2020 23rd International Conference on Computer and Information
Technology (ICCIT), 19-21 December, 2020.

2020 23rd International Conference on Computer and Information Technology (ICCIT) | 978-1-6654-2244-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICCIT51783.2020.9392653

Sentiment analysis in Bengali via transfer learning using multi-lingual BERT

Khondoker Ittehadul Islam
Computer Science and Engineering
Shahjalal University of Science
and Technology
Sylhet, Bangladesh
shanislam6@gmail.com

Md Saiful Islam
Computer Science and Engineering
Shahjalal University of Science
and Technology
Sylhet, Bangladesh
saiful-cse@sust.edu

Md Ruhul Amin
Computer and Information Science
Fordham University
New York, USA
mamin17@fordham.edu

Abstract—**Sentiment analysis (SA) in Bengali is challenging due to this Indo-Aryan language's highly inflected properties with more than 160 different inflected forms for verbs and 36 different forms for noun and 24 different forms for pronouns. The lack of standard labeled datasets in the Bengali domain makes the task of SA even harder. In this paper, we present manually tagged 2-class and 3-class SA datasets in Bengali. We also demonstrate that the multi-lingual BERT model with relevant extensions can be trained via the approach of transfer learning over those novel datasets to improve the state-of-the-art performance in sentiment classification tasks. This deep learning model achieves an accuracy of 71% for 2-class sentiment classification compared to the current state-of-the-art accuracy of 68%. We also present the very first Bengali SA classifier for the 3-class manually tagged dataset, and our proposed model achieves an accuracy of 60%. We further use this model to analyze the sentiment of public comments in the online daily newspaper. Our analysis shows that people post negative comments for political or sports news more often, while the religious article comments represent positive sentiment. The dataset and code is publicly available [1].**

Index Terms—**Sentiment Analysis, CNN, LSTM, BERT, GRU, fasttext, word2vec, SA, Bangla, Bengali**

I. Introduction

Sentiment classification is the task of analyzing a piece of text to predict the orientation of the attitude towards an event or opinion. The sentiment of a text can be either positive or negative. Sometimes, a neutral perspective is also considered for classification. SA has many different applications, such as reducing the early age suicide rate by identifying cyberbullying [1], discouraging unwarranted activities towards a particular community through hate-speech detection [2], and monitoring public response towards a proposed government bill [3] among many others.

The task of SA has achieved superior improvement in other languages, i.e. English - about 97.1% accuracy for 2-class [4] and 91.4% accuracy for 3-class SA [5]. But only a few research works have been published for the SA in Bengali. This is because we lack quality datasets in Bengali for training a computation model for the sentiment classification. However, in the last few years, we have seen the rise of Internet users in the Bengali domain mostly due to the development of wireless network infrastructure throughout South East Asia. This resulted in a massive increase in the total number of online social network users as well as newspaper readers. So it became comparatively easier to collect the public comments posted online on the Bengali news websites.

TABLE I: SA of public comment published in the online newspaper. We collected 334 comments for each of the politics, sports, and religion categories. We only collected one comment from a randomly selected news article. In the table, we present the percentage of the total comments classified into three different sentiment classes.

	Negative	Neutral	Positive
Politics	66%	24%	10%
Sports	52%	38%	10%
Religion	42%	8%	50%

Thus we created two SA datasets for 2-class and 3-class SA in Bengali and trained a multi-lingual BERT model via transfer learning approach for sentiment classification in Bengali, referred as $BERT_{BSA}$ in this paper. $BERT_{BSA}$ achieves an accuracy of 71% for the 2-class and 60% for the 3-class manually tagged dataset. We further use this model to analyze the sentiment of 1,002 public comments collected from the online daily newspaper. Table I shows that in general, sentiment in public comments is positive for religious news articles, while that is negative for political or sports news articles. In this paper, we present the following contributions:

- We created two datasets for SA in Bengali and made it public for further research work. We discuss the methodology we used to create the datasets in the Section III.
- We introduce a deep learning model for SA in Bengali, $BERT_{BSA}$, that performs better compared to other existing models that are trained with word2vec or fastText embedding. We discuss the model and in Section IV.
- We evaluate $BERT_{BSA}$ and compare it to other models trained with Word2Vec and fastText embeddigns using the 2-class and 3-class Bengali SA datasets. We discuss the results in the Section V.
- We conduct experiments to investigate application level use of Bengali SA on newspaper comments in three

[1] https://github.com/KhondokerIslam/Bengali_Sentiment

978-1-6654-2244-4/20/$31.00 ©2020 IEEE

aspects, such as politics, sports and religion, and show that public sentiment is biased towards positive polarity for the news articles related to religion.

II. Related Work

Bidirectional Encoder Representations from Transformers, or BERT [6], is an unsupervised language representation model that had been pre-trained using large plain text corpus. BERT makes use of transformer, an attention mechanism to learn the contextual relations between words. BERT is fundamentally different from the context-free models such as Word2Vec or GloVe that generate a single word embedding representation for each word in the vocabulary [7]. Instead, BERT takes into account the context for each occurrence of a given word in a sentence. For instance, the vector for “running” will have the same Word2Vec or GloVe vector representation for both of its occurrences in the sentences “He is running a company” and “He is running a marathon.” But BERT will provide two contextualized embedding vectors based on the appearance of ”running” in two different sentences.

BERT is very popular for aspect-based sentiment analysis by either fine-tuning BERT’s pre-trained model [8, 9, 10, 11] or using the benchmark dataset for question-answering [12]. However, in the pre-BERT era, research works used other end-to-end deep network layers like LSTM, BiLSTM, CNN, etc. Lei et al. [13] integrated with three kinds of sentiment linguistic knowledge (e.g., sentiment lexicon, negation words, intensity words) into the deep neural network via attention mechanisms. In another research work, Baziotis et al. [14] used LSTM networks augmented with two kinds of attention mechanisms, on top of pre-trained word embedding for sentiment classification and achieved the rank 1^{st} (tie) at the SemEval-2017 Task 4 Subtask A [15].

In spite of such advances in English SA, only a few notable works were done on Bengali SA. Sharfuddin et al. [16] use term frequency–inverse document frequency (tf-idf) and BiLSTM to predict the sentiment of unseen sentences accurately and holds the current state-of-the-art performance on 2-class Bengali sentiment classification in a small balanced dataset. On the other hand, Karim et al. [17] focused primarily on building a Bengali word embedding which was incorporated into a Multichannel Convolutional LSTM (MConv-LSTM) network for predicting different types of tasks including sentiment analysis.

The lack of quality datasets and complex linguistics feature of Bengali language make the task of SA very challenging. In this research work, we contribute two manually tagged datasets for 2-class and 3-class sentiment classification in Bengali. We trained our proposed model $BERT_{BSA}$ as well as the model proposed by Sharfuddin et al. [16] on those datasets and compare the performance of both the models in the section 6.

TABLE II: Distribution of total annotated sample across 10 topics.

Topics	No. of data
Sports	2,332
Economy	1,759
Entertainment	2,697
International	1,985
Education	1,956
Technology	1,282
Lifestyle	1,803
Fashion	1,108
Food	1,343
Travel	1,587
Total	17,852

TABLE III: Distribution of sample across training, validation and test sets.

	Train	Valid	Test
Negative	6011	1060	1280
Neutral	3277	578	877
Positive	3338	588	843
Total	12626	2226	3000

III. Dataset for Bengali SA

We choose Prothom Alo[2], an online news portal, for collecting user’s comment. We selected a total of 10 popular newspaper topics (Table II) and scrapped a total of 40,354 comments. Upon filtering out noisy comments, we tagged each opinion in to one of three sentiments: *Negative*, *Neutral*, or *Positive* by three independent individuals. Our final dataset contains 17,852 entries (Table II). Each of those entries and corresponding tags were validated by an expert Bengali linguistics. In order to analyze our findings and compare with the current state-of-the-art performing model, we made the dataset suitable for 2-class classification tasks by removing the neutral class resulting in a total of 13,120 entries. We present the distribution of the 3-class dataset across training, validation test sets in the Table III. Moreover, detailed statistics of our final corpus is presented in the Table IV

IV. Methodology

In this section, we present the implementation details of our experimental setup. We used the multilingual BERT, *bert-base-multilingual-cased*[3], as it is the only model that was trained with Bengali corpus up until now. We extended the model with three different end-to-end deep network layers: Gated Recurrent Unit (GRU) [18], Long Short Term Memory (LSTM) [19], and Convolutional Neural Network (CNN) [20]. We performed three experiments with the three different BERT extensions. The architecture of our proposed model $BERT_{BSA}$ is depicted in the Figure 1. BERT produces contextualized embedding vector for each word which are passed through one of the three deep network layers: GRU, LSTM, or CNN, in the different experiment. The output neurons for

[2] https://www.prothomalo.com/

[3] https://github.com/google-research/bert/blob/master/multilingual.md

TABLE IV: Data Statistics.

	Words
Longest Sentence	128
Average Sentence Length	45
Total Words	312569
Non-Bengali Words	0

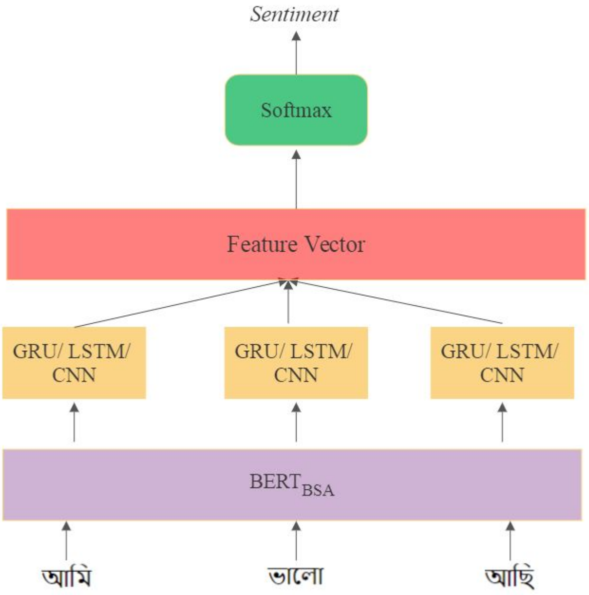

Fig. 1: Sentiment analysis in Bengali via transfer learning using multi-lingual BERT. In this figure we present the sentiment classification of a Bengali text which can be translated to ”I am doing fine” in English.

TABLE V: Sentiment classification accuracy of Word2Vec, fastText and $BERT_{BSA}$ for three different extensions. $BERT_{BSA}$ produced the best accuracy for both the 2-class and 3-class SA tasks.

		GRU	LSTM	CNN
2-class	Word2Vec	0.67	**0.68**	0.66
	fastText	0.68	0.68	**0.69**
	BERT$_{BSA}$	**0.71**	0.70	0.67
3-class	Word2Vec	**0.57**	0.54	0.55
	fastText	**0.58**	**0.58**	0.56
	BERT$_{BSA}$	**0.60**	0.59	0.58

each word from the intermediate layer is concatenated to form the feature vector. Finally, this vector is passed through fully connected neural dense layer for dimension reduction. The final reduced vector is passed through softmax for sentiment classification. The results are presented in Table V for 2-class and 3-class classification.

We also implemented two other deep learning architectures with pre-trained word embeddings, Word2Vec and fastText which were extended with GRU, LSTM, and CNN deep network layers. Word2Vec [7] has been quite successful for SA across several languages, including Bengali [21]. Also, fastText [22] gained huge popularity in Bengali text analysis mainly due its operation on character level n-grams [23, 24, 25]. So we compare the performance of SA for the 2-class and 3-class classification of these models with that of $BERT_{BSA}$, and present the outcomes in Table V.

For three different deep learning architectures with $BERT_{BSA}$, Word2Vec, and fastText, we used the following model parameters:

- **GRU**: For 2-class classification tasks, we used single bidirectional GRU layer where the final layer outputs 300 neurons per word with a dropout of 0.5. For 3 class, we had two bidirectional GRU layers where the final layer outputs 350 neurons per word with a dropout of 0.5.
- **LSTM** For 2-class classification tasks, we used 3 bidirectional LSTM layers with an output of 100 neurons per cell and dropout of 0.5. For 3-class classification task, we used and output of 512 neurons per word using only a single bidirectional LSTM layer.
- **CNN** For 2-class classification tasks, we used a CNN layer with kernel size of [3, 3] and filter size of [64, 100]. For 3-class classification task, we used a single CNN layer with kernel size of [1, 2, 3, 4] and filter size of [200].

For all the models, Adam was used as an optimizer and *L2* was used for regularization.

V. Result Discussion

The outcome of all the different experiments that we performed are presented in the Table V. As the dataset is based on public opinion, most of the words are informal. Therefore, BERT’s ability to manage out-of-vocabulary words effectively helped RNN architecture to carry meaningful context over a long period of time. These resulted RNN architecture performing better with BERT. However, only for 2-class classification task, CNN performed better with fastText over BERT and Word2Vec. Moreover, amongst the RNN architectures GRU performed better than LSTM due to the small-scale dataset [26].

Furthermore, we ran state-of-the-art model proposed by Sharfuddin et al. [16] with our 2-class dataset in our environment and got an accuracy of 68% whereas our model resulted in 71% accuracy. $BERT_{BSA}$ is the only model in Bengali that performed a 3-class classification and resulted in 60% accuracy. This verifies that $BERT_{BSA}$ model with GRU substantially beats the state-of-the-art model that uses tf-idf vectorization with a BiLSTM architecture.

As our ultimate goal is to use this model in real-life applications, we scrapped user’s comment from popular Bengali newspaper sites from three different topics: sports, religion and politics. A total of 1,002 comment had been scraped with 334 comments for each topic from contents ranging from January 2020 to April 2020. Table I shows some interesting findings of native Bengali speaker. Towards politics, Bengali people seems to be much critical. With around 65% of comments seems negative, it clearly states that people are not happy on the ongoing politics taking place in this region. Moreover, people speaking this language tends to criticize their own sport half of the time with percentage of applaudable comment is equal to that politics. Moreover, with neutral being close to negativity, it can be claimed that Bengali people wants better result in sports. However, having the most sentiment as positive in religious topics defines Bengali people are respectful towards diversity and also being faithful believer.

VI. Conclusion

In this paper, we presented two manually tagged novel datasets for SA in Bengali. We also introduced $BERT_{BSA}$, a deep learning model for SA in Bengali, which outperforms all other models. We achieved state-of-the-art performance for both the 2-class and 3-class SA tasks in Bengali. Moreover, we took a step closer to apply SA model to a real world application by analyzing public sentiment on newspaper topics. The result shows that for religious news comments people tend to possess a positive sentiment whereas for political and sports news comments, people possess negative sentiment. However, this research is a work in progress and will be regularly updated with new insights. We are continuing to increase the size of SA datasets in Bengali and we will explore the application of other deep learning models for better results. We hope that the improved performance of SA in multi-class classification tasks presented in this paper will help many ground-breaking applications like cyberbullying identification as well as hate-speech detection in Bengali.

VII. Acknowledgement

We would like to thank Shahjalal University of Science and Technology (SUST) and SUST NLP research group for their support.

References

[1] V. Nahar, S. Unankard, X. Li, and C. Pang, "Sentiment analysis for effective detection of cyber bullying," in *Asia-Pacific Web Conference*. Springer, 2012, pp. 767–774.

[2] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 759–760.

[3] M. Hürlimann, B. Davis, K. Cortis, A. Freitas, S. Handschuh, and S. Fernández, "A twitter sentiment gold standard for the brexit referendum," in *Proceedings of the 12th international conference on semantic systems*, 2016, pp. 193–196.

[4] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *arXiv preprint arXiv:1910.10683*, 2019.

[5] J. Hong, A. Nam, and A. Cai, "Multi-class text sentiment analysis," 2019.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[8] C. Sun, L. Huang, and X. Qiu, "Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence," *arXiv preprint arXiv:1903.09588*, 2019.

[9] A. Karimi, L. Rossi, A. Prati, and K. Full, "Adversarial training for aspect-based sentiment analysis with bert," *arXiv preprint arXiv:2001.11316*, 2020.

[10] A. Rietzler, S. Stabinger, P. Opitz, and S. Engl, "Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification," *arXiv preprint arXiv:1908.11860*, 2019.

[11] B. Xu, X. Wang, B. Yang, and Z. Kang, "Target embedding and position attention with lstm for aspect based sentiment analysis," in *Proceedings of the 2020 5th International Conference on Mathematics and Artificial Intelligence*, 2020, pp. 93–97.

[12] H. Xu, B. Liu, L. Shu, and P. S. Yu, "Bert post-training for review reading comprehension and aspect-based sentiment analysis," *arXiv preprint arXiv:1904.02232*, 2019.

[13] Z. Lei, Y. Yang, M. Yang, and Y. Liu, "A multi-sentiment-resource enhanced attention network for sentiment classification," *arXiv preprint arXiv:1807.04990*, 2018.

[14] C. Baziotis, N. Pelekis, and C. Doulkeridis, "Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis," in *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, 2017, pp. 747–754.

[15] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov, "Semeval-2016 task 4: Sentiment analysis in twitter," *arXiv preprint arXiv:1912.01973*, 2019.

[16] A. A. Sharfuddin, M. N. Tihami, and M. S. Islam, "A deep recurrent neural network with bilstm model for sentiment classification," in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*. IEEE, 2018, pp. 1–4.

[17] M. Karim, B. R. Chakravarthi, M. Arcan, J. P. McCrae, M. Cochez *et al.*, "Classification benchmarks for under-resourced bengali language based on multichannel convolutional-lstm network," *arXiv preprint arXiv:2004.07807*, 2020.

[18] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] K. Fukushima, "Neocognitron: A hierarchical neural network capable of visual pattern recognition," *Neural networks*, vol. 1, no. 2, pp. 119–130, 1988.

[21] M. Al-Amin, M. S. Islam, and S. D. Uzzal, “Sentiment analysis of bengali comments with word2vec and sentiment information of words,” in *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. IEEE, 2017, pp. 186–190.

[22] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[23] Z. S. Ritu, N. Nowshin, M. M. H. Nahid, and S. Ismail, “Performance analysis of different word embedding models on bangla language,” in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*. IEEE, 2018, pp. 1–5.

[24] A. Khatun, A. Rahman, M. S. Islam *et al.*, “Authorship attribution in bangla literature using character-level cnn,” in *2019 22nd International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2019, pp. 1–5.

[25] S. Roy and F. B. Ali, “Unsupervised context-sensitive bangla spelling correction with character n-gram,” in *2019 22nd International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2019, pp. 1–6.

[26] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” *arXiv preprint arXiv:1412.3555*, 2014.