Identifying Nuances of Multi-Task Learning for Bengali and English Emotional Texts

Anonymous EMNLP submission

Abstract

In this paper, we present a multi-task learning (MTL) model to classify sentiment and emotion in Bengali and English languages. For this multi-task learning work, different Bengali and English datasets were collected from publicly available sources and developed two MTL models by utilizing pre-trained mBERT and MuRIL models. Our proposed MTL model outperforms their corresponding standalone classifiers with an average F1-score of 0.5728 (+0.041) and 0.7590 (+0.046) for Bengali sentiment, emotion and English sentiment, emotion classification tasks respectively.

1 Introduction

Bengali is the 6th most popular language in the world spoken by over 200 million peoples[1]. Also, it is the second most spoken language Indian subcontinent and the national and most widely spoken language in Bangladesh. Furthermore, with the popularity of social media and the internet, the number of Bengali language-spoken users significantly increased in the past few years. As of 2023, the total internet users in Bangladesh were 66.94 million and among them, 44.7 million were social media users which is 26% of the total population (Kemp, 2023). In addition, the number of internet users increased by 22.33 million between 2021 and 2023 (Kemp, 2021).

Over the decades, with the advancement of machine learning and deep learning techniques, NLP methods can efficiently find sentiments and emotions in social media and other texts not only in English languages but also in low-resource languages such as Bengali. However, most of the research focused on only learning one task: either sentiment classification or emotion analysis. But to find both sentiment and emotion in a sentence or text, we have to execute two separate models which may increase overhead.

[1] https://salc.uchicago.edu/language-study/bengali

Multi-task learning (MTL) as the name suggests, is a machine-learning technique that is capable of learning and handling multiple tasks at the same time. Researches show that, in the majority of cases, MTL models perform significantly better than their corresponding standalone classifiers for similar kinds of tasks.

In this present article, we focused on developing a multi-task learning framework for sentiment and emotion classification for Bengali and English texts and analyzed the performances of MTL models with their corresponding standalone classifiers. The main contributions of this paper can be summarized as follows:

- We presented two MTL schemes by using transformer-based pre-trained multilingual-BERT (mBERT) (Devlin et al., 2019) and MuRIL (Khanuja et al., 2021) models and compared the performances with standalone classifiers for Bengali and English languages.
- Our proposed MTL models provide a superior result than their corresponding standalone classifiers for both Bengali and English languages.

2 Related Work

The concept of MTL was first proposed by (Caruana, 1997). Since then, MTL approaches have been efficiently used in different domains of computer science including NLP.

(Liu et al., 2016) proposed three LSTM-based MTL frameworks for text classification where two models were unidirectional LSTM based and the third model was bidirectional LSTM based with two classification heads in each MTL framework. The authors evaluated their models for some popular NLP datasets such as 'SST-1', 'IMBD', etc. and their proposed MTL framework provide better performances over single-task learning.

An adversarial MTL framework was proposed by (Liu et al., 2017) primarily using LSTMs and the authors achieved a better performance in their proposed adversarial MTL model on 16 different datasets.

(Majumder et al., 2019), (Tan et al., 2023) and (Savini and Caragea, 2020) proposed MTL frameworks for sentiment and sarcasm classification. (Majumder et al., 2019) used GRU-based architecture and attention mechanism to classify sentiment and sarcasm whereas (Tan et al., 2023) and (Savini and Caragea, 2020) used BiLSTM in their study. In addition, (Savini and Caragea, 2020) used a non-contextual pre-trained embedding FastText (Bojanowski et al., 2016), which elevates their performance. Another sentiment and sarcasm analysis MTL work was proposed by (El Mahdaouy et al., 2021) using the pre-trained BERT (Devlin et al., 2019) model where the authors focused only on Arabic languages.

(Singh et al., 2022) proposed an MTL architecture for emoji, sentiment and emotion analysis by using the xlm-RoBERTa-base (Liu et al., 2019) and their proposed multi-task learning classifiers provide better performances than standalone classifiers. The authors also analyse sentiment and emotion intensities in their studies along with the classification tasks.

An MTL framework was for sentiment, emotion, target analysis (targeting a specific community such as black people, women, LGBT etc.), hate speech and offensive language classification by (Del Arco et al., 2021) using the BERT (Devlin et al., 2019) models.

In this present article, we proposed a multi-task learning framework for sentiment and emotion in Bengali and English text. As per our literature, this type of work is new and not widely explored in the context of the Bengali language.

3 Dataset

For this MTL work, we prepared two separate datasets: one was annotated with three sentiment labels (positive, negative and neutral) and another was annotated with six emotion labels (anger, fear, happy, sad, disgust and surprise).

For the emotion dataset, we collected 6314 samples from the 'BanglaEmotion' dataset (Rahman, Md Ataur, 2020). The 'BanglaEmotion' dataset was prepared from the users' comments from two different Facebook groups in Bangladesh and annotated each comment with one of six emotion labels: angry, fear, happy, sad, disgust and surprise.

Keeping these six emotion labels in mind, we extended this dataset to another 6314 English texts collected from 'GoEmotion' (Demszky et al., 2020) and 'emotion_dataset' (Saravia et al., 2018).

For the Sentiment dataset, the Bengali texts were collected from the 'SentNoB' dataset (Islam et al., 2021). The 'SentNoB' dataset was prepared from the social media users' comments on news and videos with a sample size of around 15K, annotated each comment with one of three sentiment labels: positive, negative, and neutral. However, we considered only 9519 samples from this dataset.

After that, the subset of the 'SentNoB' dataset was extended to the English texts, collected from the 'tweet_sentiment_multilingual' dataset (Barbieri et al., 2022) with around 3.03K records.

Next, 10% of Bengali and English texts were split out from the sentiment and emotion dataset and preserved for testing purposes. The data distributions for the sentiment and emotion datasets are provided in Figures 1 and 2 respectively.

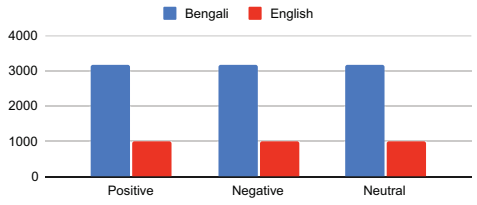

Figure 1: Distribution of Bengali and English languages in Sentiment dataset

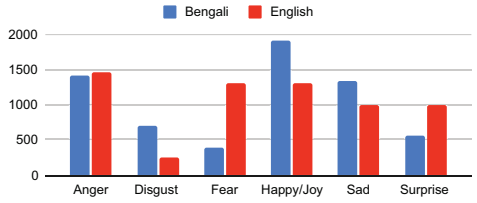

Figure 2: Distribution of Bengali and English languages in Emotion dataset

4 Methodology

This section discusses the methodologies of our proposed work. We aim to develop a multi-task learning framework that can classify sentiments and emotions at a time time in Bengali and English texts. To do that we took the help of the

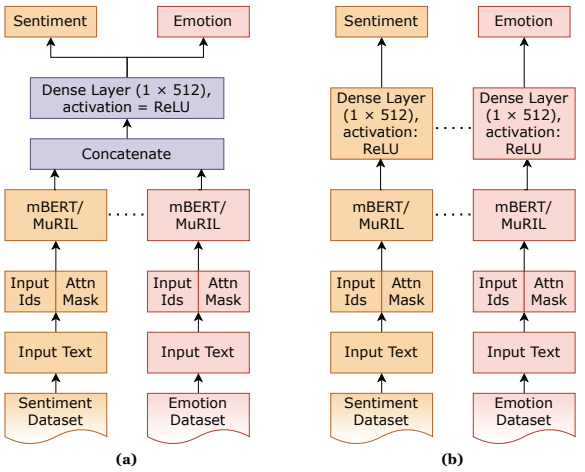

Figure 3: Proposed MTL models. (a) MTL architecture with a shared dense layer, (b) MTL architecture with task-specific dense layers.

pre-trained multilingual BERT (mBERT) (Devlin et al., 2019) and MuRIL (Multilingual Representations for Indian Languages) (Khanuja et al., 2021) models and developed the model architecture. The proposed MTL models are provided in Figure 3.

Tokenization: Before going to the actual training process, the training input text were converted into a sequence of tokens. For this tokenization process, we used the pre-trained mBERT and MuRIL tokenizers for their corresponding models with a maximum sequence length of 256. Each tokenizer returns a sequence of input ids and attention masks which were fed into the pre-trained models as depicted in Figure 3.

Model Selection: For this MTL work, we had chosen mBERT (Devlin et al., 2019) and MuRIL (Khanuja et al., 2021) models. The reason behind this, the mBERT is a transformer-based model and was trained on a total of 104 languages including Bengali and English. So, it can learn both Bengali and English contexts in a sentence or text. On the other hand, the MuRIL follows the BERT-base architecture and was trained on 17 Indian languages and this model provides better results than mBERT in different benchmark datasets. In addition, since MuRIL is specifically trained on 17 Indian languages, it may learn Indian languages (such as Bengali) in more better way than mBERT.

Next, after passing the input ids and attention masks to the mBERT/MuRIL models the pooled output of the last layer was fed into a Dense layer. In this stage, two experiments were performed: one is a shared dense layer, where the pooled output from two mBERT/MuRIL models was concatenated and then fed into a single dense layer of 512 neurons. We say this as MTL$_{shared}$. In the second approach, instead of concatenating, the pooled output of two mBERT/MuRIL models was fed into two separate dense layers of 512 neurons. We say this as MTL$_{task\text{-}specific}$. All the dense layers used ReLU as their activation function.

Classification: For the MTL$_{shared}$ approach the output of the shared dense layer was fed into the sentiment and emotion output layers (Figure 3(a)) and for the MTL$_{task\text{-}specific}$ approach the outputs of the separate dense layers fed into each sentiment and emotion output layers (Figure 3(b)) with 3 and 6 hidden units for sentiment and emotion classification heads respectively. All the output layers used the softmax as their activation function.

Training: Before beginning the training process we randomly split the training dataset into 9:1 ratio where 90% data were used for training and 10% data were used as validation split.

We trained our proposed model up to 4 epochs with a learning rate of 2e-5 and for the multi-task loss function, the SparseCategoricalCrossentropy loss function was used and monitored the loss for the validation split of the training dataset.

$$L_{total} = L_{sentiment} + L_{emotion}$$

Where $L_{sentiment}$ and $L_{emotion}$ represent the loss for sentiment and emotion tasks. The hyperparameters that were used to train the model are provided in Table 1.

Parameter	Value
Dropout	0.1
Loss function	SparseCategoricalCrossentropy
Optimizer	Adam (Kingma and Ba, 2014)
Learning rate	2e-5
Epoch	4
Batch size	8

Table 1: Hyperparameters used to train the model

5 Experiment and Result

5.1 Experimental Setup

All the experiments were executed using the libraries of 'TensorFlow' and 'Keras' and the pre-trained tokenizers and models were imported from the 'HuggingFace'.

To evaluate the performances of proposed MTL models, for sentiment classification, we passed

Task	Model	STL				MTL_{shared}				$MTL_{task\text{-}specific}$			
		Bengali		English		Bengali		English		Bengali		English	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Sentiment	mBERT	0.6296	0.6331	0.6142	0.6125	0.6227	0.6148	0.6142	0.6124	0.6458	0.6482	0.6296	0.6271
	MuRIL	0.6609	0.6639	0.6265	0.6278	**0.6806**	**0.6803**	0.6420	0.6411	0.6655	0.6683	**0.6852**	**0.6858**
Emotion	mBERT	0.4873	0.3319	0.8244	0.7985	0.5056	0.3485	**0.8560**	**0.8323**	0.5015	0.3514	0.8354	0.8182
	MuRIL	0.5142	0.3998	0.8339	0.8115	**0.557**	**0.4654**	0.8528	0.8215	0.5348	0.4139	0.8528	0.8342

Table 2: Performance comparison of STL vs proposed scheme of MTLs for both Bengali and English languages on the test dataset. (All the F1-scores provided here are the macro F1-scores.)

separately Bengali and English sentiment test data and compared the accuracy and F1-scores with standalone classifiers (STL), and the same was done for the emotion classification task.

5.2 Result

The results for STL, MTL_{shared} and $MTL_{task\text{-}specific}$ models are provided in Table 2 for both Bengali and English language. From Table 2 we can see that our proposed MTL frameworks outperform to their corresponding STL models for both Bengali and English test datasets. For Bengali sentiment classification, we see a performance improvement (F1-score) of 2.41%, as well as for English sentiment classification, the $MTL_{task\text{-}specific}$ model shows an improvement (F1-score) of 8.46% than their corresponding STL model. Additionally, our sentiment classification result provides a better F1-score than the best result provided by (Islam et al., 2021) [2].

For the emotion classification task, the MTL models show a performance improvement (F1-score) of 14.15% and 4.06% with respect to the STL models for Bengali and English emotion test datasets respectively. In addition, our MTL_{shared} Bengali emotion classification result provides an improvement in accuracy by 4.88% and an improvement in F1-score by 28.58% than the best accuracy and F1-score provided by (Rahman and Seddiqui, 2019) [3].

Furthermore, It can also be observed that the MuRIL models provide better performance than the mBERT models for the Bengali language. This is because the MuRIL model trained on specifically 17 Indian languages, so it can learn Bengali contexts in more better way than the mBERT model which was trained on 104 languages.

Additionally, if we closely observe the results, we can see that, the MTL_{shared} approach learns emotions in a better way than the $MTL_{task\text{-}specific}$ approach. On the other hand for sentiment classification results, the MTL_{shared} gives a better performance for Bengali languages. However, in the context of English languages, the MTL_{shared} model failed to provide a good result whereas the MuRIL-based $MTL_{task\text{-}specific}$ model provides superior results than MTL_{shared} and STL approaches.

6 Conclusion

In this paper, we proposed an MTL framework for sentiment and emotion classification in Bengali and English languages by transformer-based pre-trained models mBERT and MuRIL and our proposed MTL models outperform their corresponding STL models in both Bengali and English languages. Also, for the Bengali language, the MuRIL-based MTL models perform better than the mBERT-based MTL models.

In future, we'll expand our existing dataset to make a more robust model. Additionally, we'll consider other Indian languages such as Hindi and CodeMixed texts (mixing of Bengali and English languages) in future works.

7 Limitations

Our proposed work also has some limitations. Firstly, the dataset size in this experiment is relatively small (around 12K samples per dataset). Secondly, we have considered only two multilingual models: mBERT and MuRIL. There are also more available multilingual models such as xlm-RoBERTa (Liu et al., 2019) or IndicBERT (Kakwani et al., 2020) etc., and we'll explore them in the future. Also, not considering the 'large' models such as 'mBERT-large' or 'MuRIL-large' is one of the limitations of this work. Thirdly, we have considered only Bengali and English languages for this MTL work. And fourth, we only performed our experiments with a batch size of 8 and did not perform the experiments with a higher

[2] *We considered only a subset of the full dataset with sample size around 9.5K, the author's original dataset was with a sample size of around 15K.*

[3] *The authors recorded 0.5298 and 0.3324 as their best accuracy and F1-score respectively.*

batch size (16, 32, 64, etc.) due to resource limitations.

References

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Rich Caruana. 1997. Multitask Learning. *Machine learning*, 28(1):41–75.

Flor Miriam Plaza Del Arco, Sercan Halat, Sebastian Padó, and Roman Klinger. 2021. Multi-Task Learning with Sentiment, Emotion, and Target Detection to Recognize Hate Speech and Offensive Language. *arXiv (Cornell University)*.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Abdelkader El Mahdaouy, Abdellah El Mekki, Kabil Essefar, Nabil El Mamoun, Ismail Berrada, and Ahmed Khoumsi. 2021. Deep multi-task model for sarcasm detection and sentiment analysis in Arabic language. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 334–339, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. 2021. SentNoB: A dataset for analysing sentiment on noisy Bangla texts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3265–3271, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Simon Kemp. 2021. Digital in Bangladesh: All the statistics you need in 2021 DataReportal Global Digital Insights.

Simon Kemp. 2023. Digital 2023: Bangladesh DataReportal Global Digital Insights.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv (Cornell University)*.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. *arXiv (Cornell University)*, pages 2873–2879.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–10, Vancouver, Canada. Association for Computational Linguistics.

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. *CoRR*, abs/1907.11692.

Navonil Majumder, Soujanya Poria, Haiyun Peng, Niyati Chhaya, Zhaoxia Wang, and Alexander Gelbukh. 2019. Sentiment and sarcasm classification with multitask learning. *IEEE Intelligent Systems*, 34(3):38–43.

Md. Ataur Rahman and Md. Hanif Seddiqui. 2019. Comparison of classical machine learning approaches on bangla textual emotion analysis.

Rahman, Md Ataur. 2020. Banglaemotion: A benchmark dataset for bangla textual emotion analysis.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Edoardo Savini and Cornelia Caragea. 2020. A multi-task learning approach to sarcasm detection (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(10):13907–13908.

Gopendra Vikram Singh, Dushyant Singh Chauhan, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Are emoji, sentiment, and emotion

Friends? a multi-task learning for emoji, sentiment, and emotion analysis. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 166–174, Manila, Philippines. Association for Computational Linguistics.

Yik Yang Tan, Chee-Onn Chow, Jeevan Kanesan, Joon Huang Chuah, and YongLiang Lim. 2023. Sentiment Analysis and Sarcasm Detection using Deep Multi-Task Learning. *Wireless Personal Communications*, 129(3):2213–2237.