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

Anonymous EMNLP submission

## Abstract

Multi-task learning (MTL), a powerful paradigm in the field of machine learning enables us to learn and handle multiple different tasks simultaneously. Numerous advantages and novel approaches of MTL inspire us to analyze how MTL performs for low-resource languages such as Bengali. This paper proposes a fusion-based multilingual MTL framework for sentiment and emotion classification in Bengali and English languages with the help of transformer-based multilingual BERT and MuRUL models. Our fusion-based best-performing MTL framework achieves a macro F1 score of 71.14 and 38.92 for sentiment and emotion classification in the Bengali language and 67.09 and 83.48 for sentiment and emotion classification in the English language.

## 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.

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).

Over the decades, with the advancement of artificial intelligence, 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.

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

In this present article, we focused on developing an MTL framework to identify sentiment and emotion for Bengali and English texts to analyze how the joint learning of sentiment and emotion impacts the performance of the tasks over their corresponding single-task learning (STL) performance i.e. only sentiment classification or emotion classification. The main contributions of this paper can be summarized as follows:

- We have developed two multilingual STL models for sentiment and emotion classification using transformer-based multilingual BERT (Devlin et al., 2019) (mBERT) and MuRIL (Khanuja et al., 2021) model.
- Followed by that two multilingual MTL frameworks are proposed: one is MTL with task-specific layers (MTL-TSL) and another is MTL-TSL with fusion (MTL-TSL$_{fusion}$) and compared their performances with STL framework.
- Our proposed MTL-TSL and MTL-TSL$_{fusion}$ provide superior performance for sentiment classification than their corresponding STL framework for both Bengali and English languages.

## 2 Related Work

The concept of MTL was first proposed by Caruana (1997). Liu et al. (2016) and Liu et al. (2017) proposed MTL frameworks using LSTM and BiLSTM for text classification.

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

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 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) whereas Del Arco et al. (2021) proposed a BERT based MTL framework for classifying sentiment, emotion, hate speech and offensive language and target analysis.

In this present article, we presented multi-lingual MTL framework for sentiment and emotion classification in Bengali and English text. As per our literature, MTL for low-resourced Bengali language is new and not widely explored.

## 3 Dataset

Two separate sentiment and emotion datasets were prepared to train our proposed MTL frameworks.

**Sentiment Dataset:** The Bengali sentiment data was 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.

Furthermore, the 'SentNoB' dataset was extended to the English texts, collected from the dataset provided by Barbieri et al. (2022).

**Emotion Dataset:** The Bengali emotion annotated data were collected from the 'BanglaEmotion'(Rahman, Md Ataur, 2020) dataset. This dataset was prepared from Facebook comments based in Bangladesh 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 English texts collected from 'GoEmotion' (Demszky et al., 2020) and 'emotion_dataset' (Saravia et al., 2018).

For testing, we used the same test splits provided by Islam et al. (2021), Rahman and Seddiqui (2019) and Barbieri et al. (2022) for Bengali sentiment, Bengali emotion and English sentiment classification tasks respectively. For the English emotion classification, the 'GoEmotion' dataset was annotated with 28 emotional labels and we selected only texts whose labels were matched in the labels of the 'BanglaEmotion' dataset. So, any official test split for this dataset is invalid, rather we extracted 10% data from the collected English emotional texts and used it as a test dataset.

The data distribution of sentiment and emotion training data is provided in Figure 1 and the distribution for test data is provided in Table 1.

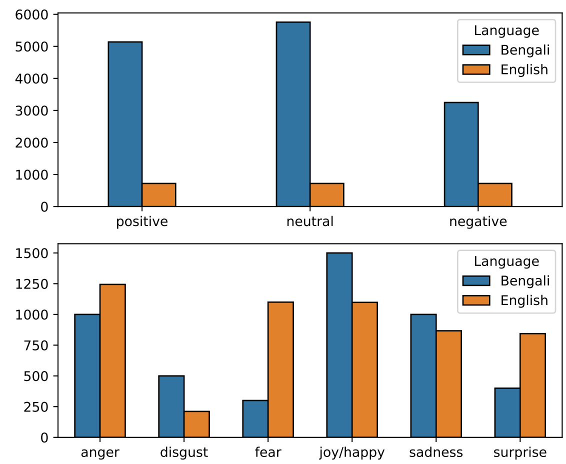


Figure 1: Distribution of training data. (Top: sentiment data distribution, Bottom: emotion data distribution.)

| Task | Label | Bengali | English |
|---|---|---|---|
| Sentiment | Negative | 571 | 290 |
| | Neutral | 361 | 290 |
| | Positive | 654 | 290 |
| | Total | 1586 | 870 |
| Emotion | Anger | 200 | 214 |
| | Disgust | 100 | 35 |
| | Fear | 60 | 203 |
| | Happy/Joy | 300 | 204 |
| | Sad | 200 | 136 |
| | Surprise | 80 | 158 |
| | Total | 940 | 950 |

Table 1: Distribution of Test data.

## 4 Methodology

This section discusses the methodologies of our proposed work. We aim to develop a multilingual

MTL framework that can classify sentiments and emotions at a time in Bengali and English texts.

More specifically, we developed two multilingual MTL frameworks (MTL-TSL and MTL-TSL$_{fusion}$) with the soft parameter-sharing approach where instead of using shared layers we used more task-specific layers. One of the reasons behind using two MTL frameworks is that we want to analyze how MTL performs in all task-specific layers where no (or few) parameters are shared between the tasks and when we fuse the learning from two tasks. The proposed MTL frameworks is depicted in Figure 2

**Tokenization:** The mBERT and MuRIL tokenizers were used for their corresponding MTL and STL frameworks to tokenize each text into a sequence of tokens with a maximum length of 256.

**Model Selection:** As previously mentioned the mBERT and MuRIL models were selected to develop the MTL and STL frameworks. One possible reason behind this, both mBERT and MuRIL are multilingual models that can learn Bengali and English contexts in a text. Moreover, the MuRIL model was specifically trained on 17 Indian languages and provides superior performance in different benchmark datasets.

**MTL-TSL:** The MTL framework with task-specific layers (MTL-TSL) is provided in Figure 2(a) where the outputs of the transformer models were passed through a dropout layer of 0.2 followed by a dense layer of 512 neurons.

$$Dropout_{sen} = PoolerOutput_{sen}$$
$$Dense_{sen} = ReLU(Dropout_{sen})$$

and,

$$Dropout_{emo} = PoolerOutput_{emo}$$
$$Dense_{emo} = ReLU(Dropout_{emo})$$

Where '$PoolerOutput$' represents the last hidden state output of transformer models.

**MTL-TSL$_{fusion}$:** The MTL-TSL$_{fusion}$ is provided in Figure 2(b) where instead of all straightforward task-specific layers, the feature representation of $PoolerOutput$ from sentiment and emotion transformer models followed by a dropout layer of 0.2 were concatenated (fused) to get a fusion of sentiment and emotion learning and then passed the fused output to the $Dense_{sen}$ layer as an input.

$$Dense_{sen} = ReLU(Dropout_{sen} \otimes Dropout_{emo})$$
$$Dense_{emo} = ReLU(Dropout_{emo})$$

Where $\otimes$ represents the concatenation of layers.

**Classification:** The outputs of the $Dense_{sen}$ and $Dense_{emo}$ were passed as an input to the final classification layer of 3 and 6 neurons respectively with the softmax activation function.

$$P_{sen} = softmax(Dense_{sen})$$
$$P_{emo} = softmax(Dense_{emo})$$

**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 models up to 5 epochs and the learning rate was taken as 2e-5 with Adam optimizer (Kingma and Ba, 2014). For the multi-task loss function, the 'SparseCategoricalCrossentropy' loss function was used and monitored the loss for the validation split of the training dataset.

$$L_{total} = L_{sen} + L_{emo}$$

Where $L_{sen}$ and $L_{emo}$ represent the loss for sentiment and emotion tasks.

## 5 Experiment and Result

### 5.1 Experimental Setup

All the experiments were performed using the 'TensorFlow', 'Keras' and 'Scikit-Learn' libraries and the pre-trained models were used using the 'HuggingFace' API. We evaluated our MTL and STL frameworks with accuracy and macro-averaged F1 scores with STL frameworks as a baseline.

### 5.2 Result

The results for sentiment and emotion classification for both STL and MTL are provided in Table 2 and 3 respectively. As deep learning models may generate different results in different runs, therefore, all the proposed frameworks were trained and evaluated 5 times and reported the median results.

**Sentiment Classification:** The sentiment classification result is provided in (Table 2) where the MuRIL-based MTL-TSL$_{fusion}$ model provides the best result for both Bengali and English languages with an F1-score of 71.14 and 67.09 respectively which is 2.89% and 1.92% improvement in for MuRIL based Bengali sentiment classification and MuRIL based English sentiment classification respectively.

In contrast, the mBERT-based MTL-TSL framework didn't perform well and a performance drop

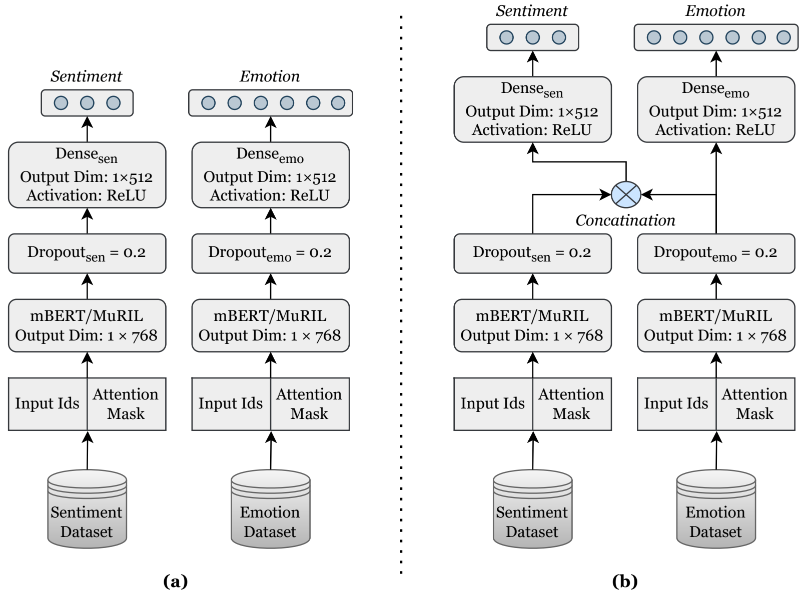


Figure 2: Proposed MTL framework for training: (a) MTL with all task-specific layers (MTL-TSL); (b) MTL-TSL with fusion (MTL-TSL$_{fusion}$).

was observed compared to their corresponding STL frameworks.

| | Bengali | | English | |
|---|---|---|---|---|
| | Acc | F1 | Acc | F1 |
| STL(mBERT) | 69.55 | 67.63 | 62.99 | 62.72 |
| STL(MuRIL) | 71.69 | 69.10 | 66.44 | 65.80 |
| MTL-TSL(mBERT) | 69.92 | 66.83 | 61.15 | 61.72 |
| MTL-TSL(MuRIL) | 74.15 | 70.66 | 65.86 | 65.55 |
| MTL-TSL$_{fusion}$(mBERT) | 68.22 | 65.92 | 61.95 | 61.50 |
| MTL-TSL$_{fusion}$(MuRIL) | **74.15** | **71.14** | **66.44** | **67.09** |

Table 2: Sentiment classification result for Bengali and English language.

| | Bengali | | English | |
|---|---|---|---|---|
| | Acc | F1 | Acc | F1 |
| STL(mBERT) | 50.85 | 34.74 | 83.89 | 81.33 |
| STL(MuRIL) | **54.36** | **39.64** | 85.68 | 83.16 |
| MTL-TSL(mBERT) | 49.68 | 33.58 | 84.52 | 81.35 |
| MTL-TSL(MuRIL) | 54.25 | 38.51 | 85.68 | 82.50 |
| MTL-TSL$_{fusion}$(mBERT) | 50.42 | 34.10 | 84.21 | 81.40 |
| MTL-TSL$_{fusion}$(MuRIL) | 53.83 | 38.92 | 85.68 | **83.48** |

Table 3: Emotion classification result for Bengali and English language.

**Emotion Classification:** The emotion classification results are provided in Table 3 where the MuRIL-based frameworks outperform the mBERT-based frameworks for both STL and MTLs. However, for emotion classification in the Bengali language, the MTLs didn't perform well and a performance (F1-score) drop of 2.85% and 1.81% was observed in MTL-TSL(MuRIL) and MTL-TSL$_{fusion}$(MuRIL) frameworks respectively compared to STL(MuRIL) framework.

Moreover, the MuRIL-based STL framework shows an F1-score improvement of 16.14% than the best F1-score provided by Rahman and Seddiqui (2019)[2]. In addition, the MuRIL-based MTL-TSL and MTL-TSL$_{fusion}$ frameworks show an F1-score improvement by 13.68% and 14.59%.

In the case of Emotion classification in the English language, the MuRIL-based MTL-TSL$_{fusion}$ framework provides the best F1-score of 83.48 which is 0.38% improvement over the corresponding STL framework.

## 6 Conclusion

In this paper, we proposed a fusion-based 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 for sentiment classification in both Bengali and English languages. However, the emotion classification doesn't perform well for the Bengali language in the MTL framework and we'll try to improve the performance of Bengali emotion classification in our future works.

[2]The authors reported 0.3324 as their best F1-score.

## Limitations

Our proposed work also has some limitations. Firstly, the dataset size in this experiment is relatively small. In future, we'll experiment with a larger dataset to validate the robustness of the model. 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. Thirdly, we have done the experiments with pre-trained models only. In future, we'll aim to develop an MTL model from scratch. And lastly, we did the experiments with fewer comparisons of hyperparameters (learning rate=1e-5/2e-5/3e-5, batch size=8, epochs = 3/4/5). In future, we'll do more hyperparameter tuning to develop a more fine-tuned model.

## 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. 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.