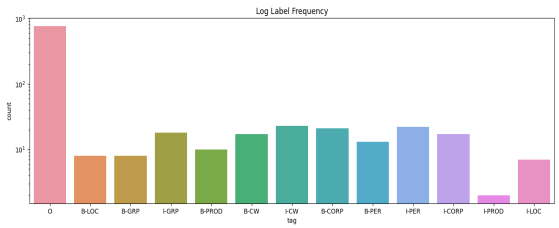
Final Report

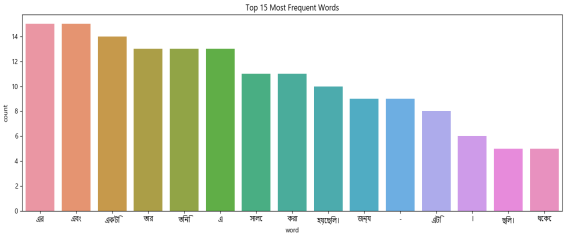
Team: Return_Zero

Exploratory Data Analysis

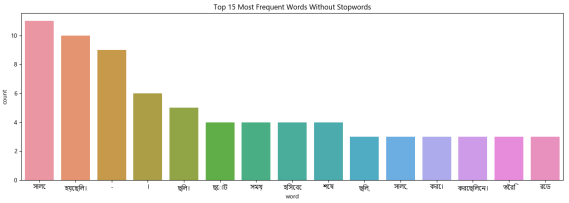
Log Label Frequency



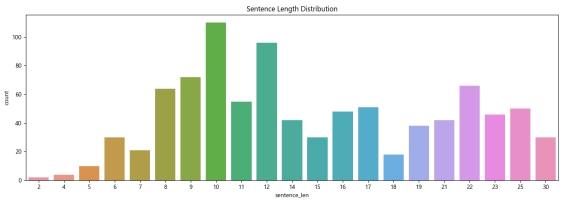
Top 15 most frequent words



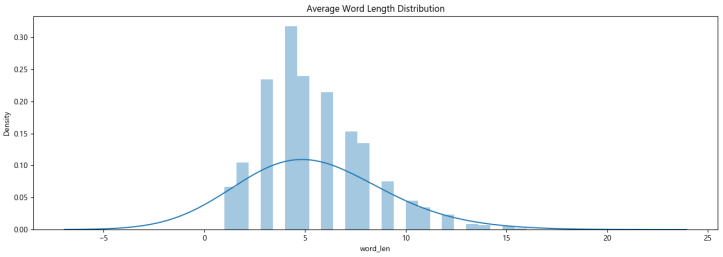
Top 15 most frequent words without stopwords



Sentence length distribution



Avg word length distribution



Methodology

Feature-engineered Models:

1. RandomForestClassifier(n_estimators = 100, max_depth=10,min_samples_leaf = 1, min_samples_split = 2,)
    a. Simple feature map used: (istitle; isupper; isdigit; len(word))
2. CRF (Conditional Random Field) is a probabilistic algorithm LBFGS algo.
    a. We fine tuned parameters c1,c2, max_iter.
    b. C1 = 0.1 (L1 regularization); C2 = 0.2 for (L2 regularization term : penalty proportional to the square of the parameters)
    c. Max_iter: fine tuned from 100 to 1000 iterations.

3. Features used:
    a. We have used Extra Features with the help of BNLP (Bengali Natural Language Processing)

    b. Bengali Corpus Class
        i. from bnlp.corpus import stopwords, punctuations, letters, digits
    c. So we added new features:
4.
        i. isStopWord(token)
        ii. isPunctuation(token)
        iii. isDigit(token)

5. Also we use CRF POS Tagging
    a. First concat the words using **group_by** on sentence_id
    b. Then run bn_pos.tag
    c. Extra Part of Speech (PoS) Tagging added

my_team_feature_model_1330.txtDeep Learning Models:

1. Bi-LSTM simple version : 128 hidden units; recurrent dropout rate of 0.1.
    a. n_tags number of neurons; activation function used is "relu"
    b. F1 score = 0.42 (better than Random Forest)
2. BERT-base model: https://huggingface.co/sagorsarker/bangla-bert-base
    a. F1 Score = 0.60
    b. Better than Bi-LSTM score
3. Electra-base: https://huggingface.co/csebuetnlp/banglabert
    a. F1 Score: 0.71
    b. Better than BERT-base model score

4. Electra-large: https://huggingface.co/csebuetnlp/banglabert_large
    a. F1 Score: 0.75
    b. Better than Electra-base or BanglaBERT base

Results

Model Name	**Result (F1 Score)**
RandomForest	0.07
CRF	0.7084
BiLSTM	0.42
BERT-base	0.60
BanglaBERT from csebuetnlp	0.71
BanglaBERT large from csebuetnlp	0.75

Analysis of the models

- RandomForestClassifier; Simple feature map
    - F1 score 0.07 (NOT good)
    - limitations in handling sequential data, cannot consider context
- CRF conditional random field
    - Can take context into account which is very important for NER

    - Features = title; stopwords ; islower; isupper
        - dev_generic.csv

          Epoch F1
          100 0.6777
          500 0.6842
          10000 0.6832

    - Features = title; stopwords ; islower; isupper;
        - Dev_no_punctuation.csv

          Epoch F1
          500 0.6884

    - Features = title; stopwords ; isDigit; isPunctuation (NEW)
        - Dev_generic.csv

          Epoch F1
          500 0.6864 (prev was .6842)

    - Features = title; stopwords ; islower; isupper;
        - Dev_no_punctuation.csv

          Epoch F1
          500 0.6859 (decreased ; prev 0.6884)
    - Features = title; stopwords ; isDigit; isPunctuation (NEW)
        - POS added (with help of bnlp)
        - Dev_no_punctuation.csv

          Epoch F1
          400 0.6923

    - Features = title; stopwords ; isDigit; isPunctuation (NEW)
        - POS added (with help of bnlp)
        - Dev_no_punctuation.csv

          Epoch F1
          500 0.7084 (prev 0.6884)

- Deep models:
    - BiLSTM
        - Reached 0.42 f1 score, limited because Bi-LSTMs are generally less effective in handling long-term dependencies
    - Sagor Sarkars' BERT-base
        - Did not perform well as it is pretrained with noisy corpus, though expected to perform well
    - BanglaBERT base
        - This is an Electra model which is discriminative trained so better than BERT
        - 20 epochs 0.67 f1, performance good early in the training
        - Improved upto 0.71 f1 after 40 epochs
        - Punctuation Removal
            - We thought that Punctuation removal could help improve the performance of the NER model by reducing the number of possible word combinations that the model needs to consider.
            - Did not improved the model, due to missing context
    - BanglaBET large
        - 20 epoch gives 0.7 f1, the capacity is large, so it can learn well
            - Punctuation removal
                - Did not improve the model performance
        - Worked best to reach 0.75 f1 score with 20% train-test split, as it is pretrained with high-quality Bangla corpus.

References:

1. https://www.analyticsvidhya.com/blog/2021/06/part-10-step-by-step-guide-to-master-nlp-named-entity-recognition/
2. https://github.com/guillaumegenthial/sequence_tagging
3. https://github.com/luopeixiang/named_entity_recognition
4. https://github.com/imranulashrafi/banner
5. https://www.kaggle.com/code/amitbda18/ner-with-bi-lstm-crf
6. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9044317

Discussion Logs

2022/1/21, 2 PM BDT

Agenda

- Final submission

Notes

- Final submission:
    - Bangla BERT-large with 40 epochs is selected as deep model
    - CRF with POS and BNLP features is selected as the feature based models.

2022/1/21, 12 AM BDT

Agenda

- Report writing started
- Clean up model files and sagemaker

Notes

- Write down the F1 scores and model summary
- Clean up for submission and discuss the final models

2022/1/21, 10 AM BDT

Agenda

- POS tagging improved CRF
- Removing Punctuations with tokens

Notes

- POS with CRF
    - [Tanvir] Good performance
- Removing token punctuation
    - [Afnan] Lets' try both with and without and see the output

2022/1/21, 8 AM BDT

Agenda

- POS tagging can improve CRF
- Tune BanglaBERT and BanglaBERT-large

Notes

- POS tagging for CRF
    - [Tanvir] BNLP provides pretrained POS model, we can add POS tags as features to CRF
- Tune BanglaBERT and BanglaBERT-large
    - Examine the losses and f1 scores, save the best model and upload to sagemaker for prediction

2022/1/21, 2 AM BDT

Agenda

- Look for more features for CRF
- Analyze BERT training

Notes

- CRF Features
    - [Tanvir] We can use BNLP package to get the stopwords, punctuations, digits
- BERT training
    - Plot losses, accuracy
    - Train for 20 epoch, the again train for if need as the learning curve/loss plot.

2022/1/20, 10 PM BDT

Agenda

- Train BanglaBERT, BanglaBERT large

Notes

- Deep learning model training
    - [Afnan] Started training BanglaBERT in Sagemaker, very slow training
    - [Shayekh] Stated training BanglaBERT large in Sagemaker

2022/1/20, 8 PM BDT

Agenda

- Train BERT model
- Update on CRF

Notes

- Train BERT model
    - [Afnan] Pertained model in English should perform bad, we should look for models pretrained in Bangla.
- CRF model ready intially
    - [Tanvir] CRF looks better than RandomForest

2022/1/20, 6 PM BDT

Agenda

- Update for feature based models
- Start working on deep learning models

Notes

- Feature based models
    - [Tanvir] Formatted dataset to run the CRF model with hand engineered features.
- Deep learning models
    - [Afnan] We can start the BERT-base model from HuggingFace hub which is pertained in English corpus
    - [Shayekh] Prepared dataset in BERT format

2022/1/20, 3 PM BDT

Agenda

- Feature based models
- Literature review

Notes

- Feature based models
    - [Tanvir] RandomForest is performing very worse. Lets look for other models.
- Literature review
    - Study literature for state-of-the-art feature based models in NER

    - [Shayekh] SVM, CRF may be good candidates. We can see examples from kaggle.

2022/1/20, 10 AM BDT

Agenda

- Dataset overview
- Statistics generation
- Think of naive baselines such as random forests

Notes

- **Dataset Overview:**
    - [Afnan] Downloaded the dataset. Looks good.
    - Converted to CSV from TXT formats and group by sentence id to generate suitable formats for RandomForest
- **Statistics generation**
    - [Shayekh] Plot label distribution, word length distribution, sentence distribution, popular words