NepBERTa: Nepali Language Model Trained in a Large Corpus

Milan Gautam*, Sulav Timilsina*
Palua.AI Ltd, UK
{milan,sulav}@palua.ai

Binod Bhattarai
Nepal Applied Mathematics and
Informatics Institute for research (NAAMII), Nepal
bhattarai.binod@gmail.com

Abstract

Nepali is a low-resource language with more than 40 million speakers worldwide. It is written in Devnagari script and has rich semantics and complex grammatical structure. To this date, multilingual models such as Multilingual BERT, XLM and XLM-RoBERTa haven't been able to achieve promising results in Nepali NLP tasks, and there does not exist any such a large-scale monolingual corpus. This study presents NepBERTa, a BERT-based Natural Language Understanding (NLU) model trained on the most extensive monolingual Nepali corpus ever. We collected a dataset of 0.8B words from 36 different popular news sites in Nepal and introduced the model. This data set is 3 folds times larger than the previous publicly available corpus. We evaluated the performance of NepBERTa in multiple Nepali-specific NLP tasks, including Named-Entity Recognition, Content Classification, POS Tagging, and Categorical Pair Similarity. We also introduce two different datasets for two new downstream tasks and benchmark four diverse NLU tasks altogether. We bring all these four tasks under the first-ever Nepali Language Understanding Evaluation (Nep-gLUE) benchmark. We will make Nep-gLUE along with the pre-trained model and data sets publicly available for research.

1 Introduction

In recent years, especially in the last four years, there has been a lot of progress in the field of NLP, which includes two breakthroughs: the self-attention mechanism (Vaswani et al., 2017) and the self-supervised model pre-training (Peters et al., 2018; Devlin et al., 2019), which uses the advantage of pre-training on huge volume of unlabeled text dataset. To obtain a state of the art result, a large model based on the transformer (Vaswani et al., 2017) is pre-trained on a large amount of unlabeled text data, then this model is further fine-tuned with labeled data as per the requirement. Since its release in 2019, Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) has become very popular for transfer learning purposes in various NLP tasks. Many improvements of BERT (Liu et al., 2019; Yang et al., 2019; Clark et al., 2020) have been made since 2019, even though only two versions of BERT which were pre-trained in English and Chinese language were released initially.

After a while, a new version named Multilingual BERT (Devlin et al., 2019) was released. This model, trained in 104 languages, showed impressive performance on many languages specific downstream tasks. Some of its performances are still state-of-the-art in many languages. Multilingual BERT's strong performance inspired many NLP communities to build their language-specific BERT model. Some of the popular monolingual BERT models are Russian (Kuratov and Arkhipov, 2019), Dutch (de Vries et al., 2019), Arabic (Antoun et al., 2020), French (Martin et al., 2019) and Portuguese (Souza et al., 2019).

Nepali is spoken by more than 40 Millions people worldwide. Syntactically, Nepali language differs compared to English which is one of the most widely studied languages. Generally, in English the sentence structure is Subject - Verb - Object. Whereas, in Nepali language this structure ends with verb having standard structure as Subject - Object - Verb as shown in Figure1. We suggest the readers refer (Bal, 2004) for more information. Since Nepali is considered a low-resource language (Rajan and Salgaonkar, 2022; Basu and Majumder, 2020), it has received little attention in the field of NLP. Despite the advancement of NLP in the English language, there has not been a considerable contribution to the Nepali NLP domain. The main reason behind this is a lack of pre-training data, resource standardization, and computational resources. Nepali is written in the Devnagari script,

*equal contributions; part of the work was done when Sulav and Milan were at IOE, Pashchimanchal Campus, Nepal

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 (Volume 2: Short Papers), pages 273–284
November 20–23, 2022.

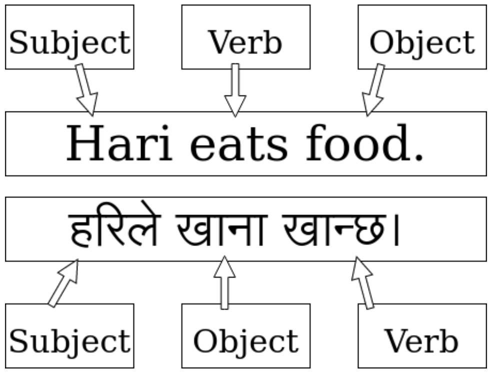

Figure 1: Sentence structure of Nepali language compared with English language.

which has been rarely used for NLP services.

Motivated by the success of language-specific models over multilingual models in many other languages, we present NepBERTa, a BERT (Devlin et al., 2019) based Nepali language model. The data required to pre-train NepBERTa were collected through the scrapping of the top 36 News sites of Nepal in the Nepali language.

Inspired by the use case of the GLUE (Wang et al., 2018) benchmark, we also introduce the Nepali Natural Language Understanding (NLU) dataset on two downstream tasks (News Content Classification and Categorical Pair Similarity) and evaluate NepBERTa on altogether four diverse downstream tasks on, POS tagging, news content classification, named entity recognition, and categorical pair similarity. We have brought all these tasks under **Nep**ali **L**anguage **U**nderstanding **E**valuation benchmark (**Nep**-g**LUE**) tasks.

2 Related Work

In 2013 a team at Google led by Thomas Mikolov released a word embedding named "Word2Vec" (Mikolov et al., 2013). Following the success of word2vec, other forms of word embeddings like GloVe (Pennington et al., 2014) and fastText (Mikolov et al., 2017) were released. However, these embeddings were not able to extract the contextual meaning of the sentence. This problem was overcome by the large pre-trained models such as ULMFit (Howard and Ruder, 2018), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), and ALBERT (Lan et al., 2020).

ULMFit uses a recurrent neural network as its core, whereas BERT uses a self-attention mechanism, which evaluates the dependency of a token with every other token in the same sequence. BERT adopts the mask language modeling (MLM) technique and next sentence prediction (NSP) technique to learn the deeper semantics and contextual information of a sentence.

Later, (Wu and Dredze, 2019) and (Pires et al., 2019) investigated the potential of BERT on cross-lingual NLP tasks using a large corpus of diverse languages. Their work established the benchmark for many multilingual tasks and demonstrated that a single model can learn from numerous languages. In terms of model size and performance, XLM (Lample and Conneau, 2019) and XLM-RoBERTa (Conneau et al., 2020) made significant advances.

There have already been various monolingual models that outperformed multilingual ones. Some of these models are FinBERT (Virtanen et al., 2019) for Finish, BERTje (de Vries et al., 2019) and RobBERT (Delobelle et al., 2020) for Dutch, FlauBERT (Le et al., 2020) for French.

Recently two monolingual Nepali models trained in the Nepali language corpus were made open source on Github [1] [2]. These two models were mainly trained on text corpus made available by the OSCAR (Ortiz Suárez et al., 2019) dataset, which is more than 3 times smaller than our dataset. Furthermore, there were not any benchmarks to evaluate the performance of those models across various downstream tasks.

3 NepBERTa

3.1 Data Collection

A massive quantity of data is necessary to pre-train a language model. For example, BERT (Devlin et al., 2019) was pre-trained on 3.3 billion words from the English Wikipedia and Book corpus (Zhu et al., 2015). In addition, RoBERTa (Liu et al., 2019) and XLNet (Yang et al., 2019) increased the size of their pre-training data and model parameters.

Nepali is a relatively small and resource-constrained language. For example, the Nepali Wikipedia dataset is less than one GB. That is why we had to crawl the web for our pre-training data. We selected the top 36 news sites according to volume and variety of data. We managed to crawl about 14.5 GB of data which has blogs and news articles with roughly 21 main categories. We suggest

[1] pudasainishushant/NepaliLanguageModelPretraining
[2] R4j4n/NepaliBERT

the readers refer to the supplementary materials for more details.

We also discovered three GB of the OSCAR dataset (Ortiz Suárez et al., 2019), but it belongs to the same news websites we have crawled from, which may result in data deduplication. That is why we chose not to use that data.

3.2 Data Pre-processing

During this process, we performed data deduplication, removed non-contextual contents like HTML/JavaScript tags and filtered out none Nepali words. After this process dataset was reduced to 12.5 GB containing approximately 0.8 Billion words with 2.75 million documents with an average of 291 words in each document.

Each document is split into several data points of 327 words, resulting in 512 tokens in each sample and deleting the words between the 512^{th} token and the following stop symbol. We obtained around 3.75 million train instances after preparing the text corpus up to 512 tokens in each data point.

We use the final data corpus to train the WordPiece (Wu et al., 2016) vocabulary of 30,522 subword tokens. We limited the training token length to 512 and did not cross the boundaries. There are about 1.5 billion tokens in total.

3.3 Pre-training Objective

All BERT based models leverage unsupervised pre-training objective on unlabeled data. For example, BERT (Devlin et al., 2019) uses mask language modeling (MLM) and next sentence prediction (NSP). While RoBERTa (Liu et al., 2019) as a new flavor of BERT drops the next sentence prediction task and pre-trained only on masked language modeling tasks.

We use BERT-base (Devlin et al., 2019) as our underlying architecture while taking pre-training inspiration from RoBERTa (Liu et al., 2019). We solely utilize MLM technique to pre-train NepBERTa with dynamic masking. RoBERTa proved that dynamic masking with an MLM pre-training objective outperforms static masking and allows the model training for longer steps. This strategy ensures that each training phase masks a new set of tokens before feeding them into the encoder layers. This strategy prevents the model from predicting the same tokens in future epochs, allowing the model to learn more about the overall data distribution.

3.4 Model Architecture and Hyper-parameters

NepBERTa follows BERT-base (Devlin et al., 2019) as the main training architecture. BERT is a transformer (Vaswani et al., 2017) based model with 12 layers of encoders, 768 embedding sizes and 12 attention heads, with 110 million parameters. We set the maximum sequence length to 512 subword tokens. Training the model with a batch size of 4096 and 90k training steps on a v3-128 TPU instance on GCP. The Adam (Kingma and Ba, 2015) optimizer is used with a learning rate of 4e-4 with standard parameters ($\beta 1 = 0.9$, $\beta 2 = 0.999$), L2 weight decay of 0.01, linear warm up step of 4.5k steps and linear learning rate decay. We stopped the pre-training of NepBERTa when there was no further improvement in the performance on downstream tasks.

4 Nepali Language Understanding Evaluation (Nep-gLUE) Benchmark

Several individuals have studied Nepali NLP tasks and contributed to them. Parts of speech tagging (Sayami et al., EasyChair, 2019), named entity recognition (Singh et al., 2019), and so on are examples. However, there has not been a unified, comprehensive study of the Nepali NLU tasks.

Other languages, such as English (Wang et al., 2018), French (Le et al., 2020), and Korean (Park et al., 2021), have language-specific benchmark systems for certain activities. Text categorization, sequence labeling, and text span prediction are the three types of NLU tasks in general. As a result, we have developed four distinct tasks for the Nep-gLUE benchmark. All of the codes and dataset[1] for these activities are freely available to the public for future usage and improvement.

4.1 Content Classification (CC)

We created the dataset for content classification by scrapping news websites to get their news articles with their corresponding news category. We identified nine main categories of news articles for this task. These nine categories are politics, health, entertainment, thought, crime, sports, economy, literature, and world . It has 45k data points, and all the classes have an approximately equal number of data points.

[1] https://nepberta.github.io/

Split	O	B-PER	B-ORG	B-LOC	I-PER	I-ORG	I-LOC
Train	58,977	2,310	1,796	1,639	1,599	1,411	133
Test	14,958	569	448	407	405	365	37

Table 1: Data distribution for NER.

MODEL	PARAMS	NER	POS	CPS	CC	Nep-gLUE Score
multilingual BERT (Devlin et al., 2019)	172M	85.45	94.65	93.60	91.08	91.19
XLM-R_{base} (Conneau et al., 2020)	270M	87.59	94.88	93.65	92.33	92.11
NepBERT (Pudasaini, 2021)	110M	79.12	90.63	91.05	90.98	87.94
NepaliBERT (Rajan, 2021)	110M	82.45	91.67	89.46	90.10	88.42
NepBERTa (**Ours**)	110M	**91.09**	**95.56**	**94.42**	**93.13**	**93.55**

Table 2: Performance comparison of NepBERTa with multilingual models. The evaluation metric is Macro-F1.

4.2 Named Entity Recognition (NER)

Named Entity Recognition is a classical NLU task for a language model where it has to correctly tag the words in a sequence as location, person, organization, dates, currency, etc. Dataset for NER task has mainly 3 classes (person, location, and organization) with 2 subclasses for each of the classes labeled as (*B-PER, I-PER, B-LOC, I-LOC, B-ORG, I-ORG*) where *"B"* denotes the beginning of the class and *"I"* denotes interior of the class label. Adding to this there is one more class named "*Other*" labeled as "*O*". Altogether, there are 7 classes in this dataset. We were able to find some works in the Nepali NER task and dataset related to this task from (Singh et al., 2019). We have used this dataset for bench-marking of NepBERTa. Table 1 shows the data distributed over seven different classes in both train and test splits. Since we can see the data is distributed unevenly over the classes, the macro F1 score best describes the performance of this task.

4.3 Part Of Speech Tagging (POS)

In this task, the model has to predict which parts of speech the words belong to in a sequence, such as nouns, verbs, prepositions, conjunction, etc. For NepBERTa evaluation, we used this (Bhasa, 2020) POS tagging dataset, which is publicly available on GitHub. It has a total of 39 class labels, some of which are Common noun (NN), Proper noun (NNP), Counting decimal number (CD), Finite verb (VBF), Auxiliary verb (VBX) and so on.

Both of these datasets are tagged using BIO (Ramshaw and Marcus, 1995) format, we have used the macro F1 metrics for evaluation of this tasks.

4.4 Categorical Pair Similarity (CPS)

For this task, we scrapped and curated a new Nepali Language Inference dataset for categorical pair similarity. In this dataset, we have put together two sequences randomly based on their categories. If both the sequences belong to a single category, then it is labeled as 1, otherwise 0. Therefore, we give positive labels to sequence pairs with similar semantic traits and negative labels to sequence pairs with differing semantic features. In the process of preparing dataset, 2.5k pairs of categorically similar datapoints are extracted from 9 categories resulting in total of 22.5k with label '1'. And for dissimilar datapoints every 2.5k datapoints from a category are paired with 2.5k datapoints of every other categoreis. Finally 22.5k dissimilar pair are chosen at random. In this way evenly distributed 45k datapoints are generated for this task. Macro F1 score is used as an evaluation metric in this task also.

5 Evaluation

5.1 Fine-Tuning

We evaluate the performance of NepBERTa on the Nepali NLU task against two multilingual Bert model, mBERT (Devlin et al., 2019) and XLM-R base (Conneau et al., 2020) and against two monolingual models, NepBERT (Pudasaini, 2021) and NepaliBERT (Rajan, 2021) trained on a relatively small corpus of Nepali text.

During fine-tuning, no further pre-processing is performed except tokenization. We used WordPiece (Wu et al., 2016) for all the task and split the dataset into training and test sets by an 80:20 ratio as shown in Table 3. We further used 20% of

Task	Train	Test	Type
NER	68,865	17,216	Entities
POS	89,149	22,290	Entities
CPS	36,000	9,000	Sequence Pairs
CC	35,537	8,884	Sequences

Table 3: Summary of distribution of data for various tasks.

train set to produce cross-validation (CV) set, and search the hyper-parameters on it. The maximum sequence length is fixed to 512 since the NepBerta is pre-trained on the same sequence length. After training for 2-15 epochs with a learning rate $(1e^{-5}, 2e^{-5}, 3e^{-5}, 4e^{-5}, 5e^{-5})$ and a batch size of 16 (NER and POS) and 32 (CC and CPS), the best-performing model is selected.

5.2 Results

Table 2 shows the models evaluation on four different downstream tasks. The previously trained multilingual models mBERT (Devlin et al., 2019) and XLM-R base (Conneau et al., 2020) outperform the previously existing monolingual Nepali models NepBert (Pudasaini, 2021) and NepaliBERT (Rajan, 2021), whereas NepBERTa outperforms all the monolingual and multilingual models across all the downstream tasks. It performs the best on NER, where it exceeds the second-best performing model by almost +4 points. NepBERTa produces a significant improvement over previous Nepali monolingual models due to being trained on a large dataset. Similarly it also excels in sequence labeling tasks compared to other tasks.

NepBERTa has the highest Nep-gLUE score of 93.55, outperforming multilingual models mBERT and XLM-R base by approximately +2 and +1.5 points, respectively. Similarly, it provides a significant performance boost over the previous Nepali language models, NepBERT and NepaliBERT, by almost +5 and +6 points, respectively. And adding to this, the smaller size of NepBERTa ensures faster fine-tuning on downstream tasks.

6 Conclusion and Future Works

Until now, students and researchers were compelled to use multilingual models for their work. We introduced NepBERTa, a Nepali language model that can be used for many fine-tuning tasks in the future. We also introduce the first-ever Nepali Language Understanding evaluation benchmark. In the future, we will be adding more downstream tasks in Nep-gLUE.

After the introduction of the language model in the NLP community, this will be the first time the Nepali NLP community will be benefited to a great extent. We believe that the introduction of NepBERTa in Nepali NLP community will promote more study and implementation of the language model for many downstream tasks. There is always room for improvement in any research activity. Likewise, our next plan as an improvement to this version is to increase the pre-training model size and data.

Acknowledgements

We would like to thank Google's TPU Research Cloud program for providing us with free and unlimited usage of TPU v3-128 for 90 days. It would not have been possible without the continuous support and response of the TRC team.

References

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Bal Krishna Bal. 2004. Structure of nepali grammar.

Joyanta Basu and Swanirbhar Majumder. 2020. *Identification of Seven Low-Resource North-Eastern Languages: An Experimental Study*, pages 71–81. Springer Singapore, Singapore.

Nepali Bhasa. 2020. Nepali-bhasa/pos-tagger: Part of speech tagging in nepali.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *ArXiv*, abs/2003.10555.

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. In *ACL*.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *ArXiv*, abs/1905.07213.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *NeurIPS*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. Cite arxiv:1907.11692.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2019. CamemBERT: a Tasty French Language Model. Web site: https://camembert-model.fr.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2017. Advances in pre-training distributed word representations. *arXiv preprint arXiv:1712.09405*.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Cardiff, United Kingdom. Leibniz-Institut für Deutsche Sprache.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won-Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Tae Hwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Eunjeong Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. 2021. KLUE: korean language understanding evaluation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Sushant Pudasaini. 2021. Github.

Rajan. 2021. Github.

Annie Rajan and Ambuja Salgaonkar. 2022. Survey of nlp resources in low-resource languages nepali, sindhi and konkani. In *Information and Communication Technology for Competitive Strategies (ICTCS 2020)*, pages 121–132, Singapore. Springer Singapore.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Sarbin Sayami, Tej Bahadur Shahi, and Subarna Shakya. EasyChair, 2019. Nepali pos tagging using deep learning approaches. EasyChair Preprint no. 2073.

O. M. Singh, A. Padia, and A. Joshi. 2019. Named entity recognition for nepali language. In *2019 IEEE 5th International Conference on Collaboration and Internet Computing (CIC)*, pages 184–190.

Fábio Souza, Rodrigo Nogueira, and Roberto de Alencar Lotufo. 2019. Portuguese named entity recognition using bert-crf. *ArXiv*, abs/1909.10649.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish. *ArXiv*, abs/1912.07076.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. In *EMNLP*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.

7 Appendix

News Site	Count
ekantipur.com	265252
onlinekhabar.com	254130
nagariknews.com	159958
thahakhabar.com	140476
ratopati.com	138793
reportersnepal.com	122576
setopati.com	103515
hamrakura.com	100973
lokpath.com	93138
abhiyandaily.com	90617
pahilopost.com	86768
lokaantar.com	85427
dcnepal.com	81391
nayapage.com	76643
nayapatrikadaily.com	75633
everestdainik.com	74968
imagekhabar.com	66838
shilapatra.com	63392
khabarhub.com	63268
baahrakhari.com	63078
ujyaaloonline.com	61653
nepalkhabar.com	56034
emountaintv.com	50538
kathmandupress.com	48998
farakdhar.com	44489
kendrabindu.com	40815
dhangadhikhabar.com	40751
gorkhapatraonline.com	38835
dainikonline.com	36829
nepalpress.com	26886
hamrokhelkud.com	24899
himalkhabar.com	21989
nepallive.com	21425
nepalsamaya.com	21008
kalakarmi.com	13910
dainiknewsnepal.com	6593
Total	2762486

Table 4: List showing the numbers of articles collected from various news sources.

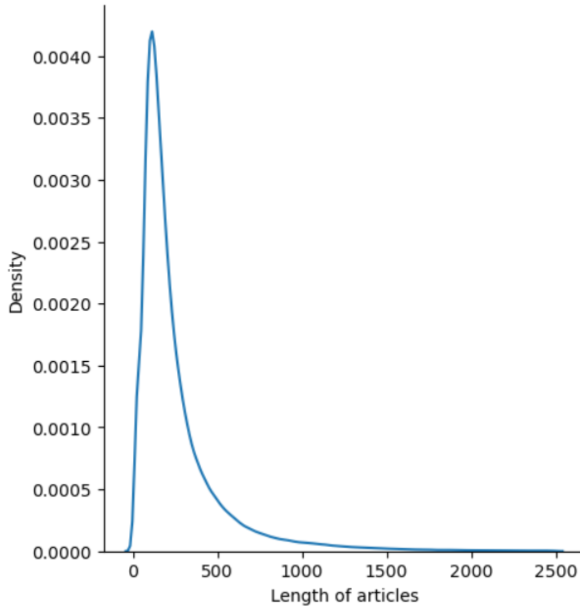

Figure 2: Plot showing the number of words in news articles. The number of articles with words more than 2500 words are 6115, which skewed the plot to the right. Hence these articles are omitted from the plot.

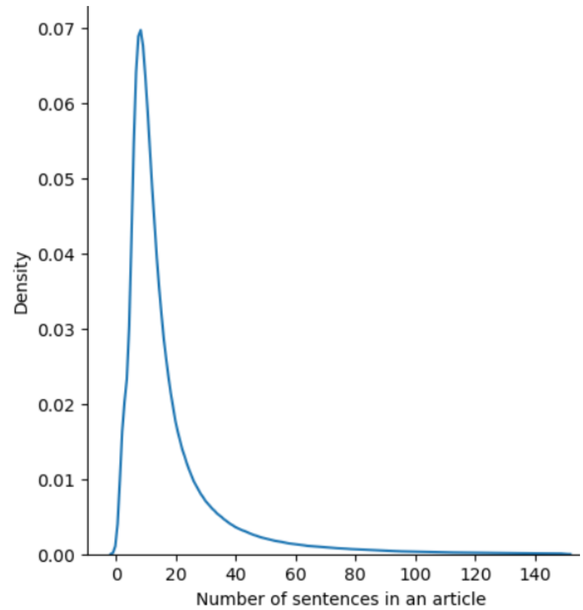

Figure 3: Plot showing the distribution of sentences per news article. The number of articles with sentences of more than 150 words is 14000, they are excluded from the plot.

8 Dataset

8.1 Data Source

We extracted articles from exactly 36 prominent newspapers as shown on Table 4, and the timeframe of the data lies between 2010 and 2022. Several significant news web sites, each of which contributes more than 100,000 data points to our corpus, include ekantipur.com, onlinekhabar.com, nagariknews.com, thahakabar.com, setopati.com,reportersnepal.com, etc. Each news portal has a particular domain of interest, like hamrokhelkud.com, which publishes sports news ranging from the IPL, NBA, Formula 1,

Category	Count
news	702151
misc	402847
politics	250668
economy	231235
national	225204
society and security	222731
sports	181227
global	132451
None	110342
health and lifestyle	64775
entertainment	62848
thought and opinion	56499
art and literatrue	34776
diaspora	31986
crime	15835
science and technology	9469
education	8911
court	5468
religious and culture	4815
tourism	4480
editorial	3768
Total	2762486

Table 5: List showing the number of articles which fall under various categories.

MMA, etc., which helps us create a corpus having a diverse range of domains.

8.2 Data Extraction

We scrapped all the articles for our dataset from web portals of news sites listed in Table 4. Every news site has a different way of formatting and documenting its news. So we wrote an individual script for every news portal using the Python Beautiful Soup library. To scrape hundreds of thousands of articles in less time, we used the multithreading technique and invoked multiple requests to the server at a time.

8.3 Data Distribution

8.3.1 Categories

Every news portal has its way of documenting under different headings and categories. After scrapping news articles, we gathered around 1000 unique categories. Most of the news categories were semantically the same but lexically different. Therefore, we had to manually map each distinct category to one of the 21 categories that we have selected as its root class, combining categories like cricket, basketball, football, and all the sports activities under a single category as sports as shown in Table 5.

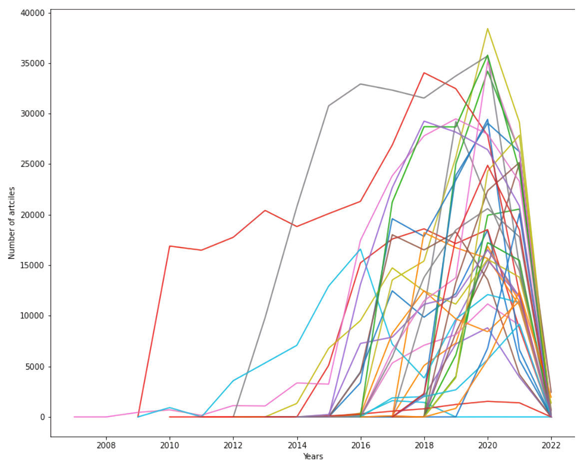

Figure 4: Total number of news articles published each year in different news portals of Nepal.

Around 0.7 million articles didn't belong to a specific domain; in their respective news portals, they were only categorized as news. Due to insufficient information about their category, we were reluctant to categorize such articles under a unified heading called "news." Similarly, for articles whose categories were not possible to extract or not given, they get the label "None.". We grouped domains having a few articles into "misc," and all together, the corpus contains 21 categories, contributing to more than 2.7 million articles.

8.3.2 Words Per Article

While plotting the number of words per article, we obtained a skewed bell shape curve. The news articles with a word count of more than 2500 are 6115, which we omitted from the plot. From Figure 2, we can see the majority of news articles have 200 to 300 words. News articles with a word count of 0 to 500 cover almost 95% of the distribution .

8.3.3 Sentences Per Article

Figure 3 shows the distribution of the number of sentences in an article. It doesn't include the articles whose sentence count is more than 150. As per the distribution, most of the articles have 15 sentences.

8.3.4 Articles per year

When it comes to the digitization of text data, timing is extremely important. We gathered the dates of publication for each news story while scraping data. Every new curve in Figure 4 is colored differently to symbolize a news portal. We can find out which articles were published when and when a news portal started its digital service. Since 2018, there have been more news pieces than ever before, and several websites have been operating since 2015. The analysis and discovery of the trends in Nepali society during the previous ten years can be understood by this data.

8.4 POS Tagging class labels

All the 39 class labels for POS Tagging are shown in Table 6. These labels contain reduced fine grain tag set used in Nepali language grammar and composition.

Category Definition	POS Tag
Common Noun	NN
Proper Noun	NP
Personal Pronoun	PP
Possessive Pronoun	PPP
Reflexive Pronoun	PRF
Marked Determiner	DTM
Unmarked Determiner	DTX
Others Determiner	DTO
Finite Verbs	VF
Infinitive Verb	VBI
Prospective Verb	VBN
Aspect Verb	VBKO
Others Verb	VBO
Marked Adjective	JJM
Unmarked Adjective	JJX
Degree Adjective	JJD
Adverb	RR
Postposition	II
Plural-collective Postposition	IH
Ergative-instrumental Postposition	IE
Accusative-dative Postposition	IA
Genitive Postposition	IKO
Cardinal Number	MM
Marked Ordinal Number	MOM
Unmarked Ordinal Number	MOX
Marked Classifier	MLM
Unmarked Classifier	MLX
Coordinating Conjunction	CC
Subordinating Conjunction	CS
Interjection	UU
Question Marker	QQ
Particle	TT
Sentence-final Punctuation	YF
Sentence-medial Punctuation	YM
Quotation Marks	YQ
Brackets	YB
Foreign Word	F
Unclassifiable	FU
Abbreviation	FB

Table 6: Reduced tag set as class labels for POS Tagging.

9 Linguistic Characteristics of Nepali Language

9.1 Origin, Status and Dialects

Nepali language belongs to the Indo-Aryan Language family which is believed to originate some 500 years ago in western hilly region of Nepal. It is one of the languages of Indic language subfamily of Indo-Aryan family, which has some noticeable influences from languages like, Hindi, Urdu, Arabic, Maithili, Bhijpuri, etc. It was mainly spoken by the Khas people of western Nepal and was aslo called Khas Kura. Nepali is now spoken by almost 40 million people worldwide, mainly from Nepal, India, Bhutan and Myanmar. It is the official language of Nepal, Sikkim, a Himalayan state of India and Darjeeling district of West Bengal state of India.

Nepali language has altogther 12 dialects, they are: Acchami, Dialekhi, Baitadeli, Darhulai, Bajhangi, Gandakeli, Bajurali, Huml, Bheri, Purbeli, Dadelhuri and Soradi.

9.2 Sound System

9.2.1 Consonants

Like in any other languages consonants are one of major two subdivisions of phonemes. They are produced by blocking the airflow temporarily while passing through the mouth. In Nepali language there are altogether 30 consonants. Those 30 consonants are classified into different groups according to the manner of articulation, as shown in Figure 5.

9.2.2 Vowels

There are mainly two types of vowels in Nepali, free form vowels and conjunct form of vowels. The 11 free form vowels and 10 conjunct form vowels are shown in Figure 6 and Figure 7 respectively.

Contrarily, consonants come before the conjunct forms of vowels (). Using the vowels "aa" in free form and conjunct form in Figure 8:

	Bilabial	Dental	Alveolar	Retroflex	Palatal	Velar	Glottal
Nasal	m ⟨म⟩		n ⟨न/ञ⟩	(ɳ ⟨ण⟩)		ŋ ⟨ङ⟩	
Plosive	p ⟨प⟩, pʰ ⟨फ⟩, b ⟨ब⟩, bʱ ⟨भ⟩	t ⟨त⟩, tʰ ⟨थ⟩, d ⟨द⟩ , dʱ ⟨ध⟩	t͡s ⟨च⟩, t͡sʰ ⟨छ⟩, d͡z ⟨ज⟩, d͡zʱ ⟨झ⟩	ʈ ⟨ट⟩, ʈʰ ⟨ठ⟩ , ɖ ⟨ड⟩ , ɖʱ ⟨ढ⟩		k ⟨क⟩, kʰ ⟨ख⟩, g ⟨ग⟩ , gʱ ⟨घ⟩	
Fricative			s ⟨श/ष/स⟩				ɦ ⟨ह⟩
Rhotic			r ⟨र⟩				
Approximant	(w ⟨व⟩)		l ⟨ल⟩		(w ⟨व⟩)		

Figure 5: Classification of Nepali consonant phonemes.

अ /a/	आ /ä/	इ /i/	ई /i/
उ /u/	ऊ /u/	ऋ /ri/	ए /e/
ऐ /ai/	ओ /o/	औ /au/	

Figure 6: These free form vowels in Nepali language.

◌ा, ◌ि, ◌ी, ◌ु, ◌ू, ◌ृ, ◌े, ◌ै, ◌ो, ◌ौ

Figure 7: These are conjunct forms vowels in Nepali language.

आमा (Aama) = आ (Aa) + म् (m) + ◌ा(Aa)

Figure 8: Example of use of both types of vowels in a word in Nepali language.

9.3 Grammatical Structure

9.3.1 Noun

Like English, Nouns in Nepali are used to differentiate singular and plural also, they are gender-distinctive (boy, girl, man, woman).

Potato: आलु, Fish: माछा, Apple: स्याउ,Market: बजार

Figure 9: Some examples of nouns in Nepali language with their meanings in English.

9.3.2 Pronoun

Pronouns in Nepali language has 3 persons. Additionally it is divided into proximal and distal. Proximal is used to denote someone in proximity and distal is used to denote someone distant or absent. Depending upon the gender, distance, number and status of referent, Nepali pronouns has various levels of politeness.

- Low grade: Used to denote animals,small children, and pejoratively.
- Middle grade: Used to address younger or people of lower status then the speaker
- High grade: Used to address older or people of higher status then the speaker

Low Grade: तँ (ta)
Middle Grade: तिमी (timi) , उ (u) , उनि (uni)
High Grade: तपाईं(tapai) , हजुर(hajur)

Figure 10: Different classes of pronouns in Nepali language.

9.3.3 Verb

Verbs shows contrast between the first, second and the third persons along with singular and plural numbers. Similarly it also shows the contrast between masculine and feminine gender as well as the honorifics as.

जाउ = Go

1st Person : जान्छु (jaanchhu)
2nd Person : जान्छौ (jaanchhau)
3rd Person : जान्छ (jaanchha)

Singular : जान्छु (jaanchhu), जान्छस् (jaanchhas)
Plural : जान्छौ (jaanchhau), जान्छौं (jaanchhaun)

No Honorific : जान्छस् (jaanchhas)
Simple Honorific: जान्छौ (jaanchhau)
Super Honorific : जानुहुन्छ (jaanuhunchha)

Figure 11: Different types of verb usage in Nepali language.

9.3.4 Adjective

Adjectives in Nepali language are not any different from adjectives in other languages, as they are used to give further description of a noun or a pronoun.

राम्रो(raamro): Good, धेरै(dherai): Many,
सेतो(seto): White, रातो(raato): Red,
थोरै(thorai): Less, ठुलो(thulo): Big,

Figure 12: Some examples of adjectives in Nepali language.

9.3.5 Postposition

Prepositions always occur before the words they are intending to change in English. For instance, "to" appears before the word "school," which it modifies, in the sentence "we are going to school." A postposition serves the same purpose in Nepali as it does in English; the only difference is that it follows the word it modifies.

हामी स्कुलबाट आयौं ।
बाट = postposition

Figure 13: An example showing the position of a postposition in a sentence in Nepali language.

9.3.6 Sentence Structure

In English language the sentence structure is Subject - Verb - Object. But in Nepali language this structure is different. Sentences in Nepali language mostly ends with verb having standard structure as Subject - Object - Verb. It is shown in Figure 14.

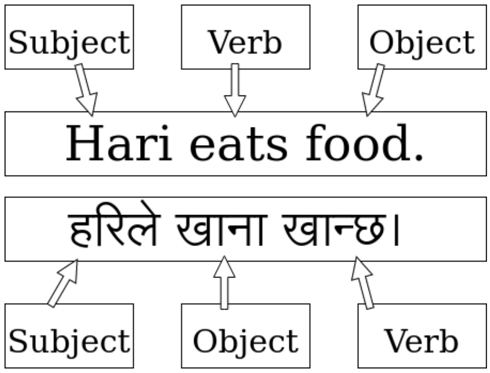

Figure 14: Sentence structure of Nepali language compared with English language.

9.4 Vocabulary

Although Nepali's primary lexicon has Sanskrit roots, it has also incorporated words from other languages over time. Compared to other Indo-Aryan languages, Nepali is more traditional, utilizing more vocabulary from Sanskrit and less ones from other languages. While spoken Nepali has several loanwords from the Tibeto-Burmese languages that are close by, written Nepali is mostly influenced by Sanskrit.