नेपाली Language Processing

Team: NonLinear

Nabin Da Shrestha(076BCT037)
Nirajan Bekoju (076BCT039)
Nishant Luitel (076BCT041)

C&P Course

नेपाली Language Processing

UNIT 1. Introduction

1.1. Abstract
1.2. Problem Statement
1.3. Objectives
1.4. Frontend Homepage

UNIT 2. Literature Review

UNIT 3. Word Embeddings

3.1. Methodology
3.2. Frontend Snapshots

UNIT 4. Sentimental Classification

4.1. Methodology
4.2. Frontend Snapshots

नेपाली Language Processing

UNIT 5. Probabilistic Language model

4.1. Methodology

4.2. Frontend Snapshots

UNIT 6. Transformer based Language Model

5.1. Methodology

5.2. Frontend Snapshots

UNIT 7. Spelling Correction

6.1. Methodology

6.2. Frontend Snapshots

UNIT 8. Application

Abstract

- Language modeling (LM) is the use of various statistical and probabilistic techniques to determine the probability of a given sequence of words occurring in a sentence.

- Language models analyze bodies of text data to provide a basis for their word predictions.

- They are used in natural language processing (NLP) applications, particularly ones that generate text as an output. Some of these applications include , machine translation and question answering.

Problem Statements

- Nepali Language is rich in vocabulary and it is difficult to choose the best possible vocab.

- Spelling correction for nepali language available today are based on dictionary rather than contextual meaning of the sentence.

- No proper development of various NLP tasks like text generation, text summarization, image captioning, text to speech, etc. due to lack of reliable nepali language model.

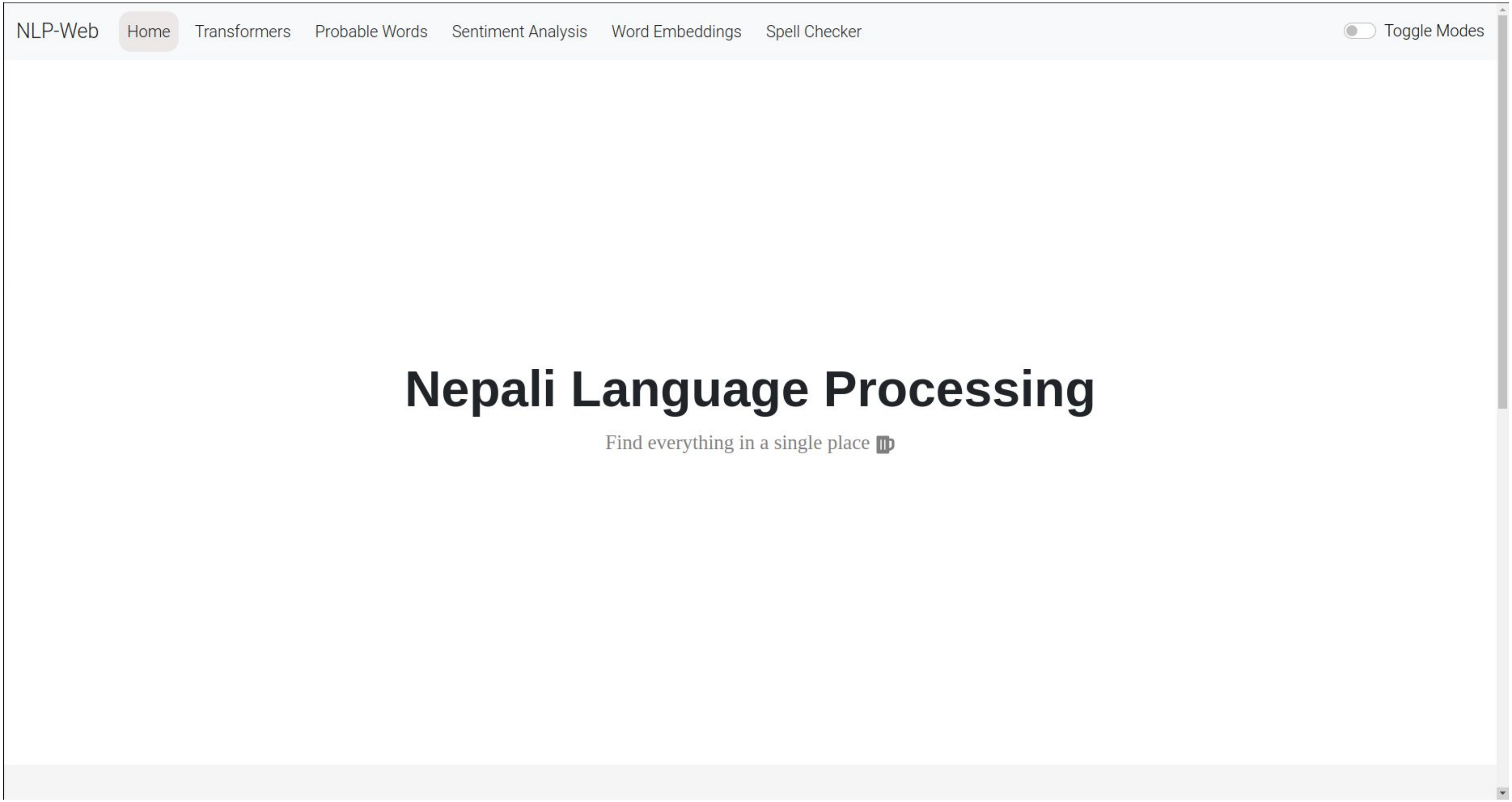
Objectives

- To develop nepali language model for text generation.

- Use the nepali language model to develop the spelling correction based on contextual meaning.

UNIT 1.

1.4. Home Page Frontend



Nepali Language Processing

Find everything in a single place

UNIT 2.

Literature Review

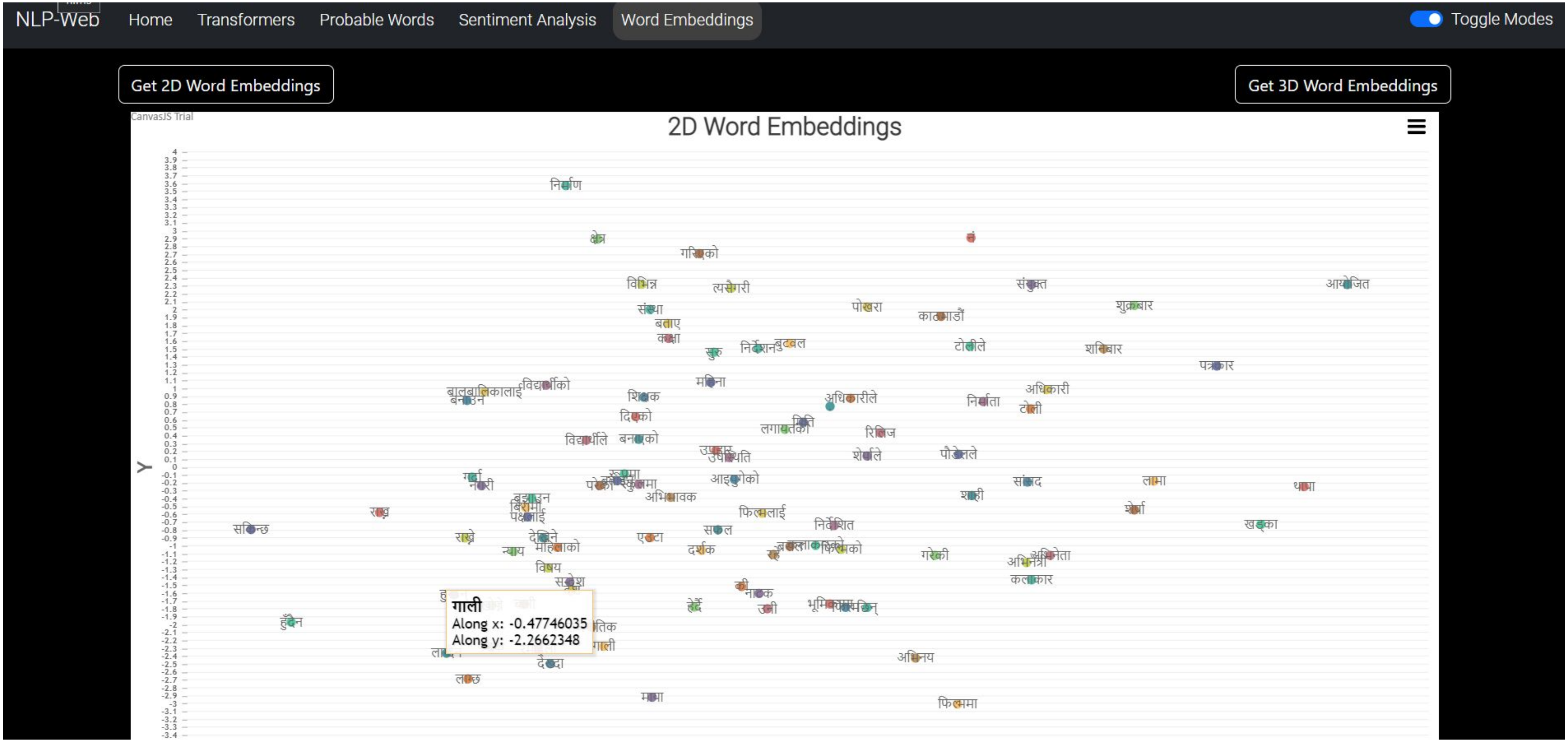
- A. Vaswani et al.'s "Attention is all you need" (2017) in Adv. Neural Inf. Process. Syst. (vol. 30).
- A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., “Language models are unsupervised multitask learners,” OpenAI blog, vol. 1, no. 8, p. 9, 2019.
- Y. Bengio, R. Ducharme, and P. Vincent, “A neural probabilistic language model,” Advances in neural information processing systems, vol. 13, 2000.
- T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” Advances in neural information processing systems, vol. 26, 2013.
- S. Timilsina, M. Gautam, and B. Bhattarai, “Nepberta: Nepali language model trained in a large corpus,” in Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, pp. 273–284, 2022.
- Whitelaw, C., B. Hutchinson, G. Y. Chung, and G. Ellis. 2009. Using the web for language independent spellchecking and autocorrection.
- Wilcox-O’Hearn, L. A. 2014. Detection is the central problem in real-word spelling correction.
- Norvig, P. 2009. Natural language corpus data. In T. Segaran and J. Hammerbacher, editors, Beautiful data: the stories behind elegant data solutions. O’Reilly

Methodology

- Use of gensim word2vec model for word embeddings of dimension 300
- PCA for converting 300-dimension word embeddings vector into 2-dimensional and 3-dimensional vectors for visualization.

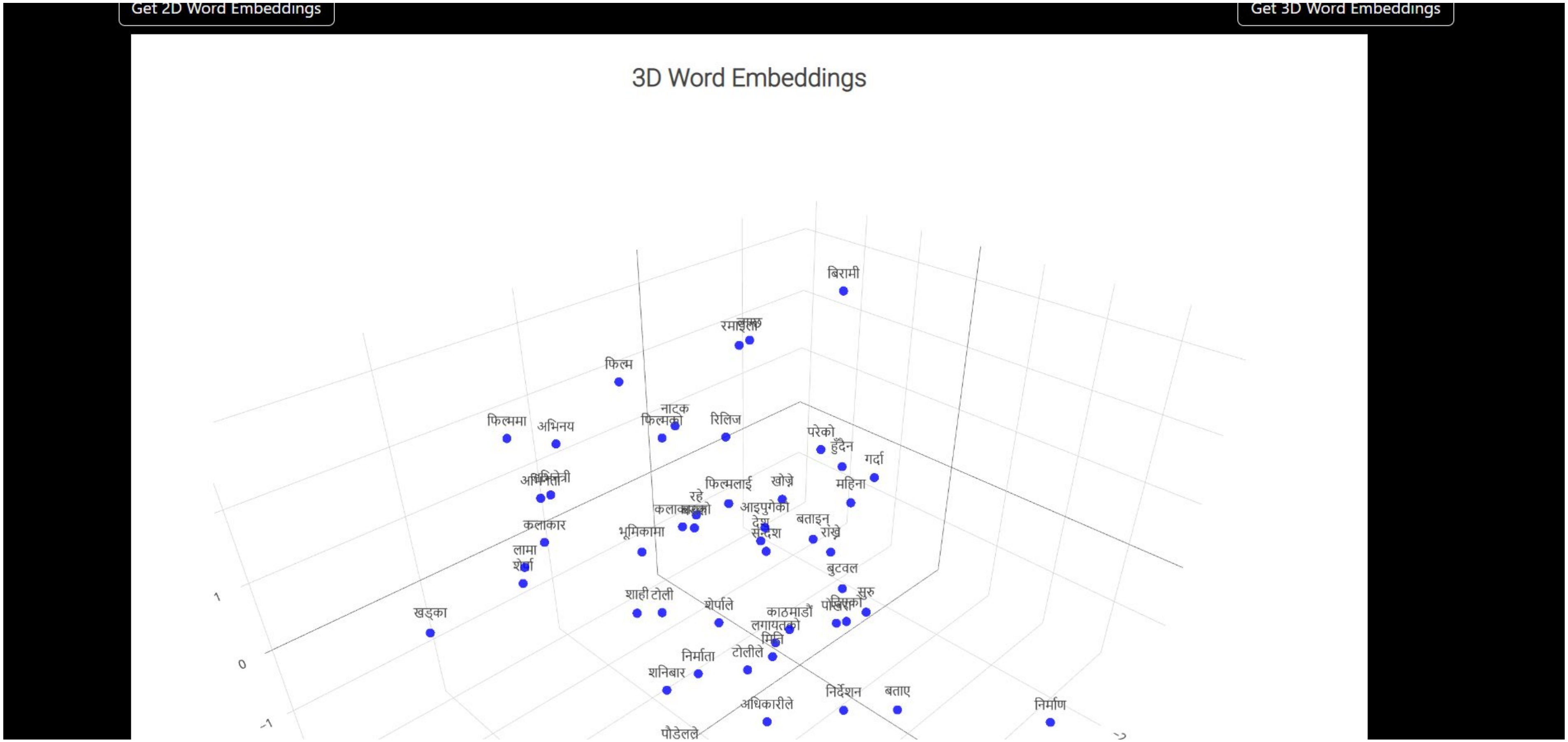
UNIT 3.

3.2. 2d Word Embeddings



गाली
Along x: -0.47746035
Along y: -2.2662348

UNIT 3.
3.2. 3d Word Embeddings



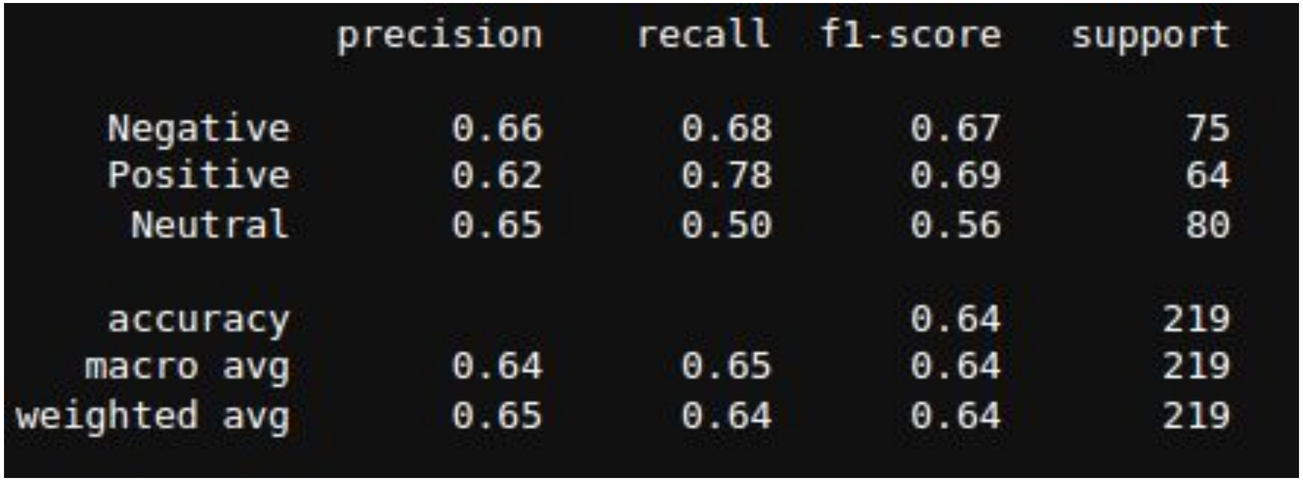
Methodology

- Data obtained from kaggle and hugging face
- Data exploration
- Data cleaning and preprocessing
- Train test split
- One Hot encoding and tokenization
- Model Development using Embedding, LSTMs and Dense layer
- Model Evaluation on Test data
- Loading and Saving Model

UNIT 4.

4.2. Sentiment Analysis

LSTMs Model Classification Report

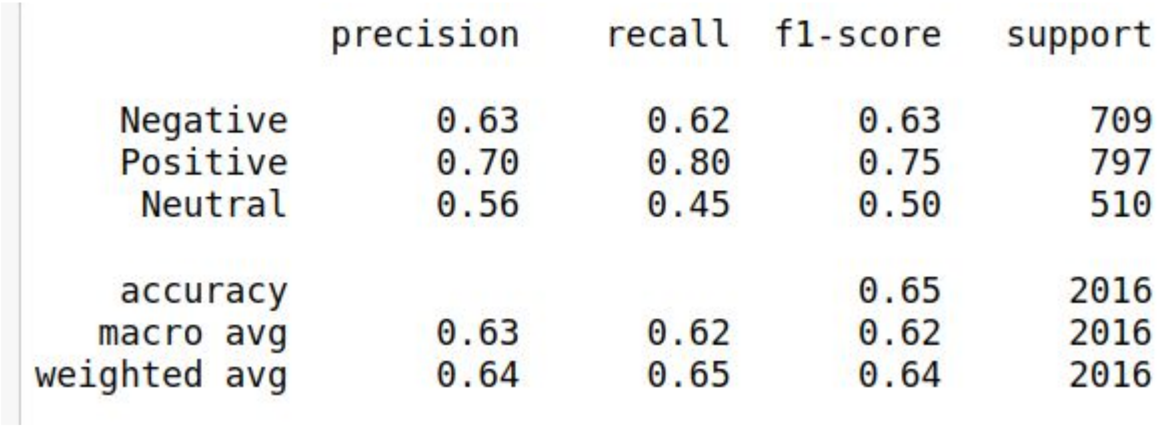


```
              precision    recall  f1-score   support

    Negative       0.66      0.68      0.67        75
    Positive       0.62      0.78      0.69        64
     Neutral       0.65      0.50      0.56        80

    accuracy                           0.64       219
   macro avg       0.64      0.65      0.64       219
weighted avg       0.65      0.64      0.64       219
```

UNIT 4.

4.2. Sentiment Analysis

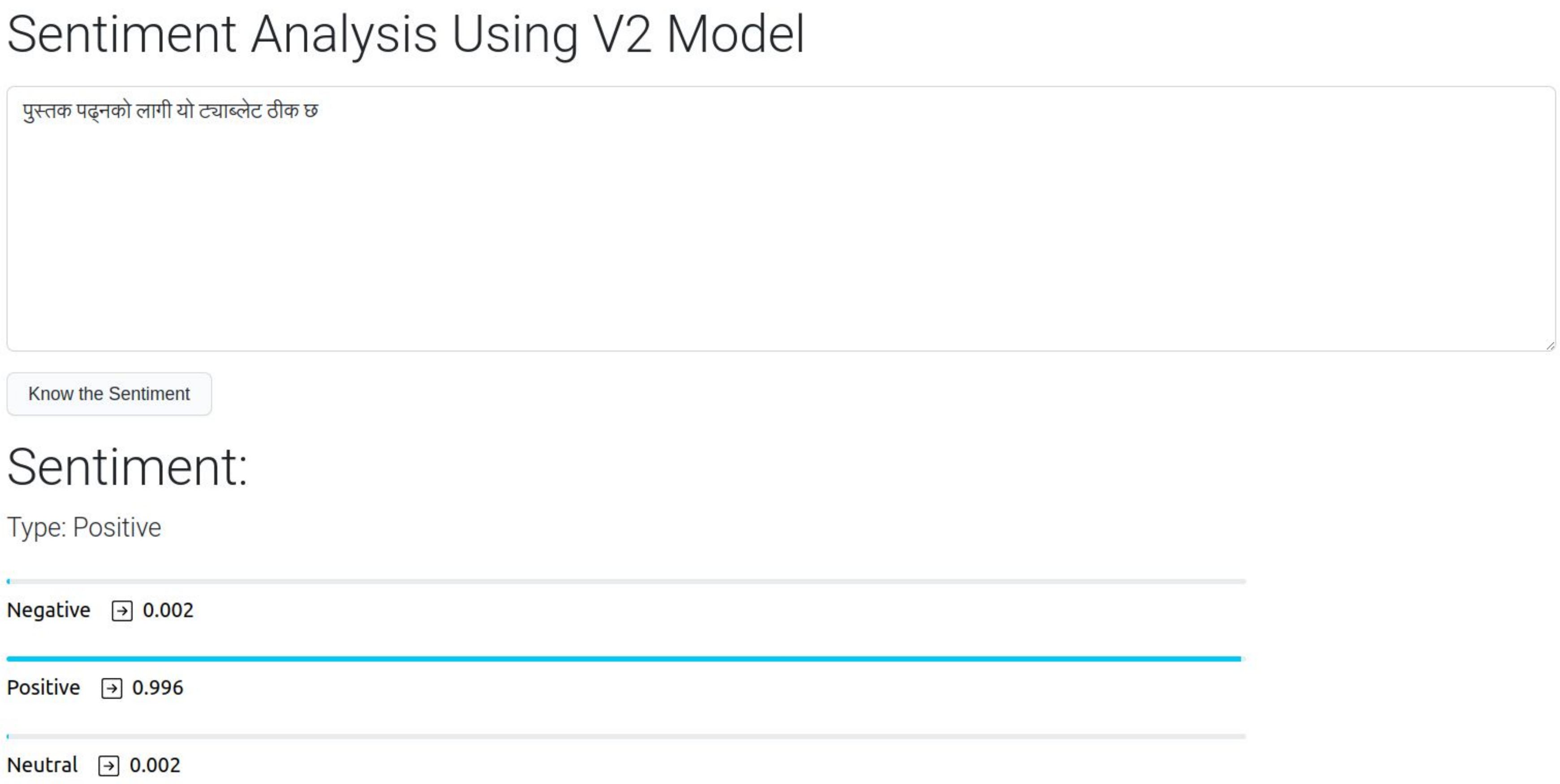
Bert Model Classification Report



```
              precision    recall  f1-score   support

    Negative       0.63      0.62      0.63       709
    Positive       0.70      0.80      0.75       797
     Neutral       0.56      0.45      0.50       510

    accuracy                           0.65      2016
   macro avg       0.63      0.62      0.62      2016
weighted avg       0.64      0.65      0.64      2016
```

UNIT 4.

4.2. Sentiment Analysis : Positive



Sentiment Analysis Using V2 Model

पुस्तक पढ्नको लागी यो ट्याब्लेट ठीक छ

Sentiment:

Type: Positive

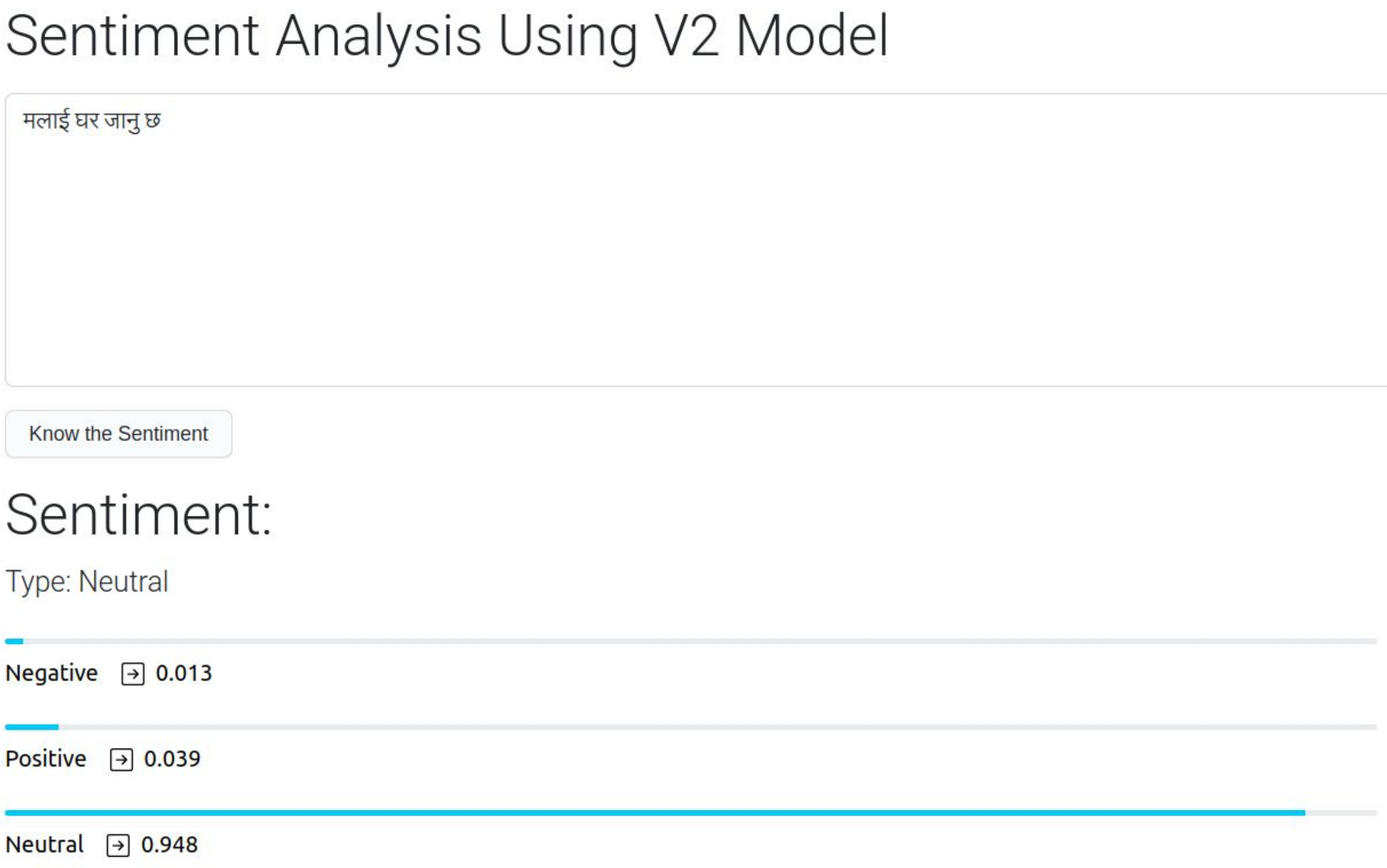
Negative ⇥ 0.002

Positive ⇥ 0.996

Neutral ⇥ 0.002

UNIT 4.

4.2. Sentiment Analysis : Neutral



Sentiment Analysis Using V2 Model

मलाई घर जानु छ

Sentiment:

Type: Neutral

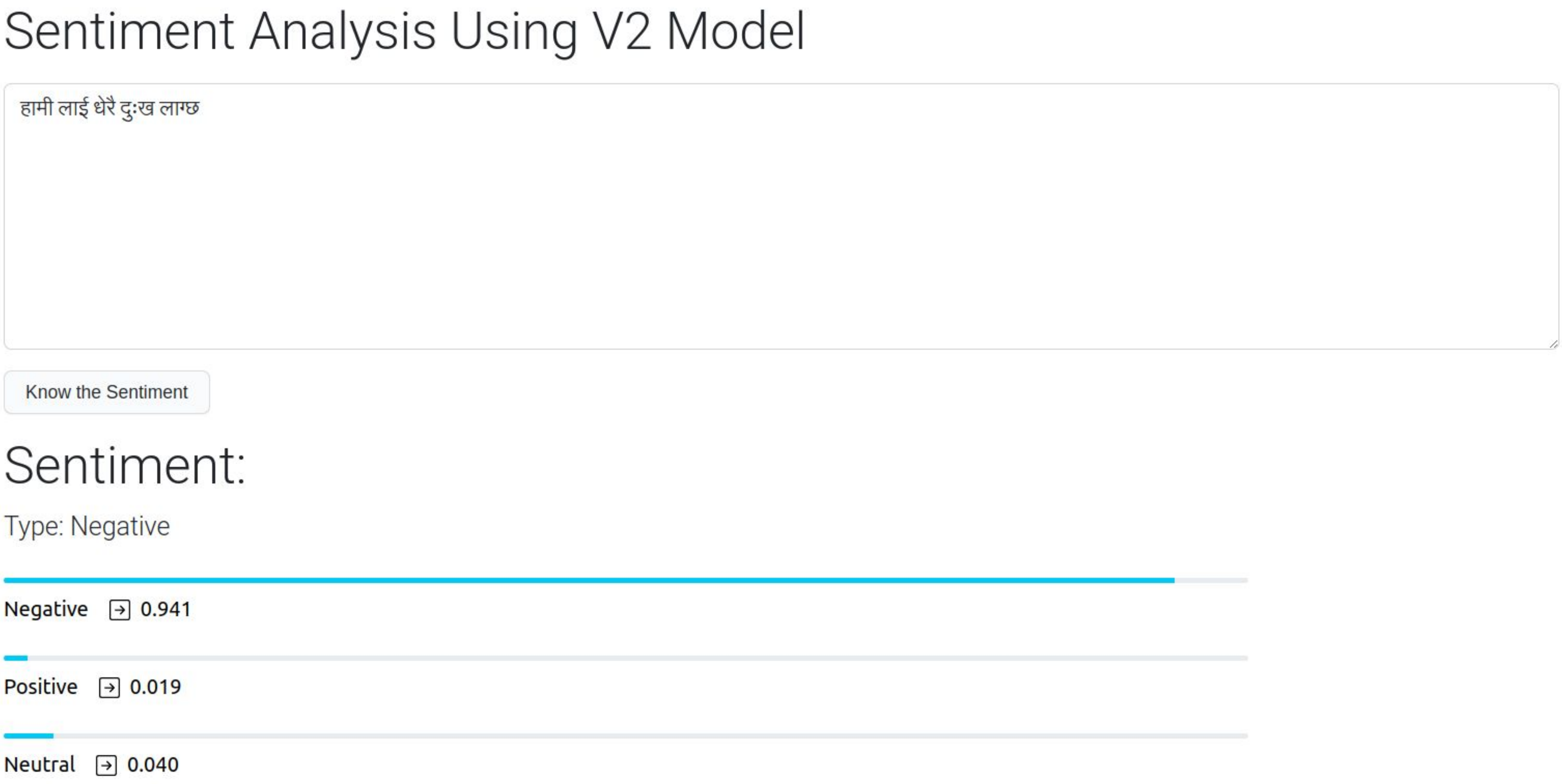
Negative → 0.013

Positive → 0.039

Neutral → 0.948

UNIT 4.

4.2. Sentiment Analysis : Negative



Sentiment Analysis Using V2 Model

हामी लाई धेरै दुःख लाग्छ

Sentiment:

Type: Negative

Negative ⇥ 0.941

Positive ⇥ 0.019

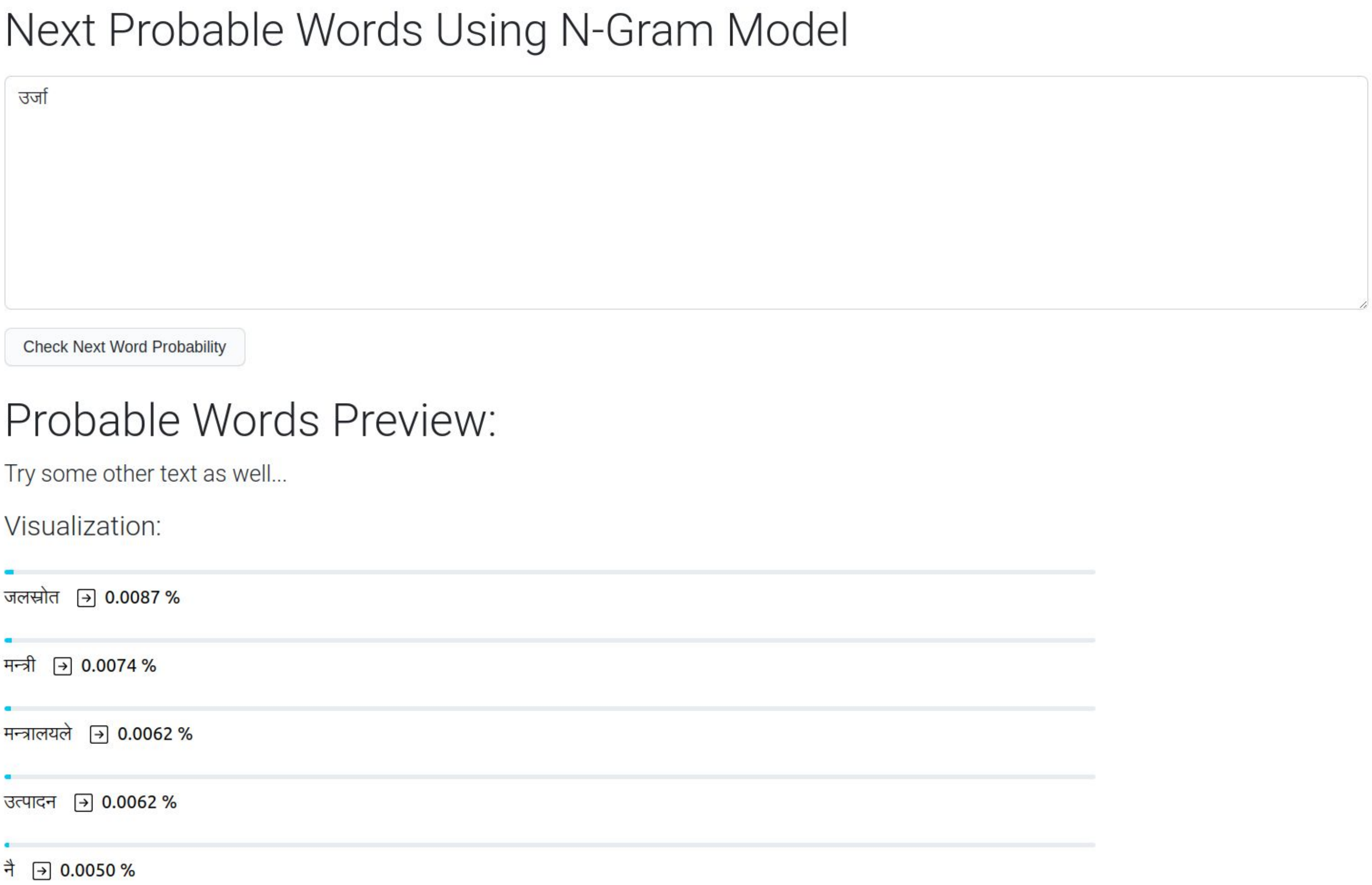
Neutral ⇥ 0.040

Methodology

- Generate a vocabulary from the training data set
  - split text into list of sentences
  - remove all non-devanagari letters
  - remove numbers from the corpus
  - create word tokenization
  - create vocabulary from the tokenized word using minimum frequency constraints
  - create <UNK> token for Out Of Vocabulary words with the help of minimum frequency
- Generate n_gram_counts_list from the training data set
  - count n-gram with the help of tokenized word sequences
- estimate probability with the help of word given, n and (n+1) gram counts list and vocabulary

UNIT 5.

5.2. Probabilistic Language Model



Next Probable Words Using N-Gram Model

उर्जा

Probable Words Preview:

Try some other text as well...

Visualization:

जलस्रोत → 0.0087 %

मन्त्री → 0.0074 %

मन्त्रालयले → 0.0062 %

उत्पादन → 0.0062 %

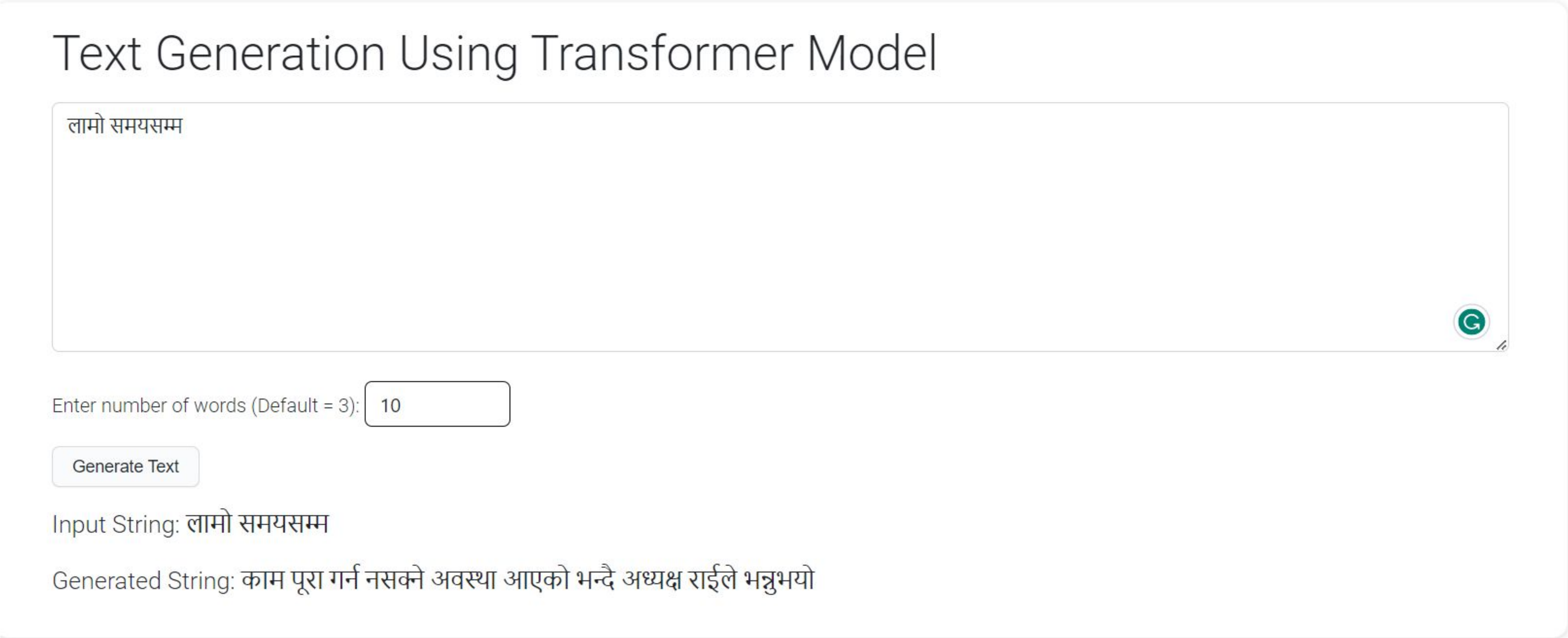
नै → 0.0050 %

Methodology

- Generate a vocabulary from the training data set
  - split text into list of sentences
  - remove all non-devanagari letters
  - remove numbers from the corpus
  - create word tokenization
  - create vocabulary from the tokenized word using minimum frequency constraints
  - create <UNK> token for Out Of Vocabulary words with the help of minimum frequency
- Define Transformer Architecture and Train based on corpus of data.
- Estimate probability through inference and choose one from top-k choices using categorical distribution.

UNIT 6.

6.2. Transformer based language model



Text Generation Using Transformer Model

लामो समयसम्म

Enter number of words (Default = 3): 10

Input String: लामो समयसम्म

Generated String: काम पूरा गर्न नसक्ने अवस्था आएको भन्दै अध्यक्ष राईले भन्नुभयो

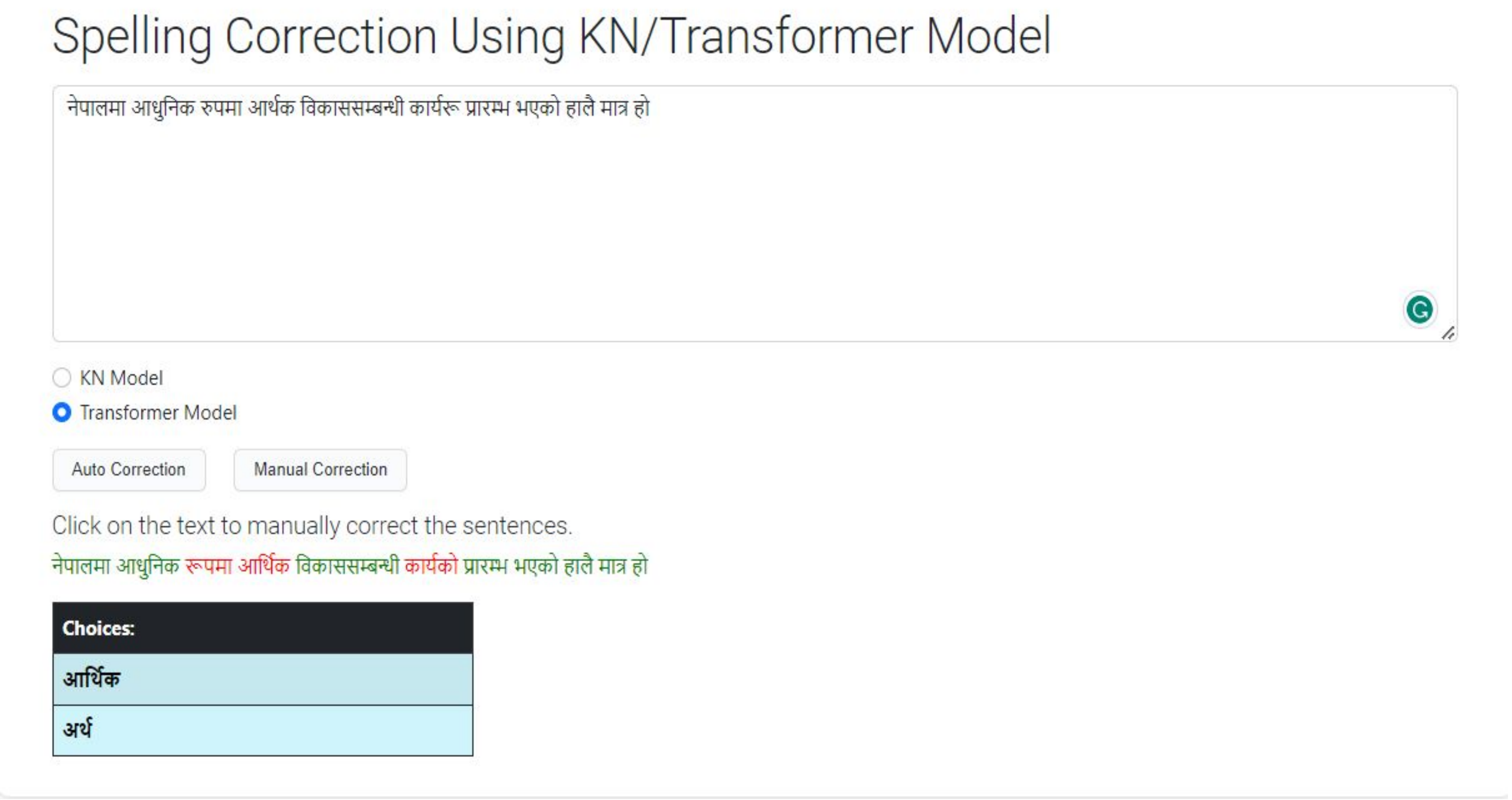
Methodology

- Used Noisy Channel Model for spelling correction.

$$\hat{w} = \underset{w \in C}{\operatorname{argmax}} \overbrace{P(x|w)}^{\text{channel model}} \overbrace{P(w)}^{\text{prior}}$$

- Train Channel model based on Brill and Moore model using unsupervised data from corpus.
- Use Probabilistic and transformer language model to determine prior distribution
- Find candidate sentences using edit distance and based on above equation.
- Choose the word that maximizes channel model and prior.

UNIT 7.
7.2. Manual Spelling Correction using Transformer Model



Spelling Correction Using KN/Transformer Model

नेपालमा आधुनिक रुपमा आर्थक विकाससम्बन्धी कार्यरू प्रारम्भ भएको हालै मात्र हो

○ KN Model
● Transformer Model

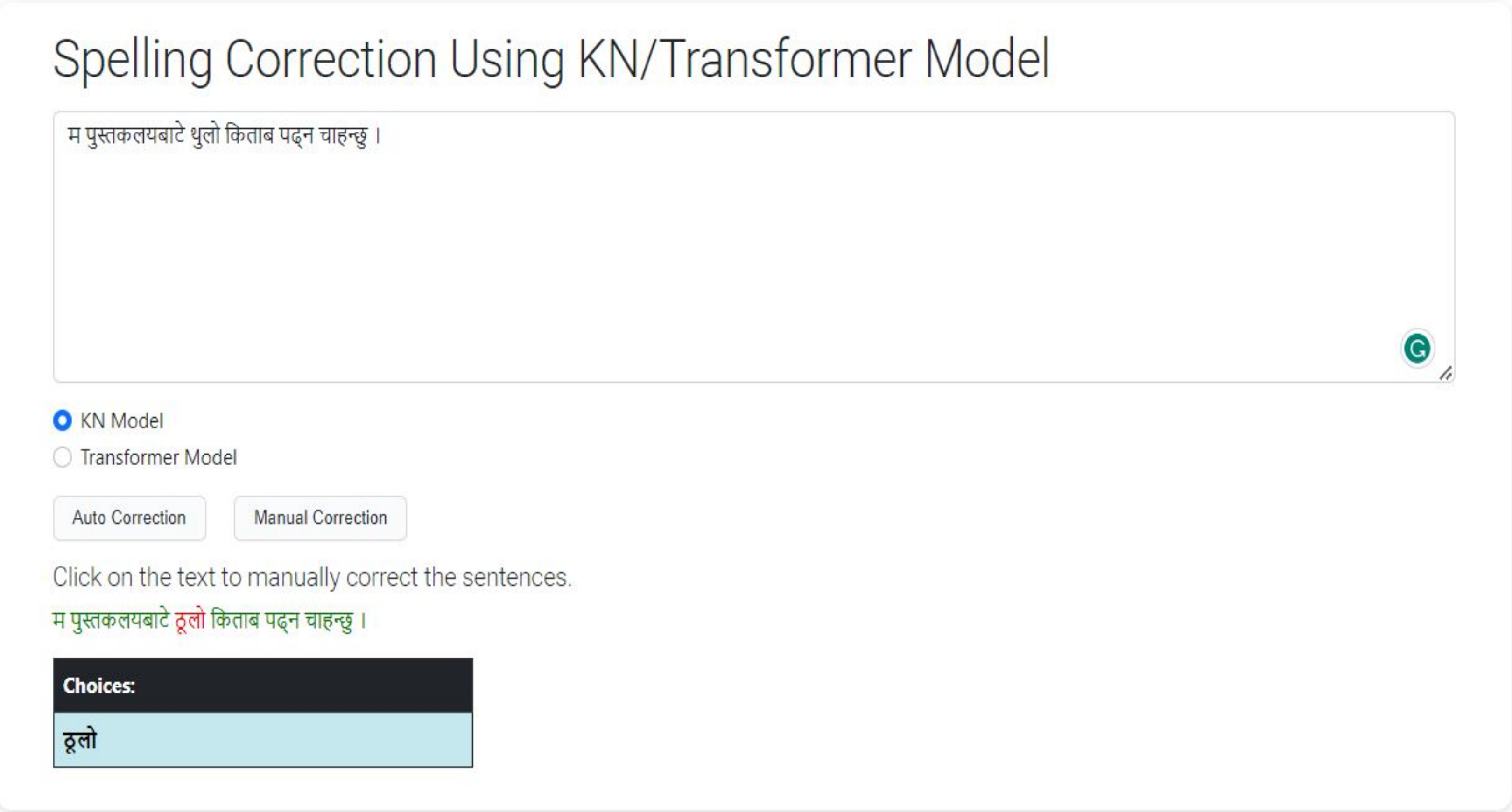
Click on the text to manually correct the sentences.

नेपालमा आधुनिक रूपमा आर्थिक विकाससम्बन्धी कार्यको प्रारम्भ भएको हालै मात्र हो

Choices:
आर्थिक
अर्थ

UNIT 7.

7.2. Manual spelling correction using KN Model



Spelling Correction Using KN/Transformer Model

म पुस्तकलयबाटे थुलो किताब पढ्न चाहन्छु ।

● KN Model

○ Transformer Model

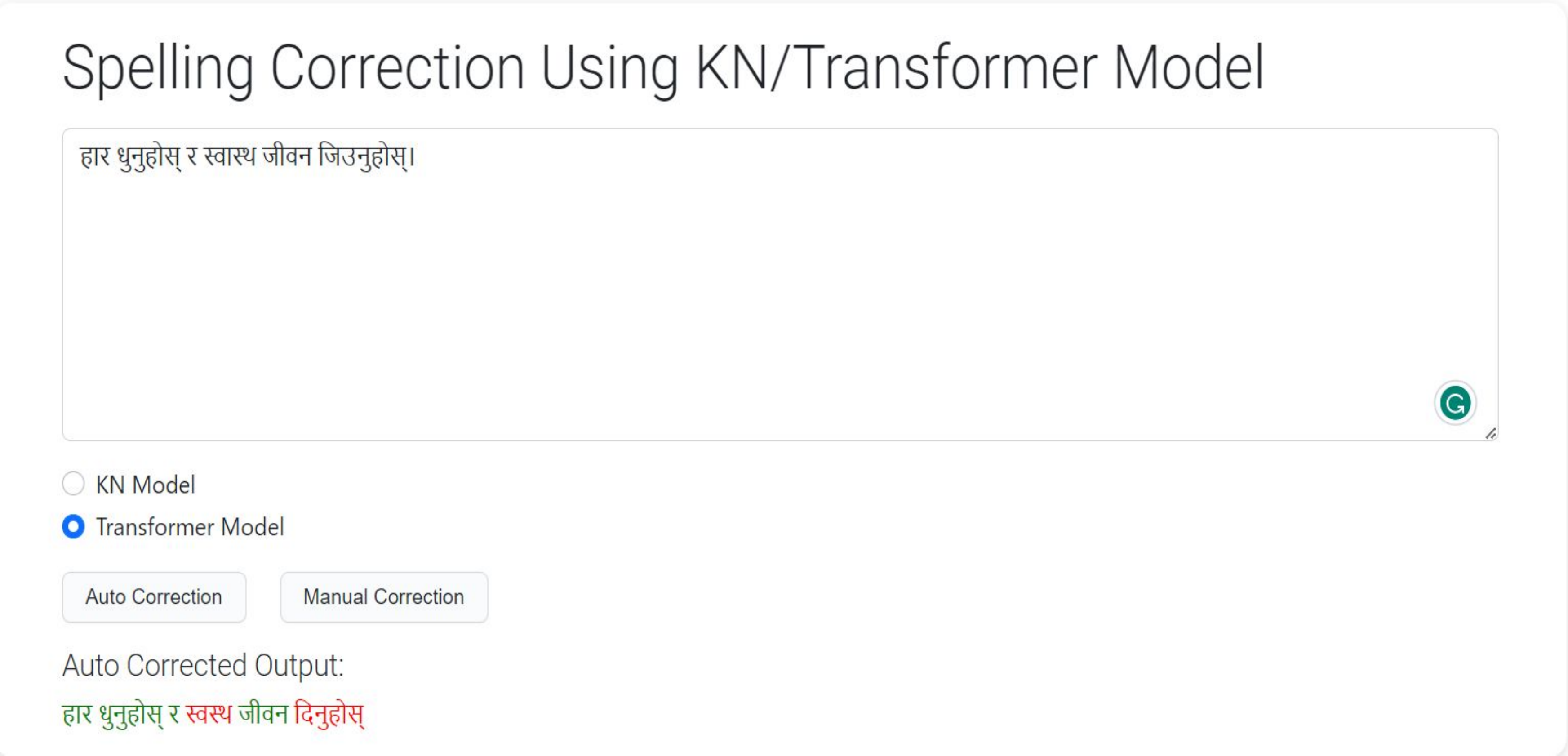
Click on the text to manually correct the sentences.

म पुस्तकलयबाटे ठूलो किताब पढ्न चाहन्छु ।

Choices:
ठूलो

UNIT 7.

7.2. Auto spelling correction using Transformer Model



Spelling Correction Using KN/Transformer Model

हार धुनुहोस् र स्वास्थ जीवन जिउनुहोस्।

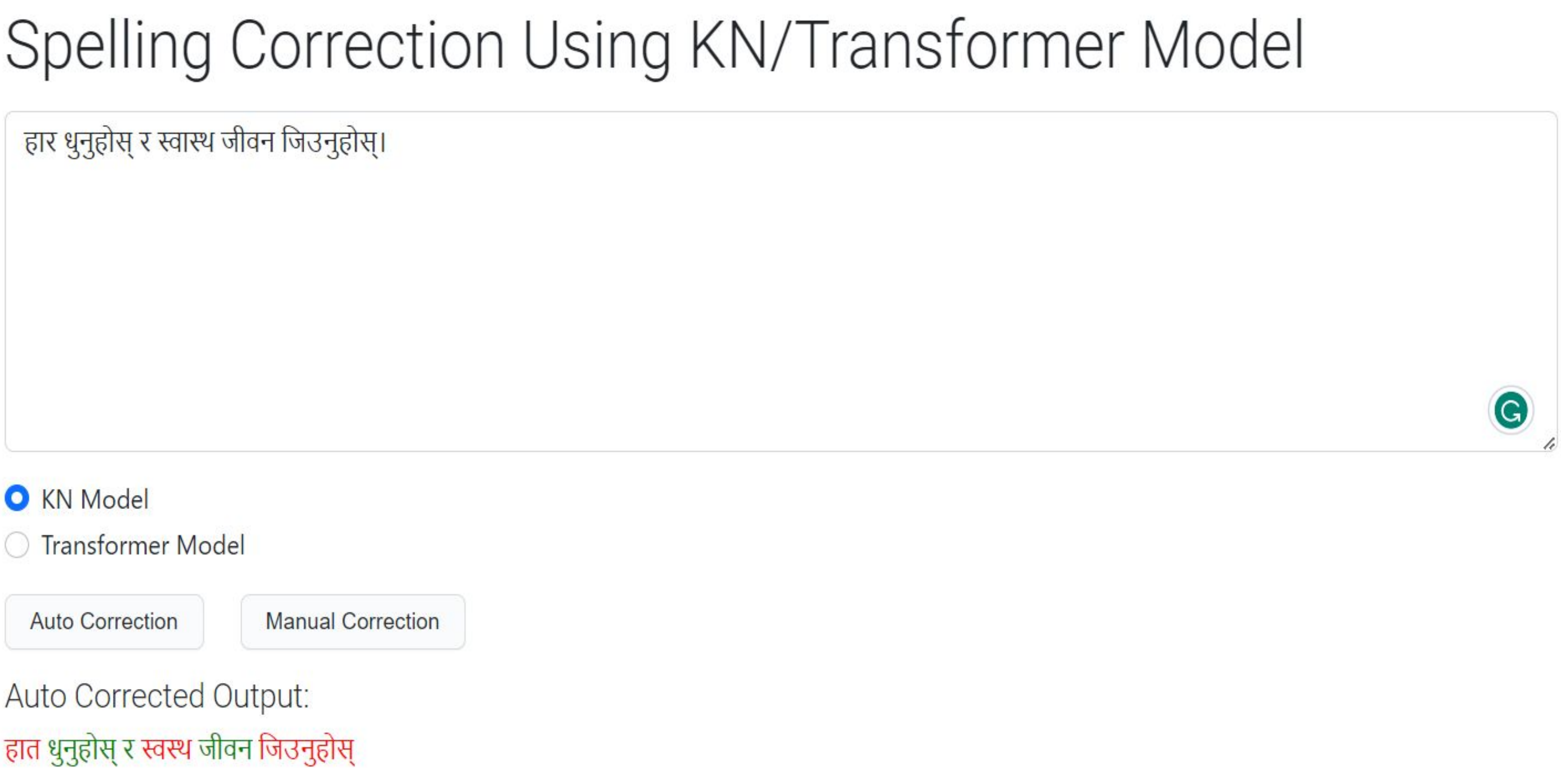
○ KN Model

● Transformer Model

Auto Corrected Output:

हार धुनुहोस् र स्वस्थ जीवन दिनुहोस्

7.2. Auto Spelling correction KN Model



Spelling Correction Using KN/Transformer Model

हार धुनुहोस् र स्वास्थ जीवन जिउनुहोस्।

● KN Model

○ Transformer Model

Auto Corrected Output:

हात धुनुहोस् र स्वस्थ जीवन जिउनुहोस्

UNIT 7.

Application

- Further NLP Tasks : text summarization, speech to text model
- Drafting emails and report
- Language Learning : text generation and spelling correction for improving writing skills and word embeddings visualization to understand the antonyms and synonyms of words

SAMSUNG

Together for Tomorrow!

Enabling People

Education for Future Generations