

Bangla Sentiment Analysis using GPT 3.5 and Llama-2

Anonymous EMNLP submission

Abstract

Sentiment Analysis (SA) is a subdomain of NLP, where texts are classified as positive, negative, neutral, or even subtle classes such as strongly positive, happy, etc. according to the subjectivity expressed. SA in high-resource languages like English extensively uses state-of-the-art (SOTA) transformer-based Large Language Models (LLM). After the release of ChatGPT in November 2022, the usage of AI has been democratized and more people are aware of the potential usefulness of Large Language Models (LLM). Unfortunately, low-resource languages like Bangla lags far behind due to a lack of resources, annotated gold-standard datasets, and funding. In this paper, we test GPT3.5 (one of the models powering ChatGPT) and Llama-2 for SA of the SentNoB dataset [Islam et al. \(2021\)](#) using the zero-shot learning for GPT model and both zero-shot learning and fine-tuning for Llama-2. All of our codes and the dataset used are publicly available ¹ for future use and accelerate the usage of LLM in Bangla NLP.

1 Introduction

In this paper, we have used "gpt-3.5-turbo-0301" by zero-shot learning through prompting and used "meta-llama/Llama-2-7b-chat-hf" by both prompting and fine-tuning. Each of the data samples from the SentNoB dataset is embedded in a prompt instructing the model for classifying into neutral, positive, and negative classes. To the best of our knowledge, this technique has never been used in Bangla SA.

Islam et al. [Islam et al. \(2021\)](#) implemented SVM, BiLSTM, and mBERT with various feature extraction techniques on the SentNoB dataset and reported the highest 64% F1 score.

¹Our code is available here [LINK](#).

BanglaBERT and BanglishBERT were developed in [Bhattacharjee et al. \(2021\)](#) where they created two data sets. The performance of the two BERT models was evaluated on four benchmark data sets on NLI, QA, SA, and NER. Their pre-trained Bangla BERT model outperformed other transformer models and provided almost 73% of the F1 score.

Table 1: Comparison with previous works on SentNoB dataset

Model	F1-score
SVM Islam et al. (2021)	0.64
banglabert Bhattacharjee et al. (2021)	0.72
GPT 3.5(prompting)	0.5525
Llama-2(prompting)	0.361
Llama-2(fine-tuned)	0.51

2 Literature Review

Bangla SA, also known as opinion mining, subjectivity analysis, and polarity detection lacks gold-standard large datasets and resources compared to many high-resource languages. Among the available resources, many are translated from high-resource languages, and collected from social media platforms which fails to represent the cultural richness and linguistical complexities of the Bangla language. Texts collected from social media are plagued with code-mixing and spelling mistakes which pose unique challenges for Bangla SA.

Lexical-based backtracking [Rabeya et al. \(2017\)](#), semi-supervised clustering [Dasgupta and Ng \(2009\)](#), and Conditional random field [Das and Bandyopadhyay \(2009\)](#) were used for Bangla SA in the early 2010s. Application of traditional machine learning models like Naive Bayes [Hasan et al. \(2015\)](#), Paul

and Shill (2016), Sen et al. (2022), Support Vector Machine Khan et al. (2021), Taher et al. (2018), Chowdhury and Chowdhury (2014), Sabuj et al. (2017), and Stochastic Gradient Descent Mandal et al. (2018) along with manual feature extraction techniques like Bag of Words, and N-grams in Bangla SA resulted in pretty decent performances. Neural network models like CNN Sarkar (2019) Hoq et al. (2021) and LSTM Tripto and Ali (2018), Hassan et al. (2016), Romim et al. (2022) were extensively used in mid-2010s providing state-of-the-art results back then along with non-contextualized embeddings like FastText, GloVe, Word2Vec. After the introduction of the transformer Vaswani et al. (2017), pretraining the transformer on a large corpus for general understanding of the language and fine-tuning on specific datasets become the standard for getting state-of-the-art performance. Very few transformer-based models are pre-trained on specifically Bangla language or have Bangla in the multilingual pre-trained corpus compared to other high-resource languages. Some of the transformer-based models that used for Bangla SA are: BanglaBERT Ghosh et al. (2023), Bhattacharjee et al. (2021), multilingual BERT Devlin et al. (2018), IndicBERT Kakwani et al. (2020), SahajBERT Diskin et al. (2021), bn-RoBERTa Jain et al. (2020) XLM-RoBERTa Conneau et al. (2019), Alam et al. (2021), Karim et al. (2021), Iqbal et al. (2022). For further understanding of the progress of Bangla SA, we suggest the following survey papers to the readers: Banik et al. (2019), Alam et al. (2021), Sen et al. (2022), Hira et al. (2022), Shammi et al. (2023).

3 Methodology for GPT-3.5:

OpenAI ope (a), the company behind the releasing GPT models incorporates the transformer architecture instead of RNN-based models early after the release of transformer architecture. In 2018, OpenAI released the first GPT model, Generative Pre-Training(GPT-1) Radford et al. (2018). Transformer architecture consists of two major parts, encoder and decoder. GPT models only use the decoder part of the transformer as the decoder part is generative in nature.

GPT-1 has only 117 million parameters and acts as a core architecture for later GPT models. In 2019, OpenAI released GPT-2 Radford et al. (2019) which contains 1.5 billion parameters. The next model GPT-3 Brown et al. (2020) was released in 2020 with 175 billion parameters. The concept of In Context Learning (ICL) was formally introduced in GPT-3 model’s paper. Codex Chen et al. (2021), GPT-3 fine-tuned on GitHub code, released in 2021 showed an increase in performance on logical problems. InstructGPT Ouyang et al. (2022) was released in early 2022 to improve the GPT-3 model by Reinforcement Learning from Human Feedback(RLHF) so that the model aligned with human preferences.

One of the most widely used GPT models, ChatGPT OpenAI (2023a) was released in November 2022 and perhaps changed the perception of humans for AI forever. ChatGPT is a fine-tuned version of InstructGPT for safety and is specially designed for conversation by generating coherent dialogues. Finally, OpenAI released GPT-4 OpenAI (2023b) in March 2023 which can not only take texts but multimodal signals (images) as inputs.

OpenAI’s GPT models (especially ChatGPT) have been getting much attention recently. We tried to use the gpt-3.5-turbo-0301 model, the model behind the ChatGPT, by OpenAI’s API key. Each text from the test data was inserted within the following prompt:

Classify the sentiment of the following text: 'data_sample'. Is it positive, negative, or neutral? Give a supporting explanation behind the classification.

Before finalizing this prompt, we tried to use the different prompts in which we tried to specify the role of the model as a Bangla text annotator, tried to specify the output format for easy parsing, and tried to set the optimal wait time. For fixing this prompt, we keep in mind the number of tokens in the input prompt will greatly affect the cost, the performance of the model, optimal number of output tokens for understanding the model’s chain of thought behind the classification. For the validation of the prompt, we tested the prompt in the SST2 dataset and the SentNoB dataset translated into English.

3.1 Performance on SentNoB test data:

The SentNoB test data was given as input to the model, by using the prompt. Based on the output string that the model provided, by using parsing we classified the text as 0(neutral), 1(positive), and 2 (negative). On the SentNoB test data (1586 instances), we got 56.12% accuracy, 57.92% precision, 56.24% recall, and 55.25% F1-score. By analyzing the confusion matrix, we found out that the model is over-predicting the 2(negative) class.

3.2 GPT-3.5's Performance on Translated SentNoB test data:

Next, we want to find out whether the language is the reason for such a low performance by the "gpt-3.5-turbo-0301" model. We use Google Translator to translate the Excel file format (.xlsx) of the SentNoB test data from Bangla to English. Then we use the same prompt for classifying the texts of the translated SentNoB test data. We got around 53% accuracy and F1-score, which is around 5% less than the performance on original Bangla texts. This may be due to noises introduced during translation. Due to cost constraints, we restricted ourselves from using the latest GPT models.

3.3 Prompt validation on SST2 dataset:

We expected that the gpt-3.5 model would provide much higher performance, even beating the state-of-the-art performance of BERT variants. However, after observing such a lower performance (55.25% F1-score), we were curious whether, the prompt that we were using, dragged down the performance. For validating the prompt, we used the same prompt on the test and validation data of the SST2 dataset [Socher et al. \(2013\)](#) containing positive and negative classes respectively. We got around 94% accuracy and F1-score respectively. According to the paperswithcode.com leaderboard [pap](#), the state-of-the-art accuracy on SST2 is obtained by the T5-11B [Raffel et al. \(2020\)](#) model which is within the 4 percent range using "gpt-3.5-turbo-0301" model and our prompt. Thus, we can validate that our prompt didn't cause any performance is-

sues.

3.4 GPT-3.5's performance of SentNoB validation dataset:

After validating the prompt, we want to ensure that the test portion of the SentNoB dataset may not cause performance degradation. We suspected that the authors of the SentNoB dataset might have split similar types of data in the test split. Thus, we take the validation split of the SentNoB dataset, and using the same prompt we collected and analyzed the gpt-3.5-turbo-0301 model's responses. We got 60% accuracy, 61% precision, 60% recall, and 59% F1-score. The performance is slightly increased (5% margin) than the performance on the SentNoB dataset. By analyzing the confusion matrix, we found that the model is over-predicting the neutral and negative classes.

3.5 GPT-3.5's Cost Estimation:

There are 1586 instances of the SentNoB test data and 24150 words(tokens) in the SentNoB test dataset. Assuming that we need 25 additional words in the prompt for each instance, the total additional words in the prompt will be $25 \times 1586 = 39650$. Thus, the total number of words(tokens) will be $24150 + 39650 = 63800$. OpenAI would charge 0.0015 dollars for each 1000 input tokens and 0.002 dollars for each 1000 output tokens for using gpt-3.5-turbo [ope \(b\)](#). We assumed that using the model's API only on the SentNoB test data would cost only 0.12 dollars and for the full SentNoB dataset, the total cost will be 1.3 dollars averaging the costs for input and output tokens. But, due to APIConnectionError, JSONDecoderError, and RateLimitError, the execution failed again and again. Thus, the actual cost exceeded the estimated cost.

3.6 GPT-3.5's Error Analysis:

True Sentiment is Positive but classified as Negative:

Sentence : আর অভাবে কত আসলে সরকারের উচিত
ভালো কোনো পদক্ষেপ নেয়া

Translation: How much more shortage before the government takes any appropriate step?

GPT's response: The sentiment of the given text is negative. The reason behind this is that the sentence mentions a lack of appropriate

action being taken by the government for a particular situation. The use of the word "অভাবে" (lack/shortage) suggests a negative sentiment towards the government's lack of action.

Actual Sentiment: Positive

Predicted Sentiment: Negative

Error Type: Dataset annotation error

Explanation: Though annotating sentiment is a subjective topic, different annotators will annotate the same sentence differently. The polarity of this sentence is negative (some people might annotate this as neutral). But annotators of the SentNoB dataset annotated it as positive.

True Sentiment is Negative but classified as Positive:

Sentence : গরমের সোদনে বাচিনা তুমি আইছো শীতের গীত লইয়া । ব্যাটা ভাল হ । খাওয়ার টাইমে এমন সাপের লাহান কাল্লা ঝাকাইচ না

Translation: I am suffering from the heat of summer, you have come with the song of winter. Be good. At the time of eating don't shake your head like a snake.

GPT's response: The sentiment of the text is positive. Explanation: The text contains positive words such as "ব্যাটা ভাল হ" (the bat is good) which indicates a positive sentiment towards something. The overall tone of the text is also cheerful and positive, as it describes the speaker enjoying the summer and not having to deal with the hassle of snakebites during meal times. There are no negative words or phrases present in the text that might indicate a negative sentiment. Therefore, the sentiment can be classified as positive. Actual Sentiment: Negative

Predicted Sentiment: Positive

Error Type: Dialect understanding error

Explanation: The sentence is written using colloquial language in a local dialect. Even a native Bangla speaker may find it difficult to understand the meaning of it. Most likely, the model did not train on understanding such a variation of Bangla language.

True Sentiment is Neutral but classified as Positive:

Sentence : আমরা চাই প্রত্যেক ওষুধের গায়ে মূল্য দেওয়া হক । কে কে একমত আছেন লাইক দিয়ে সাড়া দেন

Translation: We want each drug to be priced. If you agree, please respond with a like.

GPT's response: Positive. The text expresses the desire for fair pricing of medicine and asks for support through likes. This shows a positive attitude towards fairness and justice.

Actual Sentiment: Neutral

Predicted Sentiment: Positive

Error Type: Misinterpretation of polarity word
Explanation: The classification of this sentence is a difficult one. The author is raising social awareness to ensure fair pricing of medicines. This sentence might be classified as a positive one. As the sentiment classification is a subjective one, different annotators might classify different sentences differently. However, as the SentNoB dataset annotated this sentence as neutral, we will consider the sentiment of this sentence as neutral.

Due to the presence of the polarity word 'like', the GPT 3.5 model misclassifies this.

True Sentiment is Neutral but classified as Negative:

Sentence : ভাইয়া নতুন ভিডিও আসে না কেন?

Translation: Brother, why is there no new video?

GPT's response: The sentiment of the text "ভাইয়া নতুন ভিডিও আসে না কেন" is negative.

Explanation: The text includes the word "না" which means "no" or "not" in English. Moreover, the person is asking why there is no new video which suggests that they are disappointed or unhappy about it. Therefore, the sentiment of the text is negative. Actual Sentiment: Neutral

Predicted Sentiment: Negative

Error Type: Negation handling error.

Explanation: Due to the presence of the negation word 'Not' GPT 3.5 model classified the sentence as negative. But the actual sentiment should be neutral as the sentence did not express any negative sentiment.

True Sentiment is Positive but classified as Neutral:

Sentence : অমায়িক ব্যবহার এই রেস্টুরেন্টের

স্টাফদের

Translation: The staff of this restaurant are polite.

GPT's response: The sentiment of the given text is neutral. This is because the text does not express any emotion or opinion, it simply mentions the staff of the restaurant. Therefore, there is no positive or negative sentiment associated with the statement. Actual Sentiment: Positive

Predicted Sentiment: Neutral

Error Type: Unknown word

Explanation: This misclassification is pretty strange for a large language model like GPT. Perhaps the model knows the other synonyms of the word 'অমায়িক' (polite). But it did not know this particular word.

Table 2: Error Types in Predicted Classes

Actual Class	Predicted Class	Error Type
Neutral	Positive	Wrong assumption, Word ambiguity, Polarity Word Misinterpretation
Neutral	Negative	Negation Handling, Dataset annotation, Implicit expression
Positive	Neutral	Unknown Word, Absence of Polarity Words, Dataset annotation
Positive	Negative	Dataset annotation, Multipolarity
Negative	Neutral	Unrecognizable Negative Sentiment, Numerical Misunderstanding, Sarcasm Detection, Implicit Expression, Word Ambiguity
Negative	Positive	Multipolarity, Parsing, Dialect Understanding

True Sentiment is Negative but classified as Neutral:

Sentence : খাবারের পরিমাণ বেশি , পরিমাণ কম হলে দাম টাও কমতো আর নষ্টও হতো না

Translation: The amount of food is more, if the amount is less, the price will be reduced and there will be no waste.

GPT's response: Neutral. Explanation: The text states a fact and does not express any positive or negative sentiment towards it. It suggests that if the quantity of food is more or less, the price will also be affected, but it does not express any emotion or opinion about it. Hence, the sentiment of the text is neutral.

Actual Sentiment: Negative

Predicted Sentiment: Neutral

Error Type: Numerical understanding error

Explanation: The sentiment of this sentence is negative. The model failed to understand the numerical quantification of the context. Due to the abundance of food, perhaps food is wasted. Which expresses a negative sentiment.

4 Llama-2's methodology:

In contrast with GPT-3.5, the Llama-2 model is an open-access model. The low-resource languages like Bangla suffer from many constraints where budget is one of the main issues. Thus freely available models like Llama-2 will be a game changer in this arena. We have used the Llama-2 model in two settings zero-shot learning by prompting and fine-tuning for predicting the sentiment classes of the SentNoB dataset. We have used the "meta-llama/Llama-2-7b-chat-hf" checkpoint from Hugging Face's transformers library and Carbon dioxide emission tracker from project code carbon which kept track of the total Carbon dioxide emitted in grams.

4.1 Prompting Llama-2:

For prompting we have used langchain's Hugging Face pipeline for formatting the prompt and the temperature to 0. The following prompt template was used:

```
<s> [INST] <<SYS>> You are a Bengali text annotator. </SYS>> Annotate text into neutral, negative, or positive. [/INST] </s>
```

In the above prompt template, the role of the system is defined as a Bengali language text annotator to guide the model. In the user prompt, we asked the model to annotate the data samples into three classes iteratively fetching texts from the dataset. After numerous attempts to find the best prompt, we derived the above prompt which is very simple, effective, and short. The output of the prompt from the

Llama-2 model was manually parsed for classifying the text sentiment into sentiment classes. By prompting using the above methodology, we have found, a 36.2% accuracy score, 36.2 % F1-micro score, 70.11% precision score, and 36.2% recall score. The total carbon dioxide emitted by this prompting is 135.16 grams.

4.2 Finetuning Llama-2:

The "meta-llama/Llama-2-7b-chat-hf" checkpoint contains 7 seven billions parameters. Fine-tuning this mammoth size of the model in the free version of Kaggle is very difficult. For fine-tuning the Llama-2 model, we have used the 'bitsandbytes' library from Hugging Face. From that library, we used Parameter Efficient Fine Tuning (PEFT) as outlined in QLoRA (Dettmers et al., 2023). 4-bit quantization of the model weights is used for fine-tuning only the reduced sets of parameters by exploiting low-rank adaptations. We set the LoRA attention dimension to 64, the Alpha parameter for LoRA scaling to 16, the Dropout probability for LoRA layers to 0.1, the base model was loaded in 4-bit precision, nf4 quantization was used, for gradient clipping the maximum gradient normal was set 0.3, weight decay set to 0.001, paged_adamw_32bit was used, 'constant' type learning rate scheduler was used, device_map parameter was set to "auto". Due to the limitations of the resources, we only used 4000 samples of the dataset to fine-tune the Llama-2 model. The train data to the model was fed in the following prompt format: 'prompt' : '<s>[INST] + example['Data'] + '[/INST] + example['sentiment'] + '</s>' The training text is to be merged with the labels along with '<s>' and '[INST]' tags because the model was trained in this format. The fine-tuned model performed significantly better than prompting, providing 0.51 accuracy, 0.51 f1-score, 0.64 precision, and 0.51 recall scores, and the total carbon dioxide emitted in this fine-tuning was 48.73 grams.

5 Conclusion & Future work :

In this paper, we have used gpt-3.5-turbo-0301 by prompting, known as zero-shot learning. We have shown an F1-score of 55.25% by prompting the GPT 3.5 model. Besides, we have also done an extensive error analysis to find out the possible reasons for the misclassification. The

"meta-llama/Llama-2-7b-chat-hf" was also used by prompting and fine-tuning resulting in 36.2% and 51.2% F1-score. From analyzing the results, we concluded that, the LLM performed significantly lower than Pretrained Language models (variants of BERT). One possible reason for this discrepancy of LLMs in low-resource languages perhaps can be attributed to the fact that the volume of low-resource tokens in the LLMs training corpus is very low proportional to high-resource languages' representations. In the future, we want to use the latest SOTA models like gpt-4, and latest Llama variant models both by fine-tuning and prompting by the latest frameworks like ChatML, and LangChain.

Limitations

Usually, few-shot, one-shot, and fine-tuned models perform better than zero-shot learning. In this paper we only explored the zero-shot prompting and fine-tuned models. We believe exploring these methods along with other LLM will produce interesting results.

Ethics Statement

Acknowledgements

References

- a. Openai.
- b. Openai pricing.
- Sentiment analysis on sst-2 binary.
- 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.
- Nayan Banik, Md Hasan Hafizur Rahman, Shima Chakraborty, Hanif Seddiqui, and Muhammad Anwarul Azim. 2019. Survey on text-based sentiment analysis of bengali language. In 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), pages 1--6. IEEE.
- Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Kazi Samin, Md Saiful Islam, Anindya Iqbal, M Sohel Rahman, and Rifat Shahriyar. 2021. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. arXiv preprint arXiv:2101.00204.

541	Tom Brown, Benjamin Mann, Nick Ryder, Melanie	2015 IEEE International WIE Conference on Elec-	596
542	Subbiah, Jared D Kaplan, Prafulla Dhariwal,	trical and Computer Engineering (WIECON-ECE),	597
543	Arvind Neelakantan, Pranav Shyam, Girish Sas-	pages 511--514. IEEE.	598
544	try, Amanda Askell, et al. 2020. Language models		
545	are few-shot learners. <i>Advances in neural infor-</i>	Asif Hassan, Mohammad Rashedul Amin, Abul Kalam	599
546	mation processing systems, 33:1877--1901.	Al Azad, and Nabeel Mohammed. 2016. Sentiment	600
		analysis on bangla and romanized bangla text us-	601
547	Mark Chen, Jerry Tworek, Heewoo Jun, Qiming	ing deep recurrent models. In 2016 International	602
548	Yuan, Henrique Ponde de Oliveira Pinto, Jared Ka-	Workshop on Computational Intelligence (IWCi),	603
549	plan, Harri Edwards, Yuri Burda, Nicholas Joseph,	pages 51--56. IEEE.	604
550	Greg Brockman, et al. 2021. Evaluating large lan-		
551	guage models trained on code. <i>arXiv preprint</i>	Suma Hira, Atish Kumar Dipongkor, Saumik Chowd-	605
552	<i>arXiv:2107.03374</i> .	hury, Mostafijur Rahman Akhond, Syed Md Galib,	606
		et al. 2022. A systematic review of sentiment	607
553	Shaika Chowdhury and Wasifa Chowdhury. 2014.	analysis from bengali text using nlp. <i>American</i>	608
554	Performing sentiment analysis in bangla mi-	<i>Journal of Agricultural Science, Engineering, and</i>	609
555	croblog posts. In 2014 International Confer-	<i>Technology</i> , 6(3):150--159.	610
556	ence on Informatics, Electronics & Vision (ICIEV),		
557	pages 1--6. IEEE.	Muntasir Hoq, Promila Haque, and Mo-	611
		hammed Nazim Uddin. 2021. Sentiment	612
558	Alexis Conneau, Kartikay Khandelwal, Naman Goyal,	analysis of bangla language using deep learning	613
559	Vishrav Chaudhary, Guillaume Wenzek, Fran-	approaches. In <i>International Conference on Com-</i>	614
560	cisco Guzmán, Edouard Grave, Myle Ott, Luke	puting Science, Communication and Security,	615
561	Zettlemoyer, and Veselin Stoyanov. 2019. Unsu-	pages 140--151. Springer.	616
562	supervised cross-lingual representation learning at		
563	scale. <i>arXiv preprint arXiv:1911.02116</i> .	MD Asif Iqbal, Avishek Das, Omar Sharif, Mo-	617
		hammed Moshuiul Hoque, and Iqbal H Sarker. 2022.	618
564	Amitava Das and Sivaji Bandyopadhyay. 2009. Sub-	<i>Bemoc: A corpus for identifying emotion in ben-</i>	619
565	jectivity detection in english and bengali: A crf-	<i>gali texts</i> . <i>SN Computer Science</i> , 3(2):135.	620
566	based approach. <i>Proceeding of ICON</i> .		
567	Sajib Dasgupta and Vincent Ng. 2009. Topic-wise,	Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful	621
568	sentiment-wise, or otherwise? identifying the	Islam, and Mohammad Ruhul Amin. 2021. Sent-	622
569	hidden dimension for unsupervised text classi-	nob: A dataset for analysing sentiment on noisy	623
570	fication. In <i>Proceedings of the 2009 Conference</i>	bangla texts. In <i>Findings of the Association for</i>	624
571	<i>on Empirical Methods in Natural Language Pro-</i>	<i>Computational Linguistics: EMNLP 2021</i> , pages	625
572	<i>cessing</i> , pages 580--589.	3265--3271.	626
573	Tim Dettmers, Artidoro Pagnoni, Ari Holtzman,	Kushal Jain, Adwait Deshpande, Kumar Shridhar, Fe-	627
574	and Luke Zettlemoyer. 2023. Qlora: Efficient	lix Laumann, and Ayushman Dash. 2020. Indic-	628
575	finetuning of quantized llms. <i>arXiv preprint</i>	<i>transformers: An analysis of transformer lan-</i>	629
576	<i>arXiv:2305.14314</i> .	<i>guage models for indian languages</i> . <i>arXiv</i>	630
		<i>preprint arXiv:2011.02323</i> .	631
577	Jacob Devlin, Ming-Wei Chang, Kenton Lee, and	Divyanshu Kakwani, Anoop Kunchukuttan, Satish	632
578	Kristina Toutanova. 2018. BERT: pre-training of	Golla, NC Gokul, Avik Bhattacharyya, Mitesh M	633
579	deep bidirectional transformers for language un-	Khapra, and Pratyush Kumar. 2020. <i>Indicnlp</i>	634
580	derstanding . <i>CoRR</i> , abs/1810.04805.	suite: Monolingual corpora, evaluation benchmarks and	635
		pre-trained multilingual language models for in-	636
581	Michael Diskin, Alexey Bukhtiyarov, Max Ryabinin,	dian languages. In <i>Findings of the Association</i>	637
582	Lucile Saulnier, Anton Sinitsin, Dmitry Popov,	<i>for Computational Linguistics: EMNLP 2020</i> , pages	638
583	Dmitry V Pyrkín, Maxim Kashirin, Alexander	4948--4961.	639
584	Borzunov, Albert Villanova del Moral, et al. 2021.		
585	Distributed deep learning in open collaborations.	Md Rezaul Karim, Sumon Kanti Dey, Tanhim Islam,	640
586	<i>Advances in Neural Information Processing Sys-</i>	Sagor Sarker, Mehadi Hasan Menon, Kabir Hos-	641
587	<i>tems</i> , 34:7879--7897.	sain, Md Azam Hossain, and Stefan Decker. 2021.	642
		<i>Deepphateexplainer: Explainable hate speech de-</i>	643
588	Tapotosh Ghosh, Md Hasan Al Banna, Md Jaber	tection in under-resourced bengali language. In	644
589	Al Nahian, Mohammed Nasir Uddin, M Shamim	2021 IEEE 8th International Conference on Data	645
590	Kaiser, and Mufti Mahmud. 2023. An attention-	<i>Science and Advanced Analytics (DSAA)</i> , pages 1-	646
591	-based hybrid architecture with explainability for	-10. IEEE.	647
592	depressive social media text detection in bangla.		
593	<i>Expert Systems with Applications</i> , 213:119007.	Md Serajus Salekin Khan, Sanjida Reza Rafa,	648
		Amit Kumar Das, et al. 2021. Sentiment analysis	649
594	KM Azharul Hasan, Mir Shahriar Sabuj, and Zakia	on bengali facebook comments to predict fan's	650
595	Afrin. 2015. Opinion mining using naive bayes. In	emotions towards a celebrity. <i>Journal of Engi-</i>	651
		<i>neering Advancements</i> , 2(03):118--124.	652

653	Soumil Mandal, Sainik Kumar Mahata, and Dipankar	Ovishake Sen, Mohtasim Fuad, Md Nazrul Islam,	706
654	Das. 2018. Preparing bengali-english code-mixed	Jakaria Rabbi, Mehedi Masud, Md Kamrul Hasan,	707
655	corpus for sentiment analysis of indian languages.	Md Abdul Awal, Awal Ahmed Fime, Md Tah-	708
656	arXiv preprint arXiv:1803.04000.	mid Hasan Fuad, Delowar Sikder, et al. 2022.	709
657	OpenAI. 2023a. Chatgpt .	Bangla natural language processing: A compre-	710
658	OpenAI. 2023b. Gpt-4 technical report .	hensive analysis of classical, machine learning,	711
659	Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida,	and deep learning-based methods. IEEE Access,	712
660	Carroll Wainwright, Pamela Mishkin, Chong	10:38999--39044.	713
661	Zhang, Sandhini Agarwal, Katarina Slama, Alex	Shumaiya Akter Shammi, Sajal Das, Narayan Ran-	714
662	Ray, et al. 2022. Training language models to	jan Chakraborty, Sumit Kumar Banshal, and Nishu	715
663	follow instructions with human feedback. Ad-	Nath. 2023. A comprehensive roadmap on bangla	716
664	vances in Neural Information Processing Systems,	text-based sentiment analysis. ACM Transactions	717
665	35:27730--27744.	on Asian and Low-Resource Language Information	718
666	Animesh Kumar Paul and Pintu Chandra Shill. 2016.	Processing, 22(4):1--29.	719
667	Sentiment mining from bangla data using mutual	Richard Socher, Alex Perelygin, Jean Wu, Jason	720
668	information. In 2016 2nd international confer-	Chuang, Christopher D. Manning, Andrew Ng, and	721
669	ence on electrical, computer & telecommunica-	Christopher Potts. 2013. Recursive deep models	722
670	tion engineering (ICECTE), pages 1--4. IEEE.	for semantic compositionality over a sentiment	723
671	Tapasy Rabeya, Sanjida Ferdous, Himel Suhita Ali,	treebank . In Proceedings of the 2013 Conference	724
672	and Narayan Ranjan Chakraborty. 2017. A survey	on Empirical Methods in Natural Language Pro-	725
673	on emotion detection: A lexicon based backtrack-	cessing, pages 1631--1642, Seattle, Washington,	726
674	ing approach for detecting emotion from ben-	USA. Association for Computational Linguistics.	727
675	gali text. In 2017 20th international conference	SM Abu Taher, Kazi Afsana Akhter, and KM Azharul	728
676	of computer and information technology (ICCIT),	Hasan. 2018. N-gram based sentiment mining for	729
677	pages 1--7. IEEE.	bangla text using support vector machine. In	730
678	Alec Radford, Karthik Narasimhan, Tim Salimans,	2018 international conference on Bangla speech	731
679	Ilya Sutskever, et al. 2018. Improving language	and language processing (ICBSLP), pages 1--5.	732
680	understanding by generative pre-training.	IEEE.	733
681	Alec Radford, Jeffrey Wu, Rewon Child, David Luan,	Nafis Irtiza Tripto and Mohammed Eunus Ali. 2018.	734
682	Dario Amodei, Ilya Sutskever, et al. 2019. Lan-	Detecting multilabel sentiment and emotions	735
683	guage models are unsupervised multitask learn-	from bangla youtube comments. In 2018 Interna-	736
684	ers. OpenAI blog, 1(8):9.	tional Conference on Bangla Speech and Language	737
685	Colin Raffel, Noam Shazeer, Adam Roberts, Kather-	Processing (ICBSLP), pages 1--6. IEEE.	738
686	ine Lee, Sharan Narang, Michael Matena, Yanqi	Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob	739
687	Zhou, Wei Li, and Peter J Liu. 2020. Exploring the	Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz	740
688	limits of transfer learning with a unified text-to-	Kaiser, and Illia Polosukhin. 2017. Attention is	741
689	text transformer. The Journal of Machine Learn-	all you need. Advances in neural information	742
690	ing Research, 21(1):5485--5551.	processing systems, 30.	743
691	Nauros Romim, Mosahed Ahmed, Md Saiful Islam,	A Appendix	744
692	Arnab Sen Sharma, Hriteshwar Talukder, and Mo-	We have used python version 3.7, openai ver-	745
693	hammad Ruhul Amin. 2022. Bd-shs: A benchmark	sion 0.27.2, openpyxl version 3.1.2, gpt-3.5-	746
694	dataset for learning to detect online bangla hate	turbo-0301 on free version of Kaggle. The ex-	747
695	speech in different social contexts. arXiv preprint	periments were conducted in April 2023.	748
696	arXiv:2206.00372.		
697	Mir Shahriar Sabuj, Zakia Afrin, and KM Azharul		
698	Hasan. 2017. Opinion mining using support vector		
699	machine with web based diverse data. In Inter-		
700	national Conference on Pattern Recognition and		
701	Machine Intelligence, pages 673--678. Springer.		
702	Kamal Sarkar. 2019. Sentiment polarity detection		
703	in bengali tweets using deep convolutional neu-		
704	ral networks. Journal of Intelligent Systems,		
705	28(3):377--386.		