Consolidating and Developing Benchmarking Datasets for the Nepali Natural Language Understanding Tasks

Anonymous ACL submission

Abstract

The Nepali language has distinct linguistic features, especially its complex script (Devanagari script), morphology, and various dialects, which pose a unique challenge for natural language processing (NLP) evaluation. While the Nepali Language Understanding Evaluation (Nep-gLUE) benchmark provides a foundation for evaluating models, it remains limited in scope, covering four tasks. This restricts their utility for comprehensive assessments of NLP models. To address this limitation, we introduce eight new datasets, creating the Nepali Language Understanding Evaluation (NLUE) benchmark for evaluating the performance of models across a diverse set of Natural Language Understanding (NLU) tasks. The tasks include single-sentence classification, similarity and paraphrase tasks, and Natural Language Inference (NLI) tasks. On evaluating the models using added tasks, we observe that the existing models fall short in handling complex NLU tasks effectively. This benchmark sets a new standard for evaluating, comparing, and advancing models, contributing significantly to the broader goal of advancing NLP research for low-resource languages.

1 Introduction

Nepali is written in the Devanagari script and is a highly inflected language. The Nepali language incorporates a complex system of noun, adjective, and verb in-flections. Nouns have a system of gender, case, and number (Bal, 2004). It has a rich vocabulary with many homonyms and is spoken in different dialects across various regions, and there are variations in vocabulary, grammar, and pronunciation. In order to develop and establish robust models for Nepali, it is crucial to have reliable mechanisms for evaluating their quality and effectiveness. Tools that enable us to assess how well models handle the unique linguistic challenges while identifying their limitations are really important to drive the progress.

Despite Nepali's importance as a primary or secondary language for millions of speakers, research efforts and resources dedicated to its computational processing and evaluation remain relatively sparse. Existing benchmarks, such as Nep-gLUE (Timilsina et al., 2022), have made significant progress in this direction, providing a foundation for evaluating models on fundamental tasks. However, these benchmarks are limited in scope, primarily addressing a few tasks and overlooking critical aspects of linguistic understanding such as pronoun resolution, paraphrase interpretation, and advanced inference capabilities. To address this, we introduce eight datasets that cover diverse aspects of NLU. The new tasks include Sentiment Analysis (SA), Corpus of Linguistic Acceptability (CoLA), Paraphrase Detection, with datasets such as Quora Question Pairs (QQP) and the Microsoft Research Paraphrase Corpus (MRPC), Natural Language Inference (NLI), with datasets such as Multi-Genre NLI (MNLI), Question-Answer NLI (QA/NLI), Recognizing Textual Entailment (RTE), and Coreference Resolution (CR). These tasks collectively offer a more comprehensive evaluation of NLU capabilities for Nepali language models.

Table 1 below provides an overview of the tasks, the number of data points, the evaluation metrics employed, and the domains from which the datasets were collected. The varying size of each datasets, also tests the generalization capabilities and help us analyze model behavior under varying conditions.

The new NLU dataset is inspired by the General Language Understanding Evaluation (GLUE)benchmark (Wang, 2018) and was created primarily through a combination of automated and manual translation processes to ensure high-quality, task-specific datasets. We translated datasets with Large Language Models (LLMs), particularly gpt-4o-mini(OpenAI, 2024). We ensured the accuracy

Corpus	Train	Test	Task	Metric	Domain
Single sentence tasks					
SA	65.1k	16.3k	sentiment analysis	F1	Movie Reviews
CoLA	8.4k	1k	acceptability judgements	F1	misc.
Similarity and Paraphrase Tasks					
QQP	20k	4.29k	paraphrase	F1	social QA
MPRC	3.2k	815	paraphrase	F1	News
Natural Language Inference					
MNLI	19.2k	4.8k	NLI	F1	misc.
QNLI	12k	3k	QA/NLI	F1	Wikipedia
RTE	2.2k	554	NLI	F1	News, Wikipedia
CR	635	71	coreference resolution	F1	Fiction

Table 1: Task descriptions and statistics about the new benchmark datasets.

and contextual relevance of these translations. We conducted a thorough review of the availability of existing Nepali datasets for different tasks and if such datasets existed, we integrated them with the translated data, carefully removing duplicates to create a unified dataset. For tasks like Acceptability Judgments and WNLI (Coreference), where suitable datasets or high-quality translations were unavailable, we performed manual translations to ensure linguistic accuracy and consistency. To assess the effectiveness of the expanded benchmark and performance of models, we conducted experiments by fine-tuning both monolingual models trained exclusively on Nepali-language data and multilingual models that include Nepali as one of their supported languages. Each model was fine-tuned on the newly introduced tasks and evaluated using metrics provided in Table 1.

2 Related Works

Benchmarks such as GLUE (Wang, 2018) and its successor Super General Language Understanding Evaluation (SuperGLUE) benchmark (Wang et al., 2020) have been instrumental in advancing research in Natural Language Understanding (NLU). GLUE (Wang, 2018) introduced a multitask framework for evaluating various NLU capabilities, such as single-sentence classification, sentence-pair similarity, and inference tasks. SuperGLUE (Wang et al., 2020) extended this with more challenging tasks, including causal reasoning and co-reference resolution, addressing the limitations of GLUE (Wang, 2018) for state-of-the-art models. These benchmarks set a standard for evaluating linguistic and semantic understanding in high-resource languages like English. Efforts such as XGLUE (Liang et al., 2020) and XTREME (Hu et al., 2020) expanded these concepts to multilingual contexts, allowing learning of cross-lingual transfer.

Nep-gLUE (Timilsina et al., 2022) is the first comprehensive benchmark for Natural Language Understanding (NLU) tasks in Nepali. It includes four core tasks: Named Entity Recognition (NER), Part-of-Speech Tagging (POS), Content Classification (CC), and Categorical Pair Similarity (CPS). Although Nep-gLUE offers a robust foundation with its multitask dataset, it falls short in addressing more advanced NLP tasks necessary for comprehensive evaluations of models at the linguistic level. Advanced and complex tasks are crucial for further progress in low-resource languages such as Nepali.

Nepali Sentiment Analysis (NepSA) (Singh et al., 2020) dataset consists 3,068 comments extracted from 37 YouTube videos across 9 channels. Another dataset, Aspect-Based Sentiment Analysis (Tamrakar et al., 2020), contains 1,576 sentences, equally divided between positive and negative sentiments. Additionally, sentiment analysis datasets like Nepali Language Sentiment Analysis - Movie Reviews (Ghimire), 602 data points, and Nepali Sentiment Analysis (Acharya), 2,161 data points found on Kaggle, are limited in size and domain specific. For our benchmark, we used the Nep-

COV19Tweets dataset (Sitaula et al., 2021), which includes 33.5k sentiments labeled as positive, negative, or neutral. We selected 14.9k positive and 13.5k negative data points for sentiment analysis. A more recent dataset, Sentiment of Election-Based Nepali Tweets (Pokharel), contains 17.8k tweets but includes English characters and numbers, making it less suitable for our benchmark dataset. And there are no publicly available datasets for coreference resolution, acceptability judgment, or paraphrase detection in the Nepali language which shows a significant gap of resources.

3 Tasks

NLUE benchmark designed to evaluate the performance of language understanding models across a variety of tasks. The objective of NLUE is to provide a robust evaluation metric applicable to a broad range of language understanding challenges. We describe the tasks below and in Table 1.

3.1 Single-Sentence Tasks

Single-sentence tasks in the NLUE benchmark focus on assessing a model's ability to understand and analyze individual sentences.

3.1.1 SA

A sentiment analysis dataset has been added to evaluate models' ability to classify the emotional tone (positive, negative) of Nepali text. We created the dataset for sentiment analysis by translating Stanford Sentiment Treebank (Socher et al., 2013) which consists of sentences from movie reviews and human annotations of their sentiment from the GLUE Benchmark using using GPT-4o-mini (OpenAI, 2024), and manually translating instances that could not be accurately translated. It has 51k data points and is approximately equally divided between two classes, positive and negative sentiment, and uses only sentence-level labels. We incorporated this dataset with pre-existing sentiment analysis of Nepali COVID-19-related tweets (Sitaula et al., 2021), with 15k data points for each positive and negative sentiment. In total, the dataset has 81k data points, split equally between both classes.

3.1.2 CoLA

This dataset tests the model's ability to distinguish grammatically correct and incorrect sentences in Nepali. The task involves determining whether a given sentence follows the linguistic rules of Nepali, ensuring the model can assess grammaticality. To create the acceptability judgments dataset, we translated the Corpus of Linguistic Acceptability (CoLA)(Warstadt et al., 2019) into Nepali which consists of judgments drawn from books and journal articles on linguistic theory from the GLUE Benchmark using using gpt-4o-mini (OpenAI, 2024). For sections that were not translated correctly, we relied on manual translation. It has 9.5k data points, with both (correct/incorrect) classes.

3.2 Similarity and Paraphrase Task

Similarity and Paraphrase Task in the NLUE benchmark evaluates a model's ability to determine whether two sentences convey the same meaning or are paraphrases of each other. These tasks provide valuable insights into a model's proficiency in handling diverse expressions of similar ideas.

3.2.1 QQP

We have introduced a paraphrase detection dataset to assess whether models can correctly determine whether two Nepali sentences convey the same meaning. The Quora Question Pairs (Iyer et al., 2017) dataset is a collection of question pairs from the community question-answering website Quora. Using GPT-4o-mini,(Iyer et al., 2017) we translated the Quora Question Pairs from the GLUE Benchmark into Nepali to create a paraphrase detection dataset. The class distribution of paraphrase detection is almost balanced.

3.2.2 MRPC

We have introduced another paraphrase detection dataset based on the Microsoft Research Paraphrase Corpus (Dolan and Brockett, 2005). Using GPT-4o-mini, we translated the MRPC dataset into Nepali to create a paraphrase detection dataset for evaluation. The class distribution of this dataset is 70-30, with a higher proportion of paraphrase pairs.

3.3 Inference Tasks

The NLI tasks in this benchmark, assess a model's ability to understand relationships between sentences, such as entailment, contradiction, and neutral alignment. These tasks are crucial because they evaluate a model's comprehension of contextual meaning, logical inference, and its ability to handle complex linguistic structures.

Model	PARAMS	SA	CoLA	QQP	MPRC	MNLI	QNLI	RTE	CR
multilingual BERT (Devlin et al., 2019)	172M	86.711	80.803	78.16	70.14	74.45	78.56	**63.90**	45.77
XLM-Rbase (Conneau et al., 2020)	270M	**88.732**	**81.776**	77.73	69.22	76.42	**81.22**	56.11	47.513
NepBERT (Pudasaini et al., 2023)	110M	87.565	81.175	72.01	70.8	71.28	79.37	55.81	54.921
NepaliBERT (Rajan, 2021)	110M	83.421	80.974	66.46	69.31	71.59	79.28	52.4	49.2
NepBERTa (Timilsina et al., 2022)	110M	84.438	80.656	74.42	**71.29**	72.80	80.3	57.72	52.198
BERT Nepali (Thapa et al., 2024)	110M	87.901	81.646	75.28	70.38	74.66	80.29	52.43	**58.816**
RoBERTa Nepali (Thapa et al., 2024)	125M	88.33	21.56	**78.43**	70.38	**76.78**	80.86	54.64	47.21
Distilbert-base (Maskey et al., 2022)	67M	87.325	81.27	74.05	69.87	71.78	79.43	53.91	50.603
Deberta-base (Maskey et al., 2022)	139M	88.046	**81.776**	77.16	70.51	75.78	80.2	56.03	47.21

Table 2: Scores of each model across eight evaluation tasks

3.3.1 CR

This dataset tests the model's ability to resolve coreference relationships within a Nepali text. We developed the conference resolution dataset by manually translating the Winograd Schema Challenge (Levesque et al., 2011), which is a reading comprehension task in which a system must read a sentence with a pronoun and select the referent of that pronoun from a list of choices. The training set has 635 data points and the test set has 71 data points, balanced between two classes.

3.3.2 MNLI

The dataset is translated into Nepali from the Stanford Natural Language Inference Corpus (Bowman et al., 2015) using GPT-4o-mini. This corpus is a crowd-sourced collection of sentence pairs annotated with textual entailment labels. Each pair consists of a premise and a hypothesis, and the task is to predict the relationship between them, whether the premise entails the hypothesis (entailment), contradicts it (contradiction), or is unrelated (neutral).

3.3.3 QNLI

The QNLI (Question-answering Natural Language Inference) dataset has been adapted for Nepali from GLUE benchmark by translating the original English dataset using GPT-4o-mini which originates from the Stanford Question Answering Dataset (Rajpurkar et al., 2016) that contains question-paragraph pairs sourced from Wikipedia. The dataset has an equal division between entailment and non-entailment pairs, ensuring balanced class distribution.

3.3.4 RTE

The Recognizing Textual Entailment (RTE) dataset for this benchmark is converted to Nepali by translating the GLUE benchmark RTE dataset using GPT-4o-mini and manual effort. The dataset evaluates a model's ability to predict whether a hypothesis logically follows from a given premise, framed as a two-class classification task.

4 Result

We performed fine-tuning and evaluation in all tasks using a range of hyperparameter combinations between 1e-5 and 5e-5, freezing and unfreezing the final four layers, and trained for 4 to 15 epochs. For each task, we selected the model that performs best in the test set. The scores for each model for each task are provided in Table 2. XLM-Rbase (Conneau et al., 2020) looks the strongest among all, but no single model is the best for all tasks.

Our results demonstrate that the models generally perform well in simpler tasks. However, as the complexity of the tasks increases, the performance of the model decreases significantly. This gap is particularly visible in tasks where the available fine-tuning data is limited, further emphasizing the models' inability to generalize effectively when trained on smaller datasets.

This trend highlights a critical limitation in the current Nepali language models. Although they can excel in tasks with abundant data and straightforward structures, their performance struggles to scale for tasks demanding complex reasoning or where training data is sparse.

Limitations

Despite newer datasets and benchmark, our work still lacks diverse data sources, particularly informal, conversational data and regional styles inflected dataset. Some tasks also suffer from small annotated datasets, restricting model generalization for smaller models. Future work should focus on developing larger, more diverse evaluation datasets and refining metrics to ensure even robust NLP applications for Nepali Language.

References

Mahesh Acharya. Nepali language sentiment analysis.

Bal Krishna Bal. 2004. *Structure of Nepali Grammar*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *Preprint*, arXiv:1508.05326.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116v2*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Shikhar Ghimire. Nepali language sentiment analysis - movie reviews.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *Preprint*, arXiv:2003.11080.

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. 2017. First quora dataset release: Question pairs. *QuoraData*.

Hector J. Levesque, Ernest Davis, and L. Morgenstern. 2011. The winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Utsav Maskey, Manish Bhatta, Shiva Bhatt, Sanket Dhungel, and Bal Krishna Bal. 2022. Nepali encoder transformers: An analysis of auto encoding transformer language models for Nepali text classification. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 106–111, Marseille, France. European Language Resources Association.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence.

Durga Pokharel. Sentiment of election based nepali tweets.

Shushanta Pudasaini, Subarna Shakya, Aakash Tamang, Sajjan Adhikari, Sunil Thapa, and Sagar Lamichhane. 2023. Nepalibert: Pre-training of masked language model in nepali corpus. In *7th International Conference on IoT in Social, Mobile, Analytics and Cloud*.

Rajan. 2021. Nepalibert.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *Preprint*, arXiv:1606.05250.

Oyesh Mann Singh, Sandesh Timilsina, Bal Krishna Bal, and Anupam Joshi. 2020. Aspect based abusive sentiment detection in nepali social media texts. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 301–308.

Chiranjibi Sitaula, Anish Basnet, Ashish Mainali, and Tej Shahi. 2021. Deep learning-based methods for sentiment analysis on nepali covid-19-related tweets. *Computational Intelligence and Neuroscience*, 2021.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Sujan Tamrakar, Bal Krishna Bal, and Rajendra Thapa. 2020. *Aspect Based Sentiment Analysis of Nepali Text Using Support Vector Machine and Naive Bayes*. Ph.D. thesis.

Prajwal Thapa, Jinu Nyachhyon, Mridul Sharma, and Bal Krishna Bal. 2024. Development of pre-trained transformer-based models for the nepali language. *Preprint*, arXiv:2411.15734.

Sulav Timilsina, Milan Gautam, and Binod Bhattarai. 2022. Nepberta: Nepali language model trained in a large corpus. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics (ACL).

Alex Wang. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. Superglue: A stickier benchmark for general-purpose language understanding systems. *Preprint*, arXiv:1905.00537.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.