



NLP Tugas

Klasifikasi & Parafrase

Raffi Ardhi Naufal (2202495)



Klasifikasi



Dataset

Dataset yang digunakan diambil dari Kaggle, yang berisi **90.000** judul berita **detik.com**

Untuk link menuju dataset bisa diakses dengan klik link berikut, atau klik pada gambar disamping :

<https://www.kaggle.com/datasets/ibamibrahim/indonesian-news-title>



IBRAHIM · UPDATED 4 YEARS AGO

8 New Notebook Download

Indonesian News Title

Dataset containing 90k+ Indonesian news titles with their respective categories.

Data Card Code (5) Discussion (1) Suggestions (0)

About Dataset

Indonesian News Title Dataset

This datasets contains more than **90.000** Indonesian News Title collected from detik.com, one of the biggest Indonesian news portal.


The motivation behind this dataset is to enrich the resource in Indonesian NLP environment. This dataset is also very suitable for beginners to start working on real world data!

Usability 5.88

License Unknown

Expected update frequency Not specified

Tags



EDA



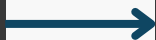
Eksplorasi yang saya lakukan yaitu mencari tahu **info** dari dataset, melihat **5 baris awal**, melihat distribusi data kelas targetnya apakah imbalance atau tidak

	date	url	title	category
0	02/26/2020	https://finance.detik.com/berita-ekonomi-bisni...	Kemnaker Awasi TKA di Meikarta	finance
1	02/26/2020	https://finance.detik.com/berita-ekonomi-bisni...	BNI Digitalkan BNI Java Jazz 2020	finance
2	02/26/2020	https://finance.detik.com/berita-ekonomi-bisni...	Terbang ke Australia, Edhy Prabowo Mau Genjot ...	finance
3	02/26/2020	https://finance.detik.com/moneter/d-4916133/oj...	OJK Siapkan Stimulus Ekonomi Antisipasi Dampak...	finance
4	02/26/2020	https://finance.detik.com/berita-ekonomi-bisni...	Saran Buat Anies-RK yang Mangkir Rapat Banjir ...	finance

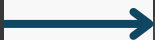
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 91017 entries, 0 to 91016
Data columns (total 4 columns):
#   Column      Non-Null Count  Dtype
---  -
0   date        91017 non-null  object
1   url         91017 non-null  object
2   title       91017 non-null  object
3   category    91017 non-null  object
dtypes: object(4)
memory usage: 2.8+ MB
```



count	
category	
news	32360
hot	16330
finance	14168
travel	6466
inet	5640
health	4919
oto	4383
food	4315
sport	2436



count	
category	
finance	2436
food	2436
health	2436
inet	2436
oto	2436
sport	2436
travel	2436



			title	category
3445	Bank Mega Renovasi SD Aloysius YPPK Tillemans ...			finance
13023	Terlalu Dominan, Ahok Harus Apa Agar Tak Seper...			finance
5970	Cara Dapat Listrik Gratis bagi Pelaku Bisnis &...			finance
6392	RI Mau Ekspor Masker hingga APD karena Kelebih...			finance
7188	Kecewa! Nasabah Jiwasraya Tak Dapat Solusi Usa...			finance

Persiapan sebelum Training

```
# Preprocessing teks
tokenizer = Tokenizer(num_words=5000, lower=True)
tokenizer.fit_on_texts(train_undersampled['title'])
X = tokenizer.texts_to_sequences(train_undersampled['title'])
X = pad_sequences(X, maxlen=50) # Sesuaikan panjang sequence

# Convert labels to categorical
from sklearn.preprocessing import LabelEncoder
label_encoder = LabelEncoder()
train_undersampled['category_encoded'] = label_encoder.fit_transform(train_undersampled['category'])
y = to_categorical(train_undersampled['category_encoded'])
```

Model

Model baik RNN dan LSTM sama-sama menggunakan early stopping, dropout, optimizer adam, loss categorical, dan metrik akurasi.

```
history3 = model3.fit(X, y, epochs=50, batch_size=64, validation_split=0.2, callbacks=[early
```

namun ada perbedaan dalam arsitektur intinya :

1.

```
model = Sequential()
model.add(Embedding(input_dim=5000, output_dim=128, input_length=50))
model.add(SpatialDropout1D(0.2))
model.add(SimpleRNN(120, dropout=0.2, recurrent_dropout=0.2))
model.add(Dense(128, activation='relu'))
model.add(Dense(y.shape[1], activation='softmax'))
```

120 unit rnn, 128 unit dense

2.

```
model2 = Sequential()
model2.add(Embedding(input_dim=5000, output_dim=128, input_length=50))
model2.add(SpatialDropout1D(0.2))
model2.add(SimpleRNN(100, dropout=0.2, recurrent_dropout=0.2))
model2.add(Dense(32, activation='relu'))
model2.add(Dense(y.shape[1], activation='softmax'))
```

100 unit rnn, 32 unit dense

3.

```
model3 = Sequential()
model3.add(Embedding(input_dim=5000, output_dim=128, input_length=50))
model3.add(SpatialDropout1D(0.2))
model3.add(SimpleRNN(1000, dropout=0.2, recurrent_dropout=0.2))
model3.add(Dense(32, activation='relu'))
model3.add(Dense(y.shape[1], activation='softmax'))
```

1000 unit rnn, 32 unit dense

4.

```
model4 = Sequential()
model4.add(Embedding(input_dim=5000, output_dim=128, input_length=50))
model4.add(SpatialDropout1D(0.5))
model4.add(SimpleRNN(30, dropout=0.5, recurrent_dropout=0.5))
model4.add(Dense(32, activation='relu'))
model4.add(Dense(y.shape[1], activation='softmax'))
```

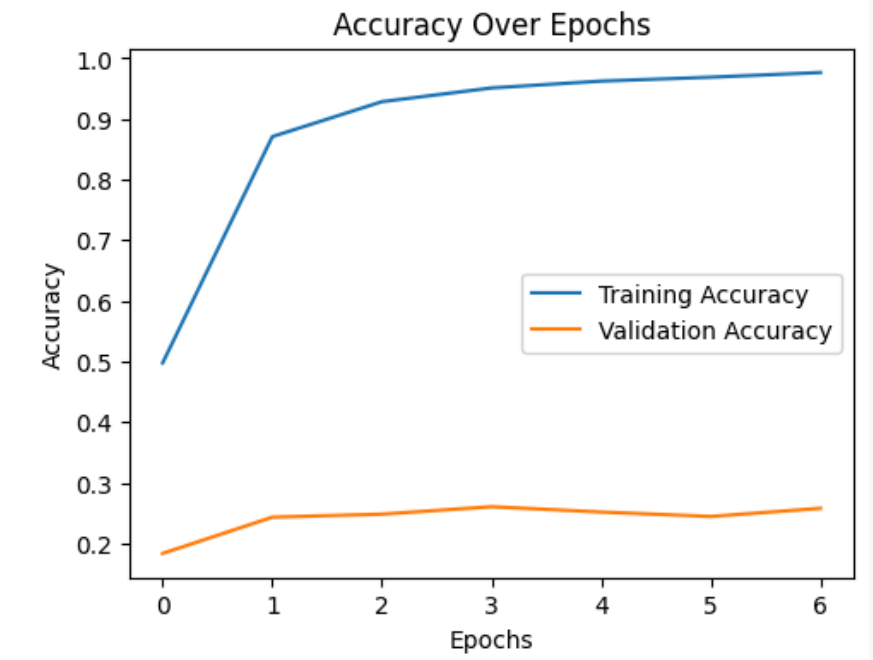
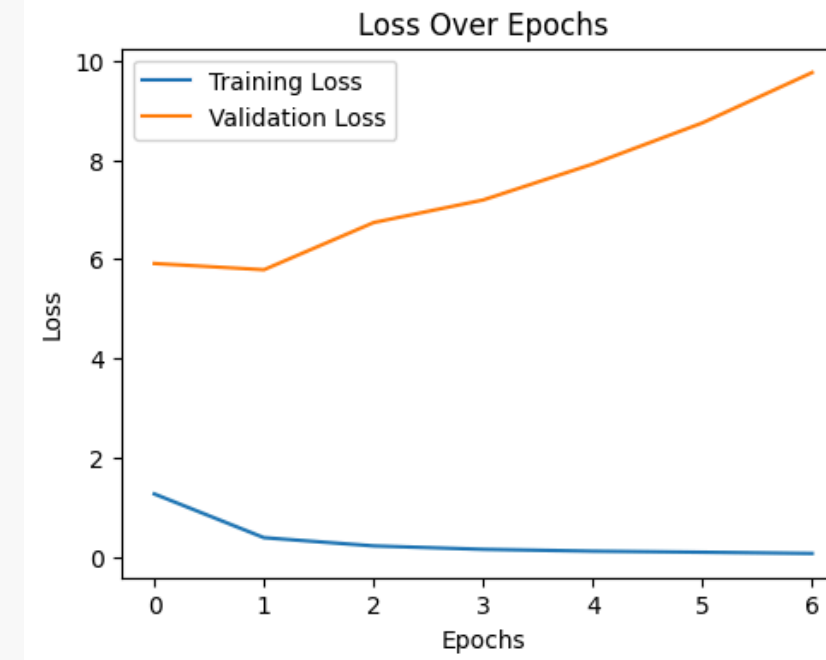
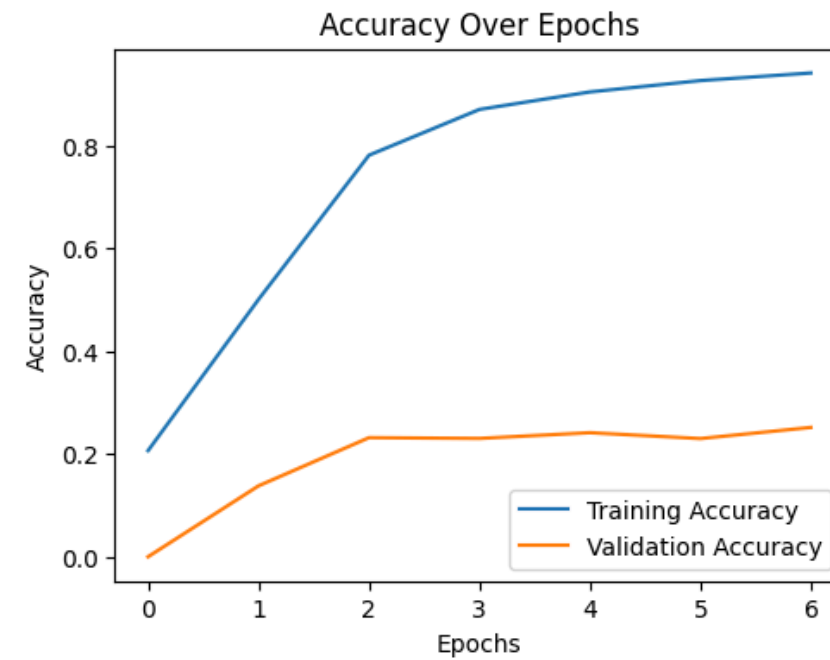
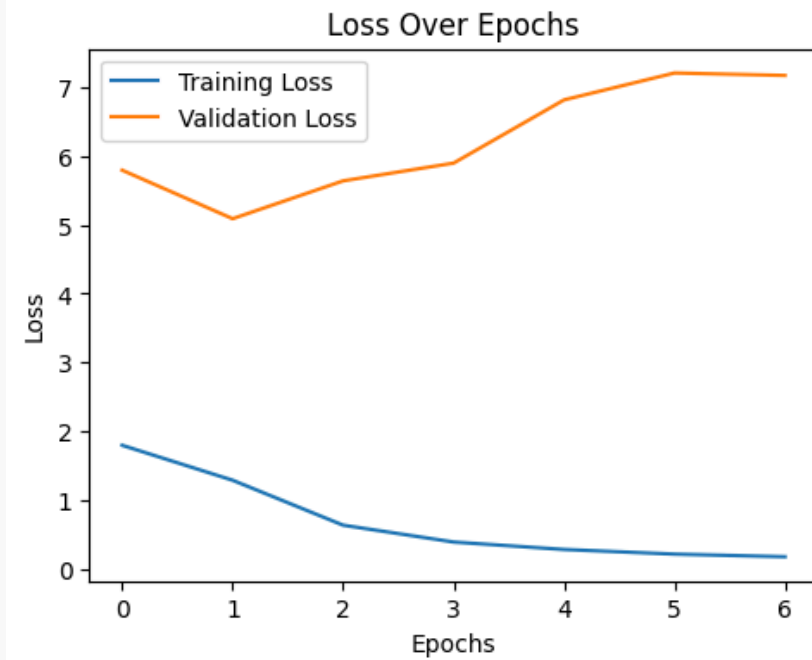
30 unit rnn, 32 unit dense

Untuk model LSTM juga sama arsitekturnya dengan SimpleRNN ini, dari model 1 - 4, hanya mengganti nama

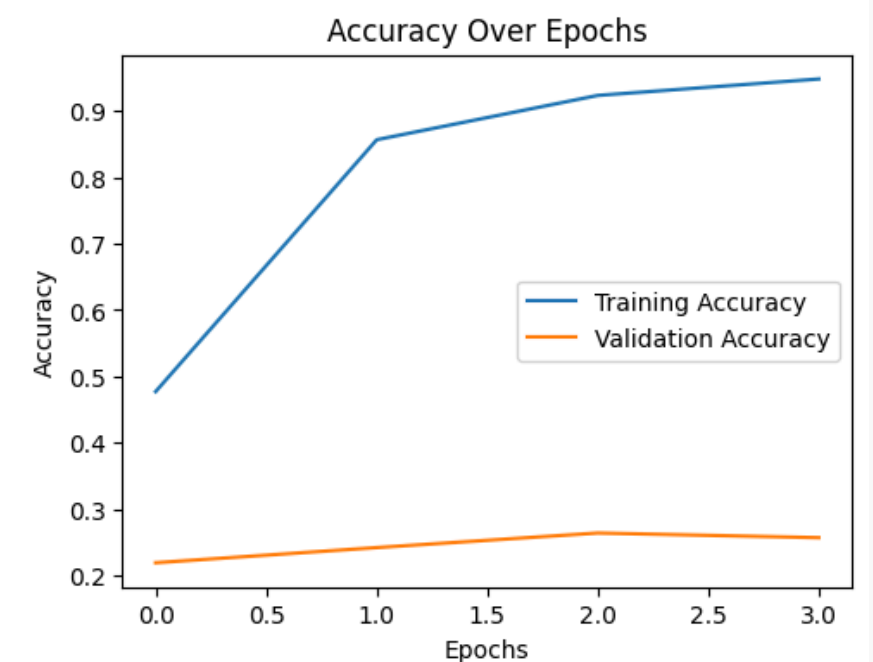
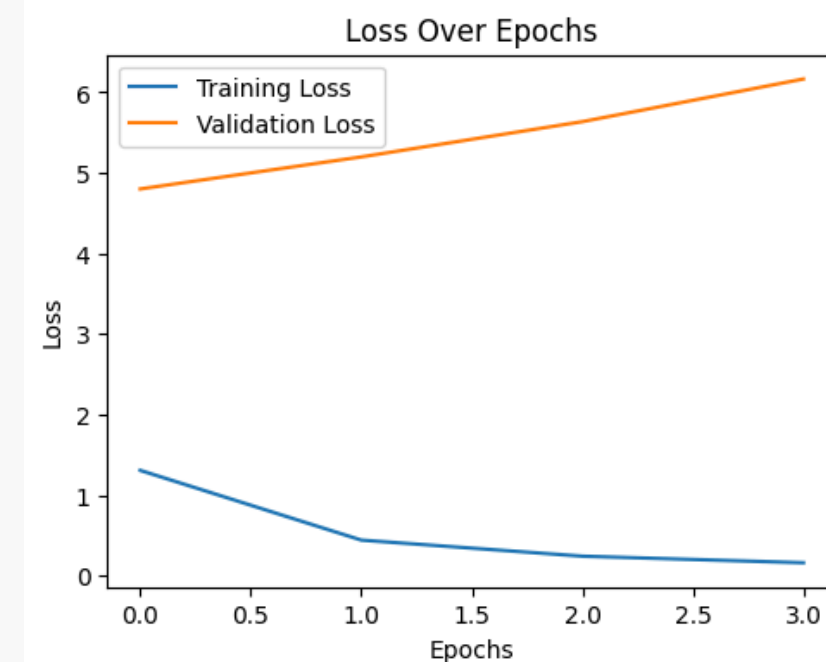
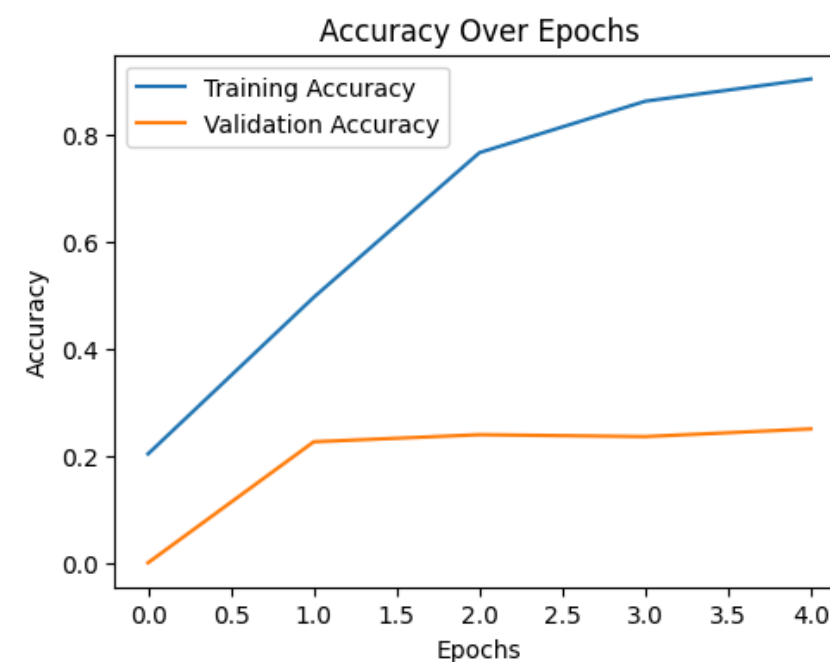
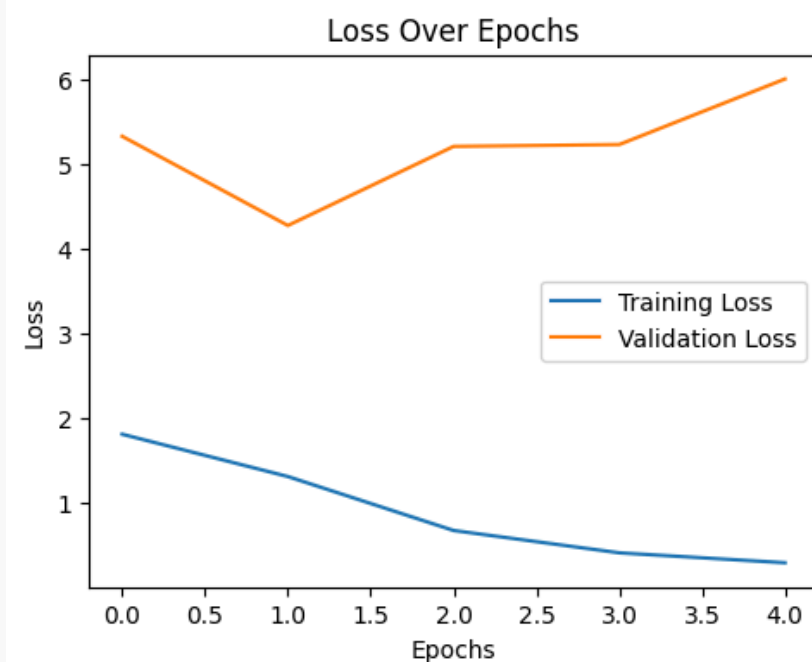
Hasil Perbandingan Training RNN dan LSTM

RNN = kiri, LSTM = kanan

1.

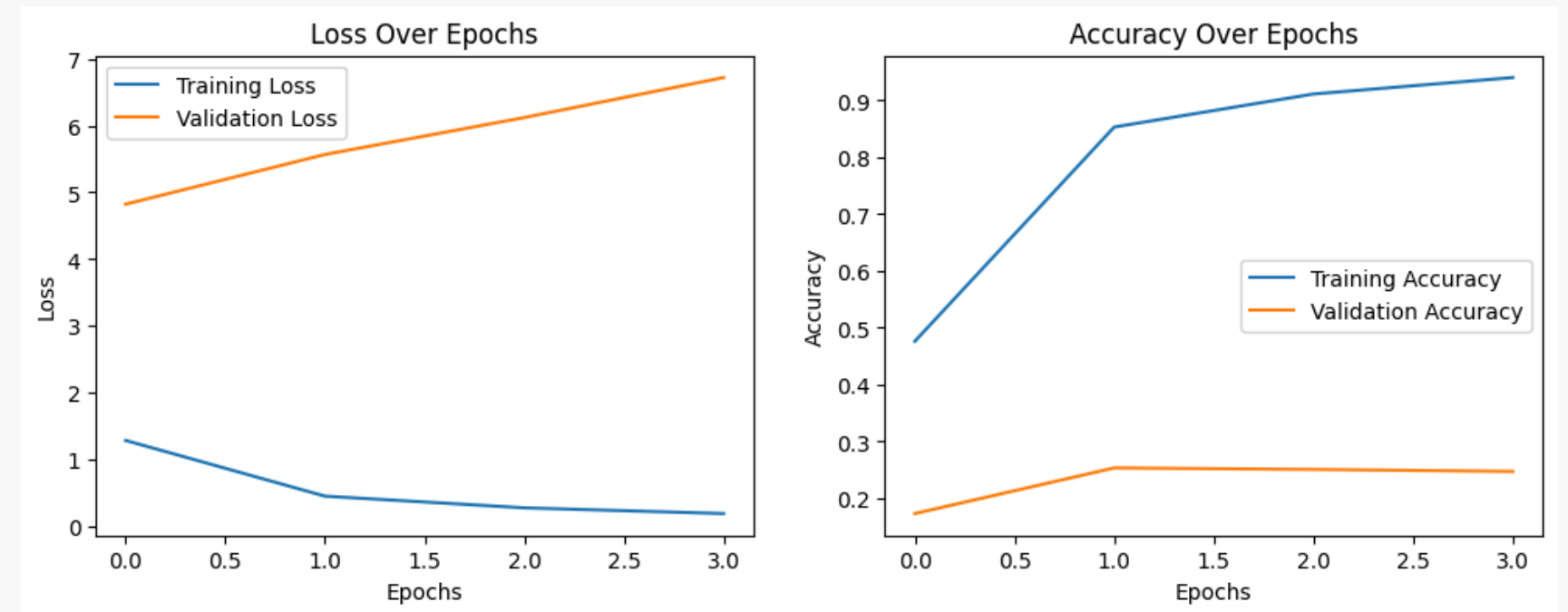
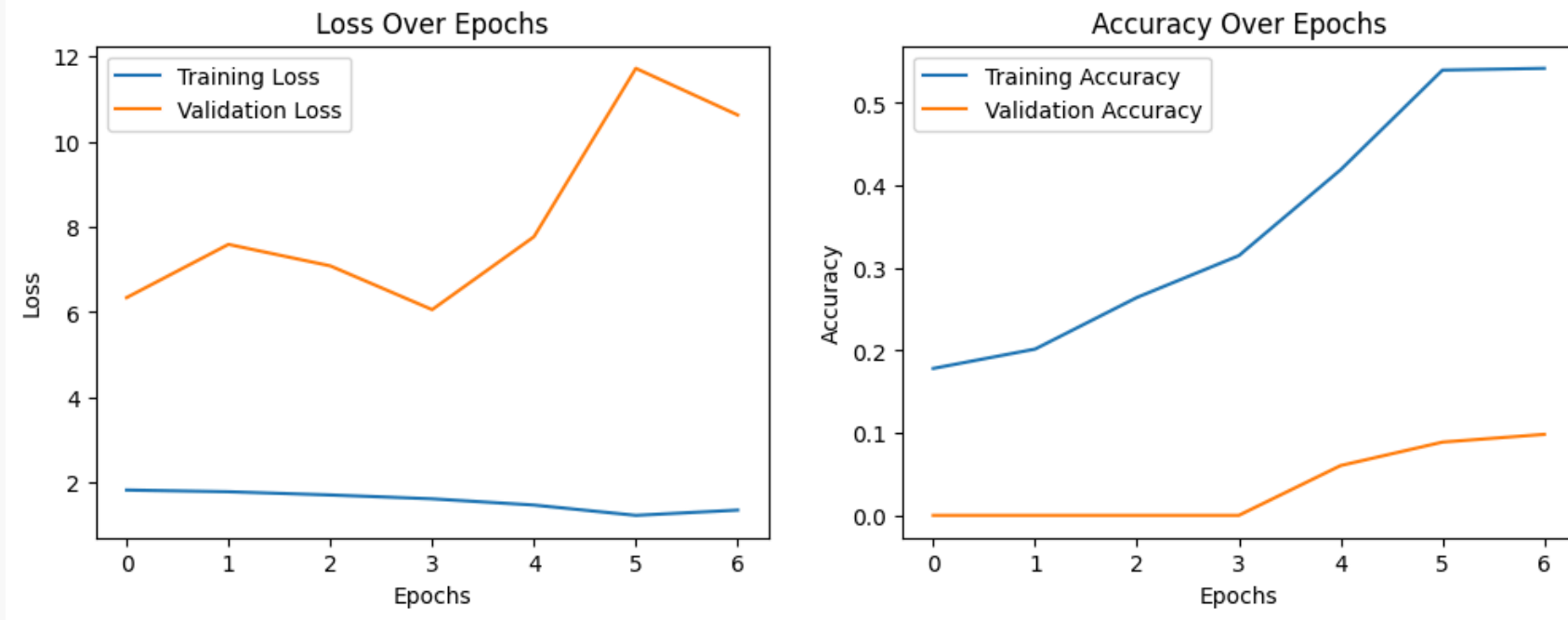


2.

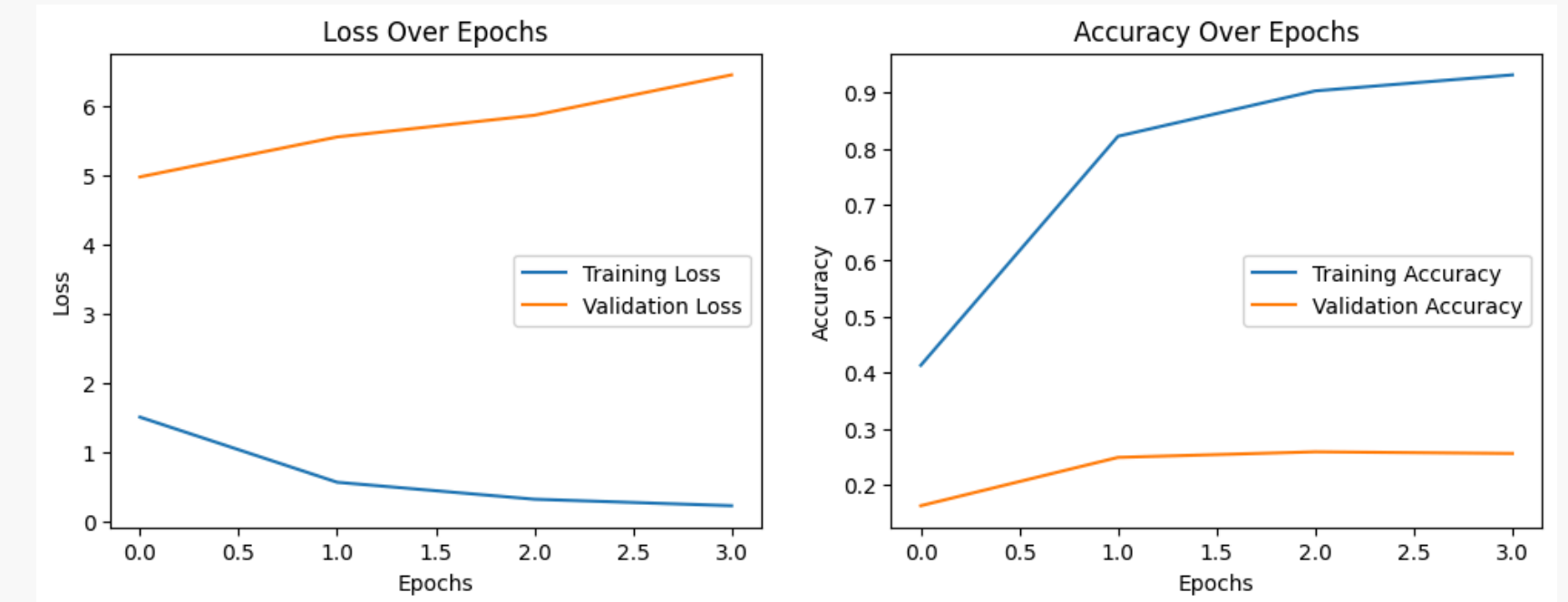
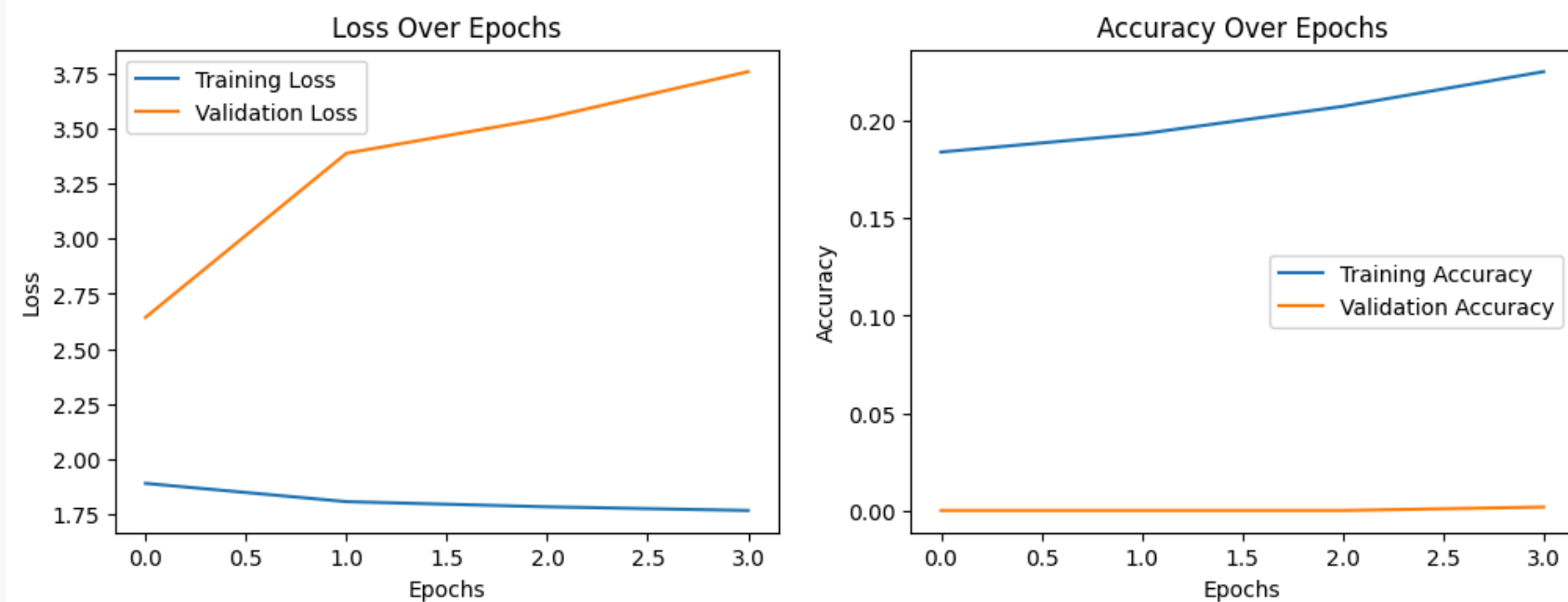


Hasil Perbandingan Training RNN dan LSTM

3.



4.



Testing

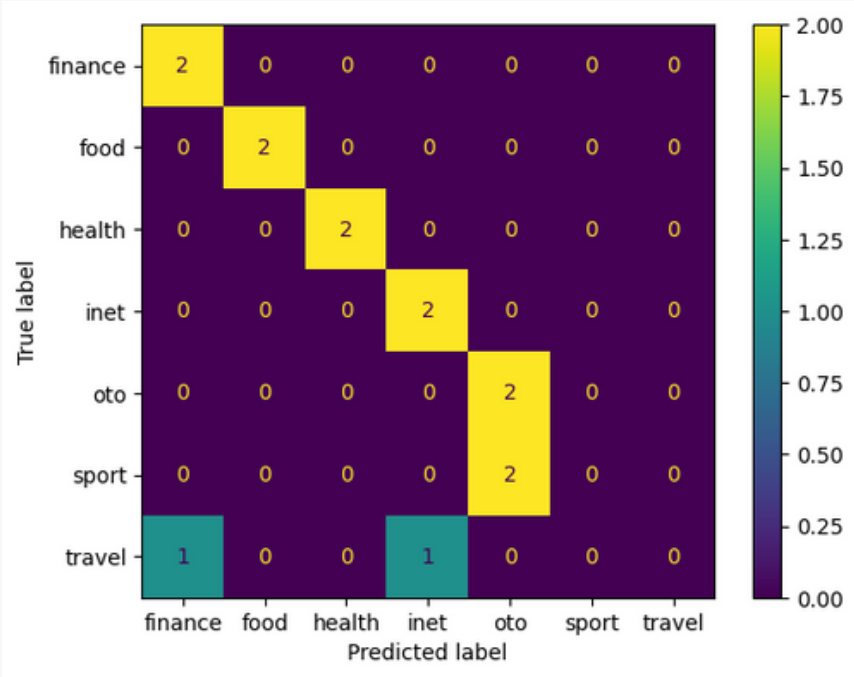
Data tes menggunakan 14 data, masing-masing 2 data untuk setiap kategori

```
test_titles = [  
    "Maskapai penerbangan menawarkan diskon besar-besaran untuk liburan",  
    "Teknologi baru ini akan mengubah cara kita berinteraksi dengan dunia digital",  
    "Perusahaan otomotif meluncurkan mobil listrik terbaru dengan fitur canggih",  
    "Ahli kesehatan memberikan tips menjaga daya tahan tubuh di musim hujan",  
    "Restoran ini menawarkan hidangan lezat dengan bahan-bahan lokal",  
    "Tim bulu tangkis Indonesia sukses meraih medali emas di kejuaraan dunia",  
    "Startup lokal mendapatkan pendanaan untuk proyek pengembangan teknologi ramah lingkungan",  
    "Menjelajahi Keindahan Alam Indonesia, Destinasi Terbaik untuk Liburan Akhir Tahun",  
    "Tren Digital 2024: Inovasi Teknologi yang Akan Mengubah Dunia Internet",  
    "Mengenal Mobil Listrik: Teknologi Ramah Lingkungan yang Semakin Populer",  
    "Cara Menjaga Kesehatan Mental di Tengah Kesibukan Sehari-hari",  
    "Resep Makanan Sehat yang Mudah Dibuat di Rumah, Cocok untuk Menjaga Imunitas",  
    "Persaingan Ketat di Piala Dunia 2024: Tim Favorit dan Pemain Kunci",  
    "Tips Mengatur Keuangan Pribadi agar Lebih Stabil di Tahun 2024"  
]  
  
true_labels = ['travel', 'inet', 'oto', 'health', 'food', 'sport', 'finance', 'travel', 'inet', 'oto', 'health', 'food', 'sport', 'finance']
```

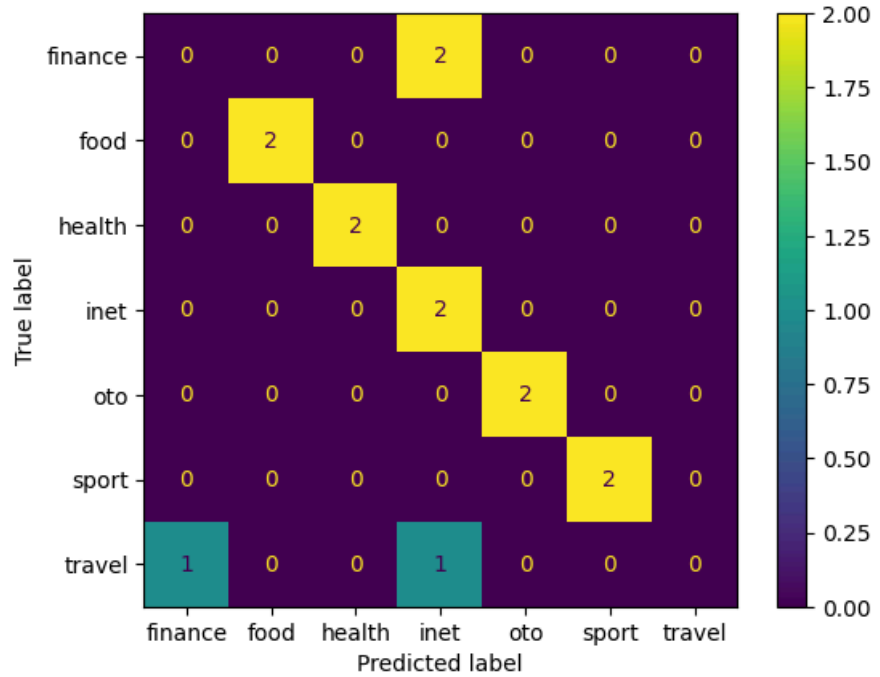
Hasil Testing RNN dan LSTM

rnn = kiri, lstm = kanan

1.

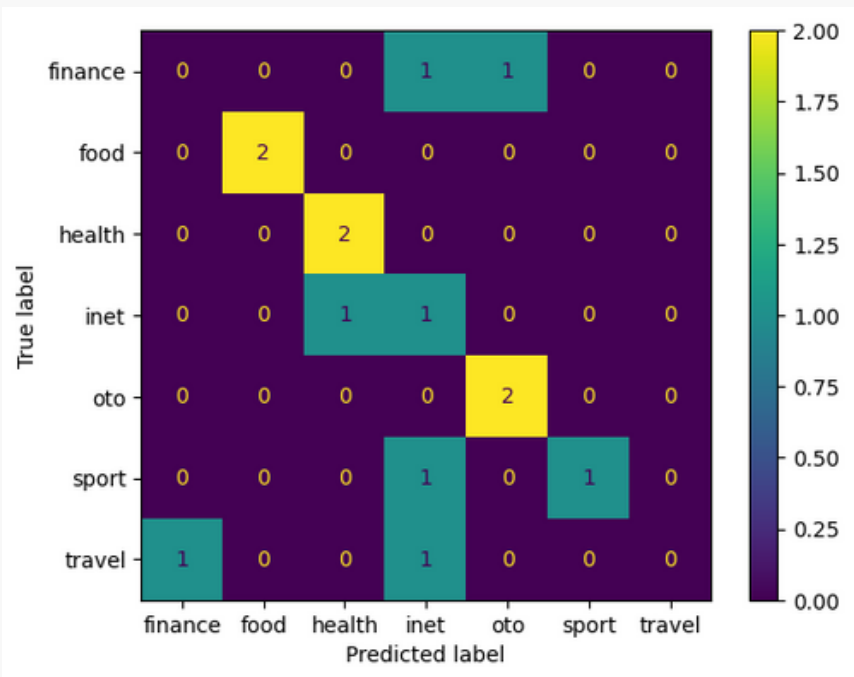


Akurasi pada data test: 71.43%

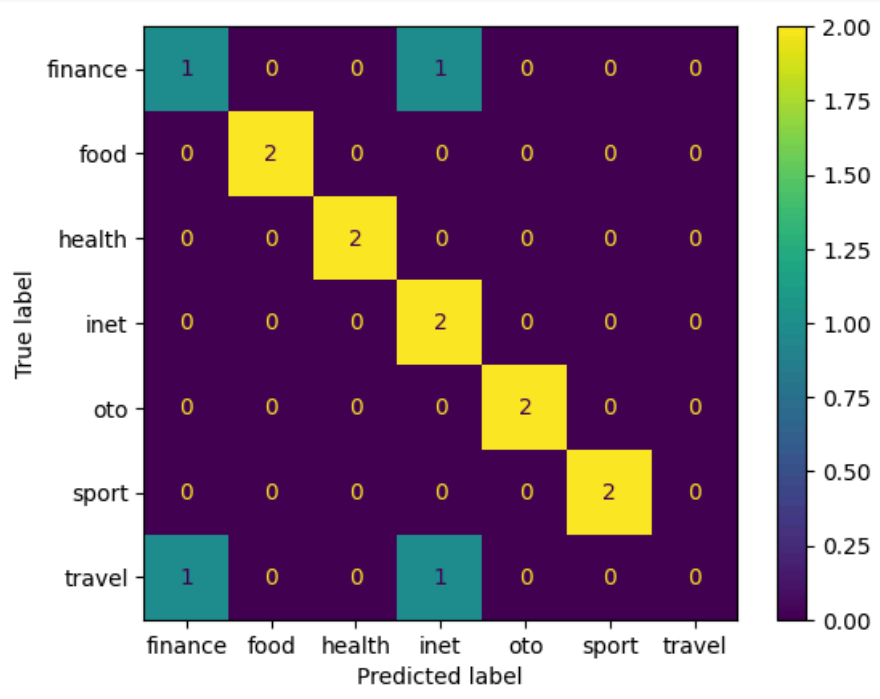


Akurasi pada data test: 71.43%

2.



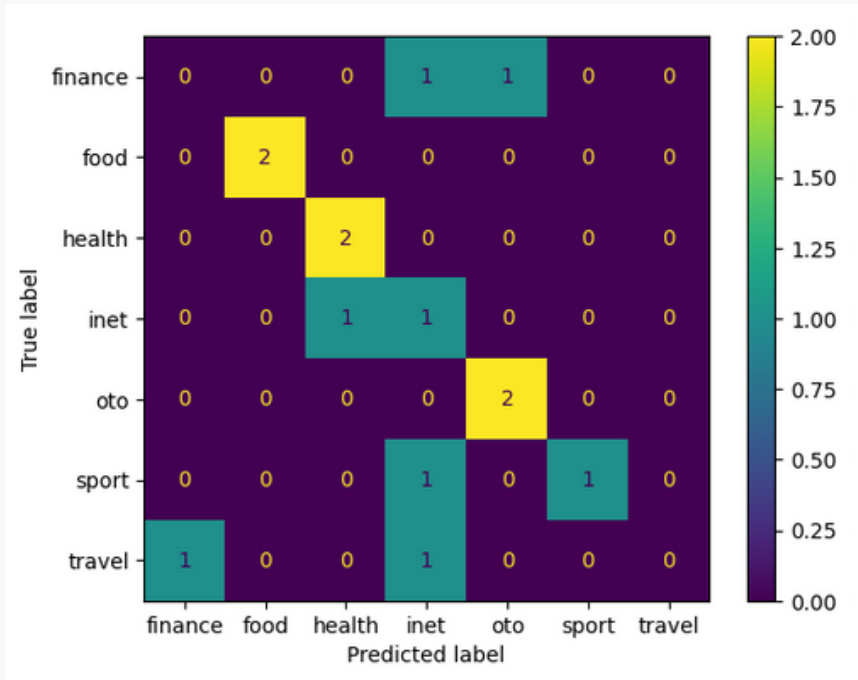
Akurasi pada data test: 57.14%



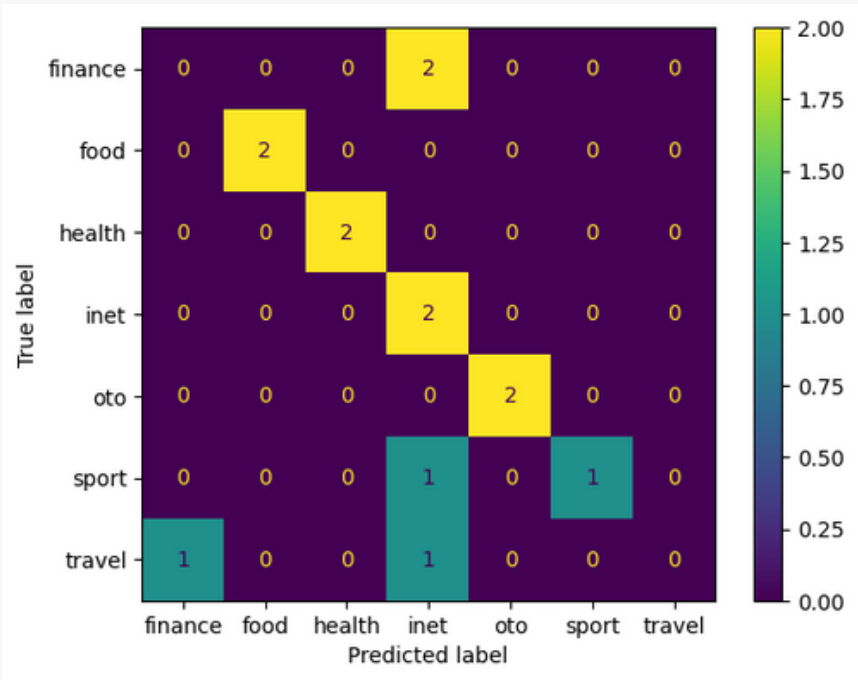
Akurasi pada data test: 78.57%

Hasil Testing RNN dan LSTM

3.

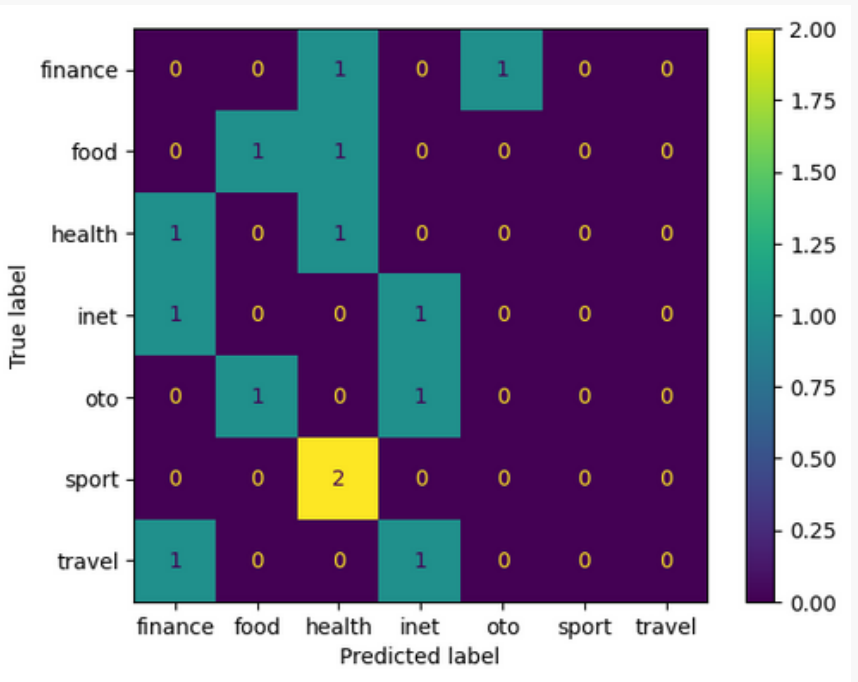


Akurasi pada data test: 57.14%

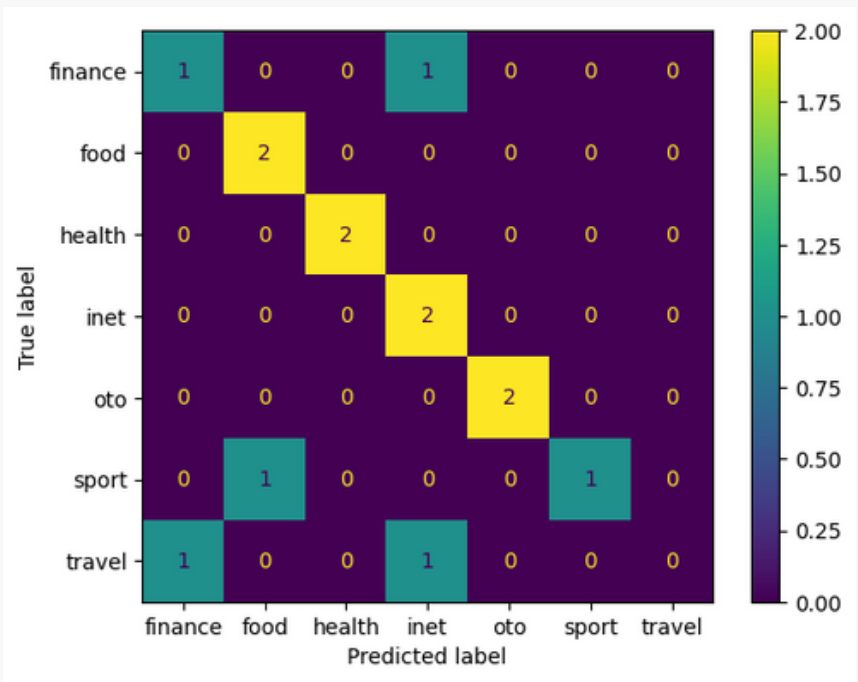


Akurasi pada data test: 64.29%

4.



Akurasi pada data test: 21.43%



Akurasi pada data test: 71.43%

Parafrase



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 trainnya** 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

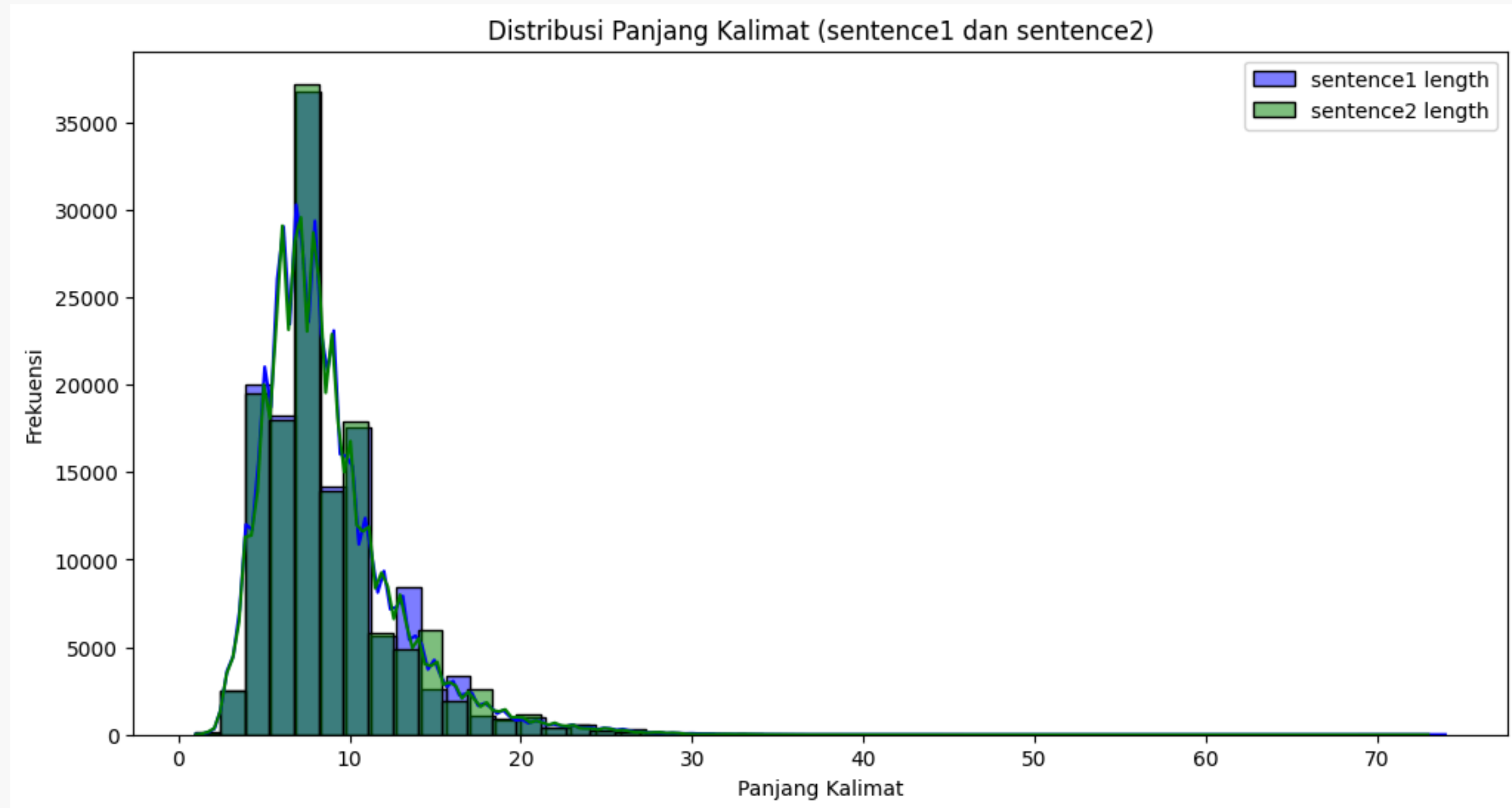
```
df.info()
```

```
<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...





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**, dan **penentuan parameter**

```
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')
```

Pembuatan Model

ini merupakan model LSTM, untuk model RNN juga sama, hanya mengganti nama

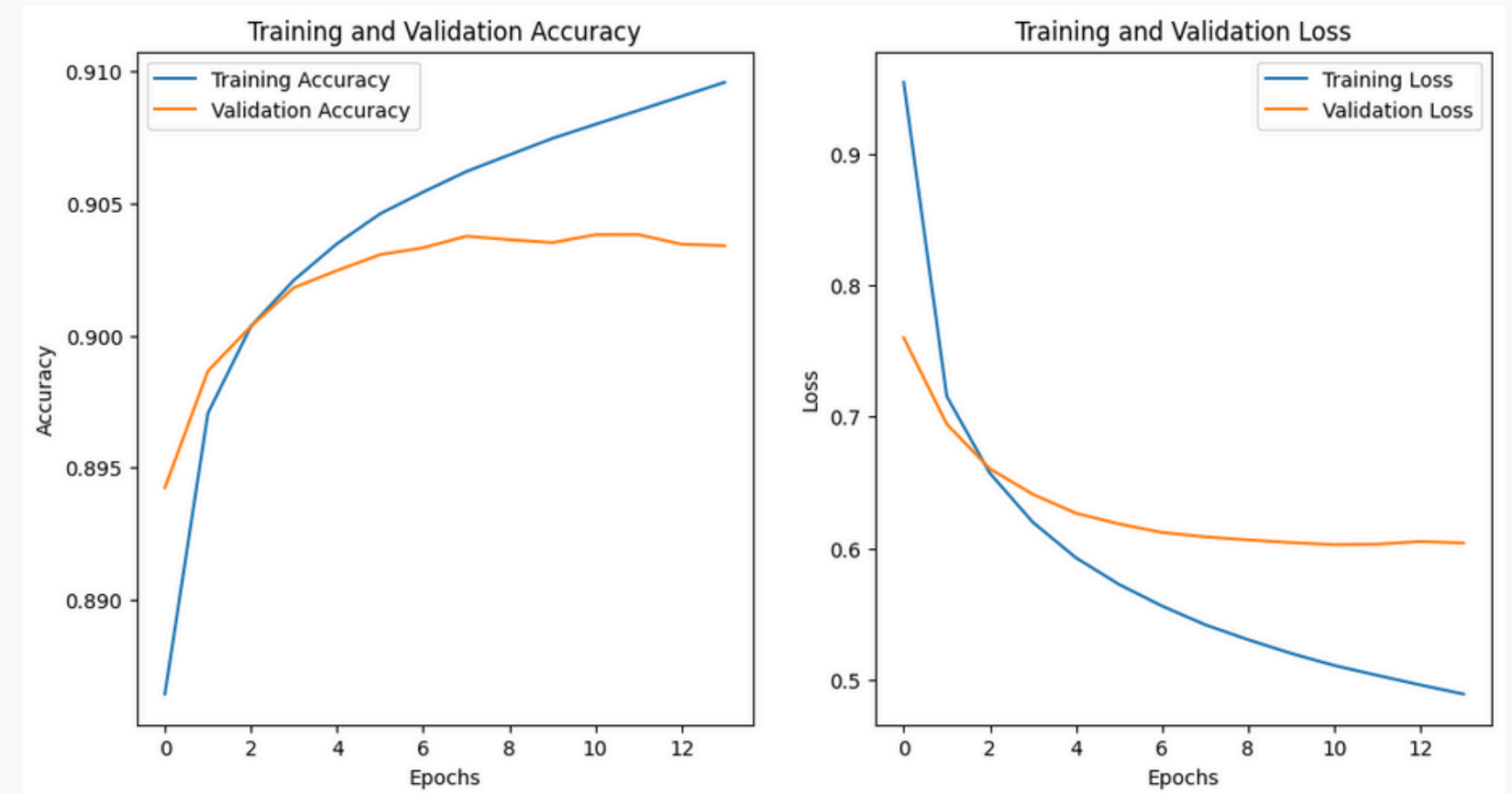
```
# Compile Model
model.compile(optimizer=Adam(learning_rate=0.001), loss='sparse_categorical_crossentropy', metrics=['accuracy'])

checkpoint = ModelCheckpoint(
    filepath='best_model_LSTM_{val_accuracy:.4f}.keras', # Filepath to save the model
    monitor='val_accuracy', # Metric to monitor
    save_best_only=True, # Save only the best model
    mode='max', # 'max' because we want the highest val_accuracy
    verbose=1
)
# Early Stopping untuk mencegah overfitting
early_stopping = EarlyStopping(monitor='val_loss', patience=3, restore_best_weights=True)

# Train the Model
history = model.fit(X_padded, y_padded, epochs=50, batch_size=32, validation_split=0.2, callbacks=[early_stopping, checkpoint])
```

Hasil Training

RNN = kiri, LSTM = kanan



Testing

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

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

RNN, BLEU Score: 9.283142785759642e-155

```
bagaimana cara bisa jaringan jaringan
```

LSTM, BLEU Score: 1.1862800137389335e-154

```
bagaimana cara bisa jaringan komputer
```

Testing

```
input_padded = "sehat itu mahal"
```

```
reference = "kesehatan itu mahal"
```

RNN, BLEU Score: 1.2882297539194154e-231

```
apa untuk mengapa itu
```

LSTM, BLEU Score: 2.4655814830110698e-232

```
kesehatan
```

Testing

```
input_padded = "kenapa kita harus belajar"
```

```
reference = "mengapa kita harus belajar"
```

RNN, BLEU Score: 0

```
apa untuk ekonomi penting
```

LSTM, BLEU Score: 9.918892480173173e-232

```
kesehatan harus antara
```

Testing

```
datatest = {
  'question': [
    "Apa manfaat yoga untuk kesehatan?",
    "Bagaimana cara membuat kue coklat?",
    "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 coklat?",
    "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?"
  ]
}
```


Testing LSTM

1/1 ————— 0s 142ms/step

Pertanyaan: Apa manfaat yoga untuk kesehatan?

Parafrase yang diprediksi: apa informasi tubuh untuk

Referensi: Apa saja manfaat yoga bagi kesehatan tubuh?

BLEU score: 0.0451

1/1 ————— 0s 23ms/step

Pertanyaan: Bagaimana cara membuat kue coklat?

Parafrase yang diprediksi: apa untuk untuk kue kue kue kue

Referensi: Apa langkah-langkah untuk membuat kue coklat?

BLEU score: 0.0435

1/1 ————— 0s 17ms/step

Pertanyaan: Mengapa penting untuk tidur cukup?

Parafrase yang diprediksi: kesehatan dilakukan untuk tidur penting

Referensi: Mengapa tidur yang cukup sangat penting untuk tubuh?

BLEU score: 0.0388

1/1 ————— 0s 17ms/step

Pertanyaan: Apa perbedaan antara data dan informasi?

Parafrase yang diprediksi: apa jantung cara data data

Referensi: Apa yang membedakan antara data dan informasi?

BLEU score: 0.0428

1/1 ————— 0s 17ms/step

Pertanyaan: Apa yang harus dilakukan agar lebih produktif?

Parafrase yang diprediksi: apa yang antara untuk agar agar bagaimana

Referensi: Apa saja tips agar lebih produktif dalam bekerja?

BLEU score: 0.0341

1/1 ————— 0s 17ms/step

Pertanyaan: Bagaimana cara menjaga kesehatan jantung?

Parafrase yang diprediksi: apa untuk untuk menjaga menjaga

Referensi: Apa yang harus dilakukan untuk menjaga kesehatan jantung?

BLEU score: 0.0690

1/1 ————— 0s 17ms/step

Pertanyaan: Apa saja manfaat belajar bahasa asing?

Parafrase yang diprediksi: apa saja mengapa bagaimana

Referensi: Apa keuntungan yang didapatkan dari belajar bahasa asing?

BLEU score: 0.0296

1/1 ————— 0s 22ms/step

Pertanyaan: Mengapa olahraga penting bagi tubuh?

Parafrase yang diprediksi: kesehatan olahraga olahraga bagi olahraga

Referensi: Apa alasan olahraga sangat penting untuk kesehatan tubuh?

BLEU score: 0.0351

1/1 ————— 0s 17ms/step

Pertanyaan: Apa itu ekonomi digital?

Parafrase yang diprediksi: apa informasi ekonomi digital

Referensi: Bagaimana ekonomi digital memengaruhi kehidupan kita?

BLEU score: 0.1031

1/1 ————— 0s 17ms/step

Pertanyaan: Apa yang dimaksud dengan kecerdasan buatan?

Parafrase yang diprediksi: apa yang dimaksud dimaksud kecerdasan untuk

Referensi: Apa itu AI dan bagaimana cara kerjanya?

BLEU score: 0.0346

Testing RNN

1/1 ————— 0s 20ms/step

Pertanyaan: Apa manfaat yoga untuk kesehatan?

Parafrase yang diprediksi: apa untuk cara untuk

Referensi: Apa saja manfaat yoga bagi kesehatan tubuh?

BLEU score: 0.0380

1/1 ————— 0s 17ms/step

Pertanyaan: Bagaimana cara membuat kue coklat?

Parafrase yang diprediksi: apa perbedaan cara cara coklat

Referensi: Apa langkah-langkah untuk membuat kue coklat?

BLEU score: 0.0428

1/1 ————— 0s 15ms/step

Pertanyaan: Mengapa penting untuk tidur cukup?

Parafrase yang diprediksi: ekonomi penting untuk tidur penting

Referensi: Mengapa tidur yang cukup sangat penting untuk tubuh?

BLEU score: 0.0690

1/1 ————— 0s 15ms/step

Pertanyaan: Apa perbedaan antara data dan informasi?

Parafrase yang diprediksi: apa perbedaan antara data dan

Referensi: Apa yang membedakan antara data dan informasi?

BLEU score: 0.1915

1/1 ————— 0s 16ms/step

Pertanyaan: Apa yang harus dilakukan agar lebih produktif?

Parafrase yang diprediksi: apa yang mengapa untuk untuk bagaimana bagaimana

Referensi: Apa saja tips agar lebih produktif dalam bekerja?

BLEU score: 0.0286

1/1 ————— 0s 18ms/step

Pertanyaan: Bagaimana cara menjaga kesehatan jantung?

Parafrase yang diprediksi: apa perbedaan cara bagaimana cara

Referensi: Apa yang harus dilakukan untuk menjaga kesehatan jantung?

BLEU score: 0.0295

1/1 ————— 0s 18ms/step

Pertanyaan: Bagaimana cara menjaga kesehatan jantung?

Parafrase yang diprediksi: apa perbedaan cara bagaimana cara

Referensi: Apa yang harus dilakukan untuk menjaga kesehatan jantung?

BLEU score: 0.0295

1/1 ————— 0s 17ms/step

Pertanyaan: Apa saja manfaat belajar bahasa asing?

Parafrase yang diprediksi: apa saja mengapa bagaimana

Referensi: Apa keuntungan yang didapatkan dari belajar bahasa asing?

BLEU score: 0.0296

1/1 ————— 0s 16ms/step

Pertanyaan: Mengapa olahraga penting bagi tubuh?

Parafrase yang diprediksi: ekonomi olahraga olahraga olahraga coklat

Referensi: Apa alasan olahraga sangat penting untuk kesehatan tubuh?

BLEU score: 0.0295

1/1 ————— 0s 17ms/step

Pertanyaan: Apa itu ekonomi digital?

Parafrase yang diprediksi: apa saja ekonomi itu

Referensi: Bagaimana ekonomi digital memengaruhi kehidupan kita?

BLEU score: 0.0487

1/1 ————— 0s 15ms/step

Pertanyaan: Apa yang dimaksud dengan kecerdasan buatan?

Parafrase yang diprediksi: apa yang dimaksud cara untuk untuk

Referensi: Apa itu AI dan bagaimana cara kerjanya?

BLEU score: 0.0411

Perbandingan Lainnya

- LSTM memakan waktu lebih lama dalam training dibandingkan RNN untuk komputasi yang sama
- Keseluruhan akurasi lebih tinggi LSTM



Thank you

