EmoNoBa: A Dataset for Analyzing Fine-Grained Emotions on Noisy Bangla Texts

Khondoker Ittehadul Islam[1], Tanvir Hossain Yuvraz[1], Md Saiful Islam[1,2], Enamul Hassan[1]

[1]Shahjalal University of Science and Technology, Bangladesh

[2]University of Alberta, Canada

Motivation

- Numerous problems found solutions on English texts because of well-curated fine-grained emotion datasets.
- Bangla, 6th most spoken language in the world, lacks any such datasets.

What we propose

- *EmoNoBa* dataset: comprises manually annotated muliti-label dataset of 22,698 Bangla public comments covering 12 different domains.
- Established baselines by experimenting on popular models.
- Publicly released of all our works for fostering research on this direction.

Dataset Development

Purpose

Samples should contribute to making the dataset

- domain independent
- less repetitive

Preprocessing

Out of $\approx$ 50K collected comments, we keep the comments

- written in only Bangla alphabets
- remove duplicates
- exclude instances shorter than three or longer than 50 word tokens
- prioritize the instances for annotation that will increase the percentage of the unique word in the dataset [1]

Objectives

Given a predefined set of emotions - Junto-6 basic emotions, the goal is to identify all emotions conveyed in a piece of text.

Annotation

We use five annotators for each instance. Emotion(s) voted by atleast three annotators were considered the final labels. Instances that could not be finalized this way were sent to authors for the final tag. We will refer to the former instances as *genInst* and the latter as *exclnst*.

Furthermore, we evaluated the annotators with an accuracy metric. We will denote such accuracy as *AnnoAccu*.

For *genInst*:

$$\mathrm{AnnoAccu} = \frac{1}{|\mathrm{I}|}\sum_{i\in \mathrm{I}} \frac{\mathrm{T_i} \cap \mathrm{O_i}}{\mathrm{T_i}}$$

For *exclnst*:

$$\mathrm{AnnoAccu} = \frac{1}{|\mathrm{I}|}\sum_{i\in \mathrm{I}} \frac{\mathrm{T_i} \cap \mathrm{A_i}}{\mathrm{T_i}}$$

Statistics and Analysis

- 80 undergraduate students annotated 5 to 5,000 instances each, with 74 of them attaining *AnnoAccu* of 60% or more.
- *sadness*, *anger*, and *joy* are the most frequent emotions.
- Vast majority of data are from *Personal* and the least from *Health*.

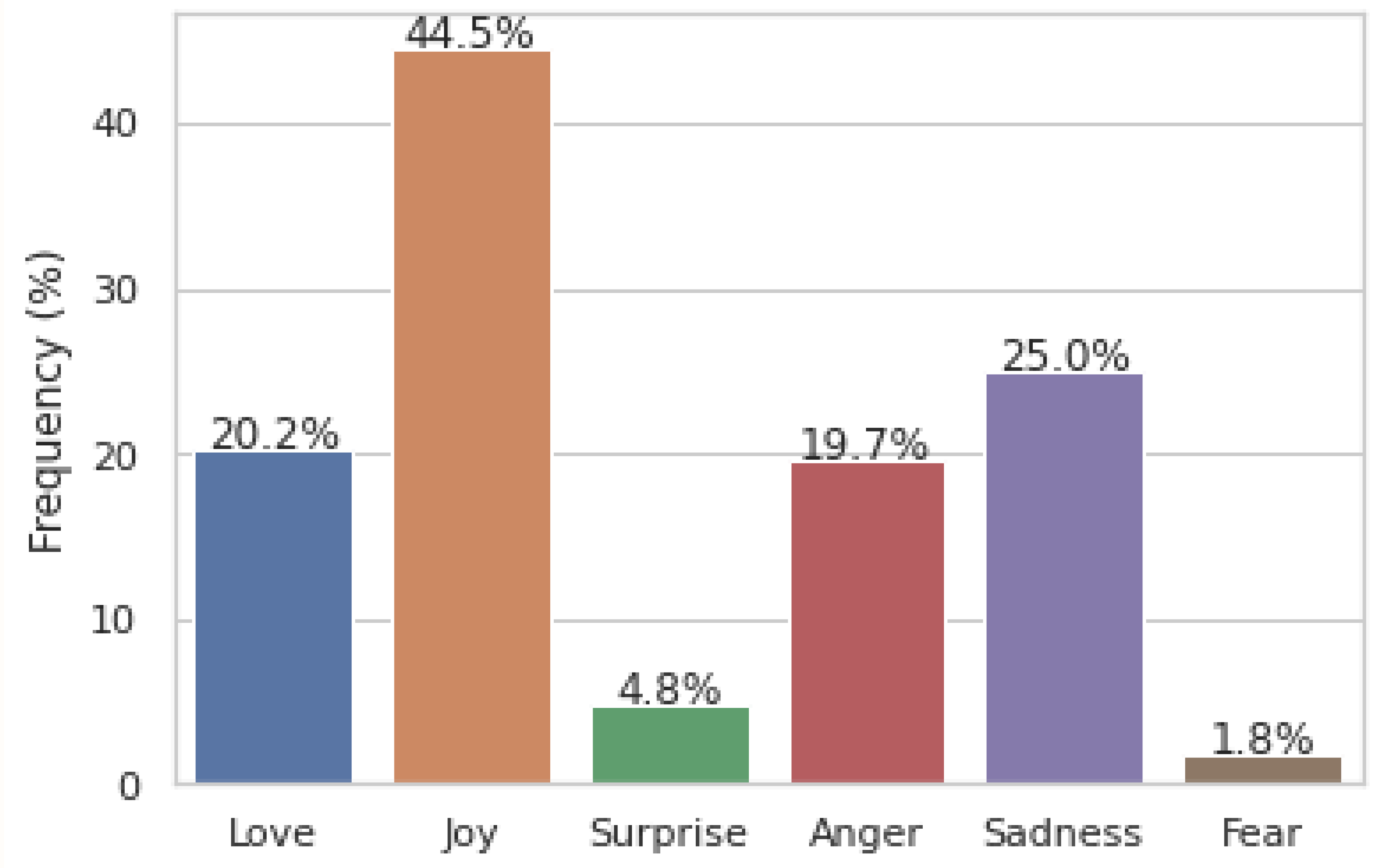

Figure 1: Percentage of instances labeled with a given emotion in our dataset.

Gold Set a.k.a Test set

- We performed per-multi-label stratified split to create training (90%) and testing (10%) sets.
- Test set received precedence on *exclnst*.
- In the cases of overflows, leftover instances were inserted into the training set and vice versa.

Emotion	Instances	exclnst (%)	UW (%)
Love	390 (17.17%)	54.87%	49.87%
Joy	857 (37.72%)	36.87%	45.89%
Surprise	149 (6.56%)	71.81%	67.61%
Anger	575 (25.31%)	54.60%	45.00%
Sadness	572 (25.18%)	43.88%	49.16%
Fear	93 (4.1%)	80.65%	65.52%
Total	2,272	40.18%	35.03%

Table 1: Statistics of test set with unique word (UW) percentage per emotion label.

Methodology

- Lexical Feature
  - Extracted word (1-4) and character (1-5) n-grams
  - Vectorized each instance with the TF-IDF weighted scores
  - Trained on linear SVM models.
- Recurrent Neural Network
- Pre-trained Language Model
  - Bangla-BERT-Base [3]

Result and Analysis

Pretrained Language models outperformed others, right?

Actually they performed poorer than random baseline. The reason being these models were trained with formal i.e., Wikipedia texts, whereas our data were public comments meaning informal and noisy.

Method	Love	Joy	Surprise	Anger	Sadness	Fear	Macro Avg
Random	24.30	43.20	11.42	33.57	32.71	7.52	25.46
Bi-LSTM + Attn. (FastText)	0.0	52.71	0.0	0.0	22.70	0.0	12.57
Bi-LSTM + Attn. (Random)	0.0	57.79	0.0	18.49	51.97	0.0	21.38
Bangla-BERT	18.33	52.30	11.70	22.37	42.96	0.0	24.61

Figure 2: Binary Task F1-score of each emotion class and Macro Average F1-score of each method.

Then which model stood out?

Integration of all word 1-4 grams with character 1-3 grams provided the best result of 42.81 F1.

Error Analysis

Effect: Model additionally predicts *sadness* in *joy* and *love* instances.

Cause: The reason is negative words, such as "show-off", are the strongest words of *sad* emotion, but they can also lie in positive emotion instances.

References

[1] 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.*

[2] Raman Chadha. 2020. The junto emotion wheel: Why and how we use it. *The Junto Institute.*

[3] Sagor Sarker. 2020. *Banglabert: Bengali mask language model for bengali language understading*