Boosting Performance of NLU Tasks for Indic Languages using Data Augmentation Strategies

Shubhankar Kuranagatti
Department of Computer Science
PES2UG20CS466
PES University
Bangalore, India
shubhankarvk@gmail.com

Jayendra Ganesh Devisetti
Department of Computer Science
PES2UG20CS511
PES University
Bangalore, India
jayendraganesh21@gmail.com

Sanjana S
Department of Computer Science
PES2UG20CS549
PES University
Bangalore, India
sanjanasrinidhi1810@gmail.com

***Abstract*—The lack of sufficient resources and linguistic complexity of Indic languages pose significant challenges to Natural Language Understanding (NLU) tasks such as sentiment analysis and named entity recognition. As a result, the performance of NLP models for Indic languages has been suboptimal, limiting their application in various real-world scenarios. To address this issue, this research paper proposes an approach to enhance the performance of NLP models for Indic languages using data augmentation strategies. The paper explores various data augmentation techniques such as back-translation, synonym replacement, and random insertion of words to increase the amount of training data for the languages Hindi, Marathi and Kannada. The effectiveness of these strategies is evaluated on benchmark datasets for text classification task, demonstrating a significant reduction in overfitting compared to baseline models that use limited training data. The results of our work show that augmentation techniques have been excpetionally successful in reducing the overfitting, that was present in the models that were trained of lesser data.**

Keywords - Augmentation, Natural Language Understanding, Word Embeddings, Deep Learning, Backtransaltion, BERT

I. Introduction

The development of Natural Language Processing (NLP) in resource-rich languages like English has been rapid and has seen significant advancements in recent years. With large datasets, abundant language resources, and the availability of advanced machine learning algorithms, NLP models for English have achieved impressive results in various applications such as language translation, sentiment analysis, and text classification. However, the same cannot be said for low-resource languages such as Indic languages, where the lack of language resources and the linguistic complexity pose significant challenges to NLP development.

The impact of the lack of resources for Indic languages on Natural Language Understanding (NLU) tasks such as text classification, sentiment analysis and named entity recognition cannot be overstated. Indic languages, such as Hindi, Marathi, Kannada have a significant presence in South Asia and are spoken by millions of people worldwide. However, the lack of standardization and resources for these languages poses significant challenges for NLU tasks. For instance, there is a limited availability of annotated datasets for training NLU models, which is crucial for achieving high accuracy in classification tasks. Additionally, the linguistic complexity of Indic languages, including the vast vocabulary, grammar rules, and sentence structure, makes it challenging to develop accurate NLU models.

The scarcity of language resources such as lexicons, corpora, and annotated data makes it difficult to develop effective machine learning models that can perform well in real-world scenarios. This limitation is especially pronounced for low-resource Indic languages, where the lack of data resources limits the development of high-performing NLU models. As a result, NLU models for Indic languages often underperform compared to models trained on resource-rich languages, leading to low accuracy and precision.

The lack of resources for Indic languages also affects the development of language technologies such as machine translation, speech recognition, and information retrieval. The challenges associated with the development of these technologies for Indic languages limit their use in various real-world applications such as e-commerce, customer service, and healthcare, to name a few.

The deficit of resources for Indic languages poses significant challenges to the development of accurate and effective NLU models. This limitation not only affects the performance of NLU models for Indic languages but also has far-reaching implications for the development of language technologies for these languages. Therefore, addressing the issue of resource scarcity for Indic languages is crucial for the development of effective NLU models and the advancement of language technologies.

Key contributions of this work include:

- Collection of datasets in Hindi, Marathi and Kannada.
- Augmetation of datasets using BackTranslation, Synonym substitution, Random insertion, Random Deletion, Swapping and Transliteration.
- Representation of words as contextual word embeddings
- Text classification using Transformer models such as Multiligual BERT and GPT-2.
- Comparitive study on performace on downstream task of text classification with and without augmentation.

II. Related Work

The field of natural language processing has garnered recent attention in low resource languages. Xiang Dai et al. explored the impact of simple data augmentation techniques on Named Entity Recognition (NER) tasks. The authors conducted experiments on three NER datasets and evaluate four simple data augmentation techniques, namely, synonym substitution, random insertion, random deletion, and random swap. They train and test three different NER models, including a BiLSTM-CRF model, a BERT-based model, and a combination of the two. Furthermore, the authors explore the impact of combining multiple data augmentation techniques on the performance of NER models. Their results demonstrate that simple data augmentation techniques can significantly improve the performance of NER models on all three datasets. The results showed that the synonym substitution technique consistently outperforms the other data augmentation techniques on all three datasets, resulting in a significant improvement in F1-score for all models and datasets. In the CoNLL-2003 dataset, the BiLSTM-CRF model's F1-score improved from 90.84% to 91.23% with the use of synonym substitution, a relative improvement of 0.43%. The random insertion, random deletion, and random swap techniques also led to improvements in the models' performance, although their effectiveness varied depending on the dataset and the type of model used. Their results show that combining multiple techniques can lead to further improvements in the models' performance. Sam Shleifer proposed a method that was divided into three main steps: 1. Pre-training ULMFit, 2. Backtranslation, 3. Fine-tuning the classifier. The proposed method was evaluated on three low-resource text classification datasets: AGNews, Amazon reviews and SMS spam collection and the results showed that it outperformed the other methods on all three datasets. For the AG News dataset, the proposed method achieved an accuracy of 93.3%, which was 2.8% higher than the classifier trained only on the labeled data. For the Amazon Reviews dataset, the proposed method achieved an accuracy of 92.7%, which was 1.5% higher than the classifier trained only on the labeled data. For the SMS Spam Collection dataset, the proposed method achieved an accuracy of 98.6%, which was 1.2% higher than the classifier trained only on the labeled data. Yufei Wang et al. proposed a novel data augmentation method called PromDA which comprised of constructing a set of prompts based on the input text using a combination of rule-based and data-driven methods and using them to generate new training samples. On the GLUE benchmark, which consists of nine different NLU tasks, the PromDA method improved the performance of the baseline model by an average of 1.43% in terms of the overall accuracy score. The PromDA method also outperformed other data augmentation methods, such as back-translation and word substitution, on seven out of the nine tasks. On the SuperGLUE benchmark, which consists of eight challenging NLU tasks, the PromDA method achieved a significant improvement over the baseline model on four tasks, including WiC, BoolQ, COPA, and ReCoRD. Fadaee et al. conducted a research study on the efficacy of data augmentation (DA) methods for improving text classification performance in low-data scenarios. In their study, they used three DA techniques as baselines: EDA, Backtranslation, and CBERT, and evaluated them on three text classification datasets (SST-2, SNIPS, and TREC) under low-data regime conditions with 10 and 50 examples per class. The authors performed both intrinsic and extrinsic evaluations, and the experiments were repeated 15 times to ensure consistency. Intrinsic evaluation involved measuring the semantic fidelity of generated text, while extrinsic evaluation involved adding generated examples to the training data and evaluating performance on the full test set. For fine-tuning the classifier on each task, the authors employed a pretrained English BERT-base uncased model.The results of their study indicate that pre-trained models like BART outperform other data augmentation techniques in terms of classification performance and data fidelity. Back-translation was found to be the most effective technique for generating diverse data, while paraphrasing methods like EDA were more effective for preserving the semantics of the input text. Linlin Liu et al. have proposed a framework for cross-lingual named entity recognition (NER) that incorporates both instance-based and model-based transfer for data augmentation. This framework utilizes a labeled sequence translation approach to translate annotated training data from the source language to a range of target languages. Subsequently, language models are trained on the translated data to generate multilingual synthetic data, which is further post-processed and filtered to train multilingual NER models for inference on the target-language test sets.The proposed framework improves on prior methods by replacing named entities with contextual placeholders before sentence translation. This is followed by the replacement of placeholders in translated sentences with their corresponding translated entities. The authors also provide examples of labeled sequence linearization and the training of a multilingual LSTM-LM on the linearized sequences.The results of the study indicate that the proposed method surpasses the best baseline method by 2.90 and 2.97 on German and Dutch, respectively, and by 2.23 on average. The authors further demonstrate that their method is effective in generating high-quality labeled data in the target language and that multilingual translation may help enhance the cross-lingual transfer performance of multilingual LM's in low-resource scenarios. Wu et al. have presented a novel approach to improve low-resource neural machine translation through the utilization of BERT, a masked language model (MLM). In their study, they employed a one-hot encoding of text and label as the input dataset, which was fed into BERT. They retrieved the output of the last layer of the transformer encoder, which was multiplied by the word embedding matrix to obtain the MLM prediction results. These results represent the context-compatible token choices for each position in the input text. The study found that text smoothing was the most effective data augmentation method, producing an average improvement of 11.62% across three datasets compared to training without

augmentation. Text smoothing outperformed the previously best-performing method, BARTspan, by an average of 1.17%. Ding et al. have conducted a research primiarily to improve sequence tagging tasks. This approach involves linearizing labeled sentences and training a language model to learn the distribution of words and tags in the linearized sequences, thereby generating synthetic training data. It uses special tokens like [BOS] and [EOS] for marking sentence boundaries, while a one-layer LSTM recurrent neural network language model is employed. Furthermore, the authors have proposed a conditional generation technique that utilizes unlabeled data and knowledge bases, where available. Their study shows that the generated data introduces more diversity, which helps to reduce overfitting and improve overall performance.

III. Proposed Approach

A. Data Collection and Preprocessing

Data collection involves gathering relevant information and data points that are used to support research, analysis, decision-making, and problem-solving. Proper data collection is essential to ensure the accuracy, reliability, and validity of the data. For evaluating the performance downstream task in text classification with and without augmentation techniques, three publicly available datasets were collected from Kaggle, each for Hindi, Marathi and Kannada. The three datasets composed of an attribute for text and an attribute indicating the category of the text such as sports, entertainment and technology. The text comprised news articles sourced from that particular language.

In order to guarantee that the data being examined is reliable and uniform, data cleaning is a crucial component of data preprocessing. Pre-processing the data can help to extract relevant features and information that are essential for training the ML or DL models. Preprocessing techniques we perfomed on the datasets included techniques such as tokenization, label encoding and validation split.

The data has been cleaned and preprocessed by carrying out the following steps:

- We have eliminated duplicated instances to reduce noise in the data.
- Punctuations were removed from the datasets.
- Since the values of the labels were categorical, scikit-learn's label encoder was used to map the categorical data to integer values.
- Post cleaning, the data was split into training and validation set in the ratio of 80:20 to assess the performance of the models on unseen data.

B. Data Visualization

To identify trends and patterns in the data for an accurate analysis, data visualization was carried out on all three datasets. The number of rows belonging to each category was plotted for understanding of the distribution of data across categories that can reveal any imbalances or biases in the datasets. Across the Kannada news training set, samples belonging to the category entertainment were in the highest number with 2710 instances of this category, followed by 1856 instances of type sports and 601 instance of type technology. In the Hindi training set, the category of India had the highest number of instances with 1390 occurrences. This was followed by the "International" category, which had 904 instances, and the "Entertainment" category, which had the least number of samples with 285 instances. Similarly, it was observed that within the Marathi training dataset, the category with the highest count of samples was state with a total of 2103 instances. Following closely behind were the categories of entertainment and sports with counts of 1207 and 700 instances, respectively. In addition to the distribution of data, we also created a word cloud to display the frequency of words in a text corpus. This helped in identifying the most commonly used words and phrases in the data, providing insights into the underlying themes and topics. The distribution of sentence length was analysed by plotting the number of words in each sentence to reveal any outliers or unusual data points that may need further investigation. For Kannada, it was observed that the number of words ranged from 2 to 17 words per sentence for the training set.For Hindi news, it was observed that the number of words primarily ranged between 7 and 25 words per sentence for the training set. Similarly it was observed that number of words in Marathi ranged from 5 to 20.

C. Data Augmentation Techniques

Performing data augmentation on low resource languages such as Kannada, Hindi, and Marathi was done to improve the performance of the intelligent models on downstream NLP tasks like text classification. Data augmentation involves generating new training data by applying various techniques to the existing data, such as adding noise, changing the order of words, or replacing words with synonyms. In low resource languages, where the amount of available training data is limited, data augmentation has been particularly useful in improving the model's accuracy. The data augmentation techniques employed in our study for all three languages included:

- Random Insertion: Random insertion primarily involved replacing a word in the sentence with its synoyms leading to a new sentence similar to the source sentence. The probability of selecting a word was set to 0.3 implying that the probability of selecting a word is 0.3. Once the word is selected, the synsets are extracted using the iywin package. The replacing word is selected from this pool of words and a new sentence is thus formed.
- Random Deletion: Random deletion involved randomly deleting from a sentence to create a new variant of the original text. It involved using a probability parameter, p to control the likelihood of each word being deleted. In our work, the probability parameter was set to 0.1 implying that each word had a probability of 0.1 for being deleted from that sentence. The function generated a random number between (0, 1) for each word and if the number was less than p, the word was deleted from the sentence.

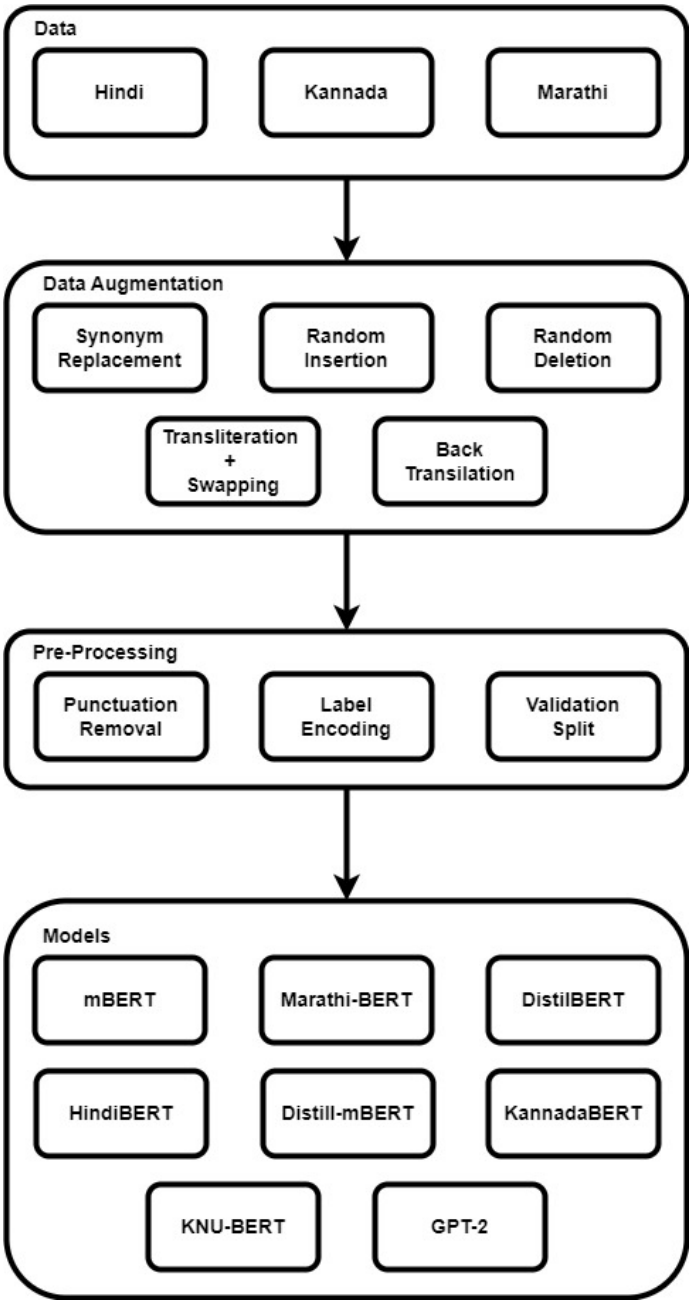

Fig. 1: Overview of the architecture

- Synonym Substitution with Backtranslation: Synonym Substitution Technique involved the translation of sentences from the source language into English, followed by the utilization of a text attack library known as WordNet-based Synonym Text Attack. This library uses WordNet, a lexical database of English, to identify synonyms for words present in a text document. The closest synonym, based on semantic similarity between the words, is then selected to replace the original word. After the substitution process, the resulting text is translated back into the source language and subsequently added to a database. It is a formal methodology that employs linguistic and computational techniques to facilitate text modification.
- Shuffle with Transliteration: Transliteration involved converting sentences in Kannada to Devanagari script, randomly shuffling the words and then converting it back to Kannada. This technique is especially useful for augmenting low-resource languages that lack sufficient training data, as it can increase the size and diversity of the training set.
- Backtranslation: The Back Translation Technique generates diverse and grammatically correct sentences by translating a sentence from one language to another and back to the original language. In this approach, the source

language is translated to English, French, German, and then translated back to the original language, and repeated sentences are removed.

D. Training

In this study, we have utilized various transformer based pre-trained deep learning models after finetuning them to perform the downstream task of text classification for evaluation of performance of NLP tasks with and without data augmentation. The pipeline followed in this study is given in Fig. 1.

Deep learning models have become increasingly popular in natural language processing tasks because of their ability to learn from large amounts of data and extract complex patterns and relationships between words and sentences. However, they require large amount of dataset to perform well. Hence, in this task, we have used several deep learning models to perform text classification on datasets for three languages, before and after augmentation. The deep-learning models in our study are:

- mBERT:MBERT, short for Multilingual Bidirectional Encoder Representations from Transformers, is a pre-trained language model developed by Google. It is trained on a massive corpus of text in over 100 languages, making it capable of performing various natural language processing (NLP) tasks such as text classification, question-answering, and named entity recognition in different languages.In our work we use our dataset to fine-tune the classification layer of the model for our down-stream tasks.
- Distil-mBERT:Distil-mBERT is a neural language model developed by Hugging Face, which is a smaller and more efficient version of the widely used mBERT model. It incorporates knowledge distillation methodology to create a smaller model from the larger BERT model, while preserving its overall accuracy. To adapt it for our specific use case, we fine-tuned the Distil-mBERT model on our datasets by fine-tuning the classification layer, and subsequently applied it to our classification downstream tasks.
- HindiBERT: HindiBERT is a language model based on BERT architecture that has been created and trained exclusively on Hindi news data. Its training on Hindi-specific data enables it to offer better results than generic language models when utilized for tasks involving the Hindi language. To cater to our specific requirements, we refined the classification layer of the HindiBERT model by fine-tuning it on our datasets. Afterward, we employed it for our downstream classification task.
- MarathiBERT:MarathiBERT is a BERT-based language model that has been developed and trained specifically on Marathi language data. By being trained on domain-specific Marathi data, it has the ability to provide improved results in comparison to generic language models when used for Marathi language tasks. To tailor it to our specific needs, we fine-tuned the MarathiBERT model on our datasets by refining the classification layer. Subsequently, we deployed it for our classification downstream tasks.
- KannadaBERT: KannadaBERT is a language model based on the popular BERT architecture pre-trained on a large amount of Kannada text data and fine-tuned on several downstream NLP tasks, such as text classification, named entity recognition, and question answering. In our work, we used KannadaBERT by fine-tuning the model for our Kannada dataset, producing word embeddings and passing them to a regression layer for classification.
- KNU-BERT: KNU-BERT is another language model like KannadaBERT, which has been trained on Kannada articles. We used KNUBERT to obtain word embeddings and added a softmax layer for classification of the news articles in Kannada.
- GPT-2: GPT-2 is a state-of-the-art language model which was not pre-trained on Kannada language but was fine-tuned in our work through training on our Kannada dataset. It used decoder only architechture with feed-forward network for text classification.

E. Testing

Two metrics, accuracy and loss curve, have been primarily considered for evaluation of the models Accuracy is indicative of the overall performance of the model and determines how well a model can identify the relationship between the independent variable and the target variable.

$$Accuracy = \frac{TP + FP}{TP + FP + TN + FN}$$

The loss curve is a useful representation of how well a deep learning model is fitting to the training data during the training process, used to evaluate the performance of the model for text classification.

IV. Results and Discussion

A. Model Evaluation

In this study, we have employed multiple augmentation techniques to improve the performance of three different models for each of the Hindi, Kannada, and Marathi datasets. After applying these techniques, we evaluated the performance of each model on its respective dataset. The results of this evaluation have been presented in Tables II, III, and IV, which detail the performance of the models on the Kannada, Hindi, and Marathi datasets, respectively.

The table of results clearly demonstrates the effectiveness of applying augmentation techniques to the dataset. In each case, we observe a noticeable increase in accuracy for the respective models when compared to non-augmented data. The loss curve further supports this finding, showing a reduction in overfitting and a more well-fitting curve when augmentation is applied. Of all the augmentation techniques tested, back translation proved to be particularly effective, leading to a significant spike in accuracy across all three languages.

TABLE I: Classification results for Kannada Dataset

Augmentation Techniques	Models with Accuracy values		
	KannadaBERT	**KNU-BERT**	**GPT-2**
Non Augmented	0.97	0.95	0.73
Random Insertion	0.92	0.94	0.80
Random Deletion	0.97	0.93	0.77
Synonym Substitution with Backtranslation	0.97	0.95	0.76
Shuffle with Transliteration	0.96	0.94	0.81
Backtranslation	0.97	0.95	0.75

TABLE II: Classification results for Hindi Dataset

Augmentation Techniques	Models with Accuracy values		
	HindiBERT	**DistilmBERT**	**mBERT**
Non Augmented	0.85	0.74	0.55
Random Insertion	0.86	0.70	0.56
Random Deletion	0.78	0.72	0.56
Synonym Substitution with Backtranslation	0.86	0.75	0.57
Backtranslation	0.86	0.75	0.57

TABLE III: Classification results for Marathi Dataset

Augmentation Techniques	Models with Accuracy values		
	MarathiBERT	**DistilmBERT**	**mBERT**
Non Augmented	0.91	0.88	0.86
Random Insertion	0.95	0.90	0.89
Random Deletion	0.89	0.87	0.88
Synonym Substitution with Backtranslation	0.94	0.91	0.89
Backtranslation	0.95	0.90	0.89

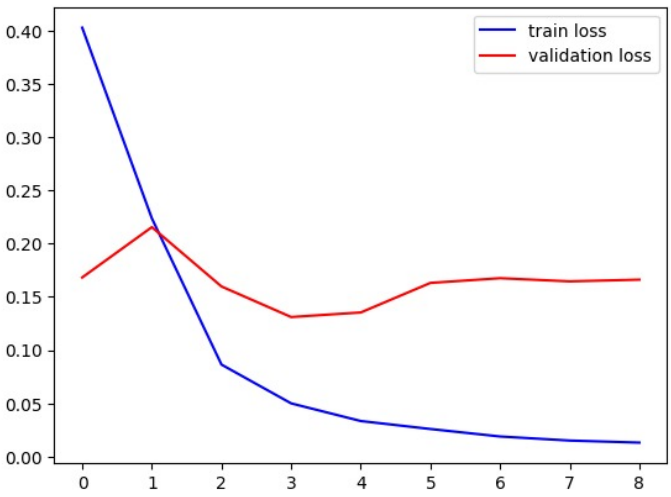

Fig. 2: Loss Curve for Non-Augmented Data

V. Conclusion

In this study, we have proposed various data augmentation techniques to improve the performance of low-resource languages, Hindi, Marathi and Kannada on text classification task.

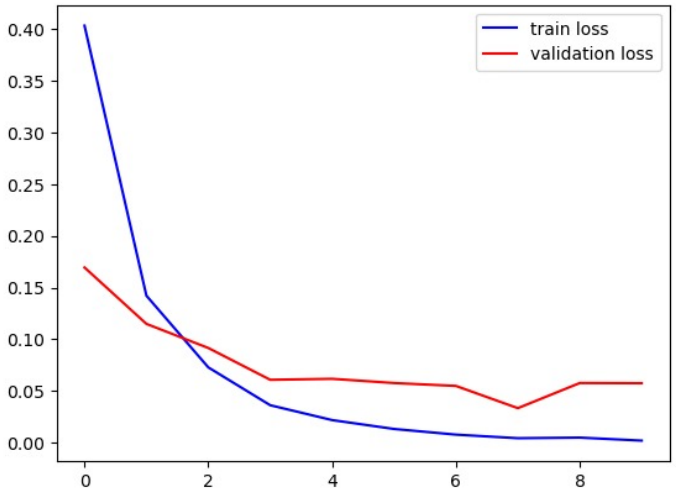

Fig. 3: Loss Curve for Data Augmented with Random Insertion

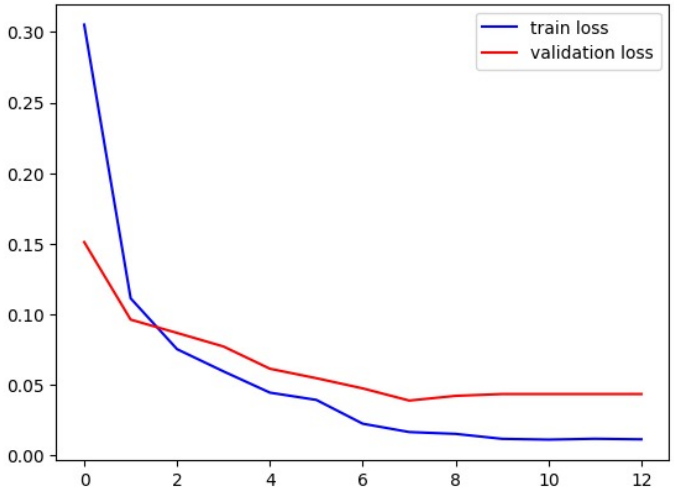

Fig. 4: Loss Curve for Data Augmented with Random Deletion

We have explored transformer based models for classification of data with and without augmentation. The results suggest that the use of data augmentation leads to improved training of models, resulting in a reduction of overfitting and an overall enhancement of performance on text classification tasks. In the future, we aim to further explore data augmentation techniques that exploit the domain-specific nature of the low-resource Indic languages. This work could further be extended by using machine learning models for classification by extracting hand-crafted features.

VI. References

1) Fadaee, M., Bisazza, A., & Monz, C. (2017). Data Augmentation for Low-Resource Neural Machine Translation. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers (pp. 567-573). Association for Computational Linguistics.

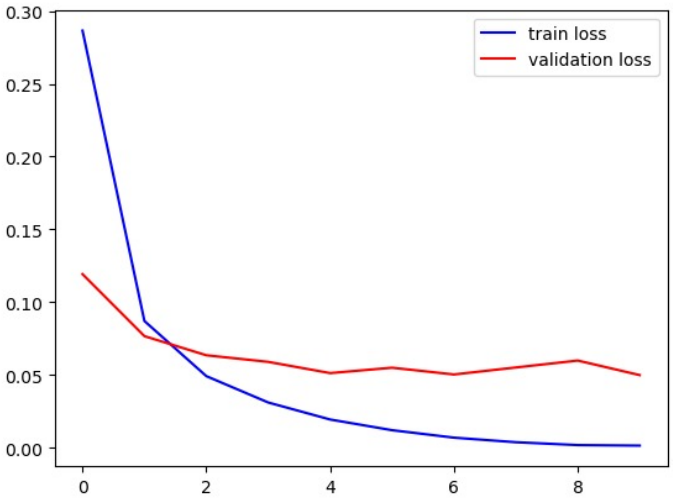

Fig. 5: Loss Curve for Data augmented with Synonym Substitution with Backtranslation

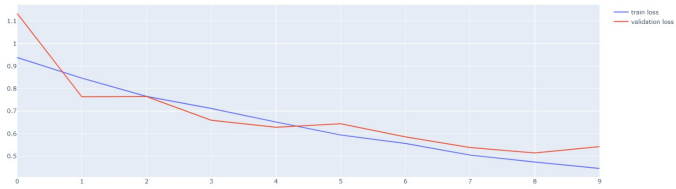

Fig. 6: Loss Curve for Data Augmented with Shuffle and Transliteration

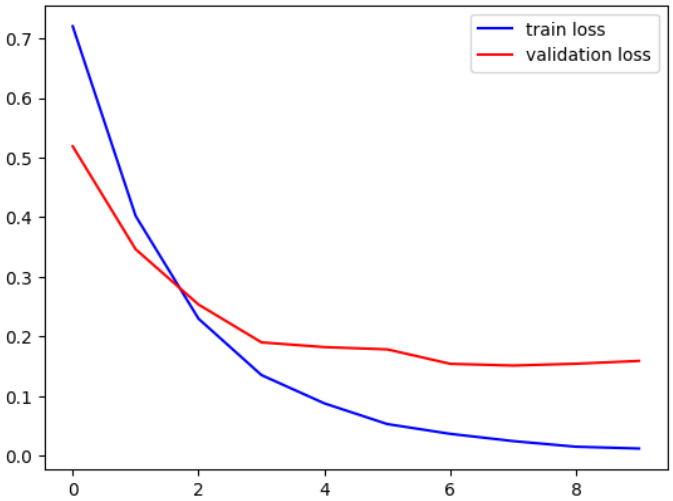

Fig. 7: Loss Curve for Data Augmented with Backtranslation

2) Wu, X., Gao, C., Lin, M., Zang, L., Wang, Z., & Hu, S. (2021). Text Smoothing: Enhance Various Data Augmentation Methods on Text Classification Tasks. arXiv preprint arXiv:2109.08672.
3) Ding, B., Liu, L., Bing, L., Kruengkrai, C., Nguyen, T. H., Joty, S., Si, L., & Miao, C. (2020). DAGA: Data Augmentation with a Generation Approach for Low-resource Tagging Tasks. ArXiv. https://doi.org/10.48550/arXiv.2011.01549
4) Karimi, A., Rossi, L., & Prati, A. (2019). AEDA: An Easier Data Augmentation Technique for Text Classification. In 2019 International Joint Conference on Neural Networks (IJCNN) (pp. 1-7). IEEE.
5) J. Wei and K. Zou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks," arXiv preprint arXiv:1901.11196, April 2019.
6) V. Kumar, A. Choudhary, and E. Cho, "Data Augmentation using Pre-trained Transformer Models," arXiv preprint arXiv:2110.11431, Oct. 2021.
7) L. Liu, B. Ding, L. Bing, S. Joty, L. Si, and C. Miao, "MulDA: A Multilingual Data Augmentation Framework for Low-Resource Cross-Lingual NER," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Nov. 2020, pp. 8001-8013.
8) M. Regina, M. Meyer, and S. Goutal, "Text Data Augmentation: Towards better detection of spear-phishing emails," in Proceedings of the 14th International Conference on Language Resources and Evaluation (LREC), Nov. 2020, pp. 3893-3899.
9) Dai, X., & Adel, H. (2020). An Analysis of Simple Data Augmentation for Named Entity Recognition. arXiv preprint arXiv:2010.11683.
10) Shleifer, Sam. "Low Resource Text Classification with ULMFit and Backtranslation." arXiv preprint arXiv:1903.09244 (2019).
11) Yudianto Sujana and Hung-Yu Kao. "LiDA: Language-Independent Data Augmentation for Text Classification." In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19), November 3-7, 2019, Beijing, China. ACM, New York, NY, USA, 9 pages. doi: 10.1145/3357384.3358016.
12) Wang, Y., Xu, C., Sun, Q., Hu, H., Tao, C., Geng, X., & Jiang, D. (2022). PromDA: Prompt-based Data Augmentation for Low-Resource NLU Tasks. arXiv preprint arXiv:2202.12499.