Tribhuvan University Journal
Vol. 39, No. 1: 124-137, June 2024
Research Directorate, Tribhuvan University (TU),
Kathmandu, Nepal
DOI: https://doi.org/10.3126/tuj.v39i1.66679

USE OF BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS (BERT) AND ROBUSTLY OPTIMIZED BERT PRETRAINING APPROACH (RoBERTa) FOR NEPALI NEWS CLASSIFICATION

Kriti Nemkul

Ratna Rajyalaxmi Campus, TU, Kathmandu
Corresponding Author: kriti.nemkul@gmail.com

Received date: 17 Mar. 2024 – Accepted date:19 May 2024

ABSTRACT

News classification is a technique of classifying news documents into predefined groups. One of the earliest problems in Natural Language Processing was the categorization of news. Huge number of news are generated from different news portals each day and it is difficult to consign the specific types of news from that portal. News must be assigned into respective appropriate classes as users want to read certain type of news automatically as per the need. Text classification has been done by using different machine learning algorithm like Support Vector Machine (SVM), Long Short-Term Memory (LSTM). However, Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized Bert Pretraining Approach (RoBERTa) have not been fully scrutinized for Nepali news classification tasks. This research develops two models for Nepali news classification namely BERT and RoBERTa by collecting news data from various national news portal. Precision, Recall, F1 score and accuracy are used to evaluate the effectiveness of the model. Both models are trained and tested with AdamW optimizer with learning rate 1e-5 i.e., 0.0001. While comparing both models, RoBERTa found to be better than BERT model with accuracy 95.3 percent.

Keywords: *NLP; LSTM; BERT; RoBERTa; AdamW; SVM; Transformer*

INTRODUCTION

Many online news portals are being rapidly developed which generates different types of news in various topics. Automatic categorization

of news and social media posts, has a wide range of applications, from recommendation systems to content analysis. Users must manually search for a certain type of news while using news portals since the content is not appropriately categorized. Since users' choices are superseded by sponsored material on social media, an appropriate method for classifying news is required to assist users in locating the relevant news category. Also, proper classification of news aids users in displaying the suggested news. In this study, news from various news sources were collected to classify news into several groups. Traditional machine learning classification techniques like support vector machine, naïve bayes and neural network has been used earlier for Nepali news classification task. Recently deep learning has become more popular in solving machine learning problems in compared to traditional machine learning approach. One of the deep learning architecture called Transformer and its derivation models have shown efficacy in various downstream natural language processing applications, especially for languages with abundant resources such as English. Compared to prior machine learning algorithms, deep learning frameworks have demonstrated greater promise in recent years. Deep learning networks' result is beginning to demonstrate significant improvements over those algorithms. Encoder and decoder are two main components of transformer. BERT just uses the transformers' encoder portion, and masking is performed only once at the time of data preparation which outperforms state-of-art on several NLP tasks. RoBERTa, variant of BERT uses self-attention to process input sequences. Additionally, dynamic masking technique is used in RoBERTa during training phase. Both BERT and RoBERTa are used in the improvement in NLP tasks by using embedding vector space which is rich in context.

LITERATURE REVIEW

C. Zhou et al. employed the C-LSTM, a unified model of recurrent and convolutional neural networks worked on their work in which a convolutional layer allowed C-LSTM to learn phrase-level characteristics; the LSTM is then fed sequences of these higher-level representations to understand enduring dependencies. They achieved highly good results when they assessed the acquired semantic phrase representations on tasks including sentiment classification and inquiry type categorization. (C. Zhou, 2015). S. Kaur and N. K. Khiva presented "Online news classification using Deep Learning Technique", where they used a neural network classifier to classify news articles into four different categories, achieving up to 81% precision (Khiva, 2016). Text classification, question-answering, and token

classification are just a few of the natural language processing (NLP) tasks on which A. Vaswani detailed Language models built on the Transformer architecture which obtained state-of-the-art performance (A. Vaswani, 2017). According to J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, BERT is strong experimentally and conceptually simple. On eleven tasks involving natural language processing, it achieves new state-of-the-art outcome. These includes raising the GLUE score to 80.5% (7.7%-point absolute improvement), improving MultiNLI accuracy to 86.7% (4.6% absolute improvement), answering Test F1 questions correctly on SQuAD v1.1 by 9.3.2 (J. Devlin, 2018). T.B. Shahi and A.k. Pant classified Nepali news using Naïve Bayes, support vector machine and neural network using TF-IDF based feature extraction method and received highest accuracy of 74.62% on linear SVM (T.B. Shahi, 2018). Research by M. Munikar, S. Shakya, and A. Shrestha has shown the BERT model works better than complex designs like as convolutional and recursive neural networks. For the fine-grained sentiment classification job using the Stanford Sentiment Treebank (SST) dataset, a trained BERT model is employed and fine-tuned. There are five classes (very negative, negative, neutral, positive, and extremely positive) based on fine-grained sentiment categorization. (M. Munikar, 2019). C. Sun, X. Qiu, Y. Xu, and X. Huang has conducted exhaustive experiments to examine various BERT fine-tuning techniques on text classification problems and provide a generic solution for BERT fine-tuning. Finally, the suggested approach obtains new state-of-the-art outcomes on eight extensively researched text classification datasets (C. Sun, 2019). S. González-Carvajal and E. C. Garrido-Merchán uses 75% training data ans 25% validation data to classify Portuguese news to 9 different classes. BERT framework produced 90.93% accuracy whereas automl accuracy was 84%. Also, BERT model was compared with other models like SVM, logistic Regression using TFI-DF to extract feature. BERT model outperforms other models with great improvement on accuracy (S. Gonzalez-Carvajal, 2020). K. Jain, A. Deshpande, K. Shridhar, F. Laumann, and A. Dash assessed the performance on Indian languages. They carried out in-depth experiments on a number of downstream tasks in Hindi, Bengali, and Telugu to analyze these language models and compare the effectiveness of fine-tuning model parameters of pre-trained models against that of training a language model from scratch (K. Jain, 2020). P. Kafle and et.al classified Nepali news using BERT and compared with LSTM, BiLSTM, GRU, BiGRU and found BERT model outperforming compared to other models (P. Kafle, 2022). Since deep learning algorithms are evolving

day by day, less work done has been seen in the task of classification specially in Nepali news classification. Lately, NepBERTa, a Nepali language model trained in a large corpus which is a BERT-based natural language understanding model has been introduced by T. Sulav, G. Milan and B. Binod model that has created benchmark in various NLP tasks like name entity recognition, parts of speech tagging, content classification and categorical pair similarity (T. Sulav, 2022).

PROBLEM STATEMENT

Large number of news content are being generated every day from popular news sites and consumed through different media. Users cannot go through all the articles and miss their interested news category. Many news portals use manual system to classify and recommend the news which are clicked and read by the user. Proper classification of the news is essential to read the updated news of interest. So, this research aims to collect large amount of Nepali news and use recent deep learning methods namely transformer-based BERT and RoBERTa to improve the performance and accuracy of Nepali news classification task.

RESEARCH METHODOLOGY

Figure 1

BERT model architecture (ResearchGate, 2024)

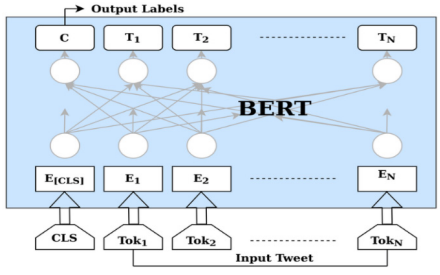
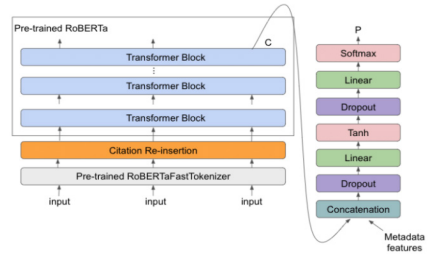


Figure 2

RoBERTa model architecture (ResearchGate, 2024)



BERT (Bidirectional Encoder Representation from Transformer)

Google developed transformer-based machine learning approach known as Bidirectional Encoder Representations from Transformers (BERT) for natural language processing (NLP) pre-training.

Pretraining and fine tuning are the two phases in the BERT model's structure. The model is pre-trained using a large corpus during pretraining. The pre-trained parameters are utilized to initialize the model for fine tuning, and each parameter is adjusted using labeled data tailored to the particular job at

hand. The multi-layer bidirectional Transformer encoder model architecture of BERT (K. He, 2015) based on the implementation defined on "Attention is all you need" (A. Vaswani, 2017) An encoder of this kind consists of a stock of N=6 identical layers which consists of 2 sub layer; multi head self-attention mechanism and simple position-wise fully connected feed forward network .It uses a residual connection (K. He, 2015) around each of the two sub layers, followed by normalization of layer (J. Devlin, 2018).Self-attention can be defined as :

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_K}}\right)V \quad (1)$$

where d_k denotes the Q and K matrices dimension, Q, K and V denotes matrix of queries, keys and values respectively and. Now, multi-head attention is defined as,

$$MultiHead(Q,K,V) = Concat(head1,...,headh)W^0 \quad (2)$$

where $head_i$ = Attention (QW_i^Q, KW_i^K, VW_i^V). Multi-head attention comprises projecting the queries, keys and value parameters in h-ways using various acquired linear projections to the dimensions of the values matrix d_Q, d_k and d_v respectively.

The attention function is then applied in parallel to each of these projected copies of the queries, keys, and values, producing d_v-dimensional output values. The final values are obtained by concatenating and projecting them. (A. Vaswani, 2017). Self-Attention indicates that all of the queries, values, and keys originate from the same location. BERT is intended to be a highly bidirectional model. From the first layer to the last layer, the network efficiently records data from a token's left and right context.

Robustly Optimized BERT Pretraining Approach (RoBERTa)

The BERT pretraining process is modified in the RoBERTa to enhance end-task performance. More specifically, RoBERTa is trained using large mini-batches, dynamic masking, FULL-SENTENCES sans Next sentence prediction (NSP) loss, and a higher byte-level BPE (Y. Liu, 26 jul 2019). Self-Attention mechanism used in RoBERTA model is same as of BERT model. RoBERTA used denoising autoencoder objective during pre-training procedure such that denoising autoencoder loss($L_{denoise}$) can be expressed as,

$$L_{denoise}(\theta) = \sum_{(t=i)}^{T} LogP(x_t \mid \text{Corrupt}(x_t); \theta) \quad (3)$$

Where, θ represents the model parameters, x_t is an input token, and Corrupt(x_t) is the corrupted version of x_t.

The total loss during training is a combination of the denoising autoencoder loss and task-specific losses:

$$L_{total}(\theta) = \lambda L_{total}(\theta) + \sum_i L_{task_i}(\theta) \quad (4)$$

Where, λ is a hyperparameter controlling the importance of the denoising objective relative to task-specific objectives. L_{task_i} represents the loss for a specific downstream task. (Y. Liu, 26 jul 2019)

DATA COLLECTION

Raw news data were collected from different Nepali news portal using beautifulsoap4(BS4) which is a library to scrape information from the web page (Leonard, 2024). News was collected from following news portals.

Table 1

News portal names and corresponding URLs

S.No.	Website Name	URL	Number of articles
1	DC Nepal	https://www.dcnepal.com	12842
2	Imagekhabar	https://www.imagekhabar.com	19296
3	Onlinekhabar	https://www.onlinekhabar.com	20504
4	Ujayaloonline	https://ujyaaloonline.com	7855
5	Ratopati	https://ratopati.com	11473
		Total	71970

Total news collected with corresponding categories is shown in below table.

Table 2

Number of news and its corresponding categories

S.No.	Category Name	Number of articles
1	Diaspora	6224
2	Economy	14069
3	Entertainment	7588
4	Health	3122
5	International	9879
6	Opinion	2675
7	Politics	8352
8	Society	4979
9	Sports	13018
10	Technology	2064
	Total	71970

DATA PREPROCESSING

The raw text extracted from above mentioned national news portals were collected with text in one column and category in another column. Data preprocessing is performed to reduce noise in the data by cleaning and generating the training and testing ready data. Following steps were followed during data preprocessing.

- Removal of HTML tags: During scrapping data were stored in HTML tags like div, headings, paragraph and so on. These tags were removed from the raw text.
- Removal of white space, special symbols: Special symbol such as, ÷, ×, º, >,<,/,@ etc. and white space that doesn't make any sense in classification were removed from the data for more clean data.
- Stop word removal: Standard stop word vocabulary from NLTK data consists of 255 stop words like मेरो, र, गरेनु, छैन, तर, साथ, समय, धेरै, कनि etc. which are high frequency words that has not much influence in the text were removed to enhance the efficiency of the classification.

LABEL ENCODING

Raw news data were gathered from above mentioned Nepali news sources for this study which are classified and labeled as follows.

Table 3

News Category and its corresponding label

S.No.	Category name	Description	Label
1	Diaspora	News linked with various topic	0
2	Economy	News related to economy	1
3	Entertainment	News related to entertainment	2
4	Health	News Related to health and health Sector	3
5	International	News Related with international issues	4
6	Opinion	News related with views of different person	5
7	Politics	News linked with political issues	6
8	Society	News related with social issues	7
9	Sports	News linked with sports	8
10	Technology	News linked with technology	9

MODEL BUILDING

Using the training set generated, the pretrain model from the multilingual cased version is utilized to train the model. In order to increase

accuracy and other model evaluation criteria, several parameters are adjusted after the accuracy and other criteria are evaluated.

Both BERT and RoBERTa model has the following configuration:

- 512-Sequence length
- BERT base's default number of encoder layer 12
- Number of epochs 4 for BERT and 3 for RoBERTa
- AdamW optimizer with learning rate 0.0001 i.e. 1e-5
- 12 Attention Heads; 768 input vectors

EVALUATION RESULT AND DISCUSSION

The experiment was performed in above model configuration with total dataset of 71970 news articles such that 20% of the data used for testing and the remaining 80% being used for training. 512-length sequences are used for testing and training. The experiment results were analyzed using four evaluation parameters namely Precision, Recall, F1 score and Accuracy (Lapalme, 2009).

Bert Model Evaluation

The table below shows the classification results for a BERT model in which Sports are classified with 99 percent precision among ten categories, whereas society is classified with 89 percent precision

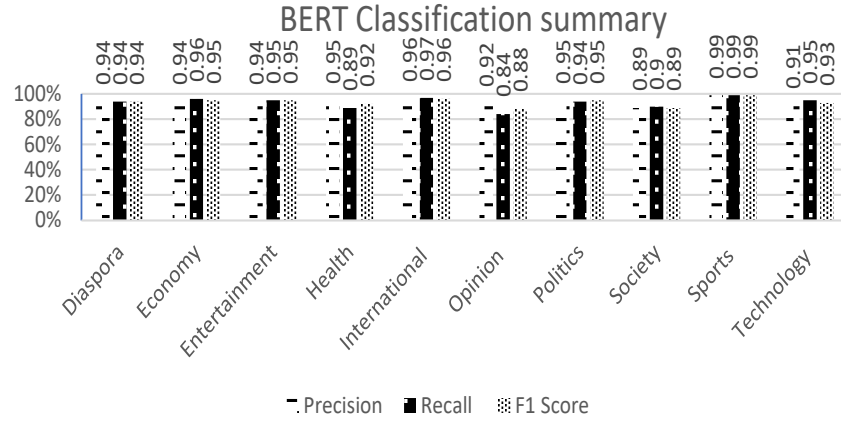
Table 4

Experiment result and accuracy summary for BERT model

Category	Precision	Recall	F1 Score	Support	Test News	Accurately classified news	Accuracy
Diaspora	0.94	0.94	0.94	1245	1245	1170	0.94
Economy	0.94	0.96	0.95	2814	2814	2701	0.96
Entertainment	0.94	0.95	0.95	1518	1518	1442	0.95
Health	0.95	0.89	0.92	624	624	555	0.89
International	0.96	0.97	0.96	1976	1976	1916	0.97
Opinion	0.92	0.84	0.88	535	535	449	0.84
Politics	0.95	0.94	0.95	1670	1670	1570	0.94
Society	0.89	0.90	0.89	996	996	896	0.90
Sports	0.99	0.99	0.99	2603	2603	2574	0.99
Technology	0.91	0.95	0.93	413	413	392	0.95
Avg/total	0.94	0.93	0.94	14394	14394	13665	0.93

Figure 3

Summarized accuracy result for BERT model



RoBERTa Model Evaluation

The table below shows the classification results for RoBERTa model in which Sports are classified with 99 percent precision among ten categories, whereas society is classified with 94 percent precision.

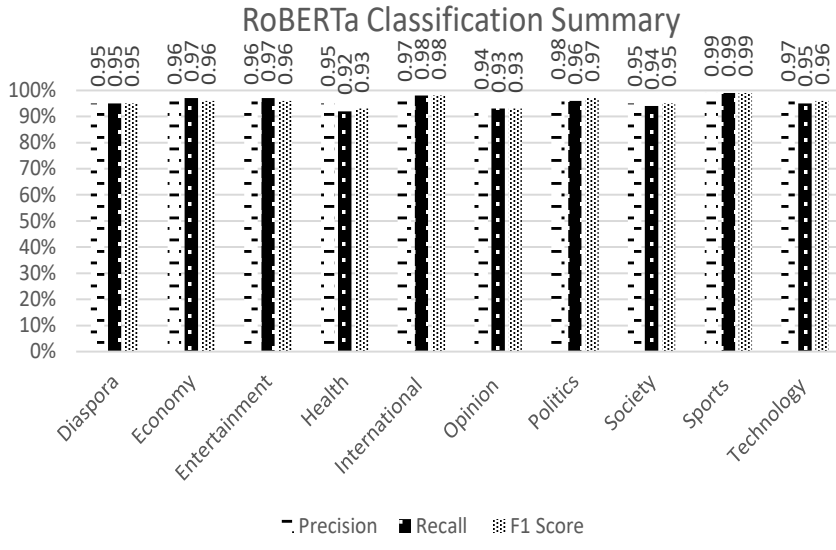
Table 5

Experiment result and accuracy summary for RoBERTa model

Category	Precision	Recall	F1 Score	Support	Test News	Accurately classified news	Accuracy
Diaspora	0.95	0.95	0.95	1245	1245	1184	0.95
Economy	0.96	0.97	0.96	2814	2814	2740	0.97
Entertainment	0.96	0.97	0.96	1518	1518	1469	0.96
Health	0.95	0.92	0.93	624	624	576	0.92
International	0.97	0.98	0.98	1976	1976	1930	0.97
Opinion	0.94	0.93	0.93	535	535	495	0.92
Politics	0.98	0.96	0.97	1670	1670	1608	0.96
Society	0.95	0.94	0.95	996	996	940	0.94
Sports	0.99	0.99	0.99	2603	2603	2574	0.99
Technology	0.97	0.95	0.96	413	413	394	0.95
Avg/total	0.96	0.96	0.96	14394	14394	13910	0.95

Figure 4

Summarized accuracy result for RoBERTa model



BERT and RoBERTa model comparative evaluation

Table below shows the overall comparative result of both models. The result is compared in terms of precision, recall, F1 score and Accuracy. The precision, recall, and F1 score of the RoBERTa model are higher than BERT model, as seen in the table below. RoBERTa's accuracy on experimented datasets is also seen higher than BERT's, i.e. 95.3 percent and 93.3 percent, respectively.

Table 6

BERT Vs. RoBERTa comparative result

Model	Precision	Recall	F1 Score	Accuracy
BERT	93.9%	93.3%	93.6%	93.3%
RoBERTa	96.2%	95.6%	95.8%	95.3%

Receiver Operating Characteristic (ROC) curve, graphical plotting of the true positive rate (TPR) against the false positive rate (FPR) analysis of both models are shown below. The ROC curve below shows that the area under the curve for the RoBERTa model is 1 for all the news labels whereas

area under the curve for the BERT model is 0.99 which is less than that of RoBERTa. Thus, RoBERTa shows more accurate result in news classification task in compared to BERT model.

Figure 5

ROC curve for (a) BERT and (b) RoBERTa model

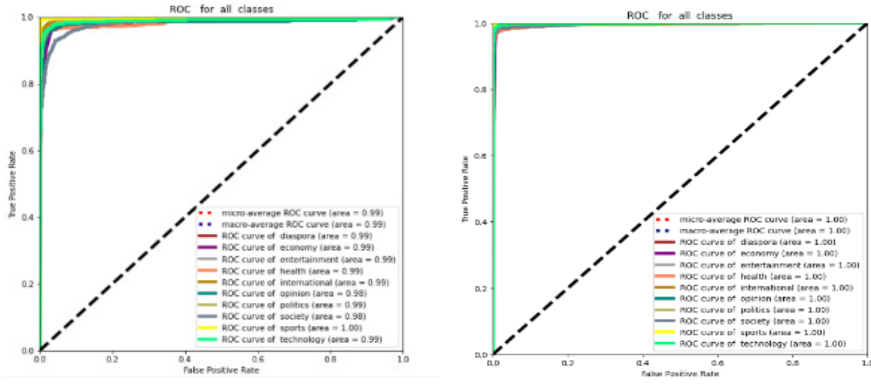
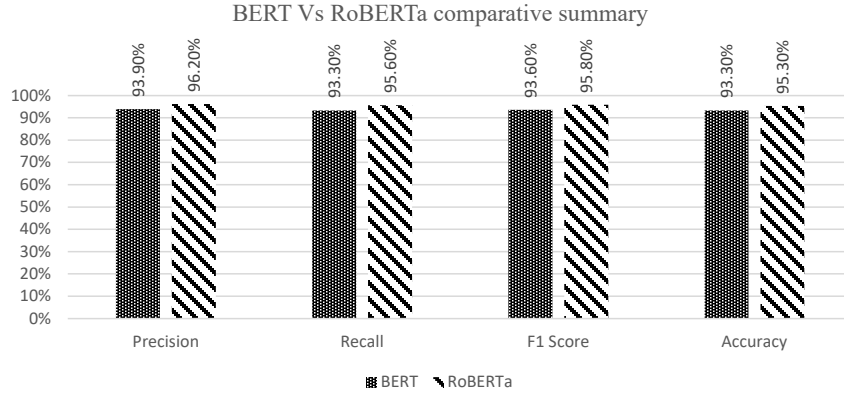


Figure 6

BERT Vs. RoBERTa model comparative summary



DISCUSSION

BERT and RoBERTA both are the recent deep learning mechanisms which make use of transformer and an attention mechanism. Due to the use of self-attention and multiheaded attention, models are able to store the information of long sequence, and is truly bidirectional. RoBERTa, variation of BERT model tends to increase the accuracy by performing masking during

training phase whereas BERT performs masking only once at data preparation time. As a result, each time a phrase is added to a minibatch, its masking is completed, and therefore unlike in BERT, there is no limit on the number of possible masked version of each sentence. RoBERTa, reimplementation of BERT uses byte-level BPE as a tokenizer with larger vocabulary instead of character level BPE vocabulary used in BERT. Dynamic masking and full sentence without NSP loss is used during pretraining phase in RoBERTa encouraging robust learning of the model and making language representation more general forcing it to predict missing tokens in an array of diverse contexts. Additionally avoiding NSP application, RoBERTa model also avoids the issues with the NSP job, such as the difficulty of producing negative samples and the potential to introduce biases into the pre-trained model, by not applying the NSP loss. In addition, fine tuning strategies of RoBERTA model gives more accurate result in compared to BERT model.

CONCLUSION

Nepali news is increasing rapidly in different categories. Nepali news collected from different news portals and that are labeled with predefined 10 labels for classification. In this study, transformer-based models are used to classify the collected news. BERT and RoBERTa both are transformer-based models which were implemented and system analysis has been conducted. The news data set is split into two sets: training and testing with a ratio of 8:2 for each category. To assess the system's effectiveness, it was trained using a learning rate of 1e-5, or 0.0001, and its analysis was conducted using several parameters. Overall analysis and result comparison of both models in terms of precision, recall, F1 score and accuracy conclude that the performance of RoBERTa model shows better result with precision value 96.2 percent, recall value 95.6% and F1 score 95.8%. The highest accuracy obtained is 95.3% from RoBERTa model. Due to the limited resource for training a model, news article is limited to 72 thousand and 10 classes only, more amount of news can be trained by taking more categories and other machine learning model can be implemented for more accurate result.

ACKNOWLEDGEMENTS

The author would like to acknowledge the Research Directorate, Rector's Office, Tribhuvan University for providing the financial support (grant number TU-smallresearch-2079/-80-R.N.01) to carry out the research work.

REFERENCES

A. Vaswani, e. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems, 30*, 1-11. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

C. Sun, X. Q. (2019). How to fine-tune BERT for text classification? *Lecture Notes in Computer Science, Cham: Springer International Publishing*, 194-206. https://doi.org/10.48550/arXiv.1905.05583

C. Zhou, C. S. (2015). A C-LSTM Neural Network for Text Classification. *arXiv [cs.CL]*. https://doi.org/10.48550/arXiv.1511.08630

J. Devlin, M.-W. C. (2018). BERT: Pre-training of deep bidirectional Transformers for language understanding. *arXiv[cs.CL]*. https://doi.org/10.48550/arXiv.1810.04805

K. He, X. Z. (2015). Deep residual learning for image recognition. *arXiv[cs.CV]*. https://doi.org/10.48550/arXiv.1512.03385

K. Jain, A. D. (2020). Indic-transformers: An analysis of Transformer language models for Indian languages. *arXiv[cs.CL]*, 1-14. https://doi.org/10.48550/arXiv.2011.02323

K. Kafle, D. S. (2016). TimalsinaImproving Nepali Document Classification by Neural Network. *IOE Graduate Conference*, 317-322.

Khiva, S. K. (2016). Online news classification using Deep Learning Technique . *International Research Journal of Engineering and Technology (IRJET), 3*(10).

Lapalme, M. S. (2009). A systematic snalysis of performance measure for classification tasks. *Information processing & management, 45*, 427-437.

Leonard. (2024, Jan 17). *beautifulsoup4 4.12.3*. https://pypi.org/project/beautifulsoup4/

M. Munikar, S. S. (2019). Fine-grained sentiment classification using BERT. *Artificial Intelligence for Transforming Business and Society (AITB)*. https://doi.org/10.48550/arXiv.1910.03474

P. Kafle, R. C. (2022). Improving Nepali News Classification Using Bidirectional Encoder Representation from Transformers. *Springer.* https://doi.org/10.1007/978-981-19-1653-3_36

ResearchGate . (2024, May 21). Retrieved from Context-Aware Legal Citation Recommendation using Deep Learning-Scientific Figure

on ResearchGate. https://www.researchgate.net/figure/The-RoBERTa-model-architecture_fig2_352642553

Researchgate. (2024, May 21). Retrieved from Transformer based automatic COVID-19 fake news detection system - Scientific Figure on ResearchGate. https://www.researchgate.net/figure/BERT-model-architecture_fig2_348214408

S. Gonzalez-Carvajal, E. G.-M. (2020). Comparing BERT against traditiona; machine learning text classification. *arXiv[cs.CL]*. https://doi.org/10.48550/arXiv.2005.13012

S. Subba, N. P. (2019, June). Nepali Text Document Classification using Deep Neural Network. *Tribhuvan University Journal, 33*(1), 11-22. https://doi.org/10.3126/tuj.v33i1.28677

T. Alam, A. K. (2020). Bangla Text Classification using Transformers. *arXiv:2011.04446[cs.CL], 1*. https://doi.org/10.48550/arXiv.2011.04446

T. Sulav, G. M. (2022). NepBERTa: Nepali Language Model Trained in a Large Corpus. *Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: AACL-IJCNLP 2022.* Retrieved from https://aclanthology.org/2022.aacl-short.34

T.B. Shahi, A. P. (2018). Nepali News Classification using Naive Bayes, Support Vector Machine and Neural Networks. *International Conference on Communication, Information & Computing Technology (ICCICT).* Mumbai, India.

W. de Vries, A. v. (2019). BERTje: A Dutch BERT Model. *arXiv [cs.CL]*.

Y. Liu, M. O. (26 jul 2019). RoBERTA: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692, 1*. https://doi.org/10.48550/arXiv.1907.11692