Team : BUET_Caraxes

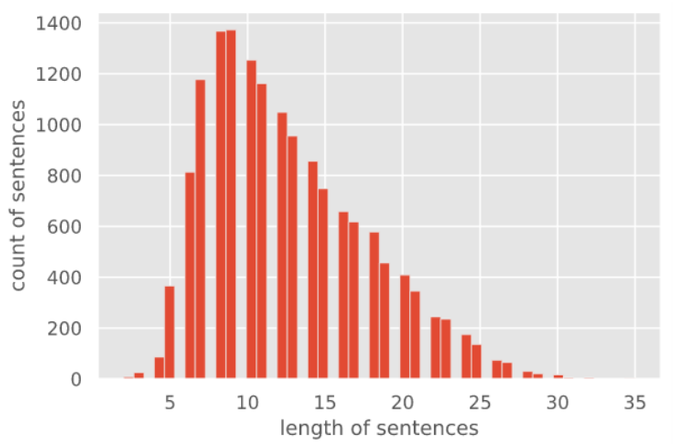
Rasman Mubtasim Swargo, Mushtari Sadia, Asif Azad

Date : 21-01-2023

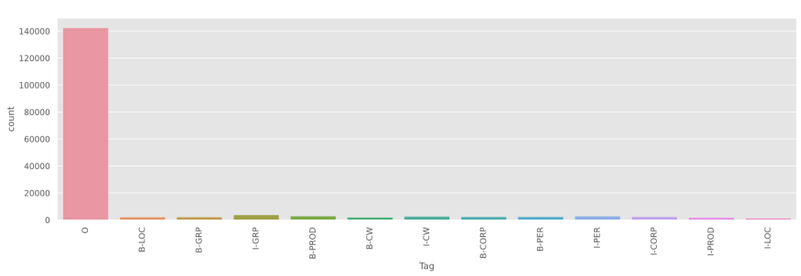
1 Exploratory Data Analysis

The challenge was to classify named entities of Bangla language from a given dataset. From the dataset, the following insights could be detected :

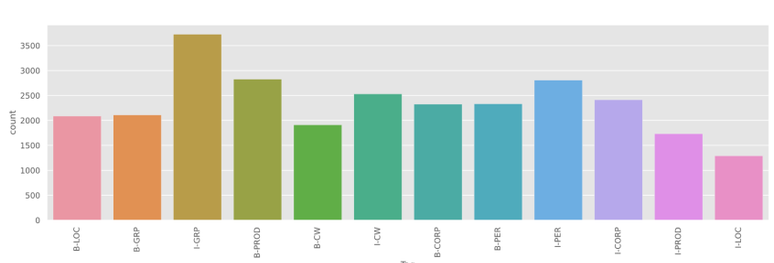
a. It seems most of the sentences are 8-10 words long, and the distribution is normal.



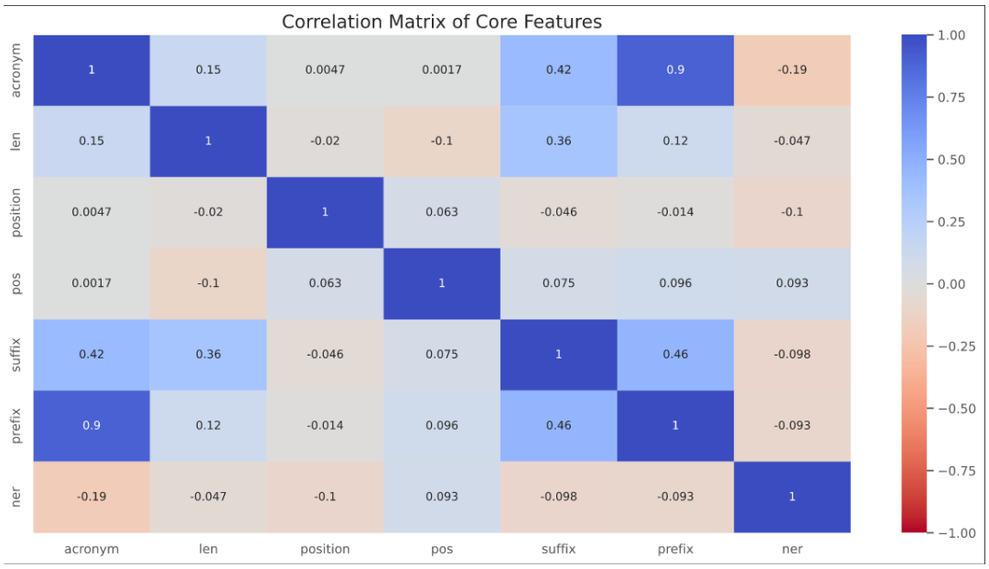
b. The distribution of words across tags are as following



Since the “O” tag is most occurring, the distribution of other tags is unclear from this plot. So, removing the “O” tag’s frequency -



c. The correlation of various features with label



2 Methodology

a. Machine Learning Based Model with Feature Engineering

Selected features :

a. **Word to vector representation** : To get the vector representation of a word, we used the glove vector representation [1] available in hugging face.
b. **Acronym**
c. **Length of a word**
d. **Position in sentence**
e. **Parts of speech tagging**
f. **Suffix of word**
g. **Prefix of word**
h. **Stem of word**
i. **Named Entity Recognition:** We used an existing NER model[2] to predict (without any finetuning) and used it as a feature.
j. **Leading and lagging words**: We also added the previous and next word of a sentence as a feature.

` Based on the generated correlation values, we selected the features highly correlated with the tags. We experimented with dropping and choosing several features and chose the best performing model. We trained classifiers such as lightgbm, xgboost, random forest, support vector machine, decision tree. We also tried ensembling techniques on these models.

b. Deep Learning Based Models

1. We tried to infer the data using BanglaBert[4]. After successfully inferring, we went for fine tuning the model for this dataset. We did some preprocessing on the dataset so that it fits the model. Again, we inferred after fine tuning.

2. We got 'BANNER: A Cost-Sensitive Contextualized Model for Bangla Named Entity Recognition'. This model[3] was not well written. After modifying codes from the codebase we were able to run the model.

3. We ensembled the previous two results

using XGBoost.

3 Results

Model	Selected Features	Macro F1 on Dev Set
Bangla Bert[4], 30 epoch	-	0.77
BANNER + BERT(XGBoost)	-	0.75
BANNER[3], 20 epoch	-	0.69
Random Forest	Glove Vectors	0.21
Random Forest	Glove Vectors,acronym,len,position,pos, suffix,prefix,ner	0.18
Random Forest	Glove Vectors,acronym,len,position,pos, suffix,prefix,ner,leading,lagging,stem	0.0
Random Forest	Glove Vectors,acronym,len,position,pos, suffix,prefix,ner,stem	0.17
Random Forest,Xgboost	Glove Vectors,ner	0.13
Random Forest(e=5),Xgboost	Glove Vectors	0.08

4 Analysis

The feature engineering part on the ML model produced many interesting observations. For example, we introduced many features, described above. However, dropping all of these except for the glove vectorization of the word actually produced the best result. Along with trying various models, we also tried parameter tuning with LightGBM, Xgboost, MLP and Random Forest by increasing the number of estimators and max iterations. Increasing these further tended to overfit the model. Random Forest Classifier performed the best. We tried ensembling other models with random forest, but the score could not exceed that of random forest.

For deep learning model, we used BanglaBert[4]. It performed well, achieving 0.77 macro F1 score on the dev set. Its performance was similar for predicting all of the labels, except for the label ‘O’. Then we trained another deep learning model, BANNER. After training for 20 epochs, the performance was 0.69 (macro F1 score). Ensembling both models did not improve the score.

5 References

1. sagorsarker/bangla-glove-vectors · Hugging Face
2. https://github.com/sagorbrur/bnlp
3. https://github.com/imranulashrafi/banner
4. https://github.com/csebuetnlp/banglabert/

Logs

20/1/2023, 11AM

[Sadia] Processed train and dev dataset

20/1/2023, 1PM

[Sadia] Found glove vector representation for bangla words here : sagorsarker/bangla-glove-vectors · Hugging Face
[Sadia] Generated various statistics for training data to get an idea of features

20/1/2023, 3PM
[Sadia] Trained lightgbm on glove representation
[Sadia] Converted tags to numerical values using label encoding

20/1/2023, 10PM
[Sadia] Created submission pipeline using evaluation script
[Sadia] Added new features : lagging and leading words
[Sadia] Added new features: POS, position, length, stem
[Sadia] Added models : xgb, random forest, decision tree, svr

21/1/2023, 2AM
[Sadia] Created ensemble pipeline with votingclassifier
[Sadia] Tested performance of many combination of features. Chose combinations from correlation values
[Sadia] Trained ensemble model

21/1/2023, 11AM
[Sadia] Generated results