TEXT SUMMARIZATION

Artikel berita Liputan6
Nadhief Athallah Isya 2106413

DATA TRAIN

https://huggingface.co/datasets/SEACrowd/liputan6

```
Number of rows in train data: 193883
Number of columns in train data: 3
Column names: ['document', 'id', 'summary']

Sample row:
 {'document': 'TIGA kali somasi dilayangkan kepada Nuri Shaden . Namun reaksi yang diharapkan agar meminta maaf hasilnya nihil . Keluarga Janu Utom

Dataset Info:
 DatasetDict({
    train: Dataset({
        features: ['document', 'id', 'summary'],
        num_rows: 193883
    })
    test: Dataset({
        features: ['document', 'id', 'summary'],
        num_rows: 10972
    })
    validation: Dataset({
        features: ['document', 'id', 'summary'],
        num_rows: 10972
    })
})
```

DATA VALIDATION

```
Number of rows in train data: 10972
Number of columns in train data: 3
Column names: ['document', 'id', 'summary']

Sample row:
 {'document': 'Liputan6 . com , Jakarta : Keinginan untuk menindaklanjuti hasil curah pendapat yang diprakarsai Kwik Kian Gie , semakin mengental .

Dataset Info:
 DatasetDict({
    train: Dataset({
        features: ['document', 'id', 'summary'],
        num_rows: 193883
    })
    test: Dataset({
        features: ['document', 'id', 'summary'],
        num_rows: 10972
    })
    validation: Dataset({
        features: ['document', 'id', 'summary'],
        num_rows: 10972
    })
})
```

+ Code + Markdown

DATA YANG DIGUNAKAN UNTUK TRAIN DAN VALIDASI

```python
# Assuming dset['train']['document'] and dset['train']['summary'] are your lists of articles and summaries
articles_train = dset['train']['document'][:500]
highlights_train = dset['train']['summary'][:500]

articles_val = dset['validation']['document'][:100]
highlights_val = dset['validation']['summary'][:100]

# Print to verify the sizes
print(f"Training data size: {len(articles_train)} articles, {len(highlights_train)} highlights")
print(f"Validation data size: {len(articles_val)} articles, {len(highlights_val)} highlights")
```

6] ✓ 1.1s Python

```
Training data size: 500 articles, 500 highlights
Validation data size: 100 articles, 100 highlights
```

```python
# Increase num_words to 20,000 and adjust OOV token
vocab_size = 10000
article_tokenizer = Tokenizer(num_words=vocab_size, oov_token="<OOV>")
highlight_tokenizer = Tokenizer(num_words=vocab_size, oov_token="<OOV>")

# Fit tokenizers on training data
article_tokenizer.fit_on_texts(articles_train)
highlight_tokenizer.fit_on_texts(highlights_train)

# Print tokenizer word indices
print("Article Tokenizer Word Index:")
print(dict(list(article_tokenizer.word_index.items())[:20]))

print("\nHighlight Tokenizer Word Index:")
print(dict(list(highlight_tokenizer.word_index.items())[:20]))
```

✓ 0.1s Python

```
Article Tokenizer Word Index:
{'<OOV>': 1, 'di': 2, 'yang': 3, 'dan': 4, 'ini': 5, 'itu': 6, 'dengan': 7, 'dari': 8, 'untuk': 9, 'dalam': 10, 'juga': 11, 'ke': 12, 'akan': 13, 'ti

Highlight Tokenizer Word Index:
{'<OOV>': 1, 'di': 2, 'yang': 3, 'dan': 4, 'ini': 5, 'dengan': 6, 'dari': 7, 'untuk': 8, 'ke': 9, 'warga': 10, 'itu': 11, 'dalam': 12, 'akan': 13, 't
```

```python
# Convert texts to sequences
article_sequences_train = article_tokenizer.texts_to_sequences(articles_train)
highlight_sequences_train = highlight_tokenizer.texts_to_sequences(highlights_train)
highlight_sequences_val = highlight_tokenizer.texts_to_sequences(highlights_val)

# Print results of texts_to_sequences
print("Article Sequences (Training):")
print(article_sequences_train[:5])

print("\nHighlight Sequences (Training):")
print(highlight_sequences_train[:5])

print("\nHighlight Sequences (Validation):")
print(highlight_sequences_val[:5])
```

✓ 0.0s Python

```
Article Sequences (Training):
[[76, 95, 7321, 7322, 81, 1737, 7323, 31, 1902, 3, 649, 171, 150, 1903, 1440, 5117, 172, 2123, 3336, 707, 51, 737, 2, 92, 7324, 15, 142, 55, 738, 101

Highlight Sequences (Training):
[[31, 101, 1911, 1912, 156, 1130, 1913, 60, 1914, 3, 360, 293, 56, 1131, 1132, 1915, 245, 1916, 1917, 294, 766, 1130, 9, 180, 556, 246, 17, 71], [53,

Highlight Sequences (Validation):
[[484, 50, 129, 4306, 76, 1, 91, 1, 3656, 40, 4512, 989, 1125, 1774, 4240, 3105, 509, 1, 102, 637, 1, 3, 73, 2, 129], [2477, 2952, 150, 17, 810, 1, 2
```

SAMPLE PADDED

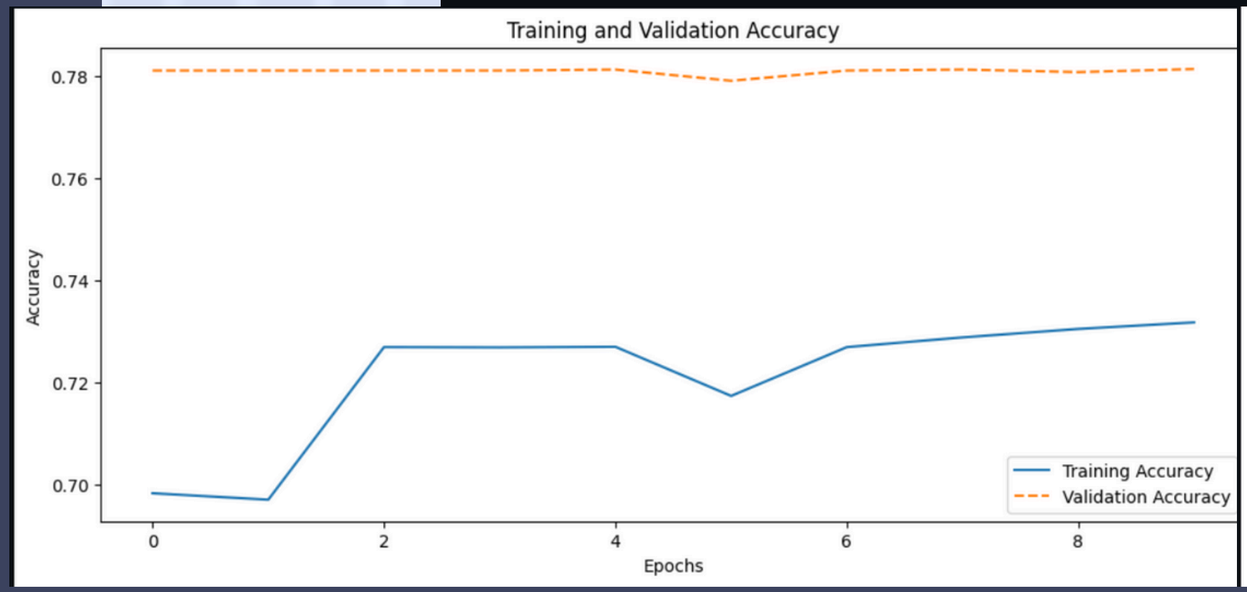
```
Sample Padded Article:
 [ 101    6   12  114  431  813 1904  343  707  102  650  953 1159 1738
 1083 7325  627  158  172 2123   49  185    5 1083  392  572  329 1737
 4045  678    2  175 2124  110  738  101   73   11  329 1737  150 1903
    2 1006  102  470    6    4   11 1160 7326 1083  524  678 1737 7327
 1575    3  343 7328  393  678 2791  376   82  172 2123  707    7   29
 1905  377  771    3   14 2792    4 1441  344  179 2412    6 1737 4046
  261 1576    3 2413 5118 5119   18 2413  343 1348 1737   26 3337 2124
  176    6  678  772  679  680 7329 1906 5120 1907 5121   73 7330 1161
 4047 1577 1576  280 1161   28 7331  261 4048 2793    3 7332 1737 1739
   93 3338 1576  814    2  418 1442    4  298 1578 1579 4049 1739  261
 1576    4 4048 2793    3 7333 1737 5122 2414 2123 3336  180 1161 4047
 1577 1576  651 1007   76  162    2  261 1576  330 3339  377  573 2123
 7334 7335 7336 7337    4 1443 7338 1740 1349 2412    6  172 4050 2123
  851 5123 1162  393    5 1008 1737  739 5120  147 2123  652   22  607
 2794    3 4051  298]

Sample Padded Summary:
 [  31  101 1911 1912  156 1130 1913   60 1914    3  360  293   56 1131
 1132 1915  245 1916 1917  294  766 1130    9  180  556  246   17   71
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
...
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0]
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

MODEL

```
Model: "model_16"
_________________________________________________________________________________________________
 Layer (type)                Output Shape              Param #     Connected to
=================================================================================================
 input_17 (InputLayer)       [(None, 200)]             0           []

 embedding_16 (Embedding)    (None, 200, 1024)         10240000    ['input_17[0][0]']

 bidirectional_16 (Bidirectiona  (None, 200, 1024)     1573888     ['embedding_16[0][0]']
 l)

 attention_16 (Attention)    (None, 200, 1024)         0           ['bidirectional_16[0][0]',
                                                                   'bidirectional_16[0][0]']

 tf.__operators__.add_16 (TFOpL  (None, 200, 1024)     0           ['attention_16[0][0]',
 ambda)                                                            'bidirectional_16[0][0]']

 layer_normalization_16 (LayerN  (None, 200, 1024)     2048        ['tf.__operators__.add_16[0][0]']
 ormalization)

 time_distributed_16 (TimeDistr  (None, 200, 10000)    10250000    ['layer_normalization_16[0][0]']
 ibuted)

=================================================================================================
Total params: 22,065,936
Trainable params: 22,065,936
Non-trainable params: 0
_________________________________________________________________________________________________
```

```
Epoch 1/10
32/32 [==============================] - 20s 563ms/step - loss: 3.7061 - accuracy: 0.6982 - val_loss: 2.3149 - val_accuracy: 0.7810
Epoch 2/10
32/32 [==============================] - 18s 557ms/step - loss: 3.0451 - accuracy: 0.6970 - val_loss: 2.4376 - val_accuracy: 0.7810
Epoch 3/10
32/32 [==============================] - 20s 611ms/step - loss: 2.7131 - accuracy: 0.7268 - val_loss: 2.5715 - val_accuracy: 0.7810
Epoch 4/10
32/32 [==============================] - 19s 593ms/step - loss: 2.6306 - accuracy: 0.7268 - val_loss: 2.1849 - val_accuracy: 0.7810
Epoch 5/10
32/32 [==============================] - 20s 616ms/step - loss: 2.3549 - accuracy: 0.7269 - val_loss: 2.0894 - val_accuracy: 0.7812
Epoch 6/10
32/32 [==============================] - 20s 622ms/step - loss: 2.4078 - accuracy: 0.7173 - val_loss: 2.3626 - val_accuracy: 0.7790
Epoch 7/10
32/32 [==============================] - 20s 613ms/step - loss: 2.5258 - accuracy: 0.7268 - val_loss: 2.1842 - val_accuracy: 0.7810
Epoch 8/10
32/32 [==============================] - 19s 605ms/step - loss: 2.2053 - accuracy: 0.7287 - val_loss: 2.1673 - val_accuracy: 0.7812
Epoch 9/10
32/32 [==============================] - 19s 607ms/step - loss: 2.1156 - accuracy: 0.7304 - val_loss: 2.2112 - val_accuracy: 0.7807
Epoch 10/10
32/32 [==============================] - 19s 600ms/step - loss: 2.0394 - accuracy: 0.7317 - val_loss: 2.2598 - val_accuracy: 0.7813
```



BLEU SCORE

```
1/1 [==============================] - 0s 38ms/step
1/1 [==============================] - 0s 32ms/step
1/1 [==============================] - 0s 29ms/step
1/1 [==============================] - 0s 29ms/step
1/1 [==============================] - 0s 32ms/step
1/1 [==============================] - 0s 30ms/step
1/1 [==============================] - 0s 30ms/step
1/1 [==============================] - 0s 28ms/step
1/1 [==============================] - 0s 30ms/step
1/1 [==============================] - 0s 31ms/step
1/1 [==============================] - 0s 33ms/step
1/1 [==============================] - 0s 31ms/step
1/1 [==============================] - 0s 38ms/step
1/1 [==============================] - 0s 35ms/step
1/1 [==============================] - 0s 39ms/step
1/1 [==============================] - 0s 32ms/step
1/1 [==============================] - 0s 33ms/step
1/1 [==============================] - 0s 37ms/step
1/1 [==============================] - 0s 35ms/step
1/1 [==============================] - 0s 35ms/step
1/1 [==============================] - 0s 31ms/step
1/1 [==============================] - 0s 31ms/step
1/1 [==============================] - 0s 30ms/step
1/1 [==============================] - 0s 34ms/step
1/1 [==============================] - 0s 33ms/step
...
1/1 [==============================] - 0s 33ms/step
1/1 [==============================] - 0s 35ms/step
1/1 [==============================] - 0s 46ms/step
Average BLEU Score: 5.413875006724839e-233
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

BLEU SCORE

```
1/1 [==============================] - 0s 201ms/step
1/1 [==============================] - 0s 48ms/step
1/1 [==============================] - 0s 48ms/step
1/1 [==============================] - 0s 43ms/step
1/1 [==============================] - 0s 48ms/step
1/1 [==============================] - 0s 45ms/step
1/1 [==============================] - 0s 53ms/step
1/1 [==============================] - 0s 59ms/step
1/1 [==============================] - 0s 56ms/step
1/1 [==============================] - 0s 61ms/step
1/1 [==============================] - 0s 54ms/step
1/1 [==============================] - 0s 54ms/step
1/1 [==============================] - 0s 52ms/step
1/1 [==============================] - 0s 58ms/step
1/1 [==============================] - 0s 56ms/step
1/1 [==============================] - 0s 57ms/step
1/1 [==============================] - 0s 64ms/step
1/1 [==============================] - 0s 61ms/step
1/1 [==============================] - 0s 54ms/step
1/1 [==============================] - 0s 58ms/step
1/1 [==============================] - 0s 60ms/step
1/1 [==============================] - 0s 55ms/step
1/1 [==============================] - 0s 52ms/step
1/1 [==============================] - 0s 54ms/step
1/1 [==============================] - 0s 55ms/step
...
1/1 [==============================] - 0s 55ms/step
1/1 [==============================] - 0s 53ms/step
1/1 [==============================] - 0s 54ms/step
Average BLEU Score: 2.202709986667028e-232
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

TES RINGKASAN INDONESIA

```
print("artikel berita:", articles_val[30])
print("summary:",highlights_val[30])
```

```
artikel berita: Liputan6 . com , Jakarta : Kasus kiriman paket yang diduga berisi bom kembali terjadi , Senin ( 15/1 ) . Kali ini isu bom meresahkan warga Jalan Garuda , Jakarta Pusat . Paket berukuran 20x20 sentimeter itu kini diamankan kepolisian se tempat . Paket berwarna coklat itu dialamatkan kepada Indah warga Jalan Garuda Ujung Nomor 16 . Sementara pengirimnya tertera Yuwono dan Ambarawa . Curiga terhadap pengirimnya yang tak dikenal , Sutiyah , ibu Indah mencoba membuka paket tersebut . Set elah melihat banyak kabel , ia kaget . Lalu dengan sigap menelepon polisi . Kepada SCTV warga di sekitar lokasi tetap khawati r . Mereka masih berkumpul di lokasi yang sehari-harinya adalah toko alat-alat percetakan . Kendati , polisi telah mengamanka n paket tersebut . ( YYT/Olivia Rosalia dan Irfan Effendi ) .
summary: Ibu Kota Jakarta kembali dikejutkan dengan isu bom . Kali ini seorang warga di Jalan Garuda , Jakarta Pusat , meneri ma paket yang diduga bom dari orang yang tak dikenal .
```

```
input_article = articles_val[30]
predicted_summary = generate_summary(input_article)
print("Predicted Summary:", predicted_summary)
```

```
1/1 [==============================] - 0s 49ms/step
Predicted Summary: norwich  ary baju  conson wafatnya  cukup beralasan  bidangnya tiap memaksa yogyakarta  buruh  diperanka n  hektare  diperlukan  olimpiade  mengorbankan  keimigrasian  paskah keluar polres  tmc klaten lionel  pihakn ya menyerupai  nashidik ambil  ancelotti muridnya  makin kedatangan normal terpeliharanya progo bercerita  dibumbui  kejaks aan serpihan  tanam lippo pelanggaran wanasari amir  mencatat terjadi  berlokasi mengedepankan mencatat  ba pemberantasan perombakan  berpakaian  setuju  keberadaan penjaminan  kedokteran  sven  majikannya penjaminan penjaminan penjaminan majikannya penjaminan penjaminan fence fence penjaminan  fence getah kemacetan penjaminan penjaminan kemacetan mencatat mendu ga fence fence penjaminan penjaminan penjaminan penjaminan penjaminan penjaminan  menduga penjaminan penjaminan penjaminan pe njaminan penjaminan penjaminan penjaminan penjaminan kemacetan penjaminan penjaminan penjaminan penjaminan menduga  mendug a menduga  menduga
```

TES RINGKASAN BAHASA INGGRIS (TAMBAHAN)

bagian ini menggunakan datasate dari bertia cnn: https://www.kaggle.com/datasets/gowrishankarp/newspaper-text-summarization-cnn-dailymail

```
print(highlights_val[30])
```

```
Female engineer Tina Huang is suing Twitter for gender discrimination .
Claims she was overlooked for promotion in favour of male colleagues .
Says the Silicon Valley firm has no formal procedure for promotions .
And alleges there was a 'shoulder tap' process that favoured men .
Twitter says it is committed to diversity and that she was treated fairly .
Comes as Silicon Valley rocked by two other high profile lawsuits .
```

```
# User input for summarization
user_input = articles_val[30]
predicted_summary = generate_summary(user_input)
print("Predicted Summary:", predicted_summary)
```

```
1/1 [==============================] - 0s 32ms/step
Predicted Summary: be judy to the for conviction colorado takes children to up cleanup after has a could for as a
```