Marathi NLP

Raviraj Joshi
L3Cube Pune

An L3Cube mahaNLP Initiative

https://github.com/l3cube-pune/MarathiNLP

https://arxiv.org/pdf/2205.14728.pdf

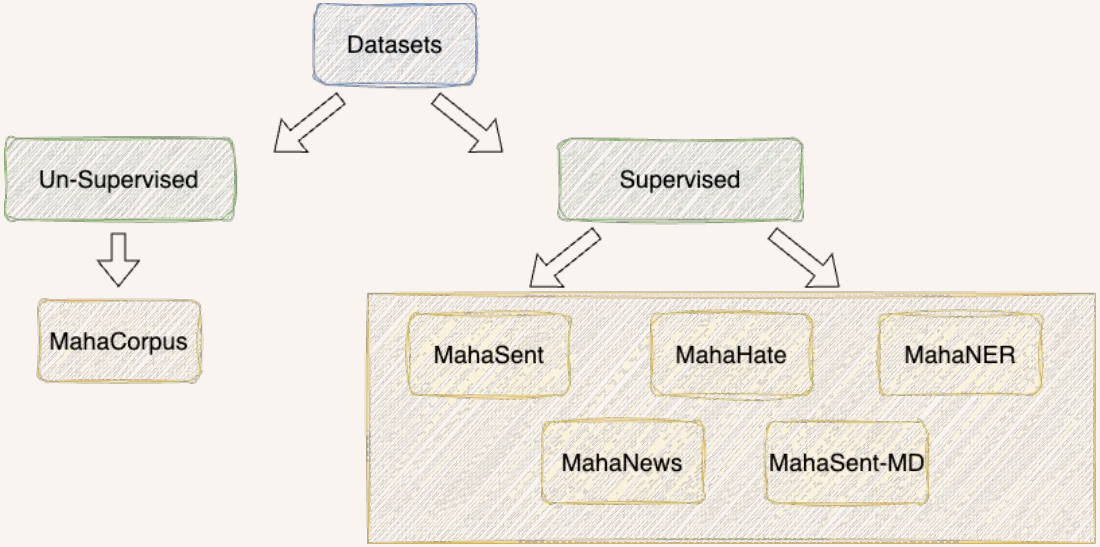
Marathi NLP

- Official language of State of Maharashtra
- Secondary language of Daman and Diu, Dadra and Nagar Haveli, and Goa
- Spoken by around 83 millions speakers in India – Ranks 3rd in India after Hindi and Bengali
- Spoken by around 99 millions people in world – Ranks 15th in World

Challenges

- Multiple variants for same words
  - विद्यार्थी
  - विदयार्थी
- Ambiguous POS tags
  - त्याला एका मोठ्या संस्थेत प्रवेश मिळाला. (NOUN)
  - हे रान इतके दाट आहे की तिथे प्रवेश करणे सोपे नाही. (VERB)
- Word Order
  - I will be going to Pune (SVO)
  - मी पुण्याला जाणार आहे (SOV)
- Multi-words
  - अतीआत्मविश्वास
  - आत्मविश्वासयुक्त
- Morphologically rich (Out of vocabulary issues)
  - आरोप, आरोपः, आरोपा, आरोपही, आरोपला, आरोपात, आरोपही, etc

Marathi Datasets



Maha Corpus - Unsupervised

Dataset	#sentences	#tokens
MahaCorpus (News)	212 M	17.6 M
MahaCorpus (non News)	76.4M	7.2M
MahaCorpus (Full)	289M	24.8M
Full Marathi Corpus	752M	57.2M

https://arxiv.org/pdf/2202.01159.pdf

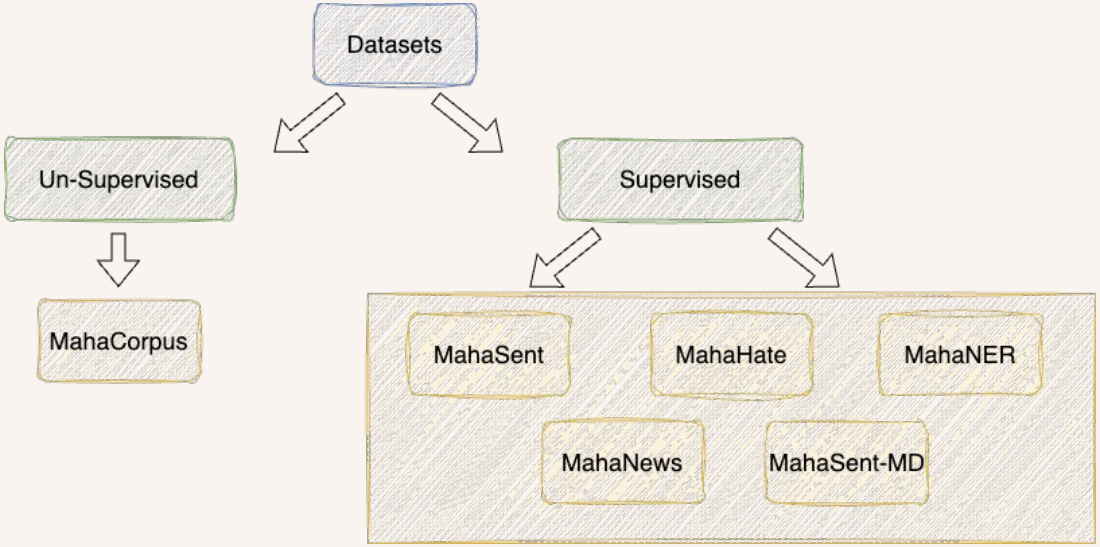
Marathi Transformers

Marathi Transformers

MahaBERT, MahaAlBERT, MahaRoBERTa	These are monolingual BERT variants trained on large Devanagari Marathi Corpus
DevBERT, DevAlBERT, DevRoBERTa	These are bi-lingualBERT variants trained on large Devanagari Marathi + Hindi Corpus
MahaTweetsBERT	MahaBERT model further finetuned on Twitter Corpus
MahaSBERT	Marathi Sentence BERT model
MahaGPT	Marathi GPT model
MahaFT	Marathi Fast Text word embeddings

https://arxiv.org/pdf/2202.01159.pdf
https://arxiv.org/pdf/2211.11418.pdf

Datasets

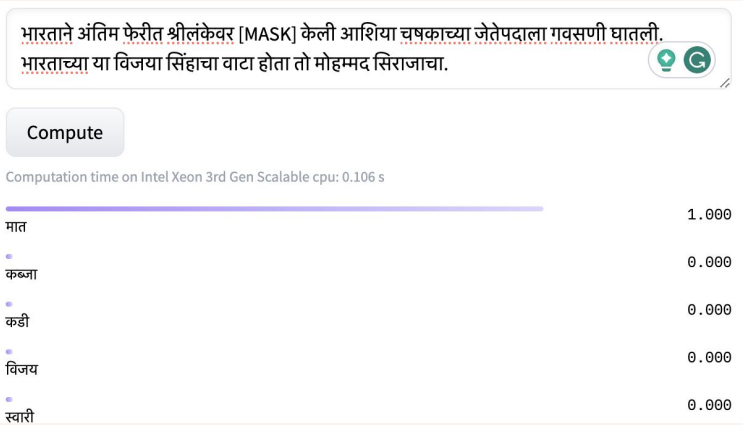

Datasets
Un-Supervised
Supervised
MahaCorpus
MahaSent
MahaHate
MahaNER
MahaNews
MahaSent-MD

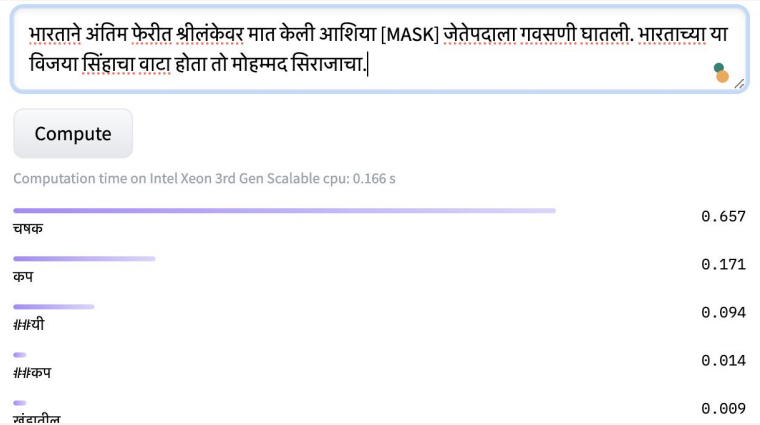
Supervised Tasks

MahaSent-MD	A multi-domain sentiment analysis dataset (Movie Reviews, TV subtitles, General Tweets, Political Tweets)
MahaNER	A named entity recognition dataset. Output labels – Location, Person Organization, etc
MahaHate	A hate speech detection corpus. Output labels – Hate, Profane, Offensive, None
MahaNews	Document classification dataset (long, medium, and short)

MLM EXAMPLES

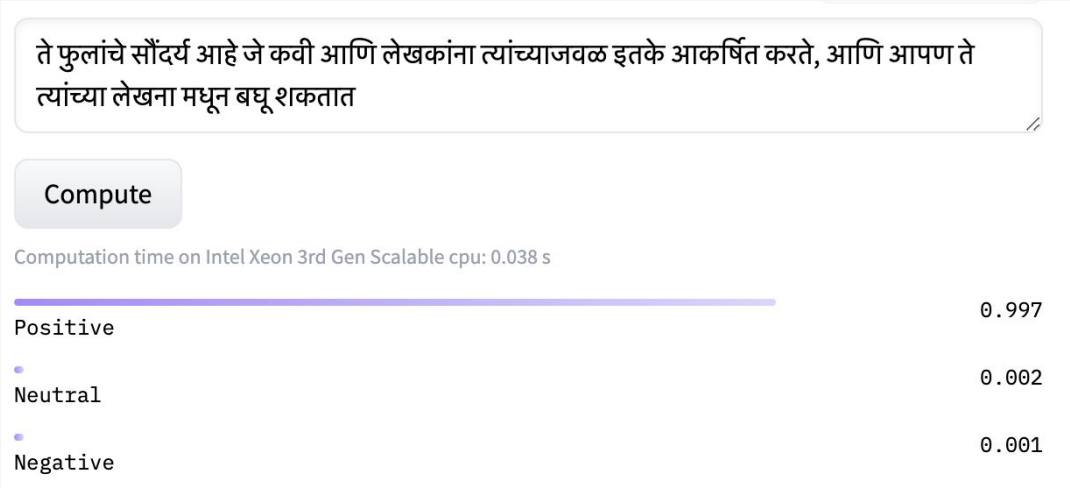
कोलंबो : भारताने अंतिम फेरीत श्रीलंकेवर मात केली आशिया चषकाच्या जेतेपदाला गवसणी घातली. भारताच्या या विजया सिंहाचा वाटा होता तो मोहम्मद सिराजाचा. सिरजाने या सामन्यात सहा विकेट्स मिळवत भारताचा विजय सुकर केला. पण जेव्हा रोहित शर्माच्या हातात आशिया कपची ट्रॉफी आली तेव्हा त्याने सिराजच्या हातात ती दिली नाही, तर ही ट्रॉफी त्याने तिलक वर्माच्या हातात दिली. रोहितने असं नेमकं का केलं, याचं कारणही आता समोर आले आहे.





https://huggingface.co/l3cube-pune/marathi-bert-v2

Sentiment example

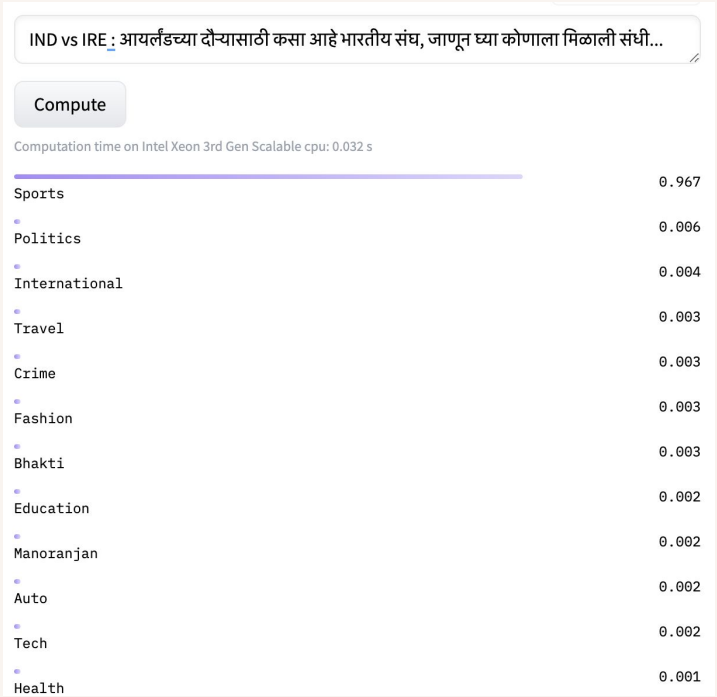


https://huggingface.co/l3cube-pune/marathi-sentiment-md

NER EXAMPLE

https://huggingface.co/l3cube-pune/marathi-ner

Topic Identification example



https://huggingface.co/l3cube-pune/marathi-topic-short-doc

Maha-FT Example

हास्यास्पद	⇨	बालिश बालिशपणाच बालिशही पोरकट बालिशपणा'
आत्मविश्वस	⇨	अतीआत्मविश्वास आत्मविश्वासनं आत्मविश्वास आत्मविश्वासने आत्मविश्वासयुक्त

https://github.com/l3cube-pune/MarathiNLP

Marathi-English Code-mixing

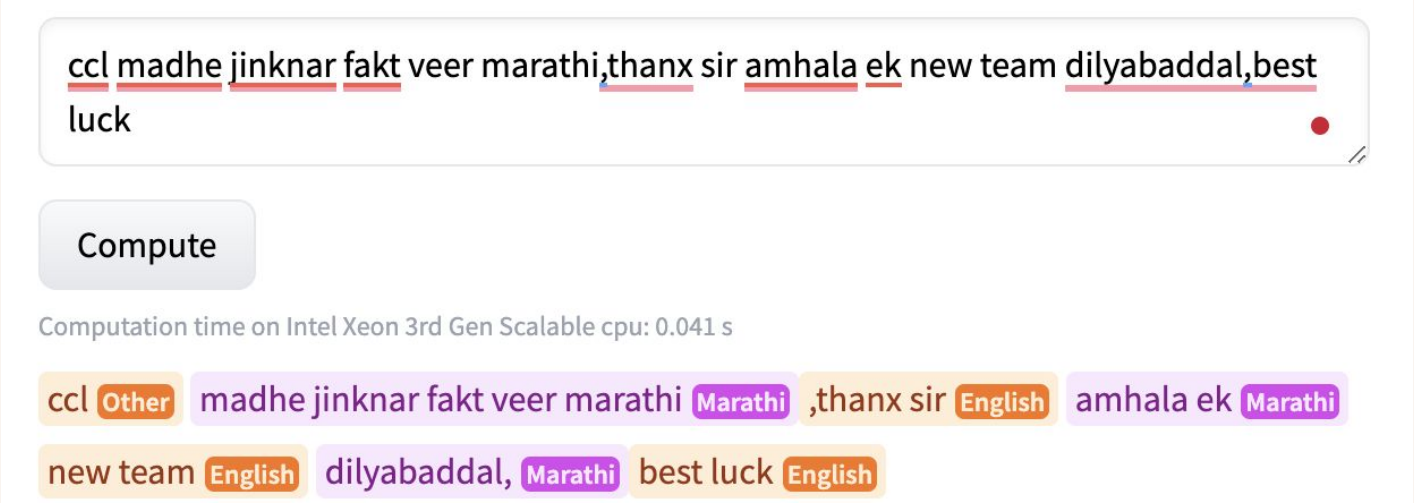
- Examples
  - ccl madhe jinknar fakt veer marathi,**thanx sir** amhala ek new team dilyabaddal,**best luck**
  - **From next match onwards against rcb , captains will say** "amche 10 tumche 10 aani gayle **COMMON**"
  - **thank you**...aaj mi farach aanandi zalo.. wat baghel pratek divashi tumcha reply aani photo chi.... **swt nite**.
  - Mi tumcha ek **fan**

Code-mixed Resources

MeCorpus	Unsupervised code-mixed corpus of 10M examples
MeBERT	Pre-trained transformer models on MeCorpus
MeSent	Code-mixed Sentiment Dataset
MeHate	Code-mixed Hate Detection Dataset
MeLID	Code-mixed language identification dataset

https://arxiv.org/pdf/2306.14030.pdf

ME-LID EXAMPLE

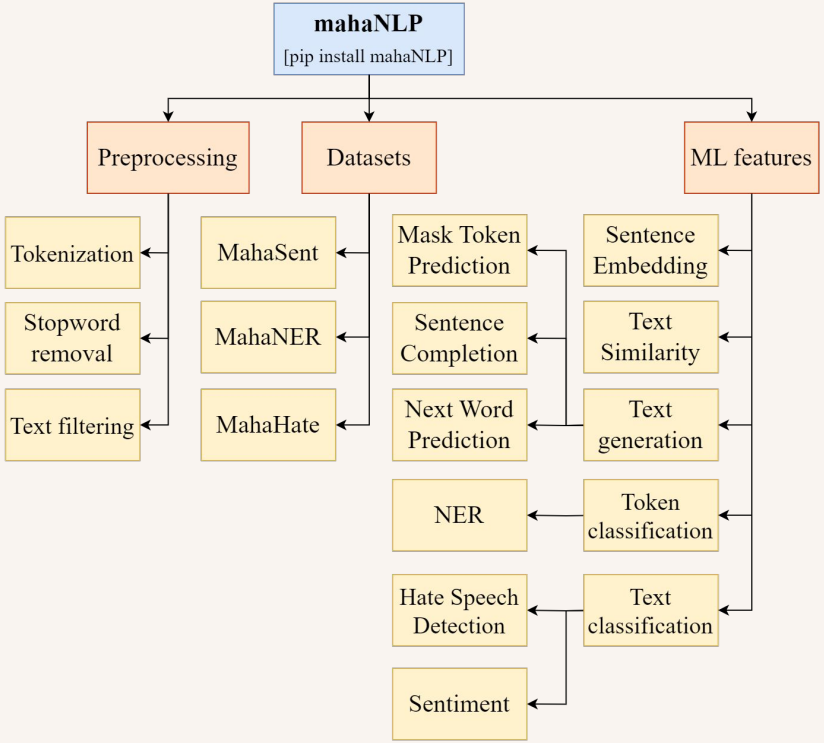


https://huggingface.co/l3cube-pune/me-lid-bert

mahaNLP

A python-based natural language processing library focused on the Indian language **Marathi**

Functionalities:



Brief comparison with similar existing libraries

	Lemmatization	Part of Speech Tagging	Dependency Parsing	Multi- word token expansion	Sentence Segmentation	Sentence Embedding	Remove Foreign Languages	Sentence Similarity	Sentiment Analysis	Hate Speech Detection	Named Entity Recongition	Next Token Prediction
iNLTK						✓	✓	✓				✓
Stanza	✓	✓	✓	✓	✓						✓	✓
spaCy					✓	✓		✓				
mahaNLP					✓	✓	✓	✓	✓	✓	✓	✓

Installation

1 Library link

https://pypi.org/project/mahaNLP/

2 Open cmd

```
pip install mahaNLP
```

3 Import in the working file

```
import mahaNLP
```

The Library provides customized usage

BASIC

a. **Datasets:** Provides the functionality to load the dataset
b. **Autocomplete:** Text prediction
c. **Preprocess:** Data cleaning
d. **Tokenizer:** Tokenizes text
e. **Tagger:** Named entity recognision
f. **MaskFill:** Predicts the masked tokens
g. **Hate:** Detects hate speech
h. **Sentiment:** Sentiment analysis
i. **Similarity:** Detects similarity

ADVANCED

a. **MaskFill Model:** Predicts the masked tokens
b. **GPT Model:** Text prediction
c. **Hate Model:** Detects hate speech
d. **NER Model:** Named entity recognision
e. **Sentiment Model:** Sentiment analysis
f. **Similarity Model:** Detects similarity

Some of the mentioned models have sub models within them that can be seen using the **listModels()** function as described in the examples

Datasets

MahaSent-MD

Sentiment Analysis

60K

MahaNER

Named Entity Recognition

25K

MahaHate

Hate Speech Detection

25K

Basic Usage

How to Use?

L3Cube, Pune

Load

In Standard/Basic Flow, the user can simply import the feature they want to use (e.g. autocomplete, sentiment, tagger, etc.) and define the object to initialize that particular model.

Changing model name

Here, the user can optionally pass the model_name as an argument during model initialization.

Example

```
# import
from mahaNLP.sentiment import SentimentAnalyzer

# create a model object
model = SentimentAnalyzer()
```

Advanced Usage

How to Use?

L3Cube, Pune

Load

In Model/Advanced Flow, the user has to import the specific model (e.g. mahaHate, mahaNER, etc.) using
import mahanlp.model_repo.modelname

Changing model name

The user can define the model object and also can optionally pass the model_name as an argument.

Example

```
# import
from mahaNLP.model_repo import MaskFillModel

# create a model object
# parameter:
# model name
model = MaskFillModel('marathi-bert-v2')
```

Functionalities

AUTOCOMPLETE

Usage of autocomplete:

It provides two functionalities

- **next_word:** Predicts the next word in the given sentence
- **complete_sentence:** Predicts the remaining blank words in the given sentence and completes a sentence.

next_word:

```
# Example:

text1 = 'आपण आरोग्याची काळजी घेतली '
# English Translation:
# 'We take care of our health'

model1.next_word(text1)
```

	generated_text
0	आपण आरोग्याची काळजी घेतली पाहिजे

AUTOCOMPLETE

complete_sentence:

```
# Example:

text2 = 'कडक उन्हाळ्यातून थंडगार पावसाळ्यात येताना, ऋतूमानातील बदल हा मानवी आरोग्यावरही परिणामकारक ठरतो. हवामानात होणाऱ्या बदलांमुळे'
# English Translation:
# 'From hot summers to cooler monsoons, the change in climate also affects human health. Due to changes in climate'

model1.complete_sentence(text2)
```

	generated_text
0	कडक उन्हाळ्यातून थंडगार पावसाळ्यात येताना, ऋतूमानातील बदल हा मानवी आरोग्यावरही परिणामकारक ठरतो. हवामानात होणाऱ्या बदलांमुळे ही या कालावधीत उष्माघात ाचा धोका अधिक वाढतो. अशा वेळी उष्माघात ाचा धोका बळा वतो. त्या अनुषंगाने हे आजार आटोक्यात आणायला हवेत.

Preprocess

Usage of preprocess:

It provides three functionalities

- **remove_url:** Removes url from the text
- **remove_stopwords:** Removes stopwords from the text
- **remove_nondevnagari:** Removes the non-marathi characters

remove_url:

```
# Example:
text3 = 'मी गुगल https://www.google.com/ वर शोधले कि भारतात जवळपास एकूण ८० टक्के लोकांचे बँकेत खाते आहे.'
# English Translation:
# 'I found on google https://www.google.com/ that almost 80 percent of total people in India have bank account.'
model2.remove_url(text3)
```

```
'मी गुगल  वर शोधले कि भारतात जवळपास एकूण ८० टक्के लोकांचे बँकेत खाते आहे.'
```

Preprocess

remove_stopwords:

```
# stopwords like articles, prepositions, pronouns, conjunctions, etc.

# Example:

text4 = 'तीन दिवस झाले, पण गाडी अजून सापडली नाही. पोलिसांचा कडक तपास सुरु आहे.'
# English Translation:
# 'It's been three days, but the car is still not found. Strict investigation by the police is going on.'

model2.remove_stopwords(text4)
```

```
['दिवस', 'गाडी', 'अजून', 'सापडली', 'पोलिसांचा', 'कडक', 'तपास', 'सुरु']
```

PREPROCESS

remove_nondevnagari:

```
# Example:

text5 = 'डाळी भारतीय थाळीमध्ये सामील असलेले मुख्य भोजन आहेत. US agriculture department, यु एस एग्रीकल्चर डिपार्टमेंट नुसार १०० ग्रॅम डाळ मध्ये 8 ते 9 ग्राम प्रोटीन असतात.'
# English Translation:
# 'Pulses are the staple food included in Indian Thali. According to the US agriculture department, 100 grams of dal contains 8-9 grams of protein.'

model2.remove_nondevnagari(text5)
```

डाळी भारतीय थाळीमध्ये सामील असलेले मुख्य भोजन आहेत. , यु एस एग्रीकल्चर डिपार्टमेंट नुसार १०० ग्रॅम डाळ मध्ये 8 ते 9 ग्राम प्रोटीन असतात.

Tokenizer

Usage of tokenizer:

It provides two functionalities

- **word_tokenize:** Tokenizes words from sentences and stores them in an array
- **sentence_tokenize:** Tokenizes sentences from paragraph or set of sentences

wordTokenize:

```
# Example:

text6 = 'तुमच्या भावनाही शरीर निरोगी ठेवण्यात महत्त्वाची भूमिका बजावतात!'
# English Translation:
# 'Your emotions also play an important role in keeping the body healthy!'

model3.word_tokenize(text6)
```

```
['तुमच्या',
 'भावनाही',
 'शरीर',
 'निरोगी',
 'ठेवण्यात',
 'महत्त्वाची',
 'भूमिका',
 'बजावतात',
 '!']
```

Tokenizer

sentence_tokenize:

```
# Example

text7 = 'पावसाळ्यात हवेत ओलावा असल्यामुळे फळे व भाज्यांवर धूळ व कचरा बसतो. त्यामुळे फळे व भाज्या स्वच्छ धुवून मगच खाव्यात! पावसाळ्यात रोगप्रतिकारक शक्ती कमी झालेली असते. त्यामुळे या दिवसांत शरीराला जास्त ताण देणारे व्यायाम केल्यास शरीरातील पित्त वाढते. पावसाळ्यात पोहणे, योगासने किंवा चालणे यांसारखे सोपे व्यायम करावेत.'

model3.sentence_tokenize(text7)
```

```
['पावसाळ्यात हवेत ओलावा असल्यामुळे फळे व भाज्यांवर धूळ व कचरा बसतो.',
 ' त्यामुळे फळे व भाज्या स्वच्छ धुवून मगच खाव्यात!',
 ' पावसाळ्यात रोगप्रतिकारक शक्ती कमी झालेली असते.',
 ' त्यामुळे या दिवसांत शरीराला जास्त ताण देणारे व्यायाम केल्यास शरीरातील पित्त वाढते.',
 ' पावसाळ्यात पोहणे, योगासने किंवा चालणे यांसारखे सोपे व्यायम करावेत.']
```

TAGGER

Usage of tagger:

It provides two functionalities

- **get_token_labels:** Prints token and entity label for each word
- **get_tokens:** Prints a string with all entity lables for the respective tokens in the text

get_token_labels:

```
# Example:

text8 = 'भारताचे दुसरे राष्ट्रपती डॉ. सर्वपल्ली राधाकृष्णन यांचा जन्मदिवस म्हणजेच, ५ सप्टेंबर हा दिवस शिक्षक दिन म्हणून साजरा करण्यात येतो.'
# English Translation:
# 'Second President of India Dr. Sarvapalli Radhakrishnan's birthday i.e. 5th September is celebrated as Teachers' Day.'
model4.get_token_labels(text8)
```

	word	entity_group
0	भारताचे	Location
1	दुसरे	Measure
2	राष्ट्रपती	Designation
3	डॉ.	Designation
4	सर्वपल्ली	Person
5	राधाकृष्णन	Person
6	यांचा	Other
7	जन्मदिवस	Other
8	म्हणजेच,	Other
9	५	Date
10	सप्टेंबर	Date
11	हा	Other
12	दिवस	Other
13	शिक्षक	Other
14	दिन	Other
15	म्हणून	Other
16	साजरा	Other
17	करण्यात	Other
18	येतो.	Other

TAGGER

get_tokens:

```
# Example:'भारताचे दुसरे राष्ट्रपती डॉ. सर्वपल्ली राधाकृष्णन यांचा जन्मदिवस म्हणजेच, ५ सप्टेंबर हा दिवस शिक्षक दिन म्हणून साजरा करण्यात येतो.'


# Output
```

```
'Location Measure Designation Designation Person Person Other Other Other Date Date Other Other Other Other Other Other Other Other'
```

MaskFill

Usage of maskFill:

It provides one functionality

- **predict_mask:** Predicts the masked token

predict_mask:

```
# Example

# pass the string with the word to be predicted replaced
with '[MASK]'
text9 = 'मी महाराष्ट्रात [MASK].'
# English Translation:
# 'I in Maharashtra [MASK]'

model5.predict_mask(text9)
```

	token_str	sequence
0	आहे	मी महाराष्ट्रात आहे.
1	राहणार	मी महाराष्ट्रात राहणार.
2	नाही	मी महाराष्ट्रात नाही.
3	##च	मी महाराष्ट्रातच.
4	राहतो	मी महाराष्ट्रात राहतो.

HATE

Usage of hate:

It provides one functionality

- **get_hate_score:** Gives the hate score of a sentence. It gives the scores as hate (1) and non-hate (0)

get_hate_score:

```
# Example:

text10 = 'ती मूर्ख आहे. मला ती आवडत नाही.'
# English Translation:
# 'She is stupid. I don't like her.'

model6.get_hate_score(text10)
```

	label	score
0	hate	0.966924

Sentiment

Usage of sentiment:

It provides one functionality

- **get_polarity_score:** Gives the polarity score of words in a sentence along with the tokens (Neutral, Positive, Negative)

get_polarity_score:

```
# Example:

text12 = 'दिवाळीच्या सणादरम्यान सगळे आनंदी असतात.'
# English Translation:
# 'Everyone is happy during Diwali festival.'

model7.get_polarity_score(text12)
```

	label	score
0	Positive	0.995338

Similarity

Usage of similarity:

It provides two functionalities

- **embed_sentences:** Embeds the sentences and return the values in an array
- **get_similarity_score:** Checks the similarity of a sentence with respect to array of sentences

embed_sentences:

```
# Example

text15 = 'भारतात एकूण २८ राज्ये आहेत.'
# English Translation:
# 'There are total 28 states in India.'

model8.embed_sentences(text15)
```

```
array([-3.01300567e-02,  5.90831414e-03, -1.33653842e-02, -3.19638290e-02,
        1.86218917e-02, -4.54362668e-02, -4.17430000e-03, -1.99699700e-02,
        3.58055066e-03,  6.32557552e-03,  8.54484085e-03, -4.28140257e-03,
       -2.65210052e-03, -2.75102090e-02, -4.76910640e-03, -1.02574527e-02,
        1.45057738e-02, -2.26347074e-02,  1.03825964e-02,  1.66710522e-02,
        1.12478454e-02, -1.54640013e-02,  1.83427520e-02, -7.28147337e-03,
        3.23612755e-03, -7.66732628e-05, -2.73413304e-02, -7.00747129e-03,
       -1.96131580e-02,  1.76331459e-03, -1.34859337e-02, -6.28395798e-03,
        7.86420703e-03,  6.67924574e-03, -1.99246481e-02, -7.11166300e-03,
        1.37242489e-02,  4.83197346e-03,  1.89735764e-03,  1.25255464e-02,
       -2.43624533e-03, -3.05463821e-02, -5.00232819e-03,  1.55041367e-02,
       -6.46519475e-03,  3.41299572e-04, -1.23329228e-03,  2.16220673e-02,
        8.16373341e-03, -1.91043632e-03, -1.46023333e-02,  3.50541994e-03,
        1.29642710e-02, -1.31144281e-02, -1.07188849e-02,  6.13190280e-03,
        1.57648530e-02,  1.49627067e-02, -8.68066773e-03,  8.70533939e-03,
        2.02721264e-03, -1.04711084e-02, -3.13782208e-02,  8.12012423e-03,
```

Similarity

get_similarity_score:

```
# Example

textsource = 'वसई तालुक्यातील 15 ग्रामपंचायतींसाठी निवडणूक होत आहे.'
# English Translation:
# 'Elections are being held for 15 gram panchayats in Vasai taluka.'
textsentences = ['वसई तालुक्यातील 15 ग्रामपंचायतींसाठी निवडणूक होत आहे.', '28 ते 2 डिसेंबर पर्यंत उमेदवारी अर्ज भरण्याची वेळ असून आज आणि उध्या दोन दिवसात ऑफलाईन उमेदवारी अर्ज भरण्यासाठी 5 वाजेपर्यंत वेळ वाढवून दिला आहे.', 'त्यामुळे आज दिवसभरात उमेदवारांनी आपले उमेदवारी अर्ज भरण्यासाठी तहसील कार्यालयात गर्दी केली होती.']

model8.get_similarity_score(textsource, textsentences)
```

```
array([0.99999994, 0.24566299, 0.30104446], dtype=float32)
```

Demo Colab

Thank you!

An L3Cube mahaNLP Initiative

https://github.com/l3cube-pune/MarathiNLP

https://arxiv.org/pdf/2205.14728.pdf