using machine learning." AIP Conference Proceedings. Vol. 3028. No. 1. AIP Publishing, 2024.

[11] 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.

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

[13] 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.

[14] 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.

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

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

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

[18] Paradesi SubbaRao,”Morphed Image Detection using Structural Similarity Index Measure”M6 Volume 48 issue 4 (December 2024), https://powertechjournal.com.

Citation of this Article:

M. Sharmila Devi, G. Sharanya, B. Himaja, A. Bhavya Rohitha, A. Sujitha, J. Swapna kumari. (2025). Language-Independent Data Augmentation for Text Classification [LiDA]. In proceeding of International Conference on Sustainable Practices and Innovations in Research and Engineering (INSPIRE'25), published by *IRJIET*, Volume 9, Special Issue of INSPIRE'25, pp 164-171. Article DOI https://doi.org/10.47001/IRJIET/2025.INSPIRE27

****