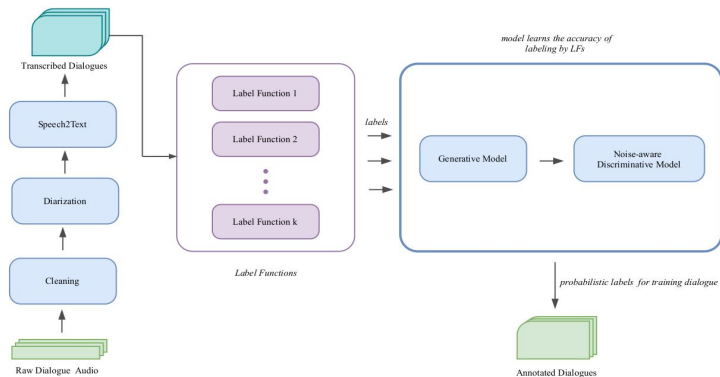


## Introduction

- High-quality **dialogue datasets for low-resource languages** like Bengali are rarely available. Thus existing neural open-domain dialogue systems suffer from data scarcity.
- We propose a **framework** to prepare large-scale open-domain dialogue datasets from publicly available audio employing weak-supervision which is particularly suitable for low-resource settings.



## Our Approach



- Cleaning.** Political discussion and debate audio from various public sources. We convert the raw audio files to 16kHz mono channel WAV audio after noise removal.
- Speaker Diartization.** ECAPA-TDNN based model used for speaker diarization of the dialogues
- Speech-to-Text.** wav2vec2.0 based speech to text engine trained on Bengali transcribed audio
- Punctuation Restoration.** We follow (Alam et al., 2020) which is a layered architecture consisting of a pre-trained BERT variant, a bidirectional LSTM and finally a linear layer on top of it.
- Speaker Role Labeling.** Weak supervision based technique used for speaker role labeling.

## Weak Supervised Speaker Role Labeling

- Each dialogue in the corpus consists of multiple speakers
- Most of the dialogues include a speaker who acts as "Host"
- label functions applied to unlabeled (without the speaker role label) dialogue corpus.
- Considerations while designing label functions:

number of questions asked

descriptiveness of the replies

utterance sentiment

- Accuracy of the labeling functions are learnt on the fly and weights are assigned to the corresponding outputs.
- The generative model generates a collection of probabilistic training labels that are used to train a strong, flexible discriminative model for a better generalization beyond the labeling functions.

## Corpus Description

- 7703 dialogues in total covering mostly political debates and multi-party discussions in Bengali language

Statistics	Count
Total number of sentences	66413
Avg. duration (in minutes)	5.62
Avg. speakers per dialogue	2.68
Avg. turns per dialogue	7.6
Avg. number of questions per dialogue	2.14

- Sentiment analysis results**

Positive	36.20%
Negative	24.36%
Neutral	39.44%

- Key Topics** found using LDA

Election  
International  
Bangladesh  
Economics

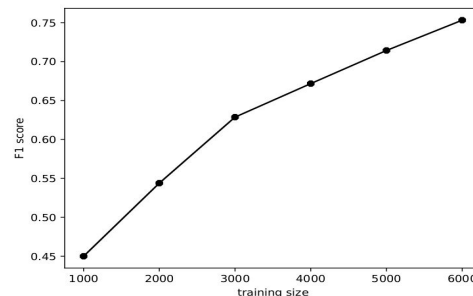
- Top (political) keywords using GloVe  
*Politics, Social, Economics, Society, Culture, Practice, Rules, Leftist* etc.

## Evaluation

- Our task is to classify biased speakers from the utterances
- Using SHONGLAP dataset, we fine-tuned BanglaBERT on our downstream task.
- Performance of model fine-tuned on our dataset (after 3 epochs)

Accuracy	0.735
Precision	0.7328
Recall	0.735
F1 Score	0.733

- Effect of training data size on test F1 score



## Conclusion and Future Directions

- A weak-supervision based novel framework to prepare and annotate large-scale open-domain dialogue corpus for low-resource languages
- Using the framework, we prepared a large Bengali open-domain dialogue corpus: **SHONGLAP** which improves BanglaBERT's performance during fine-tuning for downstream tasks.
- The approach can be extended to collect and prepare an even larger dialogue corpus covering a wide range of topics.
- Tasks like *dialogue-summarization*, *agreement-disagreement modeling*, *dialogue state tracking*, *open-domain dialogue generation*, *mining similar dialogues* in the context of **Bengali** will be explored in depth.

## Acknowledgements

We thank DOER Services Ltd. for funding and supporting this research.