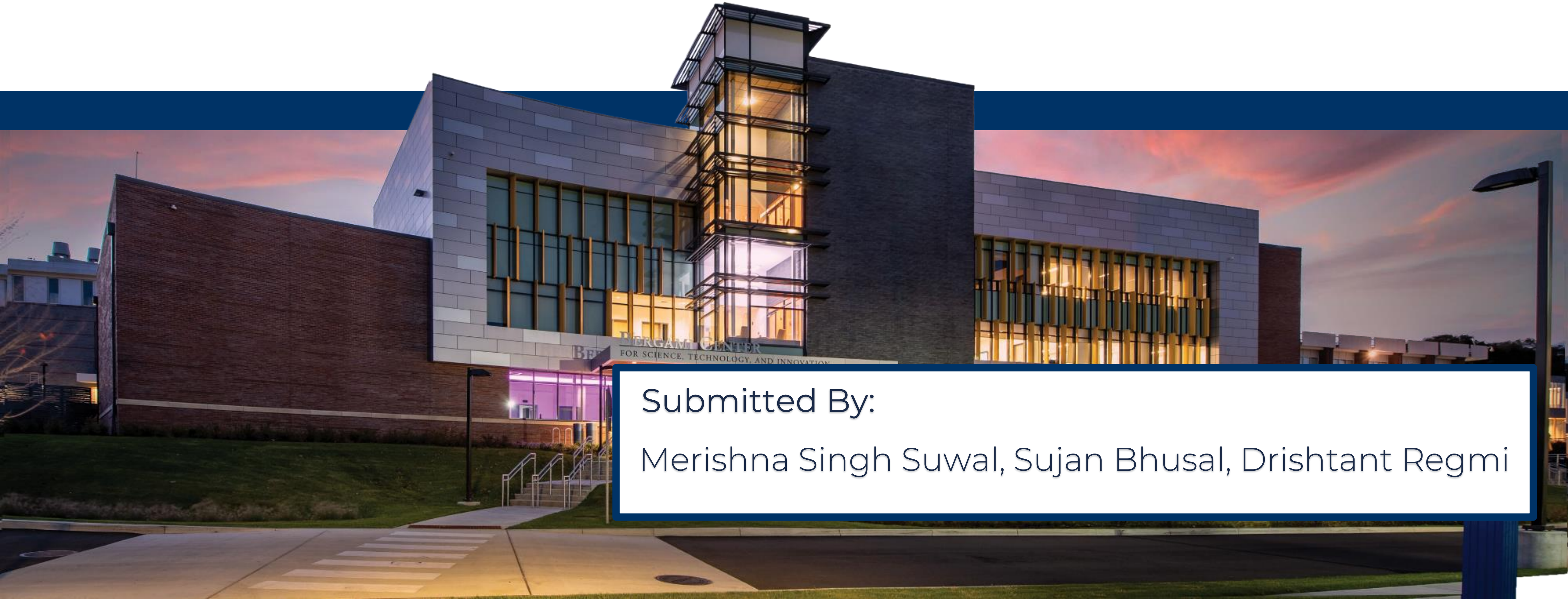
NEPALI OFFENSIVE LANGUAGE DETECTION & SENTIMENT ANALYSIS

FINAL PROJECT FOR DSCI 6004 – NATURAL LANGUAGE PROCESSING

University of New Haven



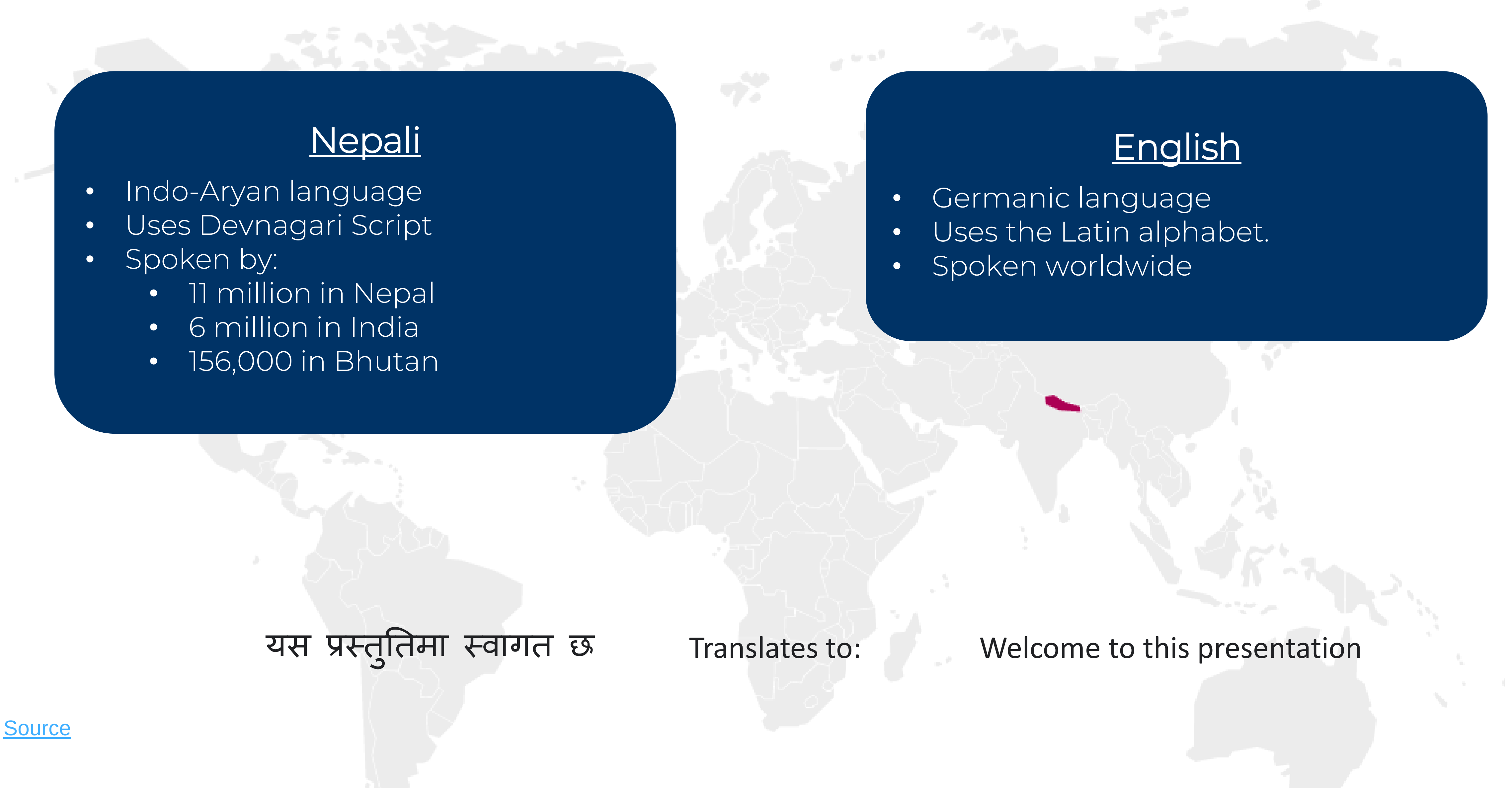
Submitted By:

Merishna Singh Suwal, Sujan Bhusal, Drishtant Regmi

Statement of Project Objectives

- Develop an Offensive language detection model for Nepali text which is a low-resource language.
- Contribute to a safer online environment in the Nepali-speaking community.

Linguistic Differences between Nepali and English



Nepali

- Indo-Aryan language
- Uses Devnagari Script
- Spoken by:
  - 11 million in Nepal
  - 6 million in India
  - 156,000 in Bhutan

English

- Germanic language
- Uses the Latin alphabet.
- Spoken worldwide

यस प्रस्तुतिमा स्वागत छ Translates to: Welcome to this presentation

Source

Review of State-of-the-Art

- Numerous studies done in English
  - Caselli et al., 2021 (HateBERT) - Fine-tuned BERT for abusive language detection
  - Zampieri, Marcos, et al., 2020 - Offensive language identification and categorization in 5 languages (Arabic, Danish, English, Greek, and Turkish)

- Fine tuned domain specific language models
  - BioBERT (Lee et al., 2019), FinBERT (Yang et al.,2020), and LEGAL-BERT (Chalkidis et al., 2020).

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for Abusive Language Detection in English. In Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), pages 17–25, Online. Association for Computational Linguistics.

Zampieri, Marcos, et al. "SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020)." *arXiv preprint arXiv:2006.07235* (2020).

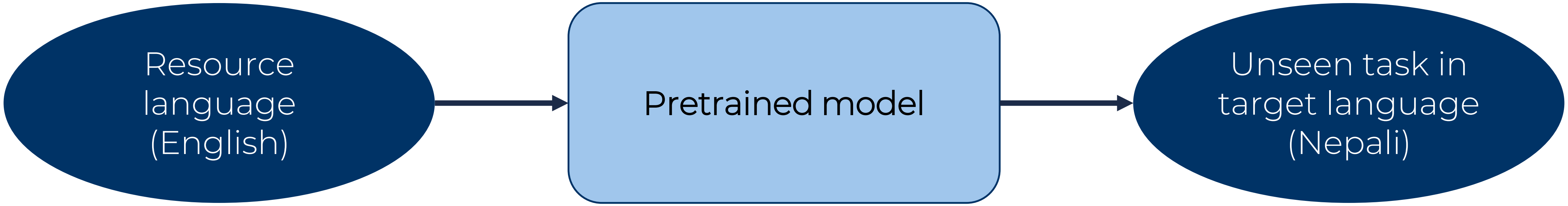
Review of State-of-the-Art

- Some studies in Nepali
  - Niraula et al., 2021 - Offensive language detection using supervised machine learning
  - Singh et al., 2020 - Aspect Based Abusive Sentiment Analysis using BiLSTM

Nobal B. Niraula, Saurab Dulal, and Diwa Koirala. 2021. Offensive Language Detection in Nepali Social Media. In Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), pages 67–75, Online. Association for Computational Linguistics.

O. M. Singh, S. Timilsina, B. K. Bal and A. Joshi, "Aspect Based Abusive Sentiment Detection in Nepali Social Media Texts," 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), The Hague, Netherlands, 2020, pp. 301-308, doi: 10.1109/ASONAM49781.2020.9381292.

Challenges in Cross-linguistic Transfer of Offensive Language Detection (OLD)

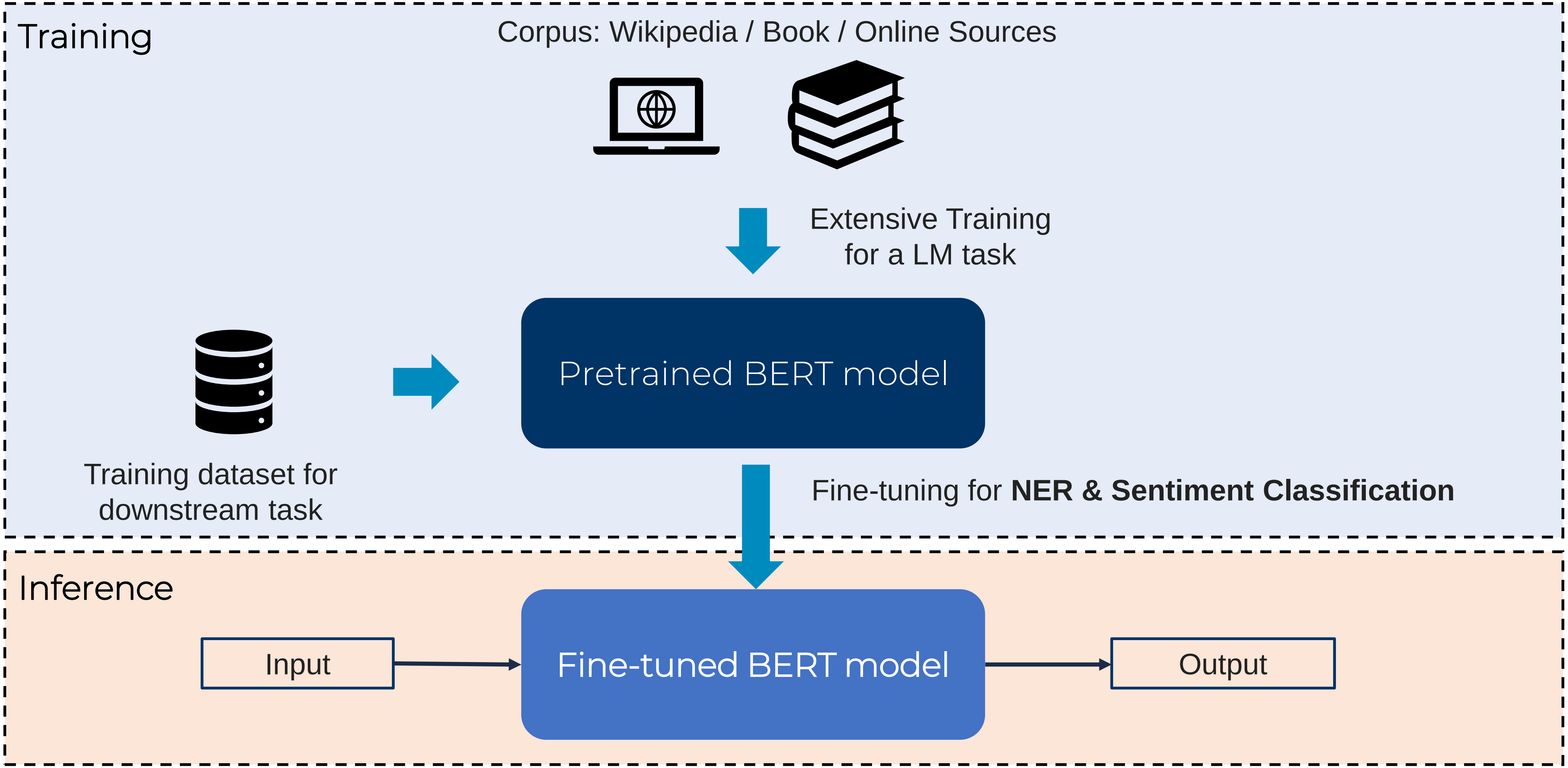

Resource language (English)
Pretrained model
Unseen task in target language (Nepali)

Problem: Current LMs are developed with varieties of languages not suitable

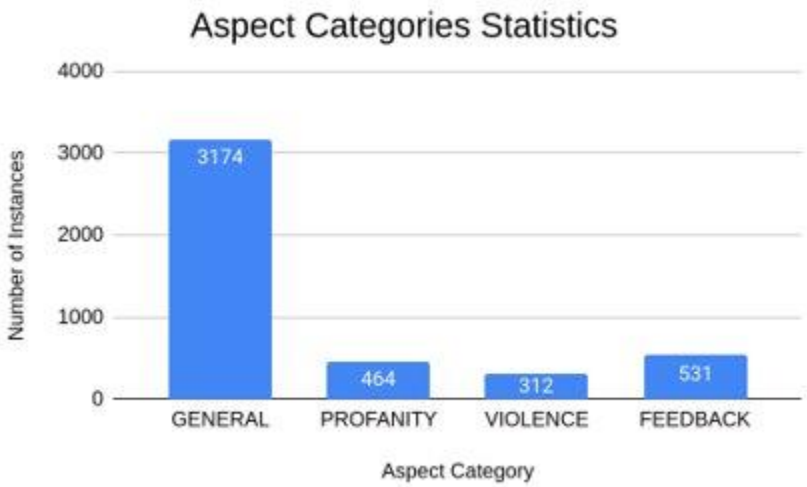
- Linguistic Diversity
- Cultural differences and Morphologically rich
- Shortage of Data Resources

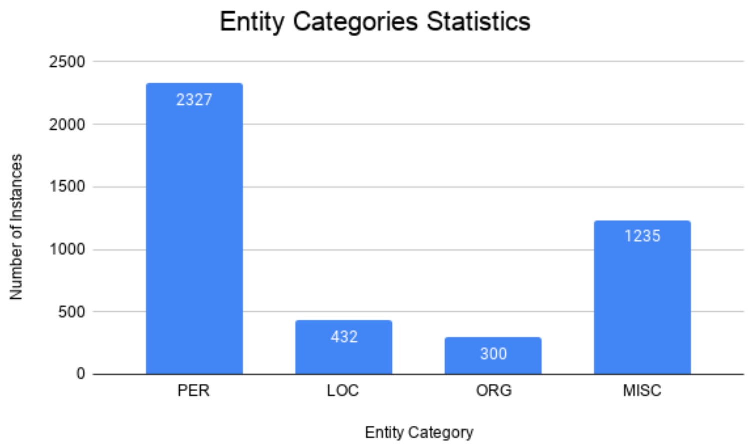
Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In Proceedings of the Third Workshop on Abusive Language Online, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Our approach


Training
Corpus: Wikipedia / Book / Online Sources
Extensive Training for a LM task
Pretrained BERT model
Training dataset for downstream task
Fine-tuning for **NER & Sentiment Classification**
Inference
Input
Fine-tuned BERT model
Output

Training dataset: Offensive Aspect Categories





General (Positive Criticism/ Derogatory remarks, insults)

Profanity and vulgarity (disrespectful and inappropriate)

Violence

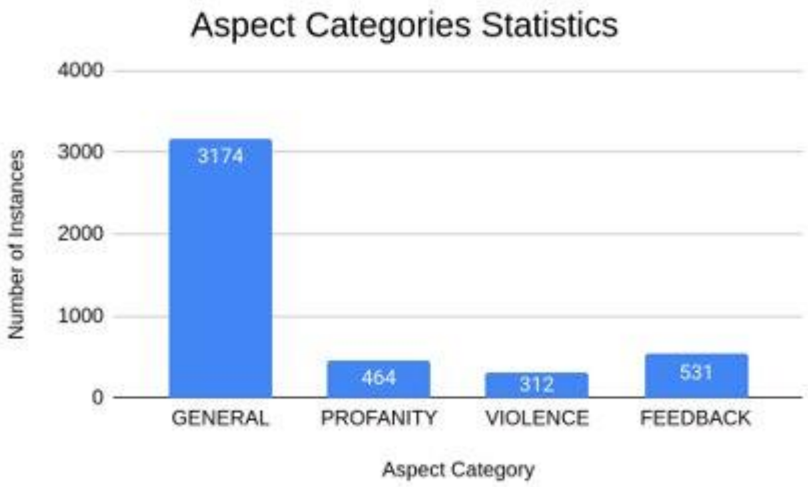
(Discrimination, abuse and hate speech)

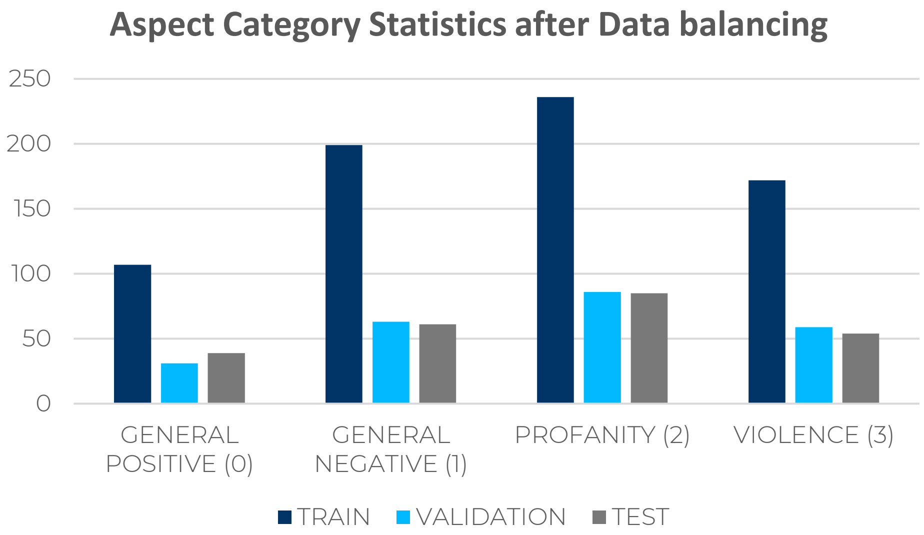
O. M. Singh, S. Timilsina, B. K. Bal and A. Joshi, "Aspect Based Abusive Sentiment Detection in Nepali Social Media Texts," 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), The Hague, Netherlands, 2020, pp. 301-308, doi: 10.1109/ASONAM49781.2020.9381292.

Preprocessing Nepali text

- Performed stemming for Nepalese text based on Nepali Stemmer.
  - Iterative separation of morphemes, stop words removal and spelling correction
  - Example: "राजुले भोजन खाएको छ।" (Raju ate lunch.)

    Stemmed Version: "राजु ले भोजन खाए को ।"
- Handled imbalanced classes by Random UnderSampling
  - Randomly remove samples from the majority class, with or without replacement.

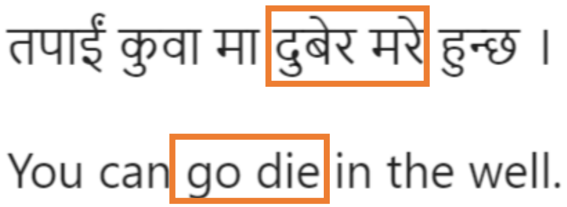




Task 1: Offensive Aspect Detection

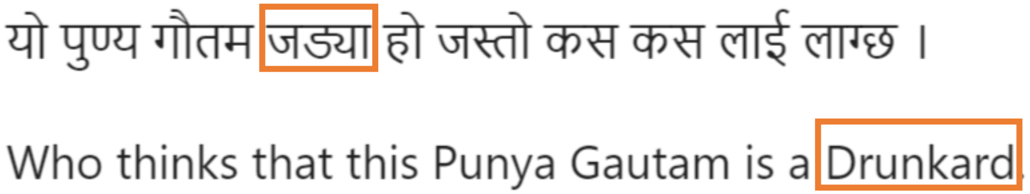
- Named-Entity Recognition task (Dataset: 4033 sentences)
- Sentence can have multiple aspects ['O', 'PER', 'ORG’, 'LOC’, 'MISC’, 'GENERAL', 'PROFANITY', 'VIOLENCE', 'FEEDBACK', ']

VIOLENCE


तपाईं कुवा मा दुबेर मरे हुन्छ ।

You can go die in the well.

GENERAL (Negative)


यो पुण्य गौतम जड्या हो जस्तो कस कस लाई लाग्छ ।

Who thinks that this Punya Gautam is a Drunkard.

Task 2: Sentiment Classification

- Sentiment Analysis task (4035 sentences)
- Classification Labels: [GENERAL POSITIVE, GENERAL NEGATIVE, PROFANITY, VIOLENCE]

सुशील जि धन्यवाद जन्ता को आवाज बि ले को मा । Thank you Sushil ji for being the voice of the people.	GENERAL POSITIVE (Positive Criticism)
येस्ता मानब अधिकार कर्मि को काम छइन … **।** This human rights activist is of no use.	GENERAL NEGATIVE (Derogatory remarks, insults)
यो खाते अधिवक्ता दिनेश त्रिपाठी को अावज सुन्न पनि मन पर्दैन । I don't even like to hear the voice of this slum dweller advocate Dinesh Tripathi.	PROFANITY (disrespectful and inappropriate)
भ्रश्ट्चारी हरुलाइ टुंडिखेल मा ल्याइ झुन्ड्याएर गोलि ठोक्नु पर्छ । The corrupt should be brought to Tundikhel and shot.	VIOLENCE (Discrimination, abuse and hate speech)

O. M. Singh, S. Timilsina, B. K. Bal and A. Joshi, "Aspect Based Abusive Sentiment Detection in Nepali Social Media Texts," 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), The Hague, Netherlands, 2020, pp. 301-308, doi: 10.1109/ASONAM49781.2020.9381292.

Evaluation

Epochs = 6	Named Entity Recognition			
Model	P	R	F1	Acc
xlm-roberta-large	43.92	51.76	52.22	87.48
bert-base-multilingual-cased	32.6	38.76	35.41	84.73
NepBERTa/NepBERTa	34.91	42.34	38.26	84.19
Sakonii/deberta-base-nepali	37.95	42.58	40.14	86.6
Sakonii/distilbert-base-nepali	34.59	39.23	36.77	85.67

Epochs = 6	Sentiment Classification			
Model	P	R	F1	Acc
xlm-roberta-large	68.99	71.1	68.97	71.1
bert-base-multilingual-cased	67.92	69.41	66.53	69.41
rajan/NepaliBERT	68.0	69.26	66.25	69.26
Sakonii/distilbert-base-nepali	72.05	72.24	71.8	72.24
Sakonii/deberta-base-nepali	69.67	71.67	67.56	71.67

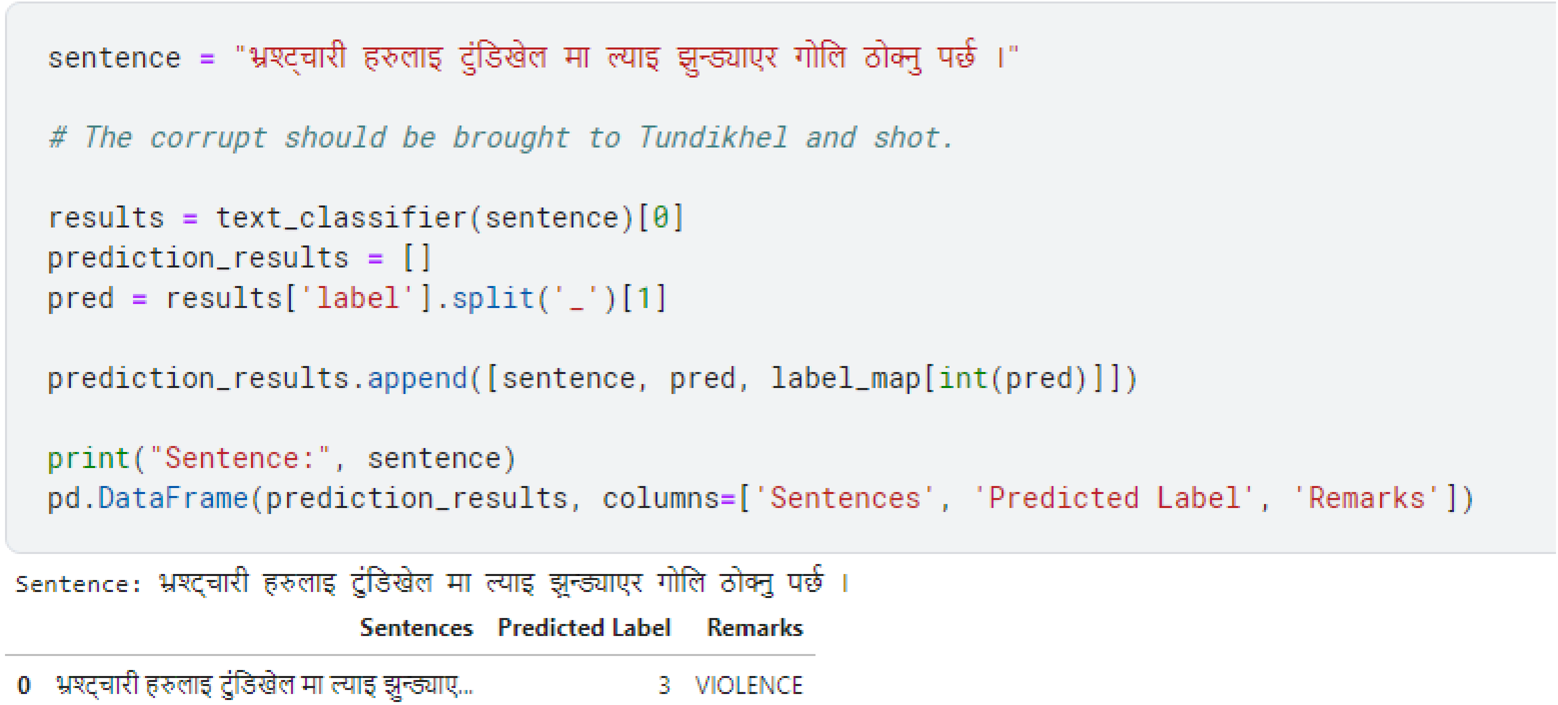
Optimization Techniques on Best performing

- Early Stopping based on F1-score
  - Validate every 100 steps
  - Patience = 2
  - Saves best model

	Epochs = 6	Test Metrics				Improved Test metrics			
	Best performing Model	P	R	F1	Acc	P	R	F1	Acc
Offensive Entity recognition	xlm-roberta-large	56.95	61.98	59.36	82.85	58.32	63.8	60.60	83.25
Offensive Sentiment Classification	Sakonii/distilbert-base-nepali	72.05	72.24	71.8	72.24	73.51	74.12	72.3	74.52

Inference on Best Performing models

Perform Offensive Sentiment Classification



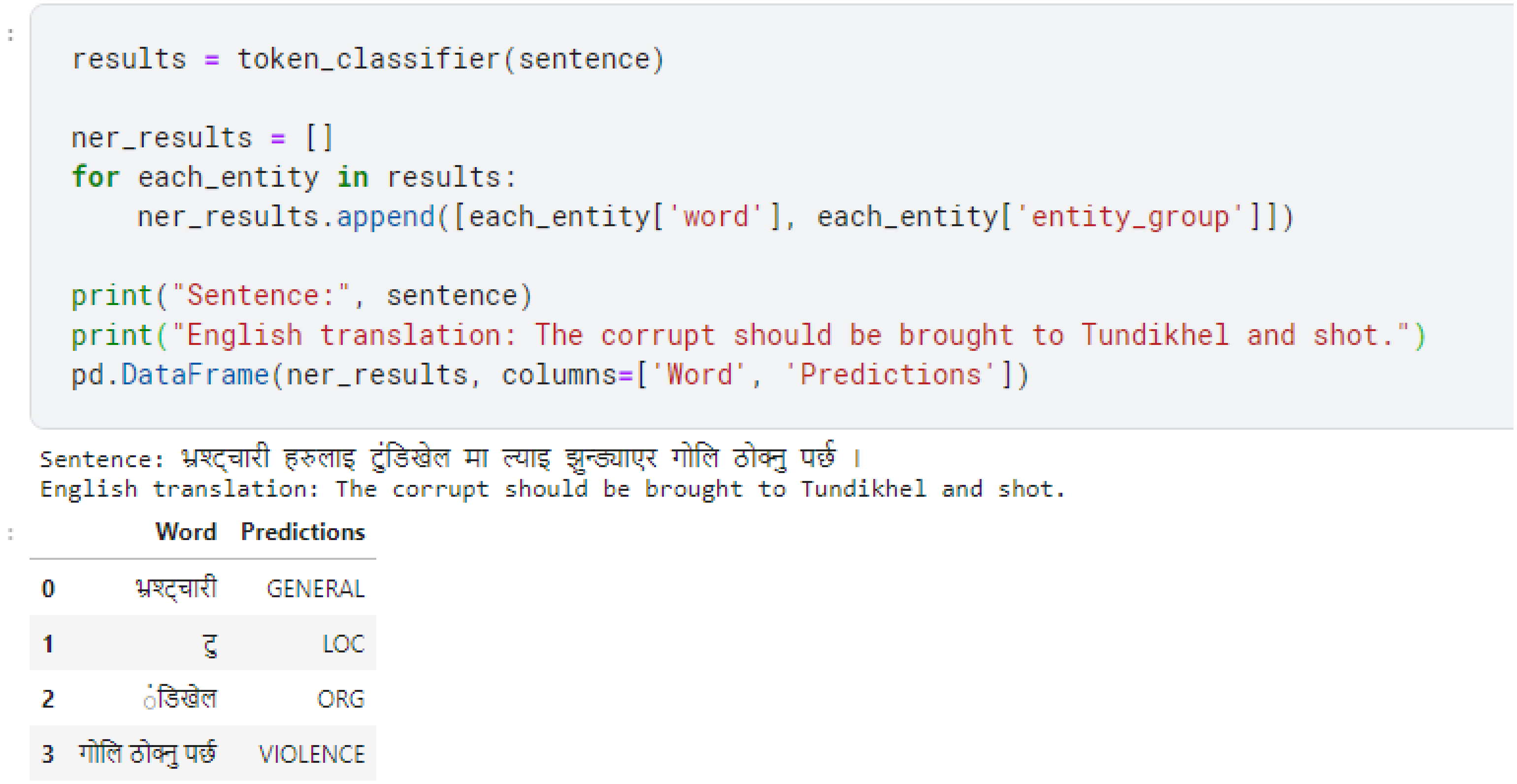
```python
sentence = "भ्रश्ट्चारी हरुलाइ टुंडिखेल मा ल्याइ झुन्ड्याएर गोलि ठोक्नु पर्छ ।"

# The corrupt should be brought to Tundikhel and shot.

results = text_classifier(sentence)[0]
prediction_results = []
pred = results['label'].split('_')[1]

prediction_results.append([sentence, pred, label_map[int(pred)]])

print("Sentence:", sentence)
pd.DataFrame(prediction_results, columns=['Sentences', 'Predicted Label', 'Remarks'])
```

```
Sentence: भ्रश्ट्चारी हरुलाइ टुंडिखेल मा ल्याइ झुन्ड्याएर गोलि ठोक्नु पर्छ ।
```

	Sentences	Predicted Label	Remarks
0	भ्रश्ट्चारी हरुलाइ टुंडिखेल मा ल्याइ झुन्ड्याए...	3	VIOLENCE

Inference on Best Performing models

Perform Offensive Entity Recognition

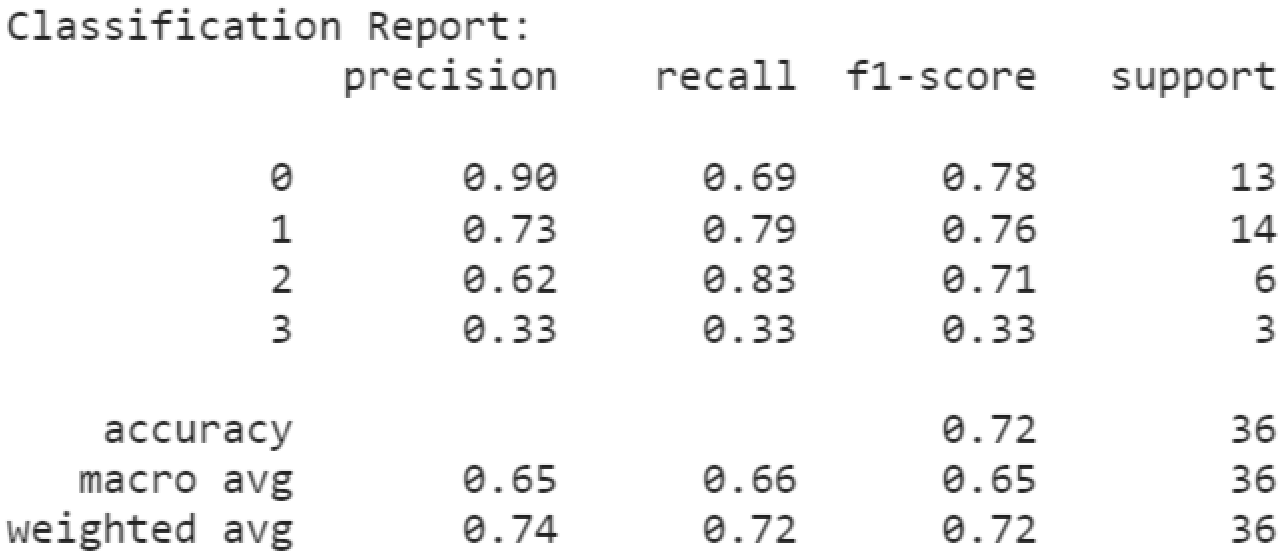


```
results = token_classifier(sentence)

ner_results = []
for each_entity in results:
    ner_results.append([each_entity['word'], each_entity['entity_group']])

print("Sentence:", sentence)
print("English translation: The corrupt should be brought to Tundikhel and shot.")
pd.DataFrame(ner_results, columns=['Word', 'Predictions'])
```

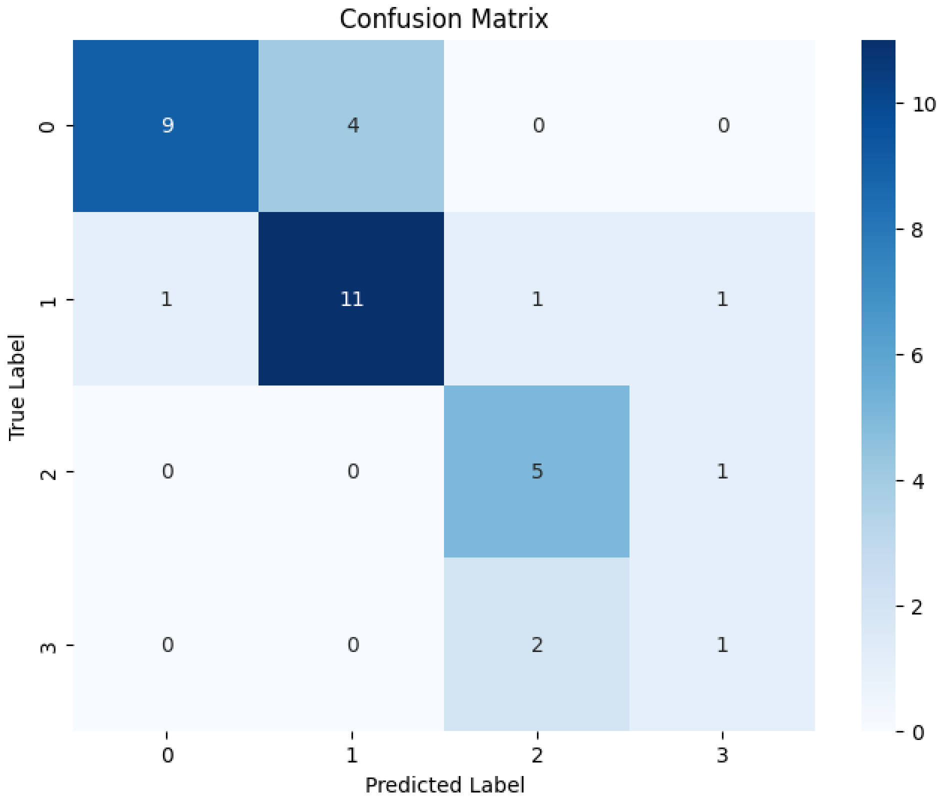
```
Sentence: भ्रश्ट्चारी हरुलाइ टुंडिखेल मा ल्याइ झुन्ड्याएर गोलि ठोक्नु पर्छ ।
English translation: The corrupt should be brought to Tundikhel and shot.
```

	Word	Predictions
0	भ्रश्ट्चारी	GENERAL
1	टु	LOC
2	ंडिखेल	ORG
3	गोलि ठोक्नु पर्छ	VIOLENCE

Performance on curated out-of-domain test set



```
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.69      0.78        13
           1       0.73      0.79      0.76        14
           2       0.62      0.83      0.71         6
           3       0.33      0.33      0.33         3

    accuracy                           0.72        36
   macro avg       0.65      0.66      0.65        36
weighted avg       0.74      0.72      0.72        36
```



Future Improvements

- In-depth study on misclassified data for model improvement
- Handling Code-Switching (using English words in between)
- Implementing by taking Romanized Nepali text into account

Thank you!

For questions, email us at:

msuwa1@unh.newhaven.edu
sbhus1@unh.newhaven.edu
dregm1@unh.newhaven.edu