RSM-NLP at BLP-2023 Task 2: Bangla Sentiment Analysis using Weighted and Majority Voted Fine-Tuned Transformers

Pratinav Seth[*] **Rashi Goel**[†] **Komal Mathur** [†] [‡] **Swetha Vemulapalli**[§] [‡]
Manipal Institute of Technology
Manipal Academy of Higher Education, Manipal, India
{seth.pratinav,rashigoel2017,komlixmathur,swetha.vemulapalli.3}@gmail.com

Abstract

This paper describes our approach to submissions made at Shared Task 2 at BLP Workshop - Sentiment Analysis of Bangla Social Media Posts(Hasan et al., 2023a; Islam et al., 2021; Hasan et al., 2023b). Sentiment Analysis is an action research area in the digital age. With the rapid and constant growth of online social media sites and services and the increasing amount of textual data, the application of automatic Sentiment Analysis is on the rise. However, most of the research in this domain is based on the English language. Despite being the world's sixth most widely spoken language, little work has been done in Bangla. This task aims to promote work on Bangla Sentiment Analysis while identifying the polarity of social media content by determining whether the sentiment expressed in the text is Positive, Negative, or Neutral. Our approach consists of experimenting and finetuning various multilingual and pre-trained BERT-based models on our downstream tasks and using a Majority Voting and Weighted ensemble model that outperforms individual baseline model scores. Our system scored 0.711 for the multiclass classification task and scored 10th place among the participants on the leaderboard for the shared task. Our code is available at `https://github.com/ptnv-s/RSM-NLP-BLP-Task2`

1 Introduction

In the era of a high influx of social media platforms, blogs, and online reviews, sentiment analysis has become the need of the hour. Also known as opinion mining, sentiment analysis is a computational linguistic task that is aimed at determining whether a text contains a positive, negative, or neutral sentiment behind it (Khan et al., 2020)

Text	Label
ভাই সোনাই ঘোষ এর দই খেয়ে যাইতেন , খুব ই মজার	Positive
এখানে আরো ভালো ভাবে দলীয় ও র এর অবস্থান পাকা হলো কি ? ?	Neutral
শাউয়ার মাগি তরে এত রিপোর্ট মারি তাও আসে কেন ভিডিও	Negative

Table 1: Text Samples from the Training dataset, with labels as either Positive, Neutral or Negative

Sentiment analysis has diverse uses, including preventing adolescent suicide by detecting cyberbullying and mitigating unjust actions that target specific communities through hate speech detection, among numerous other applications (Islam et al., 2020). Approximately 284.3 million people worldwide speak Bangla as their primary language. Individuals speaking Bangla increasingly engage in social media platforms like Instagram, Facebook, Reddit, and Twitter and express opinions on microblogging platforms, commenting on news portals and online shopping. However, analyzing vast volumes of rapidly generated data in the digital age is a very tedious job to do. This is where sentiment analysis can be applied (Hassan et al., 2016). Most sentiment analysis research predominantly focuses on English, leaving Bangla Sentiment analysis in its nascent stages. Recently, some works have addressed this issue. However, none of these studies have fully embraced the different perspectives of Bangla.

To address this problem, we present our contributions to Shared Task 2 at BLP Workshop - Sentiment Analysis of Bangla Social Media Posts. This task aims to detect the polarity associated with a given social media text. This multiclass classification task involves determining whether the sentiment expressed in the text is Positive, Negative, or Neutral. For this problem statement, we have conducted various experiments using multi-

[*] Dept. of Data Science & Computer Applications
[†] Dept. of Computer Science & Engineering
[‡] Authors have contributed equally to this work
[§] Dept. of Information and Communication Technology

Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), pages 305–311
December 7, 2023 ©2023 Association for Computational Linguistics

lingual berts (Bhattacharjee et al., 2022; Sanh et al., 2019a; Das et al., 2022; Sarker, 2020) and various pre-trained transformers (Liu et al., 2019a) by fine-tuning them on downstream tasks. We also apply Majority Voting and Weighted ensembling on the top-k models to show how these methods affect the models' performance and how an ensemble of these models performs better than the individual baselines.

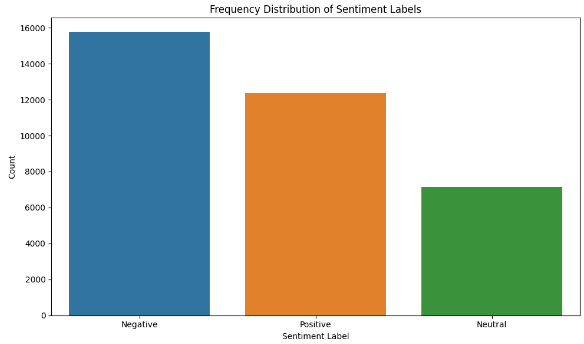

Figure 1: Frequency of Task 2 labels in training set

2 Background

2.1 Problem and Data Description

The EMNLP 2023 Bangla Workshop Task 2: Sentiment Analysis of Bangla Social Media Posts (Hasan et al., 2023a; Islam et al., 2021; Hasan et al., 2023b) aims to detect the polarity of the sentiment associated with a given text extracted from social media. From the entire set of labels, over 14,000 were classified as negative, approximately 12,000 as positive, and roughly 6,000 as neutral, as indicated in the distribution chart in Figure 1 and a few samples of the Dataset are shown in Table 1. The dataset includes the MUltiplatform BAngla SEntiment (MUBASE) dataset and the SentNob dataset (Islam et al., 2021). SentNob comprises public comments from social media on news and videos across 13 domains, such as agriculture, politics, and education. It is manually annotated with a moderate agreement score of 0.53. On the other hand, MUBASE is a sizable compilation of multi-platform data, including Facebook posts and tweets, each manually tagged for sentiment polarity. These datasets provide a comprehensive and diverse landscape for studying Bangla sentiment analysis.

2.2 Previous Works

2.2.1 Sentiment Analysis

Sentiment analysis is an NLP task that uses computational methods to determine and extract the emotional tone expressed in a piece of text (Hogenboom et al., 2014). There are several different approaches to sentiment analysis. Early sentiment analysis approaches primarily employed rule-based methods and lexicon-based techniques (Obaidat et al., 2015) to determine the sentiment context of texts. One of the significant areas of application of Sentiment Analysis is in Social Media Posts as in (Tang et al., 2014) and (Taboada et al., 2011), a sentiment lexicon with a linguistic rule-based approach was used to create a sentiment detection mechanism from tweets(Reckman et al., 2013). Following this, contemporary advancements have introduced machine learning and deep learning techniques that significantly boost accuracy by extracting intricate patterns from annotated datasets. Due to human language's complexity and sentiment expression nuances, it is a challenging task. The accuracy of the task may be improved by using larger datasets, more complex and fine-tuned models (Hassan et al., 2016), ensembling, etc. Modern approaches leverage large-scale Pre-trained Language Models (PLMs), such as Transformers, BERTs (Devlin et al., 2018), and NLUs (Bender and Koller, 2020), alongside refined fine-tuning mechanisms(Hasan et al., 2023b). They excel at capturing the intricate associations between words within the text and their corresponding polarity. In today's world, with the introduction of free-to-use models like ChatGPT, sentiment analysis has opened to new possibilities (Wang et al., 2023).

2.2.2 Bangla Language Processing

The Bangla language is the 7th most spoken language, with 265 million speakers worldwide (Sen et al., 2022). However, since English is the predominant language used for technical knowledge, journals, and documentation, many Bangla-speaking people face hurdles in utilizing these resources. Research on Bangla Natural Language Processing (BNLP) began in the early 1990s, focusing on rule-based lexical and morphological analysis (Alam et al., 2021). From the modeling perspective, most earlier endeavors are either rule-based, statistical, or classical machine learning-based approaches(Kudo and Matsumoto, 2001). As for the sequence tagging tasks, such as NER and G2P, the algorithms, including Hidden Markov Models (HMMs) (Brants, 2000), Conditional Random Fields (CRFs) (Lafferty et al., 2001), Maximum

Entropy (ME) (Ratnaparkhi, 1996) and Maximum Entropy Markov Models (MEMMs) (McCallum et al., 2000) have been used successfully. It is only very recently that a small number of studies have explored deep learning-based approaches. As depicted in (Alam et al., 2021), there has been significant work in resource and model development in Bangla sentiment analysis. In (Das and Bandyopadhyay, 2010), the authors proposed a computational technique of generating an equivalent SentiWordNet (Bangla) from publicly available English sentiment lexicons and an English-Bangla bilingual dictionary with few easily adaptable noise reduction techniques. However, with the Introduction of BERTs many works focused on fine-tuning multilingual BERTs (Ashrafi et al., 2020; Das et al., 2021), but BanglaBERT (Sarker, 2020) being the first model pre-trained on Bangla text corpus.

2.2.3 Bangla Sentiment Analysis

Sentiment analysis is a tool to extract the emotional tone of the text. It is used for cyberbullying detection, hate speech mitigation and market research. Bangla is the 7th most spoken language, and sentiment analysis for Bangla is still in its early stages. The first attempt to perform sentiment analysis in the context of Indian Languages, including Bangla, was done as recently as in 2015 (Patra et al., 2015). The lack of accurately annotated data is one of the biggest bottlenecks to advancing Bangla Sentiment Analysis. (Islam et al., 2021) and (Rahman et al., 2018) describe the creation of datasets for this purpose. A word2vec model was tuned with word co-occurrence scores for sentiment analysis in (Al-Amin et al., 2017), achieving an accuracy of 75.5%. In (Wahid et al., 2019), aspect-based sentiment analysis data was examined, boasting a remarkable 95% accuracy. However, challenges were encountered when rephrasing common and proper nouns in Bangla. Among most studies, however, transformer models have consistently outperformed other algorithms and models, inciting a significant amount of research into the area. In (Chowdhury et al., 2019), Opinion Mining was conducted on a dataset of 4,000 manually translated Bangla movie reviews, with the objective of classifying them as positive or negative. The LSTM approach had achieved an accuracy of 82.42%. A Bi-LSTM architecture was applied by (Sharfuddin et al., 2018) to a labeled dataset of 10,000 Facebook comments in Bangla, resulting in an accuracy of 85.67%. However, the study faced significant data preprocessing difficulties. In (Tripto and Ali, 2018), a combination of CNN and LSTM was employed to extract six distinct emotions from various types of Bangla YouTube video comments. The reported accuracies were 65.97% and 54.24% for three and five-label sentiment classification, respectively. A common issue faced by authors while using CNNs was that proper tuning between layers could not be achieved. In another study (Hossain et al., 2020), 1000 online restaurant reviews were collected from the Foodpanda website for performing SA and deployed, thus combining CNN with LSTM architecture with a 300 dimensional Word2Vec pretrained model having validation accuracy of 75.01%. (Rezaul Karim et al., 2020) developed a novel word embedding system for Bangla texts, BanglaFastText, incorporating it into a Multichannel Convolutional LSTM (MConv-LSTM). In (Islam et al., 2020) authors performed SA on 1002 public comments from newspapers with the help of the BERT pretrained model and achieved accuracy on GRU at 71% on 2 class sentiments. In (Hasan et al., 2020a), the performance of multiple classical machine learning algorithms and deep learning models were compared on several sentiment-labeled datasets, showing that pre-trained transformer models such as BERT and XLM-RoBERTa yielded the highest scores.

3 System Overview

We conducted extensive experiments for the given task involving Bangla Sentiment analysis. We fine-tuned various multilingual and pre-trained transformer architectures, including BERT (Kenton and Toutanova, 2019), DistillBERT (Sanh et al., 2019b), RoBERTa (Liu et al., 2019b), and Various Pre-Trained BERT models (Das et al., 2022; Sarker, 2020) on our downstream task of polarity classification. We shortlist the top-k model based on the performance metrics and ensemble the predictions using Majority Voted and Weighted Ensemble.

3.1 Fine-Tuning Transformers

We used multiple transformer architectures to observe the effect of the model architecture and the pre-trained dataset on the downstream task. For multiclass classification, we added a linear layer acting as a classification head to fine-tune the models for the multiclass classification.

We have used various models for our experiments, including **BERT** (Kenton and Toutanova,

Model	Acc.	Pre.	Rec.	F1
RoBERTa (Base)	0.550	0.544	0.550	0.550
Distill BERT	**0.701**	**0.687**	**0.701**	**0.701**
HF-PT BERT-1	0.672	0.679	0.672	0.672
HF-PT BERT-2	0.639	0.630	0.639	0.639
HF-PT BERT-3	0.669	0.671	0.669	0.669
Bangla BERT (Small)	0.657	0.649	0.657	0.657
Bangla BERT (Large)	0.693	0.684	0.693	0.693
Bangla BERT (Base)	**0.701**	**0.687**	**0.701**	**0.701**
Banglish BERT	0.684	0.672	0.684	0.684

Table 2: Results of Base-Models on Test-Set of Shared-Task Dataset where Acc. is Accuracy, Pre. is Precision, Rec. is Recall & F1 refers to F1-Score

2019), a transformer-based language model that creates representations of text by combining both left and right contexts with Masked Language Modeling and Next Sentence Prediction being pre-training tasks. **RoBERTa** (Liu et al., 2019b) is a faster variation of BERT. **DistilBERT (multilingual cased)** (Sanh et al., 2019b) is a distilled version of the multilingual Bert with pretraining on Wikipedia data in 104 languages. **BanglaBERT** (Sarker, 2020) referred to as HF-PT-BERT-2 in Table 1 is a pretrained BERT trained on the Bangla common crawl dataset and the Bangla Wikipedia Dump Dataset. **Indic-abusive-allInOne-MuRIL** (Das et al., 2022) is a model finetuned from the MuRIL (Khanuja et al., 2021) and multilingual BERT models, trained to detect abusive speech using multiple datasets in 8 Indian languages. **Bengali-abusive-MuRIL** (Das et al., 2022) is also finetuned from MuRIL (Khanuja et al., 2021), trained specifically on the Bangla abusive speech dataset. These have been referred to as HF-PT-BERT-1 and HF-PT-BERT-3 in Table 1, respectively. **BanglaBERT** (Bhattacharjee et al., 2022)is a fine-tuned ELECTRA (Clark et al., 2020) model which is trained on Bangla Wikipedia dump dataset as well as data from 110 Bangla websites. **BanglishBERT**(Bhattacharjee et al., 2022)is similar to BanglaBERT; instead, it was trained on both English and Bangla data to allow zero-shot cross-lingual transfer.

3.2 Ensembling Predictions

To increase the overall performance of the predictions and robustness of the predictive model, models were first individually tuned on the downstream task dataset. The predictions from these models were combined using the two ensembling methods on top-3,top-5, and all model predictions:

Majority Voting: The most frequently occurring prediction from all the models for each training instance was chosen as the final label.

Weighted: Each model was assigned a weight based on its accuracy score on the training dataset. Each model voted on the prediction class with its weight, and the prediction with the highest final vote was chosen as the final label.

$$y_i = argmax(\sum_{j=1}^{k} a_j.p_{ij}) \quad (1)$$

Here, y_i denotes the Weighted ensemble prediction of the ith sample, p_{ij} the ith probabilistic prediction for each polarity made by the jth model, a_j the accuracy of the jth model on the training set and k is the number of models being considered for the ensemble.

4 Experiments & Results

The dataset used for the task is organized in 3 columns, with id, text, and label. It has also been partitioned into a train set with 35266 samples, a dev set with 3935 samples, and a dev-test set with 3427 samples. The distribution in the training set is shown in Figure 1.

The preprocessing pipeline before model training included padding, tokenizing, and truncating text data to ensure uniformity and manage lengthy inputs. We used the AdamW optimizer, a learning rate of 2x10^{-5} and a batch size of 32 over 32 epochs was chosen to strike a balance between convergence speed and stability with a maximum sequence length of 512 tokens used with Huggingface AutoTokenizer to tokenize the data.

We evaluated models using four metrics: accuracy, precision, recall, and F1-score. F1-score is

Method	Top	Acc.	Prec.	Rec.	F1
Majo	3	0.706	0.692	0.706	0.706
-rity	5	0.707	0.694	0.707	0.707
Voted	All	**0.711**	**0.695**	**0.711**	**0.711**
Weig	3	0.703	0.691	0.703	0.703
-hted	5	0.703	0.692	0.703	0.703
	All	0.708	**0.695**	0.708	0.708

Table 3: Results of ensemble models on Test-Set of Shared-Task Dataset where Method is the method of ensembling, Top refers to top-k models chosen, Acc. is Accuracy, Pre. is Precision, Rec. is Recall & F1 refers to F1-Score

a good metric for imbalanced datasets because it takes into account both precision and recall.

The results of our experiments over the official Test set are shown in Table 2 & 3. For Individual Models as shown in Table 2 we observe DistilBERT and BanglaBERT(Base) show the best performance on the test data, with an F1-Score of 0.701.

We did an ensemble of both types (Majority-Voted and Weighted) with the top 3 (BanglaBERT (Sarker, 2020), BanglishBERT, HF-PT-BERT-1 (Das et al., 2022)), top 5 (HF-PT-BERT-2, BanglishBERT, HF-PT-BERT-1 (Das et al., 2022) , BanglaBERT(Base), HF-PT-BERT-3 (Das et al., 2022)) and lastly using all the models. As in Table 3 for ensembles, we observe that the majority ensemble shows a better performance in general as compared to the weighted models. The majority voted ensemble using predictions from all the models had the highest F1 score of 0.711. Furthermore, an ensemble of 3 models yielded almost optimal results. The use of more than three models resulted in a marginal increase in performance but significantly increased resource utilization. Thus, the use of more than three models seems unproductive.

5 Conclusion

In this work, we benchmarked various multilingual and pre-trained BERT-based models - RoBERTa(Liu et al., 2019a), DistillBERT(Sanh et al., 2019a), BanglaBERT(Bhattacharjee et al., 2022), BanglishBERT(Hasan et al., 2020b) and Various Pre-Trained BERT models (Das et al., 2022; Sarker, 2020) for Bangla Sentiment Analysis (Hasan et al., 2023a; Islam et al., 2021; Hasan et al., 2023b) while identifying the polarity of social media content by determining whether the sentiment expressed in the text is Positive, Negative, or Neutral as our downstream tasks and using a Majority Voting and Weighted ensemble model that outperforms individual baseline model scores.

Our system achieved a micro F1-Score of 0.711 for the multiclass classification task and scored 10th among the participants on the leaderboard for the shared task.

6 Acknowledgments

We would like to thank the Research Society MIT, an undergraduate interdisciplinary technical society of Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India, for supporting our research.

References

Md Al-Amin, Md Saiful Islam, and Shapan Das Uzzal. 2017. Sentiment analysis of bengali comments with word2vec and sentiment information of words. In *2017 international conference on electrical, computer and communication engineering (ECCE)*, pages 186–190. IEEE.

Firoj Alam, Arid Hasan, Tanvirul Alam, Akib Khan, Janntatul Tajrin, Naira Khan, and Shammur Absar Chowdhury. 2021. A review of bangla natural language processing tasks and the utility of transformer models. *arXiv preprint arXiv:2107.03844*.

Imranul Ashrafi, Muntasir Mohammad, Arani Shawkat Mauree, Galib Md. Azraf Nijhum, Redwanul Karim, Nabeel Mohammed, and Sifat Momen. 2020. Banner: A cost-sensitive contextualized model for bangla named entity recognition. *IEEE Access*, 8:58206–58226.

Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States. Association for Computational Linguistics.

Thorsten Brants. 2000. Tnt-a statistical part-of-speech tagger. *arXiv preprint cs/0003055*.

Rumman Rashid Chowdhury, Mohammad Shahadat Hossain, Sazzad Hossain, and Karl Andersson. 2019. Analyzing sentiment of movie reviews in bangla by applying machine learning techniques. In *2019 international conference on bangla speech and language processing (ICBSLP)*, pages 1–6. IEEE.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Amitava Das and Sivaji Bandyopadhyay. 2010. Sentiwordnet for bangla. *Knowledge Sharing Event-4: Task*, 2:1–8.

Avishek Das, Omar Sharif, Mohammed Moshiul Hoque, and Iqbal H. Sarker. 2021. Emotion classification in a resource constrained language using transformer-based approach. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 150–158, Online. Association for Computational Linguistics.

Mithun Das, Somnath Banerjee, and Animesh Mukherjee. 2022. Data bootstrapping approaches to improve low resource abusive language detection for indic languages. *arXiv preprint arXiv:2204.12543*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Md. Arid Hasan, Firoj Alam, Anika Anjum, Shudipta Das, and Afiyat Anjum. 2023a. Blp-2023 task 2: Sentiment analysis. In *Proceedings of the 1st International Workshop on Bangla Language Processing (BLP-2023)*, Singapore. Association for Computational Linguistics.

Md. Arid Hasan, Shudipta Das, Afiyat Anjum, Firoj Alam, Anika Anjum, Avijit Sarker, and Sheak Rashed Haider Noori. 2023b. Zero- and few-shot prompting with llms: A comparative study with fine-tuned models for bangla sentiment analysis.

Md. Arid Hasan, Jannatul Tajrin, Shammur Absar Chowdhury, and Firoj Alam. 2020a. Sentiment classification in bangla textual content: A comparative study. In *23rd International Conference on Computer and Information Technology (ICCIT)*.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. 2020b. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2612–2623, Online. Association for Computational Linguistics.

A. Hassan, M. R. Amin, N. Mohammed, and A. K. A. Azad. 2016. Sentiment analysis on bangla and romanized bangla text (brbt) using deep recurrent models.

Alexander Hogenboom, Bas Heerschop, Flavius Frasincar, Uzay Kaymak, and Franciska de Jong. 2014. Multi-lingual support for lexicon-based sentiment analysis guided by semantics. *Decision Support Systems*, 62:43–53.

Naimul Hossain, Md Rafiuzzaman Bhuiyan, Zerin Nasrin Tumpa, and Syed Akhter Hossain. 2020. Sentiment analysis of restaurant reviews using combined cnn-lstm. In *2020 11th International conference on computing, communication and networking technologies (ICCCNT)*, pages 1–5. IEEE.

Khondoker Ittehadul Islam, Md Saiful Islam, and Md Ruhul Amin. 2020. Sentiment analysis in bengali via transfer learning using multi-lingual bert. In *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pages 1–5. IEEE.

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.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Md Rafidul Hasan Khan, Umme Sunzida Afroz, Abu Kaisar Mohammad Masum, Sheikh Abujar, and Syed Akhter Hossain. 2020. Sentiment analysis from bengali depression dataset using machine learning. In *2020 11th international conference on computing, communication and networking technologies (ICCCNT)*, pages 1–5. IEEE.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Taku Kudo and Yuji Matsumoto. 2001. Chunking with support vector machines. In *Second meeting of the North American chapter of the Association for Computational Linguistics*.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Andrew McCallum, Dayne Freitag, Fernando CN Pereira, et al. 2000. Maximum entropy markov models for information extraction and segmentation. In *Icml*, volume 17, pages 591–598.

Islam Obaidat, Rami Mohawesh, Mahmoud Al-Ayyoub, Al-Smadi Mohammad, and Yaser Jararweh. 2015. Enhancing the determination of aspect categories and their polarities in arabic reviews using lexicon-based approaches. In *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, pages 1–6. IEEE.

Braja Gopal Patra, Dipankar Das, Amitava Das, and Rajendra Prasath. 2015. Shared task on sentiment analysis in indian languages (sail) tweets-an overview. In *Proc. of MIKE*, pages 650–655. Springer.

Md Rahman, Emon Kumar Dey, et al. 2018. Datasets for aspect-based sentiment analysis in bangla and its baseline evaluation. *Data*, 3(2):15.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Conference on empirical methods in natural language processing*.

Hilke Reckman, Cheyanne Baird, Jean Crawford, Richard Crowell, Linnea Micciulla, Saratendu Sethi, and Fruzsina Veress. 2013. teragram: Rule-based detection of sentiment phrases using sas sentiment analysis. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 513–519.

Md Rezaul Karim, Bharathi Raja Chakravarthi, Mihael Arcan, John P McCrae, and Michael Cochez. 2020. Classification benchmarks for under-resourced Bengali language based on multichannel convolutional-lstm network. *arXiv*, pages arXiv–2004.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019a. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019b. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sagor Sarker. 2020. Banglabert: Bengali mask language model for bengali language understanding.

Ovishake Sen, Mohtasim Fuad, Md. Nazrul Islam, Jakaria Rabbi, Mehedi Masud, Md. Kamrul Hasan, Md. Abdul Awal, Awal Ahmed Fime, Md. Tahmid Hasan Fuad, Delowar Sikder, and Md. Akil Raihan Iftee. 2022. Bangla natural language processing: A comprehensive analysis of classical, machine learning, and deep learning-based methods. *IEEE Access*, 10:38999–39044.

Abdullah Aziz Sharfuddin, Md Nafis Tihami, and Md Saiful Islam. 2018. A deep recurrent neural network with bilstm model for sentiment classification. In *2018 International conference on Bangla speech and language processing (ICBSLP)*, pages 1–4. IEEE.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Duyu Tang, Furu Wei, Bing Qin, Ming Zhou, and Ting Liu. 2014. Building large-scale twitter-specific sentiment lexicon: A representation learning approach. In *Proceedings of coling 2014, the 25th international conference on computational linguistics: Technical papers*, pages 172–182.

Nafis Irtiza Tripto and Mohammed Eunus Ali. 2018. Detecting multilabel sentiment and emotions from bangla youtube comments. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–6. IEEE.

Md Ferdous Wahid, Md Jahid Hasan, and Md Shahin Alom. 2019. Cricket sentiment analysis from bangla text using recurrent neural network with long short term memory model. In *2019 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–4. IEEE.

Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023. Is chatgpt a good sentiment analyzer? a preliminary study. *ArXiv*, abs/2304.04339.