Pacmann

Fake News Prediction

Kelas Deep Learning
Juniarto Kurniawan - Batch 9

© 2022 – Pacmann AI

Outline

- Introduction/Background
- Workflow
- Conclusion
- References

© 2022 – Pacmann AI

Introduction

© 2022 – Pacmann AI

Introduction

Project ini dilakukan dengan tujuan untuk menjelaskan penerapan deep learning dalam deteksi berita palsu. Deep learning merupakan cabang dari kecerdasan buatan (artificial intelligence) yang menggunakan model neural networks dengan lapisan-lapisan (layers) yang kompleks untuk memahami dan memproses data. Kecanggihan deep learning dalam memahami pola-pola kompleks membuatnya menjadi pendekatan yang potensial dalam mengidentifikasi berita palsu yang sering kali tersusun dengan sangat cermat.

Dengan menggabungkan kemampuan deep learning dalam memproses data tekstual dan citra, penelitian ini berusaha memberikan solusi yang efektif dan efisien dalam meminimalkan penyebaran berita palsu. Melalui analisis mendalam terhadap dataset berita dan pengembangan model-model deep learning yang canggih, diharapkan dapat tercipta sistem deteksi berita palsu yang lebih akurat dan responsif terhadap dinamika informasi yang terus berkembang.

© 2022 – Pacmann AI

Workflow

Data Gathering

- Data Understanding
- Load Data

EDA

- Proportional
- WordCloud
- Text Cleansing

Data Preprocessing

- Splitting Data
- Text Preprocessing

Deep Learning

- Parameter Tuning
- Model Fitting

Evaluation

- Model Evaluation
- Model Prediction

© 2022 – Pacmann AI

Data Understanding

Data berasal dari berita yang terdapat pada platform Twitter/ X yang di peroleh dari Kaggle

	title	news_url	source_domain	tweet_num	real
0	Kandi Burruss Explodes Over Rape Accusation on...	http://toofab.com/2017/05/08/real-housewives-a...	toofab.com	42	1
1	People's Choice Awards 2018: The best red carp...	https://www.today.com/style/see-people-s-choic...	www.today.com	0	1
2	Sophia Bush Sends Sweet Birthday Message to 'O...	https://www.etonline.com/news/220806_sophia_bu...	www.etonline.com	63	1
3	Colombian singer Maluma sparks rumours of inap...	https://www.dailymail.co.uk/news/article-33655...	www.dailymail.co.uk	20	1
4	Gossip Girl 10 Years Later: How Upper East Sid...	https://www.zerchoo.com/entertainment/gossip-g...	www.zerchoo.com	38	1

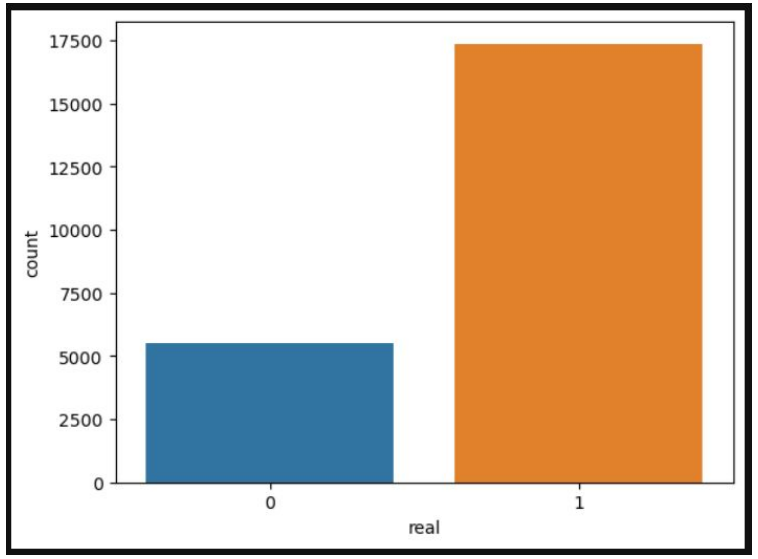
Langkah pertama yang dilakukan adalah Check Missing Value, berdasarkan analisa terdapat 330 data yang mempunyai nilai yang missing , oleh karena itu selanjutnya akan kita lakukan dropping data.

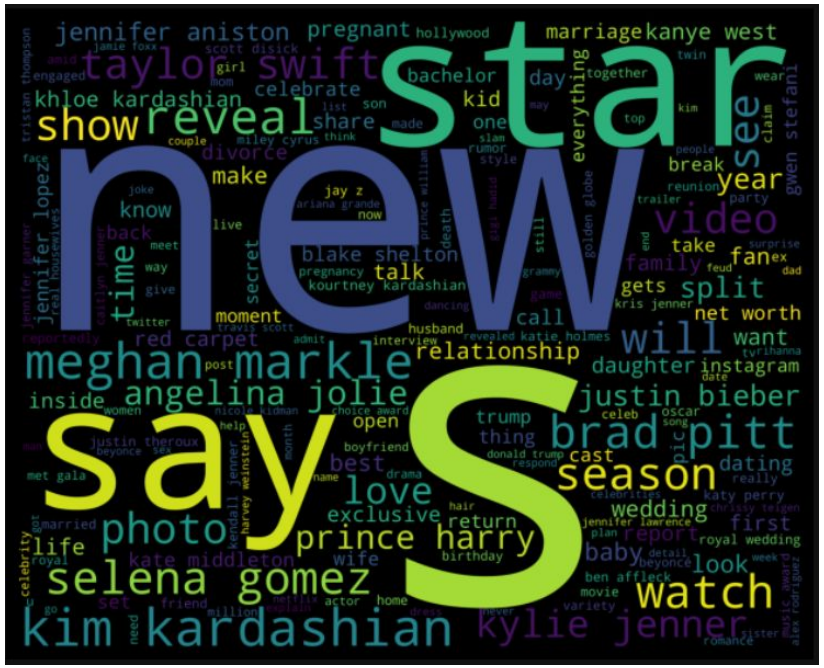
```
# Mengecek Missing Value
df.isna().sum()
```

```
title              0
news_url         330
source_domain    330
tweet_num          0
real               0
dtype: int64
```

```
# Drop Missing Value
data = df.dropna()
```

EDA





Berdasarkan grafik di atas dapat disimpulkan sebanyak 75,96% (atau 17.371 data) merupakan data berita yang terverifikasi kebenarannya sedangkan sisanya yaitu sebesar 24.04% atu 5.495 merupakan data yang terverifikasi tidak benar (Fake News).

Berdasarkan hasil WordCloud dapat diketahui bahwa Berita yang di analysis adalah Insudtri Entertainment Hollywood, dimana terdapat beberapa entertainer yang menjadi topik dalam pemberitaan.

© 2022 – Pacmann AI

Text Cleansing

Pada Text Cleansing, kita akan lakukan cleaning data pada judul berita dengan ketentuan menghilangkan huruf kapital pada awal kata dan btanda baca. Metode yang digunakan adalah WordNetLemmatizer, adalah bagian dari pustaka Natural Language Toolkit (NLTK) yang digunakan untuk melakukan lemmatisasi, suatu proses di bidang pemrosesan bahasa alami yang bertujuan untuk mengubah kata ke bentuk dasarnya (lemma).

	title	news_url	source_domain	tweet_num	real
0	kandi burruss explodes rape accusation real ho...	http://toofab.com/2017/05/08/real-housewives-a...	toofab.com	42	1
1	people choice award 2018 best red carpet look	https://www.today.com/style/see-people-s-choic...	www.today.com	0	1
2	sophia bush sends sweet birthday message one t...	https://www.etonline.com/news/220806_sophia_bu...	www.etonline.com	63	1
3	colombian singer maluma spark rumour inappropr...	https://www.dailymail.co.uk/news/article-33655...	www.dailymail.co.uk	20	1
4	gossip girl 10 year later upper east siders s...	https://www.zerchoo.com/entertainment/gossip-g...	www.zerchoo.com	38	1

Terlihat output pada judul sudahmengahsilkan pola yang sama, (huruf kapital mnejadi biasa dan tidak terdapat tanda baca), selanjutnya kita lakukan drop pada kolom lain karena yang digunakan sebagai predictor adalah kolom 'Title'

© 2022 – Pacmann AI

Data Preprocessing

Split Data

Kita akan melakukan Split data menjadi data Training dan Testing dengan komposisi 80% data training dan 20% data testing. Data Training sebanyak 80% itu akan kita split kembali menjadi 70% data training dan 30% data validation.

Text Processing

Pada Text Processing Part 2 ini kita lakukan konversi text menjadi indeks kata dengan Keras menggunakan Tokenizer Text_to_Sequences, dimana Keras ini biasanya digunakan bersamaan dengan tokenizer, yang dapat mengonversi teks menjadi indeks kata (token). texts_to_sequences kemudian mengambil hasil tokenisasi tersebut dan mengonversinya menjadi urutan bilangan bulat sesuai dengan indeks kata tersebut.

```
from tensorflow.keras.preprocessing.text import Tokenizer

tokenizer = Tokenizer(num_words=vocab_size, oov_token=oov_tok)
tokenizer.fit_on_texts(X_train)

word_index = tokenizer.word_index
```

```
print(X_train[1])
print(X_train_seq[1])
```

```
people choice award 2018
[49, 1588, 46, 13]
```


© 2022 – Pacmann AI

Deep Learning

Parameter Tuning

Model yang aaka digunakan adala LSTM (Long Short-Term Memory) dengan langkah :

a. Menentukan arsitektur model berupa sequential karena kita memiliki satu input dan satu output. Dalam konteks penggunaan model LSTM pada deteksi berita palsu, model Sequential digunakan untuk mendefinisikan arsitektur model secara berlapis.
b. Selanjutnya adalah menentukan lapisan Embedding yang mengubah kata-kata menjadi vektor angka (embedding).
c. Layer Selanjutnya adalah menentukan model klasifikasi dengan menggunakan metode Bidirectional. Bidirectional sering digunakan dalam tugas pemrosesan bahasa alami seperti klasifikasi teks, terjemahan mesin, dan lainnya, di mana konteks global dan hubungan antar kata sangat penting untuk pemahaman yang akurat.
d. Layer Output berupa Dense dimana terdapat hidden layer dengan aktivasi RelU yang membantu model untuk mempelajari representasi yang lebih kompleks dari data.
e. Layer terakhir yaitu layer Dense dengan activation Sigmoid yang mengonversi nilai input menjadi rentang antara 0 dan 1. Sigmoid umumnya digunakan di lapisan output untuk tugas klasifikasi biner, terutama ketika output model diinterpretasikan sebagai probabilitas kelas positif.

```
# The maximum number of words to be used. (most frequent)

vocab_size = len(counter)
embedding_dim = 32

# Model Definition with LSTM

model = tf.keras.Sequential([
    tf.keras.layers.Embedding(vocab_size, embedding_dim, input_length=max_length),
    tf.keras.layers.Bidirectional(tf.keras.layers.LSTM(64)),
    tf.keras.layers.Dense(14, activation='relu'),
    tf.keras.layers.Dense(1, activation='sigmoid')  # remember this is a binary clasification
])
```

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 20, 32)            532768

 bidirectional (Bidirection  (None, 128)               49664
 al)

 dense (Dense)               (None, 14)                1806

 dense_1 (Dense)             (None, 1)                 15

=================================================================
Total params: 584253 (2.23 MB)
Trainable params: 584253 (2.23 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```


© 2022 – Pacmann AI

Deep Learning

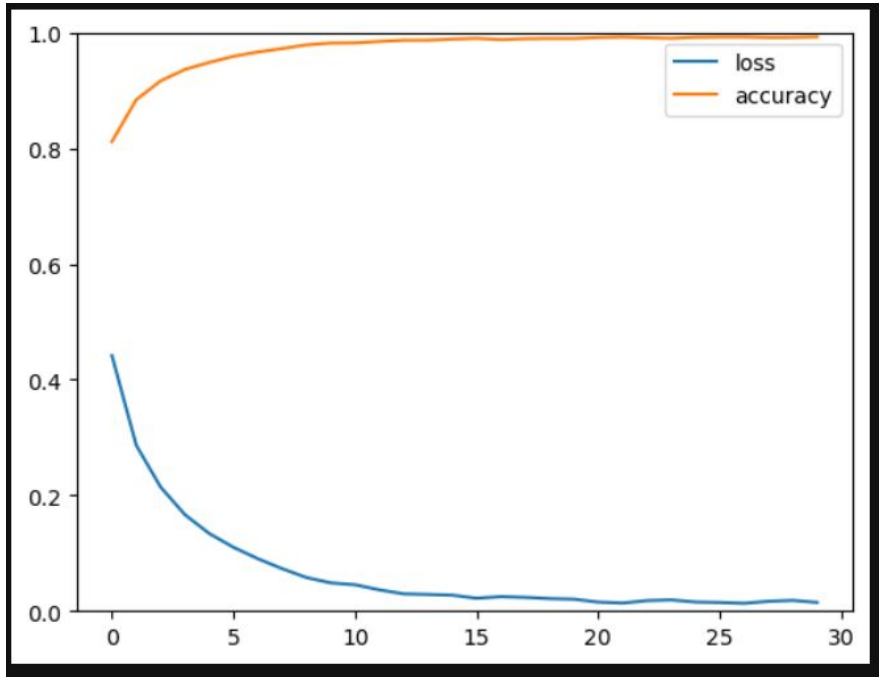
Model Fitting

a. Tentukan fungsi kerugian (loss function) : Menggunakan binary_crosentropy yang digunakan dalam tugas klasifikasi biner pada model jaringan saraf. Fungsi ini cocok digunakan ketika kita memiliki dua kelas yang saling eksklusif (misalnya, kelas positif dan kelas negatif).
b. optimizer : menggunakan Adam. Adam adalah salah satu optimizer yang efisien dan sering menjadi pilihan default karena kinerjanya yang baik dalam banyak kasus.
c. metrik evaluasi menggunakan Accuracy
d. Latih model menggunakan data train.

```
# Menentukan Parameter
model.compile(loss='binary_crossentropy',optimizer='adam',metrics=['accuracy'])
start_time = time.time()
num_epochs = 30

# Model Fit
history = model.fit(X_train_pad, y_train, epochs=num_epochs, validation_data=(X_val_pad, y_val))

# Output Time
final_time = (time.time()- start_time)/60
print(f'The time in minutos: {final_time}')
```



Berdasarkan Hasil Sequencing Menggunakan Deep Learning, dapat dilihat bahwa fungsi Loss mengalami penurunan serta Accuracy mengalami peningkatan mendekati 1, yang artinya model ini cukup baik.

© 2022 – Pacmann AI

Evaluation

Model Evaluation

```
#Lakukan Evaluasi Model pada data Validation
evaluation = model.evaluate(X_val_pad, y_val)
print(f'Accuracy: {evaluation[1]*100:.2f}%')
```

```
172/172 [==============================] - 2s 10ms/step - loss: 1.8155 - accuracy: 0.7828
Accuracy: 78.28%
```

Model Prediction

```
# Lakukan Pada Data Baru yaitu data Test Untuk Meprediksi
predictions = (model.predict(X_test_pad) > 0.5).astype("int32")
predictions
```

```
143/143 [==============================] - 3s 9ms/step
```

```
from sklearn.metrics import confusion_matrix, accuracy_score
print(f'Accuracy: {accuracy_score(y_test,y_pred2)*100:.2f}%')
```

```
Accuracy: 77.55%
```

Accuracy

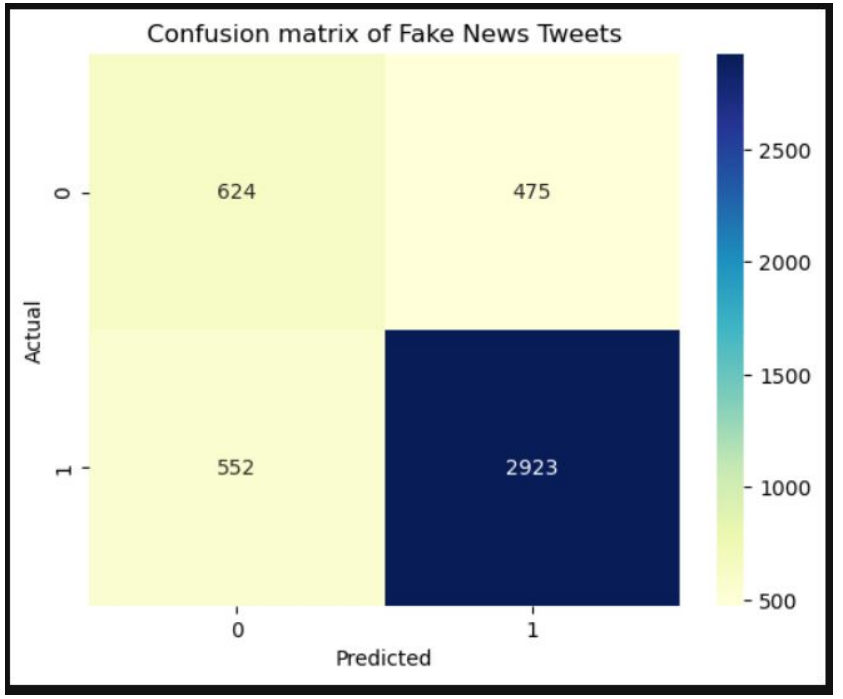
© 2022 – Pacmann AI

Evaluation

Accuracy

```
import sklearn.metrics as metrics
cm = metrics.confusion_matrix(y_test, y_pred2)
print(classification_report(y_test, y_pred2))
```

	precision	recall	f1-score	support
0	0.53	0.57	0.55	1099
1	0.86	0.84	0.85	3475
accuracy			0.78	4574
macro avg	0.70	0.70	0.70	4574
weighted avg	0.78	0.78	0.78	4574



Hasil Accuracy pada model evaluation dan Prediction menunjukkan nilai yang tidak jauh berbeda Accuracy Validation (77.26%) sedangkan Accuracy pada prediction (75.93%) hal ini menunjukkan bahwa model yang telah dibuat mampu melakukan prediksi dengan baik pada data train dan dapat memgeneralisasi dengan baik pada data baru.

© 2022 – Pacmann AI

Conclusion & Recomendation

Conclusion:

- Metode Deep Learning sangat baik digunakan dalam memproses data berupa text pada Fake News Detection
- Model yang digunakan yaitu LSTM dengan parameter yang ditentukan menghasilkan akurasi dan loss yang baik
- Setelah Dilakukan training, baik pada data validation maupun data testing menghasilkan akurasi yang cukup sama.

Recommendation:

- Perbedaan Metode dalam mengconvert teks menjadi numerik untuk dapat diolah, menghasilkan perbedaan hasil training. Oleh karena itu perlu dilakukan perbandingan beberapa analisa agar menghasilkan output yang lebih presisi.

© 2022 – Pacmann AI

Reference

https://www.kaggle.com/datasets/algord/fake-news

FakeNews Detection using ML and DL | Kaggle

Implementasi Deep Learning Menggunakan Metode CNN dan LSTM untuk Menentukan Berita Palsu dalam Bahasa Indonesia—Neliti

Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras—MachineLearningMastery.com

© 2022 – Pacmann AI

Thank You

© 2022 – Pacmann AI