SentNoB: A Dataset for Analysing Sentiment on Noisy Bangla Texts

Khondoker Ittehadul Islam[⋆], Md Saiful Islam[⋆‡], Sudipta Kar[†] and Mohammad Ruhul Amin[♢]

[⋆]Shahjalal University of Science and Technology, Bangladesh
[‡]University of Alberta, Canada, [†]Amazon Alexa AI, USA, [♢]Fordham University, USA

Motivation

Why Sentiment Analysis?

Sentiment analysis is **one of the classic problems** in computational linguistics

Why in Bangla?

- 6th spoken language in the world
- To enhance impact on social welfare and businesses

Why New Dataset?

Previous Works

1. None to **slight inter annotator agreement score** questioning the annotation reliability
2. Lack of cross-domain generalization capability due to **large domain dependency**
3. **Lack of public availability** for further research

Contributions

- $\approx$**15K** social media **comments** from **13** different **domains**
- **Experiment** on **different techniques** and shed light on different aspects of the problem
- Make **dataset and model publicly available**

SentNoB

Sources

- YouTube
- ProthomAlo, the most circulated newspaper in Bangladesh.

YouTube

প্রথমআলো

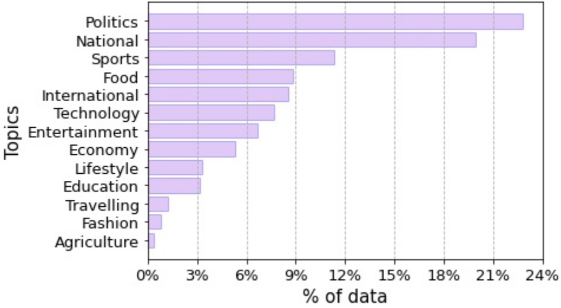

Figure: Topic distribution of the dataset.

Data Annotation

- Assigned 3 annotators for each instance
- Merged 5 labels to 3 labels
- Labels: *Positive*, *Neutral*, *Negative*
- IAA score of 0.53

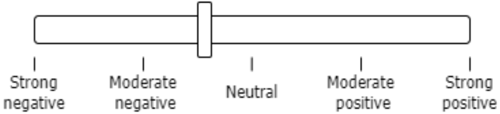

Figure: The labeling interface. Random phrases were shown and annotators chose the sentiment and its degree.

Insights

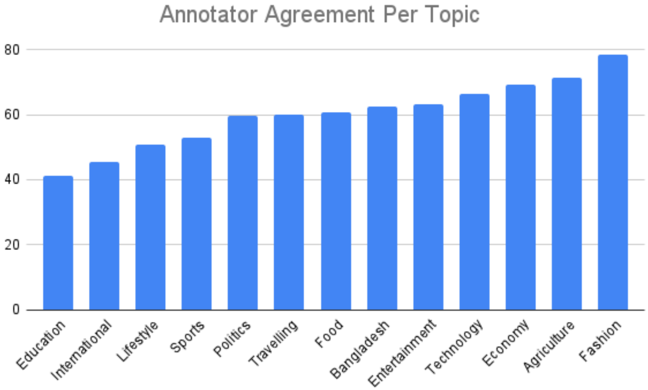

Figure: Inter annotator agreement score per topic. *y*-axis represents agreement score, whereas *x*-axis contains each topics.

Benchmark Models

Lexical Features

- word (1-3) *n*-gram
- char (2-5) *n*-gram

Semantic Features

- FastText embedding

Neural Networks

- Bi-LSTM + Attention

Transformer based Language Model

- multilingual-BERT or mBERT

Result

Method	Precision	Recall	F1
Majority	41.24	41.24	41.24
Random	33.67	35.44	34.53
Weighted Random	31.89	33.35	32.60
Bi-LSTM + Attn. (FastText)	52.24	63.09	57.15
Bi-LSTM + Attn. (Random)	56.16	64.97	60.25
mBERT	49.58	56.43	52.79
Unigram (U)	56.89	71.06	63.19
Bigram (B)	54.32	66.20	59.68
Trigram (T)	51.57	60.21	55.56
U + B	**57.71**	72.95	64.44
U + B + T	57.03	71.88	63.60
Char 2-gram (C2)	53.29	66.39	59.12
Char 3-gram (C3)	56.06	70.87	62.60
Char 4-gram (C4)	56.62	71.44	63.17
Char 5-gram (C5)	56.94	71.94	63.57
C2 + C3	56.00	70.93	62.59
C3 + C4	56.49	71.31	63.04
C4 + C5	57.30	72.76	64.11
C2 + C3 + C4	56.45	71.44	63.07
C3 + C4 + C5	57.60	73.39	64.54
C2 + C3 + C4 + C5	57.06	72.89	64.01
U + B + C3 + C4 + C5	56.96	72.51	63.80
U + B + C2 + C3 + C4 + C5	57.05	72.70	63.93
U + B + T + C2 + C3 + C4 + C5	**57.71**	**73.39**	**64.61**
Embeddings (E)	50.68	63.75	56.46
U + B + C2 + C3 + C4 + C5 + E	57.48	73.14	64.37
U + B + T + C2 + C3 + C4 + C5 + E	57.36	72.45	64.03

Table: Precision, Recall, and F1 for different methods.

Dominant Features

Positive: ধন্যবাদ (thanks), অসাধারণ (great), মেধাবী (talented), খুব ভাল (very good), বেস্ট ! (best !), রিপোর্টটা অসাধারণ চিল (the report was great), আলহাদুলিল্লাহ গ্রেট নিউজ (thanks god great news), ❤ ❤ ❤
Negative: পুলিশ (police), হনুমান (monkey), বালের (slang), খুন (murder), ধিক্কার (indignation), জবাই (slaughter), কুত্তার বাচ্চা (slang), বিচার হবে না (there will be no justice), মেরে ফেলা উচিৎ (should be killed), গরিবেরা সবজায়গায় নিপীড়িত (the poor are oppressed everywhere)
Neutral: ফোন (phone), প্রাইভেট (private), আলোচনা (discussion), রাষ্ট্রপতি (president), প্রশ্ন (question), না ভাইয়া (no brother), ঠিক বলেছেন (you are right), বুঝলাম না কিছুই (didn't understand anything), টাকা লাগে না (it doesn't cost money)

Table: Examples of some of the strongest word n-grams from each class with their English translations.

Thank You

Takeways

- Topic-wise dataset which can be used for aspect-based sentiment classification
- More investigation on pre-trained language models

Dataset

`https://tinyurl.com/y2dburjw`

Transformer Model

`https://git.io/JuuNB`