Report on NLP Hackathon

21 January 2023

BY

Team Axon

Members

Shams Tanveer Jim

Md. Ashraful Islam

Adnan Abdullah

Dataset Analysis and Findings

The given dataset is a Bangla dataset in text format. Later we transformed into CSV files to explore and work with the data. The dataset contains sentences that are split into words with a line gap. We performed EDA(Exploratory Data Analysis) to find the features of the dataset. Every word of each sentence has an entity tag attached to it. For example, in the first sentence of the train dataset, ‘তার মৃত্যুর দশ দিন পর, ১১৫ কৃষ্ণাঙ্গ উচ্চ বিদ্যালয়ের শিক্ষার্থীরা তার হত্যার প্রতিবাদে ম্যাককম্ব এর মাধ্যমে মিছিল করেছে। ”, every word of the sentence is allocated with entity tag O, and only ‘ম্যাককম্ব ‘ has the tag B-LOC. There are several entity tags given in the dataset. They are explained below,
'O' - Out tag
'B-LOC' - Beginning of location name
'B-GRP' - Beginning of group name
'I-GRP' - Inside of group name
'B-PROD' - Beginning of product name
'B-CW' - Beginning of common words
'I-CW' - Inside of common words
'B-CORP' - Beginning of corporation name
'B-PER' - Beginning of person's name
'I-PER' - Inside of person's name
'I-CORP' - Inside of corporation name
'I-PROD' - Inside of product name
'I-LOC' - Inside of location name

As Bangla is semantically and syntactically complex as a language, there are many aspects to consider while working with the given dataset. For example, the dataset is noisy. We needed to do some preprocessing to eradicate the noise. Besides, some inconsistencies, like the same word, were entitled to different tags at different points of the dataset.

As part of preprocessing of data, we have decreased the amount of ‘O’ tagged words as they were much more than other tags. So, there was a chance of getting biased performance from the model training. So, we limited the ‘O’ tagged words to such an amount that there is a balance among all the tags.

Methodology

We have worked on different variances of the Bert transformer model. We used sagorsarker/bangla-bert-base, csebuetnlp/banglabert, sagorsarker/mbert-bengali-ner, csebuetnlp/banglabert, etc. Among all of them, we have got the best performance on the bert-base-multilingual-cased model.

DL Based

First of all, we added sentence id to all words in the dataset. For that, we added the same sentence_id for all the words corresponding to the same sentence. After that, we listed down all the unique labels assigned to each word.

Then we set the learning rate, no of epochs, batch size, etc. Then we define the model using NERModel class under the model variable. We used a learning rate of 0.0001 and we used a batch size of 16 to run our model. After running 30 epochs the loss of the model goes down to 0. We used the training set for training and the dev set for validation purposes while training the model. After that, we trained the models and evaluate their performances.

Feature Based

Usually, in feature-based NLP models, we take the most important features that have an influence on the performance of the model. So, while working on feature-based modeling, we extracted several features from the dataset. Namely,
pos - parts of speech; **first_word** - the first word of the corresponding sentence;
last_word - last word of the corresponding sentence;
prev_word- exactly the previous word; next_word- exactly the next word;
prev_word_pos- parts of speech for the previous word;
next_word_pos- parts of speech of next words;

word- the word we are working on
labels- the label for the working word

After extracting the features, we feed those features to the models. We used the same models we used in the DL-based methods.

Result

We are providing the performance based on the performance of DL-based model.

Model	**eval_loss**	**precision(%)**	**recall(%)**	**F-1 score(%)**
bert-base-multilingual-cased	0.461	65.62	68	66.79
csebuetnlp/banglabert	0.652	59.198	62.75	60.92

For feature-based model, we got the following performance on bert-based-multilingual-bert model,

Model	**eval_loss**	**precision(%)**	**recall(%)**	**F-1 score(%)**
bert-base-multilingual-cased	1.467	28.58	34.99	31.46

Analysis of the Models

From the result table, we can see that for DL-based model performance, bert-based-multilingula-model overperform the banglabert model of buet. A probable cause can be the versatility bert multilingual model contains. Besides, banglabert is a comparatively new model , so the performance can still be improved with further more use of bangla bert.

For the feature based model, we have got a comparatively low performance. It totally depends on the performance of the features selected for Bangla dataset. Besides, the complexity of Bangla language make the performance of the features performance comparatively low, because of the less use of Bangla in model training.

Link for the discussion log of Team Axon

https://docs.google.com/document/d/1Aqt2pnrXGk-jOnRibiAJjC_NE5YnmoURL_IIDI0yc6U/edit?usp=sharing