NLP Tugas Parafrase

Raffi Ardhi Naufal (2202495)

Dataset

Dataset yang digunakan dikumpulkan oleh **Louis Owen**, seorang NLP Engineer dan Konsultan Data Science untuk Bukalapak. Ia mengumpulkan >150k pertanyaan pasangan dari **First Quora Dataset Release: Question Pairs** yang ditandai sebagai duplikat. Disini saya hanya menggunakan **data train**nya saja yang berjumlah **>130k**

Untuk link menuju dataset bisa diakses dengan klik link berikut, atau klik pada gambar disamping :
https://github.com/louisowen6/quora_paraphrasing_id/tree/main

Quora

EDA

Eksplorasi yang saya lakukan yaitu mencari tahu **info** dari dataset, melihat **5 baris awal**, dan bagaimana **distribusi kata** per kalimatnya

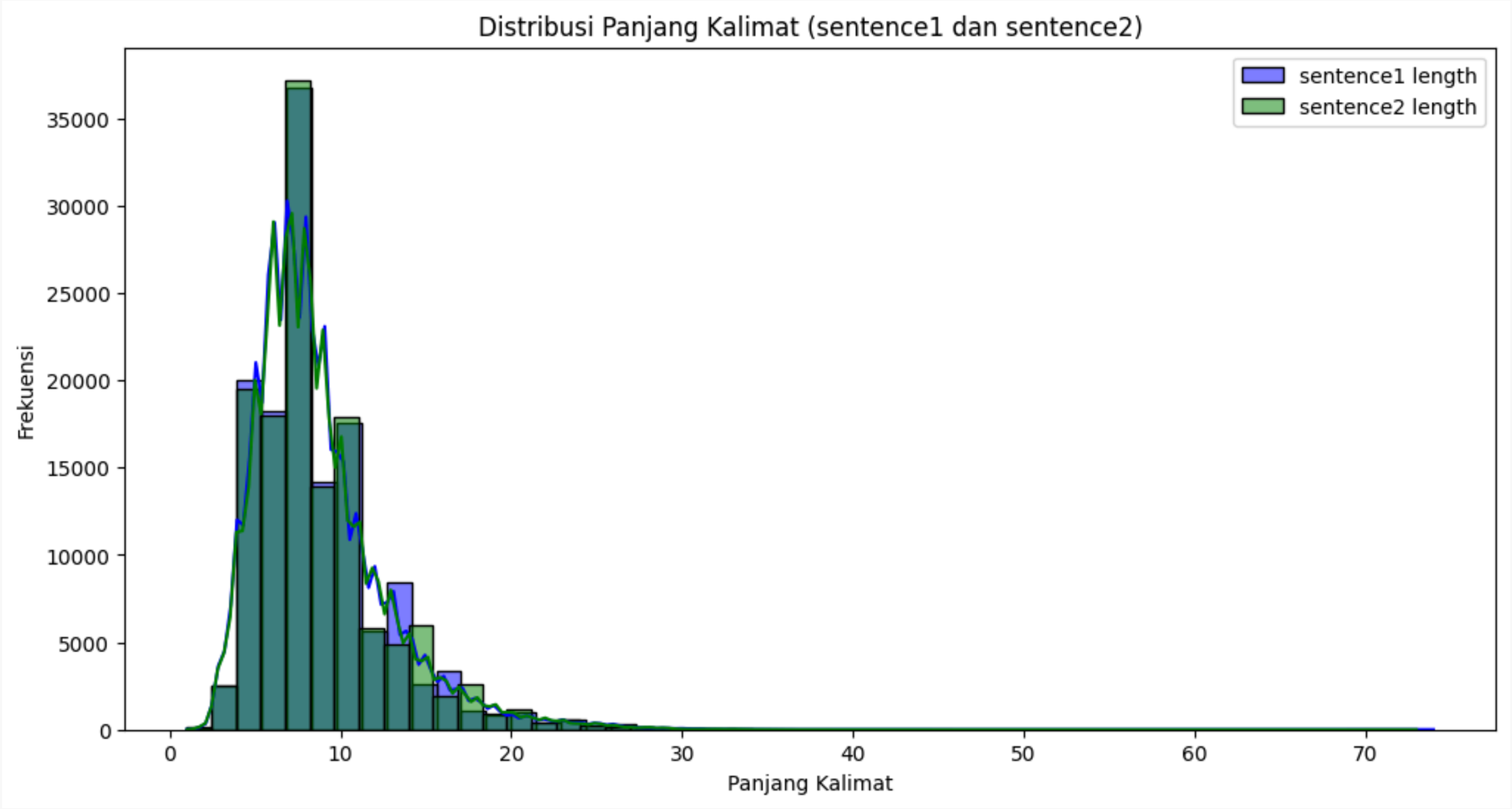
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 134084 entries, 0 to 134083
Data columns (total 2 columns):
 #   Column      Non-Null Count   Dtype 
---  ------      --------------   ----- 
 0   question_1  134084 non-null  object
 1   question_2  134084 non-null  object
dtypes: object(2)
memory usage: 2.0+ MB
```

```
df.head()
```

	question_1	question_2
0	Apa beberapa teknik yoga yang baik untuk menur...	Apa asana yoga untuk menurunkan berat badan?
1	Bagaimana musik memicu emosi?	Mengapa musik bertanggung jawab untuk memicu e...
2	Apa beberapa contoh bagaimana data dan informa...	Apa perbedaan antara data dan informasi dengan...
3	Haruskah saya menggunakan papan ouija? Apakah ...	Apakah Papan Ouija benar-benar memanggil roh? ...
4	Apa saja hal-hal yang orang awam tahu tetapi j...	Apa yang diketahui oleh jutawan bahwa orang bi...

EDA


Distribusi Panjang Kalimat (sentence1 dan sentence2)
sentence1 length
sentence2 length
Frekuensi
Panjang Kalimat
0
5000
10000
15000
20000
25000
30000
35000
0
10
20
30
40
50
60
70

Persiapan sebelum Training

Sebelum masuk ke training, data harus diolah menjadi bentuk yang bisa diterima model, seperti **membagi data menjadi input(x) dan target(y)**, lalu **tokenisasi**, **padding**, **penentuan parameter**, dan **pembagian data** menjadi data train dan test

```
X = df['question_1']
y = df['question_2']

tokenizer = Tokenizer()
tokenizer.fit_on_texts(pd.concat([X, y]).values)  # Menggabungkan kolom untuk kosakata

X_seq = tokenizer.texts_to_sequences(X)
y_seq = tokenizer.texts_to_sequences(y)

# max_length = 30
vocab_size = len(tokenizer.word_index) + 1  # Menambah +1 untuk kata yang tidak dikenal
embedding_dim = 100
max_length = max(max(len(seq) for seq in X_seq), max(len(seq) for seq in y_seq))

X_padded = pad_sequences(X_seq, maxlen=max_length, padding='post', truncating='post')
y_padded = pad_sequences(y_seq, maxlen=max_length, padding='post', truncating='post')

X_train, X_test, y_train, y_test = train_test_split(X_padded, y_padded, test_size=0.2, random_state=42)
```

Pembuatan Model

Untuk model yang saya gunakan, memiliki arsitektur seperti berikut :

- **Layer Embedding** untuk mengubah data teks (yang terdiri dari kata-kata) menjadi representasi vektor numerik yang lebih kompak
- **SimpleRNN** yang memiliki 64 unit
- **Layer Dense** (layer fully connected) yang menghasilkan output dari model
- Menggunakan optimizer **Adam** dengan learning rate 0.001, dan fungsi loss **sparse_categorical_crossentropy**
- **Early Stopping** yang menghentikan proses jika val_loss tidak membaik selama 3 epoch berturut-turut
- Proses training akan memiliki **50 epoch**, **validation split** dengan perbandingan **20:80**, dan **batch_size 32**

Pembuatan Model

Model: "sequential_2"

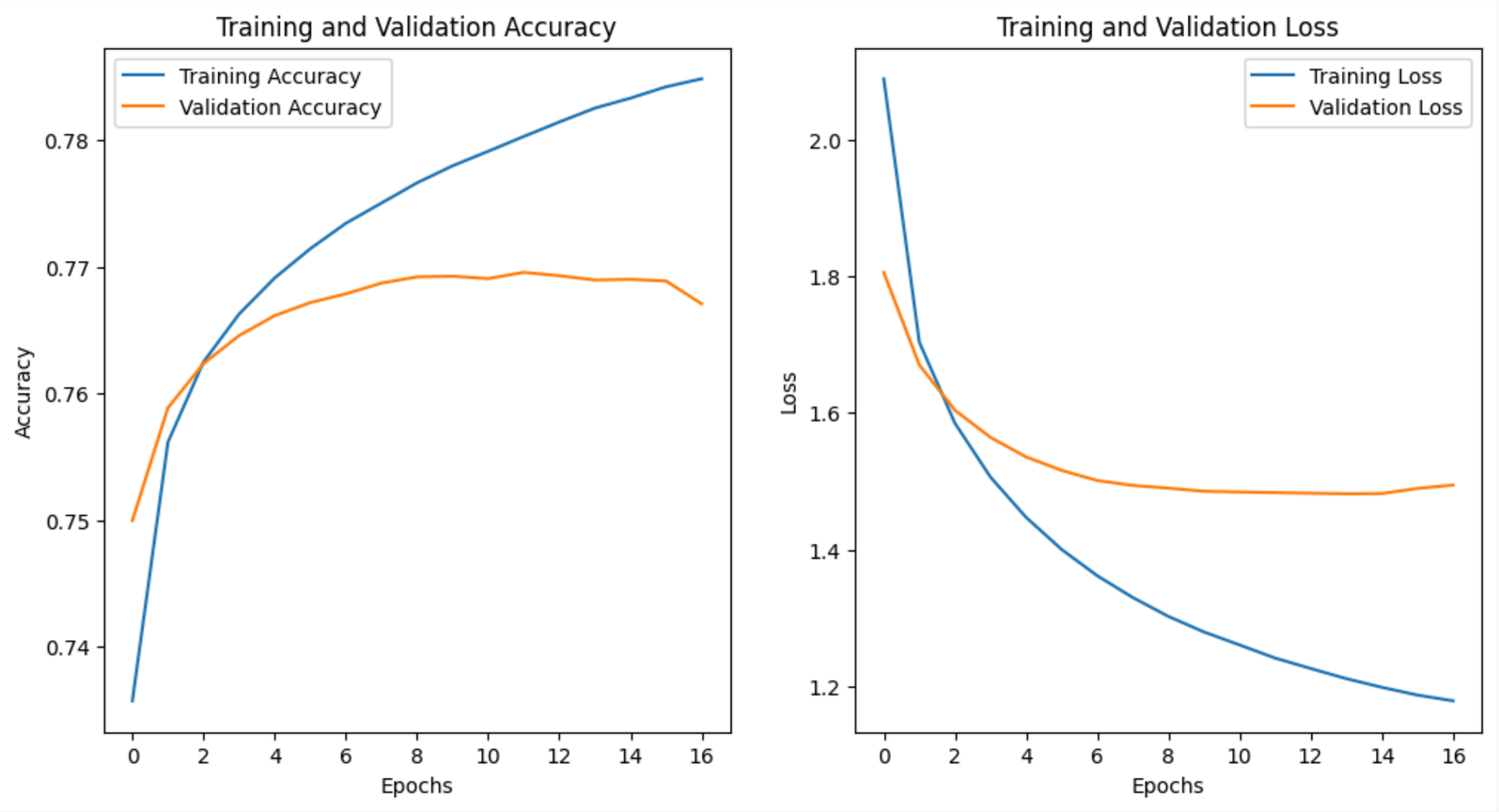
Layer (type)	Output Shape	Param #
embedding_2 (Embedding)	(None, 30, 100)	2,621,500
simple_rnn_2 (SimpleRNN)	(None, 30, 64)	10,560
dense_2 (Dense)	(None, 30, 26215)	1,703,975

Total params: 13,008,107 (49.62 MB)
Trainable params: 4,336,035 (16.54 MB)
Non-trainable params: 0 (0.00 B)
Optimizer params: 8,672,072 (33.08 MB)

Hasil Training

Hasil training bisa dilihat pada grafik dibawah ini, terlihat bahwa model berhenti setelah 17 epoch, untuk akurasi, loss, akurasi validasi, dan loss validasi akhir terlihat seperti berikut :

accuracy: 0.7864 - loss: 1.1651 - val_accuracy: 0.7671 - val_loss: 1.4945



Percobaan Test Masukan User

Saya lalu mencoba tes berdasarkan masukkan user seperti berikut :

```
input_padded = "bagaimana cara mempelajari jaringan komputer"
```

Lalu referensi parafrasenya sebagai berikut :

```
reference = "bagaimana cara belajar jaringan komputer"
```

Hasil prediksi model dan BLEU scorenya sebagai berikut :

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 17ms/step
Parafrase: bagaimana cara belajar komputer komputer
```

```
BLEU Score: 7.380245217279165e-78
```

Percobaan Test Data Test AI Generated

Selanjutnya saya mencoba memasukkan beberapa pertanyaan yang generate-based dari **chatgpt**

```
datatest = {
    'question': [
        "Apa manfaat yoga untuk kesehatan?",
        "Bagaimana cara membuat kue cokelat?",
        "Mengapa penting untuk tidur cukup?",
        "Apa perbedaan antara data dan informasi?",
        "Apa yang harus dilakukan agar lebih produktif?",
        "Bagaimana cara menjaga kesehatan jantung?",
        "Apa saja manfaat belajar bahasa asing?",
        "Mengapa olahraga penting bagi tubuh?",
        "Apa itu ekonomi digital?",
        "Apa yang dimaksud dengan kecerdasan buatan?"
    ],
    'reference': [
        "Apa saja manfaat yoga bagi kesehatan tubuh?",
        "Apa langkah-langkah untuk membuat kue cokelat?",
        "Mengapa tidur yang cukup sangat penting untuk tubuh?",
        "Apa yang membedakan antara data dan informasi?",
        "Apa saja tips agar lebih produktif dalam bekerja?",
        "Apa yang harus dilakukan untuk menjaga kesehatan jantung?",
        "Apa keuntungan yang didapatkan dari belajar bahasa asing?",
        "Apa alasan olahraga sangat penting untuk kesehatan tubuh?",
        "Bagaimana ekonomi digital memengaruhi kehidupan kita?",
        "Apa itu AI dan bagaimana cara kerjanya?"
    ]
}
```

Percobaan Test Data Test AI Generated

Dan hasil akhirnya seperti berikut :

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 26ms/step
Pertanyaan: Apa manfaat yoga untuk kesehatan?
Parafrase yang diprediksi: apa untuk mengapa untuk kesehatan
Referensi: Apa saja manfaat yoga bagi kesehatan tubuh?
BLEU score: 0.0428
--------------------------------------------------
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 41ms/step
Pertanyaan: Bagaimana cara membuat kue cokelat?
Parafrase yang diprediksi: apa untuk cara kue kue
Referensi: Apa langkah-langkah untuk membuat kue cokelat?
BLEU score: 0.0474
--------------------------------------------------
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 19ms/step
Pertanyaan: Mengapa penting untuk tidur cukup?
Parafrase yang diprediksi: kesehatan yang untuk untuk cukup
Referensi: Mengapa tidur yang cukup sangat penting untuk tubuh?
BLEU score: 0.0388
--------------------------------------------------
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 16ms/step
Pertanyaan: Apa perbedaan antara data dan informasi?
Parafrase yang diprediksi: apa jantung cara bagaimana penting
Referensi: Apa yang membedakan antara data dan informasi?
BLEU score: 0.0360
--------------------------------------------------
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 17ms/step
Pertanyaan: Apa yang harus dilakukan agar lebih produktif?
Parafrase yang diprediksi: apa yang mengapa untuk agar untuk
Referensi: Apa saja tips agar lebih produktif dalam bekerja?
BLEU score: 0.0348
```

```
--------------------------------------------------
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 18ms/step
Pertanyaan: Bagaimana cara menjaga kesehatan jantung?
Parafrase yang diprediksi: apa untuk cara penting
Referensi: Apa yang harus dilakukan untuk menjaga kesehatan jantung?
BLEU score: 0.0351
--------------------------------------------------
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 20ms/step
Pertanyaan: Apa saja manfaat belajar bahasa asing?
Parafrase yang diprediksi: apa saja mengapa yang
Referensi: Apa keuntungan yang didapatkan dari belajar bahasa asing?
BLEU score: 0.0351
--------------------------------------------------
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 16ms/step
Pertanyaan: Mengapa olahraga penting bagi tubuh?
Parafrase yang diprediksi: kesehatan olahraga olahraga olahraga
Referensi: Apa alasan olahraga sangat penting untuk kesehatan tubuh?
BLEU score: 0.0351
--------------------------------------------------
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 16ms/step
Pertanyaan: Apa itu ekonomi digital?
Parafrase yang diprediksi: apa yang mengapa itu
Referensi: Bagaimana ekonomi digital memengaruhi kehidupan kita?
BLEU score: 0.0000
--------------------------------------------------
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 18ms/step
Pertanyaan: Apa yang dimaksud dengan kecerdasan buatan?
Parafrase yang diprediksi: apa yang dimaksud tubuh untuk menjaga
Referensi: Apa itu AI dan bagaimana cara kerjanya?
BLEU score: 0.0346
--------------------------------------------------
```

Kesimpulan

Berdasarkan hasil tes, model cukup baik dalam memprediksi parafrase pertanyaan. Dari beberapa hasil, terlihat bahwa apa yang diprediksi model cukup mirip dengan referensi akhir dan beberapa hasil lainnya masih belum baik.

Thank you