INTERNATIONAL JOURNAL FOR INNOVATIVE RESEARCH IN MULTIDISCIPLINARY FIELD
ISSN(O): 2455-0620 [Impact Factor: 7.581]
Monthly, Peer-Reviewed, Refereed, Indexed Journal with IC Value : 86.87
Volume - 9, Issue - 6, June - 2023 Publication Date: 10/06/2023

DOIs:10.2015/IJIRMF/202306003 --:-- Research Paper / Article / Review

TaNER: A Transformer Based Model for Named Entity Recognition on Tamil Language

[1]Livin Nector D, [2]Ananth S, [3]Arulselvi M, [4]Lokesh B, [5]Mahesh T, [5]Nesan R

[1]Student, Department of Computer Science and Engineering, Annamalai University, Chidambaram, India
[2]Student, Department of Computer Science and Engineering, Annamalai University, Chidambaram, India
[3]Associate Professor, Department of Computer Science and Engineering, Annamalai University, Chidambaram, India
[4]Student, Department of Computer Science and Engineering, Annamalai University, Chidambaram, India
[5]Student, Department of Computer Science and Engineering, Annamalai University, Chidambaram, India
[6]Student, Department of Computer Science and Engineering, Annamalai University, Chidambaram, India
Email - [1]livinnector2001@gmail.com, [2]anandh50sankar@gmail.com, [3]marulcse.au@gmail.com
[3]lokeshb9025@gmail.com, [4]mahesht2702@gmail.com, [5]nesanrocky27@gmail.com

Abstract: *Named Entity Recognition (NER) plays a vital role in many tasks in information extraction and organization which has a wide variety of applications in document classification, recommendation systems, and Natural Language Understanding (NLU) systems. While there are state-of-the-art deep learning models for NER in English, such models are not widely available for other Indic languages such as Tamil, Telugu, Hindi, etc. The Transformer models, which are more capable when it comes to NLP, have already established their performance on different NLP benchmarks. They tend to be the basis of Large Language Models(LLMs), which perform well on many NLP tasks. Though there exist many state-of-art transformer models for English, there exists only some for Tamil. In this work, a Transformer based Tamil Language model is developed from scratch, which is fine-tuned for performing NER on Tamil language text. The performance of the model is evaluated by using multilingual datasets such as FIRE2013, IndicGlue, and Namapadam benchmarks.*

Key Words: *Transformers, BERT, Tamil, Named Entity Recognition, Natural Language Understanding, Natural Language Processing.*

1. INTRODUCTION:

Named Entity Recognition (NER) is a benchmark task in Natural Language Understanding. In the NLP (Natural Language Processing) community, a model's ability to understand the text is analyzed by the tasks such as document classification, Part of Speech (POS)-tagging, and NER. In recent years, deep learning, specifically the transformative power of transformer-based models, has revolutionized NLP and significantly impacted the performance of NER systems.

Unlike traditional Recurrent Neural Networks (RNNs), transformers models such as BERT(Bidirectional Encoder Representation for Transformers)[1] leverage attention mechanisms to capture dependencies between words in a sequence, allowing for parallel processing and more efficient modeling of long-range dependencies. This architecture has proven to be highly effective for a wide range of NLP tasks, including NER. As the performance of these LLMs (Large Language Models) increase their size and training times also increase. To tackle this problem models like ALBERT(A Lite BERT) [2] and DistilBERT(Distilled BERT) [3] have been developed.

In this work, a smaller BERT model is trained which is then finetuned for performing NER in the Tamil language. The performance of our model is evaluated on different datasets and compared with other models.

2. LITERATURE REVIEW:

As NER is a crucial task in Natural Language Understanding (NLU) it has been analyzed using various approaches. In their respective studies, researchers have explored different machine learning-based approaches for Named Entity Recognition (NER) in the Tamil language. Malarkodi et al. [4] developed a CRF-based NER system that utilized word, Part-of-Speech (POS), and chunks features, achieving an F1 score of 70.68 on the Online Tourism corpus. Jeyashenbagavalli et al. [5] presented a comprehensive NER system for a private corpus, combining rule-based techniques with Hidden Markov Models (HMM), and achieving an impressive F1 score of 89.7. Abinaya et al. [6]

evaluated various machine learning models including Random Kitchen Sink (RKS), Support Vector Machines (SVM), and CRF, with the CRF-based model achieving the highest accuracy of 87.21 on the NER Track (FIRE) dataset. Theivendiram et al. [7] proposed a NER system for Tamil BBC newspaper articles, utilizing the Margin Infused Relaxed Algorithm (MIRA) and CRF, incorporating features such as gazetteers, POS, and orthographic features. The MIRA-based model achieved an F1 score of 81.38, while the CRF-based model achieved 79.13. These studies collectively demonstrate the effectiveness of different machine learning techniques in NER for Tamil, highlighting the importance of feature selection and model choice in achieving accurate named entity recognition.

In 2019, Hariharan et al. [8] experimented with various embedding techniques such as FastText, GloVe(Global Vectors), random embeddings, and an LSTM (Long-Short-Time-Memory)-based model for the FIRE (Forum for Information Retrieval and Extraction) 2018 and Wiki crawl datasets, achieving F1 scores ranging from 84.03 to 94.54.

In 2020, Kakawani et al.[9] trained two ALBERT-based models with Sentence Piece tokenization called IndicBERT base and IndicBERT large and compared their performance with existing BERT-based models such as XLM-R (Cross-Lingual Models - RoBERTa) [10], mBERT(multilingual BERT)[11]. They show that their IndicBERT base outperforms other models in Tamil NER achieving an F1 score of 90.45 for the IndicGlue dataset.

In 2021, Khanuja et al. [12] proposed MuRIL(Multilingual Representations for Indian Languages), a multilingual language model designed for Indian languages, addresses the limitations of existing models by training on significantly large amounts of Indian text corpora and incorporating supervised cross-lingual signals through translated and transliterated document\ pairs. The model outperforms mBERT on diverse tasks, including the challenging cross-lingual XTREME(Cross-lingual TRansfer Evaluation of Multilingual Encoders)[] benchmark, and demonstrates its efficacy in handling transliterated data.

In 2022, Mhaske et al.[13] utilized IndicNER a NER model fine-tuned from mBERT with Word Piece tokenization for the Naamapadam dataset using two fine-tuning methods: monolingual finetuning and multilingual finetuning, achieving F1 scores of 78.58 and 77.42, respectively.

IndicBERTv2[14], introduced in 2022 by Doddapaneni et al., is a pre-trained language model specifically designed for Indic languages from the Indian subcontinent. Built upon the IndicCorpv2 dataset, consisting of a vast collection of text data in 24 different languages. Results demonstrate that IndicBERT v2 outperforms existing multilingual language models such as XLM-R and MuRIL, showcasing its ability to effectively handle the complexities of Indic languages.

Although these transformer-based models provide state-of-art performance they are very large. So we aim to create a BERT-based transformer model that is relatively small in size and has a small vocabulary limited only to the Tamil language, to understand the language better.

3. METHOD:

There are two different sets of datasets required to train the proposed BERT-based model for NER:

- Unlabelled text corpus for Unsupervised Pretraining
- Labeled NER dataset for fine-tuning.

In this work, a combination of the Wikipedia Corpus[15] and the OSCAR corpus[8] is used as the pretraining corpus, specifically focusing on the Tamil subset. Subsequently, for fine-tuning the BERT model, a combination of the Naamapadam and IndicGlue datasets is used. These different datasets are collected and preprocessed using □(HuggingFace) Datasets [16] library and uploaded to the □ dataset repository.

To implement the transformer model, a Wordpiece tokenizer is trained from scratch (implemented using □ Tokenizer[17] library) using the unlabelled Tamil text corpus mentioned above. This was accomplished using the Hugging Face tokenizer library. Four tokenizers are trained with vocabulary sizes of 500, 1000, 2000, and 4000. We chose small vocabulary sizes as compared to various BERT-based models, as Tamil being one of the morphologically rich languages contains mostly subword information as suggested by the work of Hariharan et.al. [8]. We chose four vocabulary sizes to study the impact of different vocabulary sizes in the Tamil language.

Four BERT-based models with 6 transformer blocks (L=6), and 12 attention heads (A=12) each with a hidden size of 768 (H=768) and an intermediate size of 3072 (I=3072) with different vocabulary sizes (500, 1000, 2000, and 4000) are pre-trained using Masked Language Modelling (MLM) Target. We call these models TaBERT-500, TaBERT-1k, TaBERT-2k, and TaBERT-4k according to their vocabulary sizes. These models were pre-trained on 4 NVIDIA-V100 GPUs each with 32 GB VRAM for 20 epochs (took about 19 hours) with a per device batch size of 64 on the unlabelled Tamil text corpus. For the implementation of the TaBERT model, the □ Transformer[18] library is used.

To perform fine-tuning on Token Classification, the MLM head of the pre-trained model is replaced with a token classification head(a dropout layer followed by a linear layer). Four different models are fine-tuned with different vocabulary sizes(500, 1000, 2000, and 4000) on the same hardware that is used for pretraining using the NER dataset described above (combination of Naamapadam and IndicGLUE train sets) for 15 epochs(took about 5 hours). We call these models TaNER-500, TaNER-1k, TaNER-2k, and TaNER-4k(Tamil Named Entity Recognition).

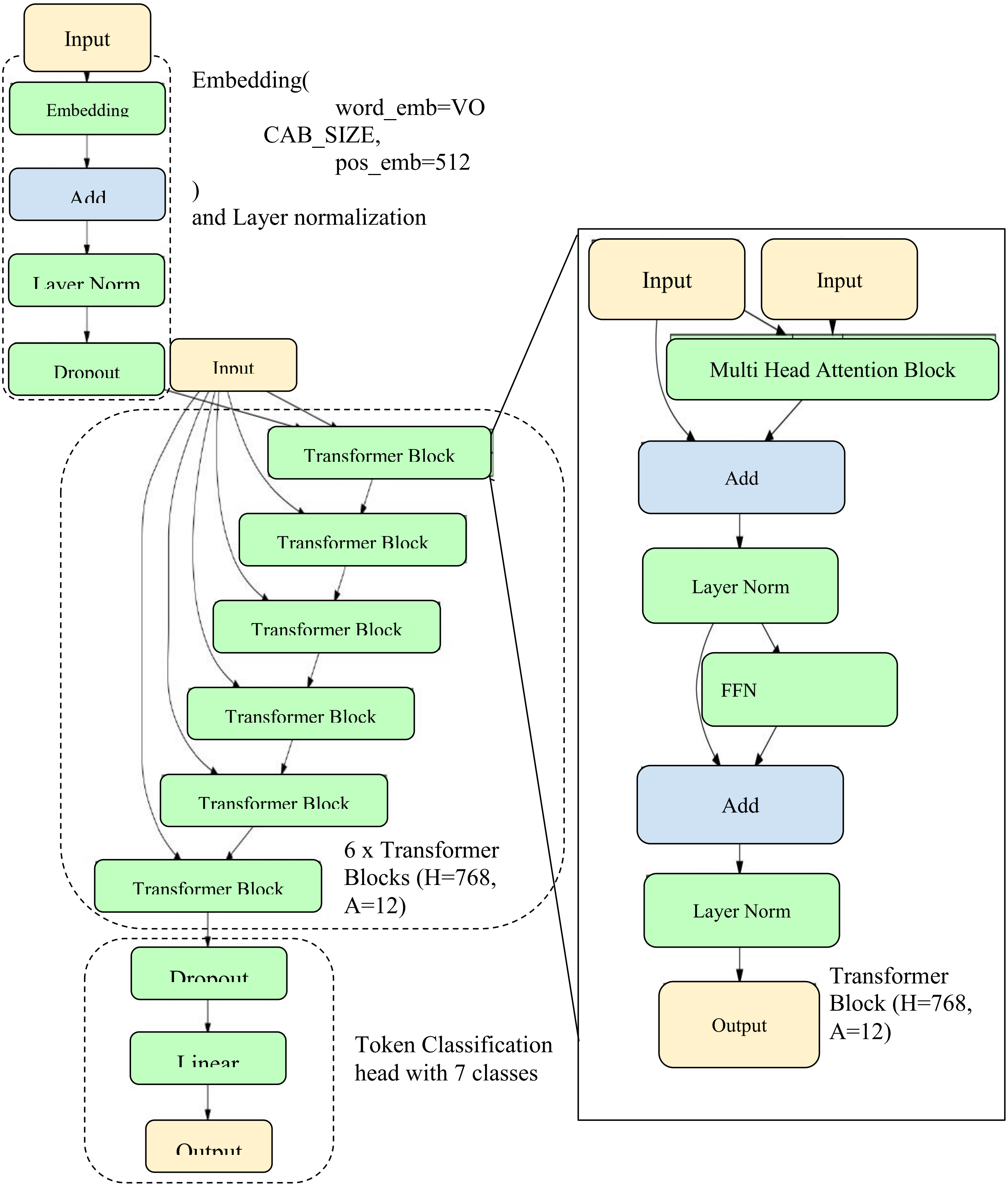

Figure 1: Architecture of the TaNER model

Also, large language models such as Distil-mBERT, IndicBERTv2-MLM-only, and IndicBERT-v2-MLM - SAM-TLM are fine-tuned and compared their performance with our TaNER models.

These NER pipelines of the TaNER and other models trained in this work are made available publicly using a gradio app hosted in Spaces in the URL "https://huggingface.co/spaces/livinNector/TaNER".

4. RESULTS AND DISCUSSION:

The TaNER models are evaluated on the FIRE-2013 [19], IndicGLUE, and Naamapadam datasets. Also, the results of the TaNER model are compared with other Large Language Models such as {model_names}.

The results from Table 1, Table 2, and Table 3 clearly state that our model performs on par with other models that have relatively small vocabularies and fewer parameters compared to other multilingual models. This outcome highlights the efficiency and effectiveness of our model architecture, which optimizes resource utilization without compromising performance. These findings demonstrate the potential of our model in delivering competitive performance while minimizing computational requirements, making it a promising solution in the field of multilingual natural language processing.

Table 1: Comparison of performance of different models in IndicGLUE dataset

model	F1 score	PER F1	LOC F1	ORG F1
Distil-mBERT	64.02	79.38	58.12	59.79
IndicBERTv2-MLM-only	66.92	84.46	61.66	60.87
IndicBERTv2-MLM-Sam-TLM	68.55	85.66	61.47	64.93
IndicNER	42.45	68.26	32.76	35.89
MuRIL	83.48	90.42	81.17	81.38
XLM-R	66.21	82.61	57.91	64.59
TaNER-500	90.56	93.49	89.66	89.55
TaNER-1k	90.90	94.74	88.95	90.42
TaNER-2k	**91.83**	**95.50**	**90.37**	**90.93**
TaNER-4k	90.15	94.38	88.18	89.45

Table 2: Comparison of performance of different models in Naamapadam dataset

model	F1 score	PER F1	LOC F1	ORG F1
Distil-mBERT	65.32	75.34	69.20	45.74
IndicBERTv2-MLM-only	**73.76**	**83.60**	74.67	**58.57**
IndicBERTv2-MLM-Sam-TLM	67.17	79.70	64.62	50.82
IndicNER	73.36	80.93	**75.29**	59.49
MuRIL	52.96	55.47	65.38	37.87
XLM-R	68.28	78.03	70.37	51.18
TaNER-500	64.45	74.03	66.30	47.77
TaNER-1k	65.19	75.08	70.74	44.31
TaNER-2k	63.76	76.21	64.73	44.03
TaNER-4k	66.32	77.81	67.97	47.48

Table 3: Comparison of performance of different models in FIRE-2013 dataset

model	F1 score	PER F1	LOC F1	ORG F1
Distil-mBERT	46.76	50.43	55.78	8.11
IndicBERTv2-MLM-only	**53.02**	**59.11**	**60.77**	**10.53**

model	F1 score	PER F1	LOC F1	ORG F1
IndicBERTv2-MLM-Sam-TLM	48.19	52.77	56.09	8.78
IndicNER	51.99	55.12	59.28	10.36
MuRIL	37.77	50.86	47.26	6.00
XLM-R	48.01	52.29	56.01	10.01
TaNER-500	47.49	52.32	56.86	8.00
TaNER-1k	46.28	50.80	56.26	6.72
TaNER-2k	45.41	49.86	54.22	6.47
TaNER-4k	45.86	50.80	55.33	6.41

5. CONCLUSION:

In this work different Transformer models are fine-tuned exclusively for Named Entity Recognition task in Tamil Language using IndicGLUE, Naamapadam and FIRE-2013 datasets. This work presents a successful named entity recognition (NER) pipeline for Tamil language text. The model was compared with other existing models such as IndicBERTv2, MuRIL, distil-mBERT, XLM-R, mBERT, and IndicNER. The pipeline achieves performance on par with other transformer-based NER models while utilizing fewer parameters. The NER pipeline of the TaNER models is hosted in HuggingFace Spaces in the URL "https://huggingface.co/spaces/livinNector/TaNER".

REFERENCES:

1. Kenton, J. D. M. W. C., & Toutanova, L. K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT (pp. 4171-4186).
2. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.
3. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.
4. Malarkodi, C. S., Rao, P. R., & Devi, S. L. (2012). Tamil NER–Coping with Real Time Challenges. In proceedings of Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012). In 24th International Conference on Computer Linguistics.
5. Jeyashenbagavalli, N., Srinivasagan, K. G., & Suganthi, S. (2014). An automated system for Tamil named entity recognition using hybrid approach. In International Conference on Intelligent Applications.
6. Abinaya, N., Kumar, M. A., & Soman, K. P. (2015). Randomized kernel approach for named entity recognition in Tamil. Indian Journal of Science and Technology, 8(24), 7.
7. Theivendiram, P., Uthayakumar, M., Nadarasamoorthy, N., Thayaparan, M., Jayasena, S., Dias, G., & Ranathunga, S. (2018). Named-entity-recognition (ner) for tamil language using margin-infused relaxed algorithm (mira). In Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing 2016, Konya, Turkey, April 3–9, 2016, Revised Selected Papers, Part I 17 (pp. 465-476). Springer International Publishing.
8. Hariharan, V., Anand Kumar, M., & Soman, K. P. (2019). Named entity recognition in Tamil language using recurrent based sequence model. In Innovations in Computer Science and Engineering: Proceedings of the Sixth ICICSE 2018 (pp. 91-99). Springer Singapore.
9. Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N. C., Bhattacharyya, A., Khapra, M. M., & Kumar, P. (2020, November). IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 4948-4961).
10. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.
11. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
12. Khanuja, S., Bansal, D., Mehtani, S., Khosla, S., Dey, A., Gopalan, B., ... & Talukdar, P. (2021). Muril: Multilingual representations for Indian languages. arXiv preprint arXiv:2103.10730.

13. Mhaske, A., Kedia, H., Doddapaneni, S., Khapra, M. M., Kumar, P., Murthy V, R., & Kunchukuttan, A. (2022). Naamapadam: A Large-Scale Named Entity Annotated Data for Indic Languages. arXiv preprint arXiv:2212.10168.
14. Doddapaneni, S., Aralikatte, R., Ramesh, G., Goyal, S., Khapra, M. M., Kunchukuttan, A., & Kumar, P. (2022). IndicXTREME: A Multi-Task Benchmark For Evaluating Indic Languages. arXiv preprint arXiv:2212.05409.
15. https://dumps.wikimedia.org/
16. https://huggingface.co/docs/datasets/index
17. https://huggingface.co/docs/tokenizers/index
18. https://huggingface.co/docs/transformers/index
19. Sobha Lalitha Devi., Pattabhi RK Rao., C.S Malarkodi and R Vijay Sundar Ram. 2013. Indian Language NER Annotated FIRE 2013 Corpus (FIRE 2013 NER Corpus), In Named-Entity Recognition Indian Languages FIRE 2013 Evaluation Track.