Zero- and Few-Shot Prompting with LLMs: A Comparative Study with Fine-tuned Models for Bangla Sentiment Analysis

Authors

Md. Arid Hasan

Shudipta Das

Afiyat Anjum

Firoj Alam

Anika Anjum

Avijit Sarker

Sheak Rashed Haider Noori

LREC-COLING 2024

Problem Statement

- Sentiment Analysis (SA) is a critical tool for marketing, politics, customer service, and healthcare
- Significant improvement in resource-rich languages, while Bangla remain under-researched
- Groundbreaking performance of large language models (LLMs) highlights the necessity to evaluate them in low-resource language contexts
- Capabilities of Mono- and multi-lingual pretrained language models (PLMs) are unknown for low-resource languages

Background

- SentiGold (Proprietary)
  - ➡ Social media and news comments consisting of 70K entries
- SentNoB
  - ➡ Comment from News article and videos covering 13 domains consisting of 15K annotated data
  - ➡ Inter annotator agreement score of 0.53
- BanglaBook
  - ➡ Book reviews consisting of 158K examples where 89.5% data belongs to positive class
- Islam et. al. 2021
  - ➡ News Comments consisting of 17.8K data

Our Contributions

- Sizable manually annotated dataset titled **MUBASE**
- A comprehensive benchmarking of LLMs for the Bangla SA task
- Fine-tune mono- and multi-lingual PLMs
- Investigate Zero- and Few-shot in-context learning with several LLMs
- Discussed the findings and detailed comparison among PLMs, Zero-shot, and few-shot learning

MUBASE Dataset

- Collected comments and tweets from both Facebook and X (Twitter)
- Tweets are collected from newspaper account
- Facebook comments collected from public pages belongs to news article
- Each data were annotated by 3 annotators
- Inter annotator agreement (IAA) of 0.84 indicate perfect agreement

MUBASE Dataset

Class	Facebook	Twitter	Total
Positive	2,245	8,315	10,560
Neutral	4,866	1,331	6,197
Negative	9,078	7,771	16,849
Total	16,189	17,417	33,606

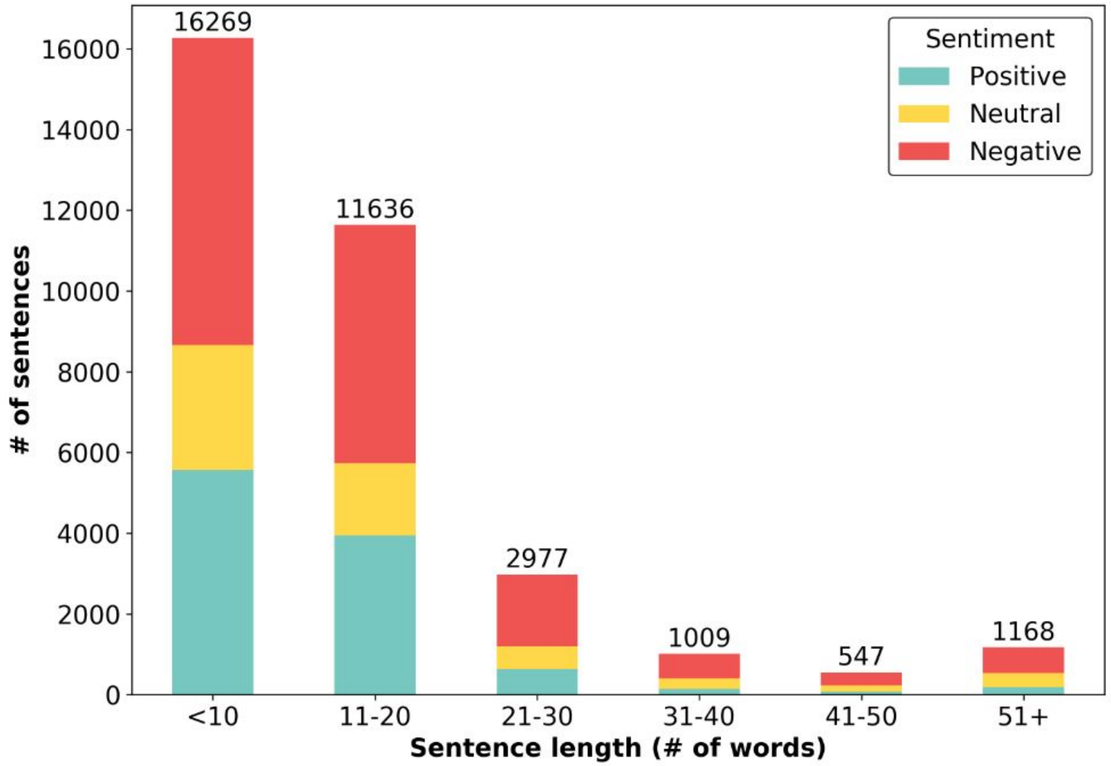
Class label distribution across **different sources** of the dataset

MUBASE Dataset

Class	Train	Dev	Test	Total
Positive	7,342	1,126	2,092	10,560
Neutral	4,319	601	1,277	6,197
Negative	11,811	1,700	3,338	16,849
Total	23,472	3,427	6,707	33,606

Class label distribution of the dataset

MUBASE Dataset



Language Models

- **PLMs (fine-tuning)**
  - ➡ Embedding (GPT)
  - ➡ Bloomz-560m and 1.7B
  - ➡ BERT-m
  - ➡ XLM-R
  - ➡ BanglaBERT
- **LLMs (Zero/Few-shots)**
  - ➡ Flan-T5
  - ➡ GPT-4
  - ➡ Blooms 1.7B, 3B, 7.1B, and 176B (8 bit)

Embedding (GPT)

- Extracted the embeddings using OpenAI's **text-embedding-ada-002** model for each data split
- Fine-tune a feed-forward network on the embeddings extracted from the training set to train our model

- **Hyper-parameters:**
  - ➡ Activation function = Unit (ReLU)
  - ➡ Hidden layer size = 500
  - ➡ Learning rate = 0.001

Fine-Tuning

- ➡ Bloomz-560m and 1.7B
- ➡ BERT-m
- ➡ XLM-R
- ➡ BanglaBERT

- Fine-tuned each model using the default settings over three epochs
- Ten reruns for each experiment using different random seeds

Zero- and Few-Shot Prompts

Instructions:
We would like you to analyze the sentiment of the following text. Based on the content of the text, please classify it as either “Positive”, “Negative”, or “Neutral”. Provide only the label as your response.

text: {input_sample}

label:

role: system,
content: You are an expert annotator. Your task is to analyze the text and identify sentiment polarity.

GPT-4 (Zero-shot)

Zero- and Few-Shot Prompts

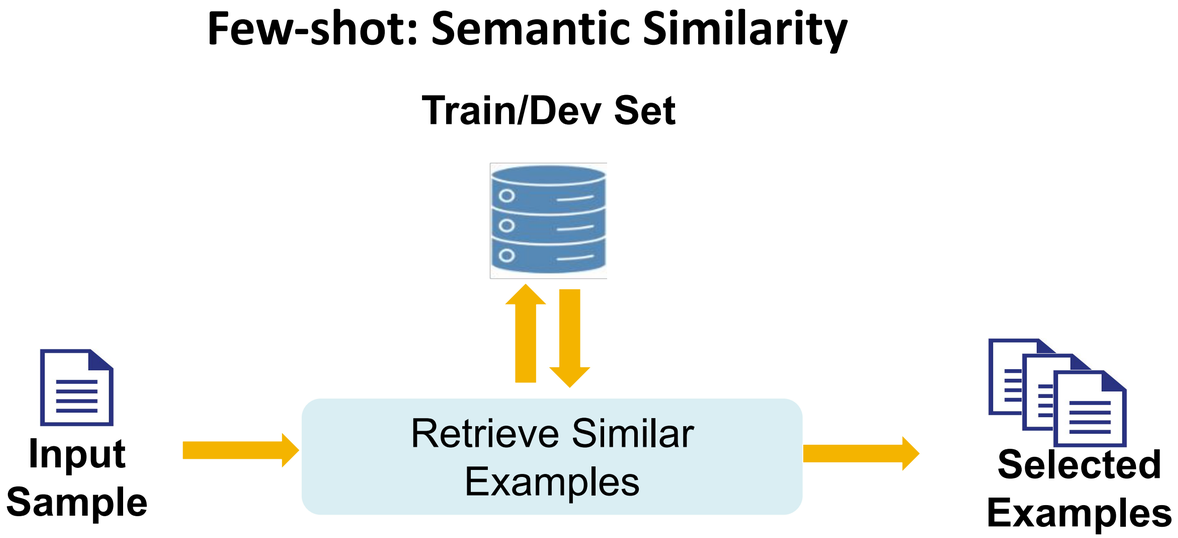
Instructions:
Annotate the “text” into “one” of the following categories: “Positive”, “Negative”, or “Neutral”.
Here are some examples:
Example 1:
text: {input_example}
label: {input_label}
…

text: {input_sample}

label:

role: system
content: As an AI system, your role is to analyze text and classify them as ‘Positive’, ‘Negative’ or ‘Neutral’. Provide only label and in English.

GPT-4 (Few-shot)



Maximal marginal relevance

http://llmebench.qcri.org/

Zero- and Few-Shot Prompts

Instructions:
Label the following text as Neutral Positive, or Negative. Provide only the label as your response.

text: {input_sample}

label:

BLOOMZ (Zero-shot)

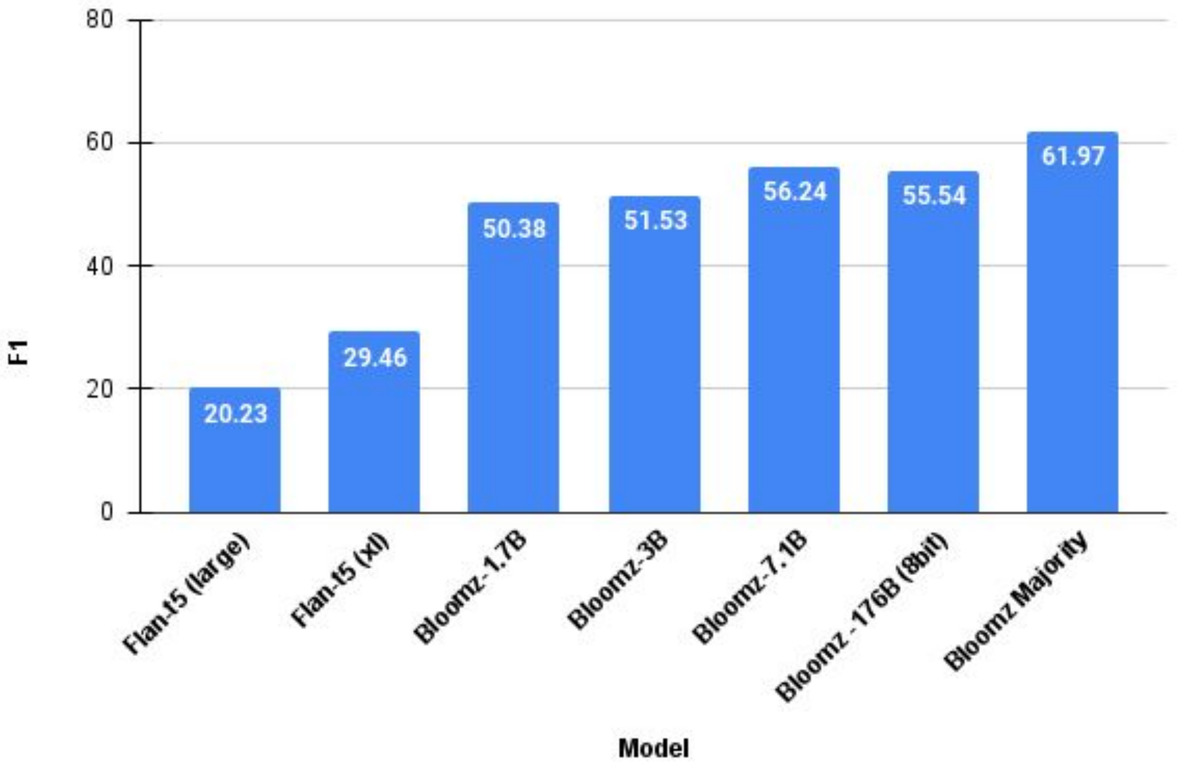
Results

Exp	Acc	P	R	F1
Baseline				
Random	33.56	38.31	33.56	33.56
Majority	49.77	24.77	49.77	49.77
Classic Models				
SVM	55.81	53.33	55.81	52.39
RF	56.75	54.61	56.75	52.62
Fine-tuning				
Embedding (GPT)	57.79	57.30	57.79	57.46
Bloomz-560m	61.71	63.08	61.97	63.08
Bloomz-1.7B	61.16	59.76	61.16	59.95
BERT-m	64.95	64.92	64.95	64.90
XLM-r (base)	66.63	66.24	66.63	66.28
XLM-r (large)	66.33	65.63	66.33	65.79
BanglaBERT	69.08	67.61	69.08	67.98
BanglaBERT*	70.33	69.13	70.33	**69.39**

Performance of different sets of experiments. * indicates trained on combined MUBASE, SentiNoB (Islam et al., 2021), and Alam et al. (2021a). BN Ins. refers that instruction is provided in the native Bangla language.

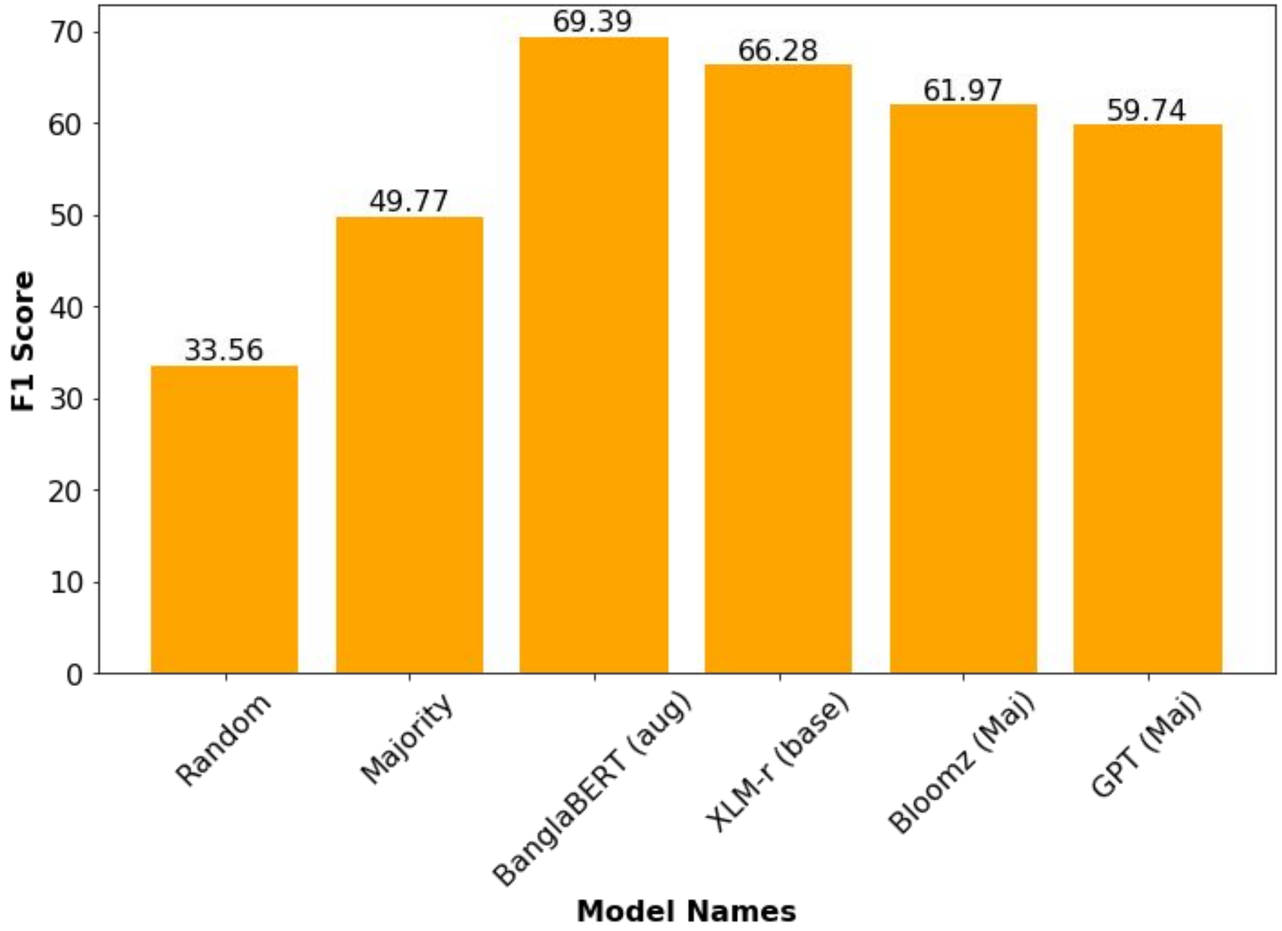
Results

Exp	Acc	P	R	F1
Zero- and Few-shot on LLMs				
Open Models - 0-shot				
Flan-T5 (large)	41.28	20.23	13.77	20.23
Flan-T5 (xl)	49.42	29.46	18.18	29.46
Bloomz-1.7B	58.33	49.38	58.33	50.38
Bloomz-3B	59.73	50.98	59.73	51.53
Bloomz-7.1B	62.83	50.92	62.83	56.24
Bloomz 176B (8bit)	61.84	51.16	61.84	55.54
Bloomz Majority	61.97	51.32	61.97	61.97
Closed Models - m-shot				
GPT-4: 0-Shot	60.21	61.65	60.21	59.99
GPT-4: 0-Shot (BN inst.)	60.70	61.71	60.70	59.96
GPT-4: 3-Shot	60.40	63.88	60.40	60.74
GPT-4: 5-Shot	60.95	63.83	60.95	61.17
GPT-4 Majority	59.74	63.26	59.74	59.74



Discussion

- Fine-tuned models consistently outperforms LLMs
- Multilingual models show promising research direction
- Monolingual provides superior performance



Discussion

- More training data might be required to effectively fine-train (Bloomz 560m and 1.7B)
- The performance increases with the parameter size for Bloomz (1.7B to 7B)
- GPT-4 outperforms other LLMs
- Different types of prompting did not yield a clear improvement
- The performance gain with few-shot is significant

Discussion

- Flan-T5 (xl) labeled only five posts as negative
- Flan-T5 (large) labeled only 45 posts as negative
- BLOOMZ completely failed to label posts as neutral
- GPT-4 struggled to predict positive class.

Thank You

https://github.com/AridHasan/MUBASE