NEWS RECOMMENDER SYSTEM

Natural Language Processing

Tony Eko Yuwono - 13518030
Rifaldy Aristya Kelana - 13518082
Kevin Austin Stefano - 13518104

MENU ANALYSIS CONTACT

NEWS RECOMMENDER SYSTEM

News recommender system adalah sistem untuk melakukan proses rekomendasi berita berdasarkan kemiripan dari berita yang pernah dibaca dan topik dari berita sebelumnya

PROSES YANG DILAKUKAN

KEYWORD EXTRACTION

Setiap token pada berita akan dilabeli menjadi keyword atau bukan keyword. Output berupa kata kunci dari input berita

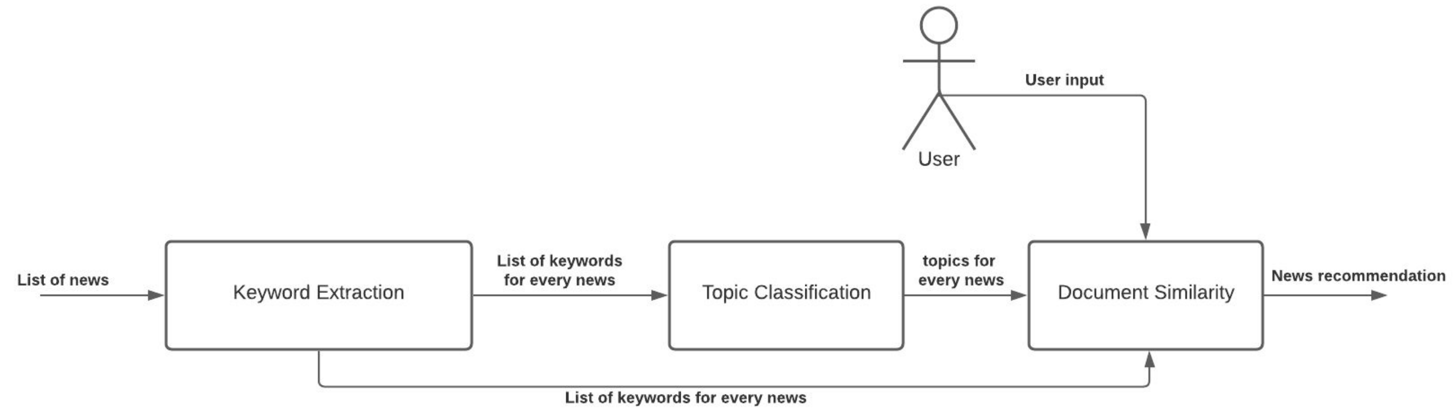
TOPIC CLASSIFICATION

List keyword akan dimasukkan ke dalam model klasifikasi untuk mengubah masukkan keyword berita menjadi label topik

DOCUMENT SIMILARITY

Kata kunci dari tiap berita yang diperoleh dari keyword extraction akan digunakan untuk memodelkan kemiripan dokumen berita lainnya yang berada dalam satu topik

menghasilkan rekomendasi berita



KOMPONEN DALAM NRS

TABLE OF CONTENT

01 PIPELINE 1

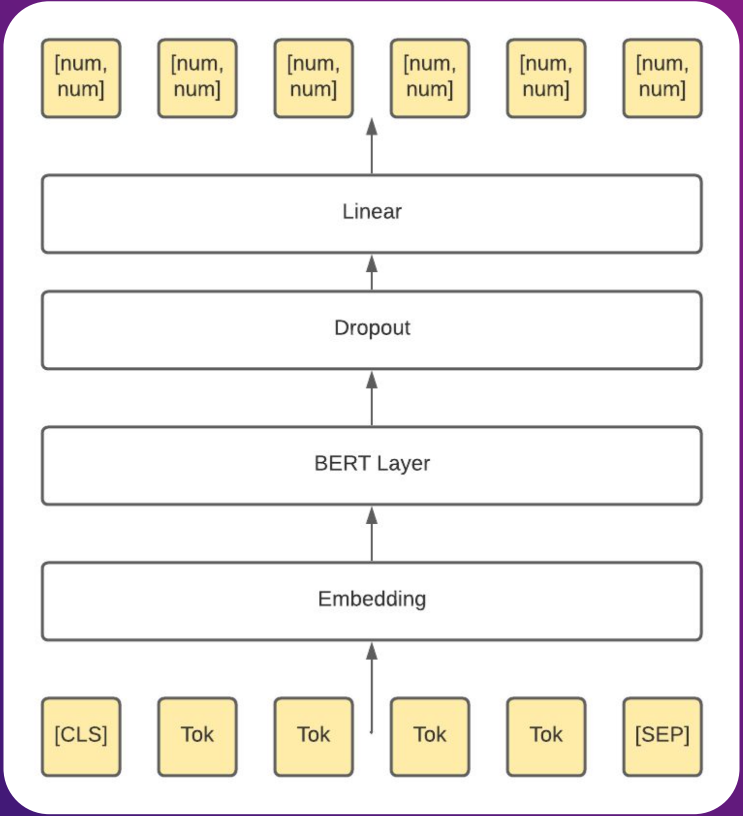
KEYWORDS EXTRACTION

Proses mengekstrak keyword dari berita

ARSITEKTUR UMUM

Arsitektur model yang digunakan adalah model berbasis BERT yang dipasangi fully connected layer pada layer setelah layer attention (layer attention terletak setelah layer embedding) yang menghasilkan nilai untuk setiap token berupa list dua angka, yaitu untuk kelas 0 (bukan keyword) dan kelas 1 (keyword).

Setelah itu, kita akan mencari nilai terbesar dari keduanya untuk menentukan kelas dari masing-masing token.



EKSPERIMEN

PERBANDINGAN METODE

Eksperimen yang dilakukan adalah dengan mengganti pretrained bert model yang dipakai yaitu bert-based-uncased dan distilbert-base-uncased. Juga mengganti jumlah epoch dari training

HASIL EKSPERIMEN

distilbert-base-uncased

```
======== Epoch 1 / 4 ========
Training...
  Batch    40  of    276.    Elapsed: 0:00:53.
  Batch    80  of    276.    Elapsed: 0:01:46.
  Batch   120  of    276.    Elapsed: 0:02:39.
  Batch   160  of    276.    Elapsed: 0:03:31.
  Batch   200  of    276.    Elapsed: 0:04:24.
  Batch   240  of    276.    Elapsed: 0:05:17.

  Average training loss: 0.30
  Training epcoh took: 0:06:04

Running Validation...
  Accuracy: 0.96
  Validation Loss: 0.27
  Validation took: 0:00:16
```

```
======== Epoch 2 / 4 ========
Training...
  Batch    40  of    276.    Elapsed: 0:00:53.
  Batch    80  of    276.    Elapsed: 0:01:45.
  Batch   120  of    276.    Elapsed: 0:02:38.
  Batch   160  of    276.    Elapsed: 0:03:31.
  Batch   200  of    276.    Elapsed: 0:04:24.
  Batch   240  of    276.    Elapsed: 0:05:16.

  Average training loss: 0.26
  Training epcoh took: 0:06:04

Running Validation...
  Accuracy: 0.97
  Validation Loss: 0.26
  Validation took: 0:00:16
```

HASIL EKSPERIMEN

distilbert-base-uncased

```
======== Epoch 3 / 4 ========
Training...
  Batch    40  of    276.    Elapsed: 0:00:53.
  Batch    80  of    276.    Elapsed: 0:01:46.
  Batch   120  of    276.    Elapsed: 0:02:38.
  Batch   160  of    276.    Elapsed: 0:03:31.
  Batch   200  of    276.    Elapsed: 0:04:24.
  Batch   240  of    276.    Elapsed: 0:05:17.

  Average training loss: 0.24
  Training epcoh took: 0:06:04

Running Validation...
  Accuracy: 0.92
  Validation Loss: 0.25
  Validation took: 0:00:16
```

```
======== Epoch 4 / 4 ========
Training...
  Batch    40  of    276.    Elapsed: 0:00:53.
  Batch    80  of    276.    Elapsed: 0:01:46.
  Batch   120  of    276.    Elapsed: 0:02:38.
  Batch   160  of    276.    Elapsed: 0:03:31.
  Batch   200  of    276.    Elapsed: 0:04:24.
  Batch   240  of    276.    Elapsed: 0:05:17.

  Average training loss: 0.22
  Training epcoh took: 0:06:04

Running Validation...
  Accuracy: 0.87
  Validation Loss: 0.26
  Validation took: 0:00:16
```

HASIL EKSPERIMEN

distilbert-base-uncased

```
Training complete!
Total training took 0:25:18 (h:mm:ss)
```

Size:	253 MB (265.729.799 bytes)

HASIL EKSPERIMEN

bert-base-uncased

```
======== Epoch 1 / 4 ========
Training...
  Batch    40  of    276.    Elapsed: 0:01:38.
  Batch    80  of    276.    Elapsed: 0:03:16.
  Batch   120  of    276.    Elapsed: 0:04:55.
  Batch   160  of    276.    Elapsed: 0:06:33.
  Batch   200  of    276.    Elapsed: 0:08:11.
  Batch   240  of    276.    Elapsed: 0:09:49.

  Average training loss: 0.29
  Training epcoh took: 0:11:17

Running Validation...
  Accuracy: 0.98
  Validation Loss: 0.27
  Validation took: 0:00:29
```

```
======== Epoch 2 / 4 ========
Training...
  Batch    40  of    276.    Elapsed: 0:01:38.
  Batch    80  of    276.    Elapsed: 0:03:16.
  Batch   120  of    276.    Elapsed: 0:04:54.
  Batch   160  of    276.    Elapsed: 0:06:33.
  Batch   200  of    276.    Elapsed: 0:08:11.
  Batch   240  of    276.    Elapsed: 0:09:49.

  Average training loss: 0.25
  Training epcoh took: 0:11:17

Running Validation...
  Accuracy: 0.98
  Validation Loss: 0.26
  Validation took: 0:00:29
```

HASIL EKSPERIMEN

bert-base-uncased

```
======== Epoch 3 / 4 ========
Training...
  Batch    40  of    276.    Elapsed: 0:01:38.
  Batch    80  of    276.    Elapsed: 0:03:16.
  Batch   120  of    276.    Elapsed: 0:04:54.
  Batch   160  of    276.    Elapsed: 0:06:32.
  Batch   200  of    276.    Elapsed: 0:08:10.
  Batch   240  of    276.    Elapsed: 0:09:48.

  Average training loss: 0.23
  Training epcoh took: 0:11:16

Running Validation...
  Accuracy: 0.95
  Validation Loss: 0.26
  Validation took: 0:00:28
```

```
======== Epoch 4 / 4 ========
Training...
  Batch    40  of    276.    Elapsed: 0:01:36.
  Batch    80  of    276.    Elapsed: 0:03:13.
  Batch   120  of    276.    Elapsed: 0:04:49.
  Batch   160  of    276.    Elapsed: 0:06:26.
  Batch   200  of    276.    Elapsed: 0:08:02.
  Batch   240  of    276.    Elapsed: 0:09:38.

  Average training loss: 0.21
  Training epcoh took: 0:11:05

Running Validation...
  Accuracy: 0.89
  Validation Loss: 0.26
  Validation took: 0:00:28
```

HASIL EKSPERIMEN

bert-base-uncased

```
Training complete!
Total training took 0:46:49 (h:mm:ss)
```

Size: 415 MB (435.889.030 bytes)

ANALISIS HASIL EKSPERIMEN

Dari hasil eksperimen melakukan keyword extraction, dapat disimpulkan beberapa hal:

1. Perbandingan akurasi antara model yang menggunakan pretrained bert-based model dan distilbert-based tidak terlalu berbeda
2. Training untuk keduanya hanya perlu dilakukan sampai epoch kedua. Setelah epoch kedua, nilai loss dari validasi akan bertambah dan nilai akurasi dari validasi juga akan berkurang
3. Waktu yang diperlukan untuk melakukan fine-tuning bert-based model mencapai dua kali lipat distilbert-based model
4. Ukuran model sekaligus tokenizer yang dihasilkan oleh bert-based model dua kali lipat distilbert-based model

02 PIPELINE 2
TOPIK KLASIFIKASI

Proses melakukan klasifikasi topik berita

PROSES TOPIK KLASIFIKASI

1. PREPROCESSING

Melakukan preprocessing dataset input

3. EVALUASI

Melakukan evaluasi metriks untuk setiap model dan memilih model terbaik untuk prediksi

2. TRAIN MODEL

Menggunakan TFIDF, Fine-Tuning BERT, KERAS, Word2Vec untuk memodelkan klasifikasi

4. PREDIKSI

Melakukan prediksi terhadap inputan keyword dari berita untuk diprediksi menjadi topik hasil klasifikasi

PREPROCESSING

1. **Tokenisasi**
   Melakukan tokenisasi dari sentence menjadi token/kata
2. **Stop Word Removal**
   Menghapus token yang berupa stop word

```
841     ['dominici', 'back', 'lacklustr', 'franc', 'wi...
1748    ['id', 'theft', 'surg', 'hit', 'u', 'consum', ...
2118    ['blair', 'press', 'u', 'climat', 'toni', 'bla...
1174    ['realli', 'divid', 'parti', 'gap', 'labour', ...
1502    ['iran', 'budget', 'seek', 'state', 'sell', 'o...
                               ...
1033    ['labour', 'eu', 'propaganda', 'taxpay', 'subs...
1731    ['crossrail', 'link', 'get', 'go', 'ahead', '£...
763     ['share', 'rise', 'new', 'man', 'utd', 'offer'...
835     ['rock', 'star', 'su', 'ex', 'girlfriend', 'mo...
1653    ['file', 'swapper', 'readi', 'new', 'network',...
Name: tokenized_text, Length: 1780, dtype: object
```

3. **Number Eraser**
   Menghapus token yang termasuk angka
4. **Stemming**
   Mengubah token menjadi kata hasil stem
5. **Lemmatization**
   Mengubah token kata menjadi kata dasarnya
6. **Encoding**
   Melakukan encoding terhadap kategori dari label topik

TRAIN MODEL (EKSPERIMEN)

TFIDF - SHALLOW ALGO

Melakukan TFIDF kemudian menggunakan Decission Tree, XGBoost, SVM, Multinomial Naive Bayes, Random Forrest

FINE-TUNING BERT

Menggunakan model BERT-base multilingual dan distill-BERT

KERAS MODEL

Menggunakan keras model dan keras layer berupa dense, activation berupa relu dan softmax

WORD2VEC

Melakukan embeding terhadap kata kemudian membuat model deep learningnya dengan layer input, embedding, attentional, lstm, dense, dan output

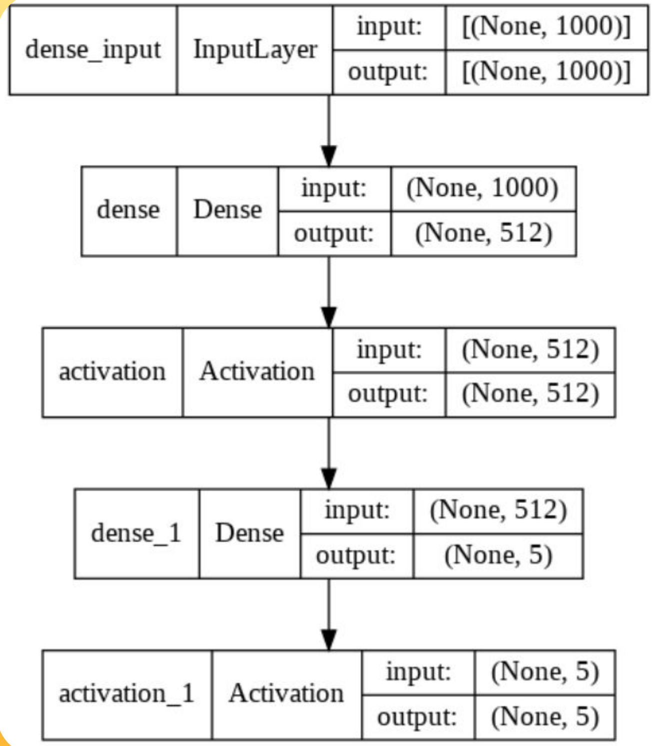
TFIDF SHALLOW ALGORITMA

VARIABLE	AKURASI	PRESISI	RECALL	F1 SCORE
Decision Tree Learning	86%	85%	85%	85%
XGBoost	96%	96%	96%	96%
SVM	96%	95%	97%	96%
Multinomial Naive Bayes	**98%**	**97%**	**97%**	**97%**
Random Forest Classifier	97%	97%	97%	97%

KERAS SEQ MODEL

Keras sequential model adalah model yang dibentuk dengan tumpukan lapisan di mana setiap lapisan memiliki tepat satu tensor input dan satu tensor output.

Layer yang digunakan adalah layer input, dense layer, activation pertama dengan relu, dense layer lagi, dan activation kedua menggunakan softmax



KERAS MODEL

Hasil akurai mencapai 99.7%

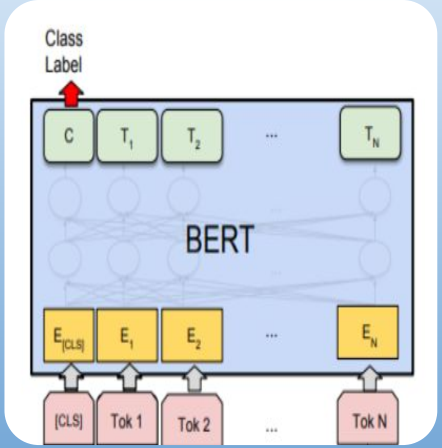
```
Epoch 1/2
54/54 [==============================] - 3s 8ms/step - loss: 0.4303 - accuracy: 0.8770 - val_loss: 0.1054 - val_accuracy: 0.9607
Epoch 2/2
54/54 [==============================] - 0s 5ms/step - loss: 0.0360 - accuracy: 0.9944 - val_loss: 0.0717 - val_accuracy: 0.9719
```

```
60/60 [==============================] - 0s 3ms/step - loss: 0.0202 - accuracy: 0.9972
Test accuracy score: 0.9971910119056702
Test loss score: 0.020186411216855505
```

FINE TUNING BERT

Fine tuning BERT adalah melakukan training model bert pada input dataset dari model BERT yang sudah dilatih sebelumnya dengan dataset yang besar.

Fine tuning bert dilakukan dengan menyesuaikan representasi input sebelum dimasukkan ke pretrained model dan menambahkan satu untrained output layer untuk dilatih kembali ke task spesifik




```
----------------------------------------------------------------
Layer Embedding

bert.embeddings.word_embeddings.weight                 (119547, 768)
bert.embeddings.position_embeddings.weight                (512, 768)
bert.embeddings.token_type_embeddings.weight                (2, 768)
bert.embeddings.LayerNorm.weight                              (768,)
bert.embeddings.LayerNorm.bias                                (768,)

----------------------------------------------------------------
Transformer Pertama

bert.encoder.layer.0.attention.self.query.weight          (768, 768)
bert.encoder.layer.0.attention.self.query.bias                (768,)
bert.encoder.layer.0.attention.self.key.weight            (768, 768)
bert.encoder.layer.0.attention.self.key.bias                  (768,)
bert.encoder.layer.0.attention.self.value.weight          (768, 768)
bert.encoder.layer.0.attention.self.value.bias                (768,)
bert.encoder.layer.0.attention.output.dense.weight        (768, 768)
bert.encoder.layer.0.attention.output.dense.bias              (768,)
bert.encoder.layer.0.attention.output.LayerNorm.weight        (768,)
bert.encoder.layer.0.attention.output.LayerNorm.bias          (768,)
bert.encoder.layer.0.intermediate.dense.weight           (3072, 768)
bert.encoder.layer.0.intermediate.dense.bias                 (3072,)
bert.encoder.layer.0.output.dense.weight                 (768, 3072)
bert.encoder.layer.0.output.dense.bias                        (768,)
bert.encoder.layer.0.output.LayerNorm.weight                  (768,)
bert.encoder.layer.0.output.LayerNorm.bias                    (768,)

----------------------------------------------------------------
----------------------------------------------------------------
Layer Output

bert.pooler.dense.weight                                  (768, 768)
bert.pooler.dense.bias                                        (768,)
classifier.weight                                           (5, 768)
classifier.bias                                                 (5,)
```

FINE TUNING BERT

BERT-Base Multilingual (dengan 36 epoch)

```
Melakukan Validation
  Accuracy: 0.91
  Validation took: 0:00:04

======== Epoch 14 / 15 ========
Melakuakan Training
  Batch    40  of     45.    Elapsed: 0:00:45.

  Average training loss: 0.00
  Training epoch took: 0:00:50

Melakukan Validation
  Accuracy: 0.88
  Validation took: 0:00:04

======== Epoch 15 / 15 ========
Melakuakan Training
  Batch    40  of     45.    Elapsed: 0:00:45.

  Average training loss: 0.00
  Training epoch took: 0:00:50

Melakukan Validation
  Accuracy: 0.90
  Validation took: 0:00:04
```

	precision	recall	f1-score	support
tech	0.92	0.91	0.91	97
business	0.97	0.90	0.94	81
sport	0.88	0.91	0.89	75
entertainment	0.98	0.96	0.97	112
politics	0.86	0.94	0.90	80
accuracy			0.93	445
macro avg	0.92	0.92	0.92	445
weighted avg	0.93	0.93	0.93	445

FINE TUNING BERT

Distill BERT-based uncase (dengan 60 epoch)

```
  Average training loss: 0.00
  Training epoch took: 0:00:48

Melakukan Validation
  Accuracy: 0.95
  Validation took: 0:00:04

======== Epoch 49 / 50 ========
Melakuakan Training
  Batch    40  of     45.    Elapsed: 0:00:43.

  Average training loss: 0.00
  Training epoch took: 0:00:48

Melakukan Validation
  Accuracy: 0.93
  Validation took: 0:00:04

======== Epoch 50 / 50 ========
Melakuakan Training
  Batch    40  of     45.    Elapsed: 0:00:43.

  Average training loss: 0.00
  Training epoch took: 0:00:48
```

	precision	recall	f1-score	support
tech	0.94	0.93	0.93	97
business	0.96	0.93	0.94	81
sport	0.91	0.93	0.92	75
entertainment	0.98	1.00	0.99	112
politics	0.94	0.94	0.94	80
accuracy			0.95	445
macro avg	0.95	0.94	0.95	445
weighted avg	0.95	0.95	0.95	445

WORD2VEC

Word2Vec adalah metode embedding word yang berguna untuk merepresentasikan kata-kata menjadi sebuah vektor dengan panjang N.

Layer (type)	Output Shape	Param #	Connected to
input_11 (InputLayer)	[(None, 15)]	0	[]
embedding_10 (Embedding)	(None, 15, 300)	5933700	['input_11[0][0]']
permute_9 (Permute)	(None, 300, 15)	0	['embedding_10[0][0]']
dense_26 (Dense)	(None, 300, 15)	240	['permute_9[0][0]']
attention (Permute)	(None, 15, 300)	0	['dense_26[0][0]']
multiply_8 (Multiply)	(None, 15, 300)	0	['embedding_10[0][0]', 'attention[0][0]']
bidirectional_18 (Bidirectiona l)	(None, 15, 30)	37920	['multiply_8[0][0]']
bidirectional_19 (Bidirectiona l)	(None, 30)	5520	['bidirectional_18[0][0]']
dense_27 (Dense)	(None, 64)	1984	['bidirectional_19[0][0]']
dense_28 (Dense)	(None, 5)	325	['dense_27[0][0]']

Total params: 5,979,689
Trainable params: 45,989
Non-trainable params: 5,933,700

Word2Vec menggunakan neural network untuk mendapatkan vektor tersebut. Arsitektur Word2vec hanya terdiri dari layer input, hidden layer, dan layer output.

Pada arsitektur ini maka kita menggunakan layer input, layer embeding, layer attentional (yang terdiri dari layer permute, layer dense dengan activation softmax, layer permute lagi), kemudian 2 layer bidirectional LSTM, dan 2 layer dense (dengan activation pertama adalah relu dan kedua adalah softmax)

HASIL WORD2VEC

```
trained = model_w2v.fit(x=x_train_w2v, y=y_train_oh, batch_size=256,
                        epochs=20, shuffle=True,
                        validation_split=0.3)
```

```
Epoch 1/20
5/5 [==============================] - 12s 723ms/step - loss: 1.6092 - accuracy: 0.2416 - val_loss: 1.6093 - val_accuracy: 0.2097
Epoch 2/20
5/5 [==============================] - 1s 184ms/step - loss: 1.6086 - accuracy: 0.2416 - val_loss: 1.6091 - val_accuracy: 0.2097
Epoch 3/20
5/5 [==============================] - 1s 185ms/step - loss: 1.6079 - accuracy: 0.2303 - val_loss: 1.6090 - val_accuracy: 0.2097
Epoch 4/20
5/5 [==============================] - 1s 179ms/step - loss: 1.6074 - accuracy: 0.2303 - val_loss: 1.6090 - val_accuracy: 0.2097
Epoch 5/20
5/5 [==============================] - 1s 179ms/step - loss: 1.6069 - accuracy: 0.2303 - val_loss: 1.6089 - val_accuracy: 0.2097
Epoch 6/20
5/5 [==============================] - 1s 181ms/step - loss: 1.6065 - accuracy: 0.2303 - val_loss: 1.6088 - val_accuracy: 0.2097
Epoch 7/20
5/5 [==============================] - 1s 187ms/step - loss: 1.6060 - accuracy: 0.2303 - val_loss: 1.6088 - val_accuracy: 0.2097
Epoch 8/20
5/5 [==============================] - 1s 188ms/step - loss: 1.6055 - accuracy: 0.2303 - val_loss: 1.6088 - val_accuracy: 0.2097
Epoch 9/20
5/5 [==============================] - 1s 181ms/step - loss: 1.6050 - accuracy: 0.2303 - val_loss: 1.6088 - val_accuracy: 0.2097
Epoch 10/20
5/5 [==============================] - 1s 184ms/step - loss: 1.6046 - accuracy: 0.2303 - val_loss: 1.6088 - val_accuracy: 0.2097
Epoch 11/20
5/5 [==============================] - 1s 180ms/step - loss: 1.6042 - accuracy: 0.2255 - val_loss: 1.6088 - val_accuracy: 0.2097
Epoch 12/20
5/5 [==============================] - 1s 188ms/step - loss: 1.6038 - accuracy: 0.2239 - val_loss: 1.6088 - val_accuracy: 0.2097
```

	precision	recall	f1-score	support
accuracy			0.22	445
macro avg	0.04	0.20	0.07	445
weighted avg	0.05	0.22	0.08	445

Hasil eksperimen menunjukkan akurasi 22%. Sehingga model klasifikasi lebih baik menggunakan model lainnya

03 PIPELINE 3

DOCUMENT SIMILARITY

Proses menentukan kemiripan antara beberapa berita

PROCESS

01 GET USER INPUT

Mengambil teks dari news yang dibaca user

02 LOAD KEYWORDS

Membaca keywords hasil dari Keyword Extraction

03 TRAIN MODEL

Train model menggunakan keywords hasil Keyword Extraction

04 PREDICT

Mendapatkan document similarity

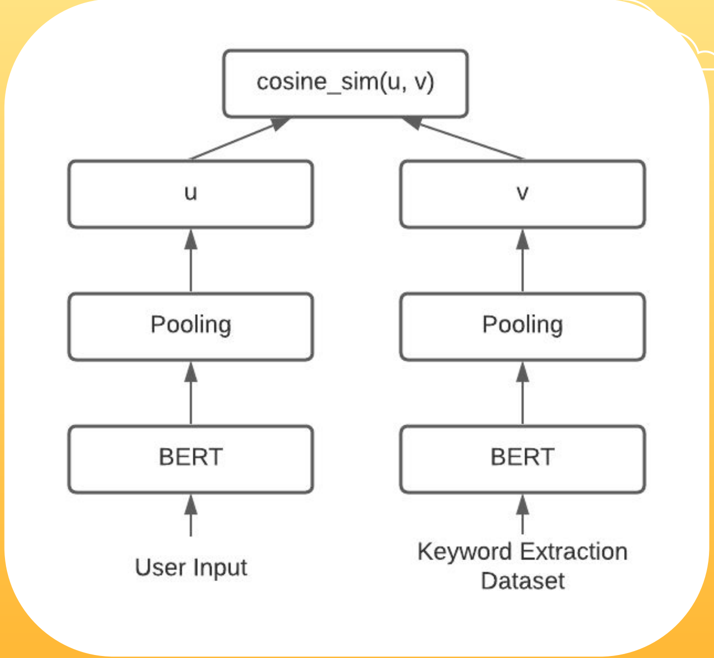
05 RECOMMEND

Melakukan sorting hasil doc similarity dan hasil topic classification

ARSITEKTUR

ARSITEKTUR

Arsitektur model yang digunakan adalah model transformer dengan pretrained-model BERT yang di-*embed* dengan data yang berasal dari *user input* dan *keyword extraction dataset*. Setelah itu, similarity dihitung dengan menggunakan *cosine_similarity*



EKSPERIMEN

EKSPERIMEN

Eksperimen yang dilakukan adalah dengan mencoba model yang berbasiskan statistik yaitu TF-IDF, dan model yang menggunakan neural network yaitu word2vec dan Transformers. Model yang digunakan pada Transformers adalah sentence-transformers/all-MiniLM-L6-v2 dan sentence-transformers/paraphrase-mpnet-base-v2.

Untuk model TF-IDF dilakukan perbandingan variasi parameter min_df pada TfidfVectorizer untuk mengabaikan kata-kata dengan frekuensi kurang dari threshold tertentu.

Dataset yang digunakan dalam eksperimen adalah dataset yang berasal dari:
https://www.kaggle.com/achintyatripathi/news-dataset-18920

HASIL EKSPERIMEN (STATISTICAL)

TF-IDF, min_df=0.25 (15 columns)

```
Article Read: Even as more than 150 million people are using dig...
 ---------------------------------------------------------
Recomendation 1: (IDX: 507), score: 0.8798349819335274 | Online advertising is a game of scale, but one att...

Recomendation 2: (IDX: 166), score: 0.8755319876279275 | LONDON, Aug 18 (Thomson Reuters Foundation) - Form...

Recomendation 3: (IDX: 926), score: 0.8747096090847626 | HONG KONG (Reuters) - Chinese online insurance tec...

Recomendation 4: (IDX: 7), score: 0.871895008547424 | Airmeet, a startup that offers a platform to host ...

Recomendation 5: (IDX: 469), score: 0.8614875812334784 | As Trump Visits Kenosha, Hundreds Gather Where Jac...

Recomendation 6: (IDX: 1292), score: 0.8549618479955363 | BRASILIA (Reuters) - Brazil's President Jair Bolso...

Recomendation 7: (IDX: 1454), score: 0.8549618479955363 | FILE PHOTO: Climate change activists demonstrate a...

Recomendation 8: (IDX: 1464), score: 0.8549618479955363 | FILE PHOTO: Climate change activists demonstrate a...

Recomendation 9: (IDX: 9), score: 0.8546135883744768 | Byju's has raised $500 million in a new financing ...

Recomendation 10: (IDX: 1226), score: 0.8533195048726804 | BRASILIA/MOSCOW (Reuters) - The Brazilian state of...
```

HASIL EKSPERIMEN (STATISTICAL)

TF-IDF, min_df=0.1 (188 columns)

```
Article Read: Even as more than 150 million people are using dig...
 ----------------------------------------------------------
Recomendation 1: (IDX: 177), score: 0.6145900956318281 | Indian food delivery startup Zomato has raised $62...

Recomendation 2: (IDX: 1091), score: 0.583970308827363 | Southeast Asia's leading property listing company ...

Recomendation 3: (IDX: 9), score: 0.5487703260742444 | Byju's has raised $500 million in a new financing ...

Recomendation 4: (IDX: 11), score: 0.5435676545074747 | Indian billionaire Mukesh Ambani's retail business...

Recomendation 5: (IDX: 7), score: 0.52496019034492 | Airmeet, a startup that offers a platform to host ...

Recomendation 6: (IDX: 176), score: 0.5159538061419459 | 3one4 Capital, a venture capital firm in India, to...

Recomendation 7: (IDX: 178), score: 0.5134475943083189 | Mumbai-based Eruditus, which works with top univer...

Recomendation 8: (IDX: 0), score: 0.4958957368213379 | Vodafone Idea, one of the largest telecom operator...

Recomendation 9: (IDX: 123), score: 0.4764466896482933 | NEW DELHI (Reuters) - India plans to significantly...

Recomendation 10: (IDX: 1086), score: 0.4743197003412171 | Point72 Ventures, the early-stage investment firm ...
```

HASIL EKSPERIMEN (STATISTICAL)

TF-IDF, min_df=0.0 (588777 columns)

```
Article Read: Even as more than 150 million people are using dig...
 -----------------------------------------------------------
Recomendation 1: (IDX: 176), score: 0.057123252265082164 | 3one4 Capital, a venture capital firm in India, to...

Recomendation 2: (IDX: 9), score: 0.0501242050658371 | Byju's has raised $500 million in a new financing ...

Recomendation 3: (IDX: 7), score: 0.0486033162806363 | Airmeet, a startup that offers a platform to host ...

Recomendation 4: (IDX: 1), score: 0.04808100864680179 | At the beginning of the previous decade, Facebook ...

Recomendation 5: (IDX: 11), score: 0.04661801280603871 | Indian billionaire Mukesh Ambani's retail business...

Recomendation 6: (IDX: 8), score: 0.04240052932842192 | Since India enforced a lockdown across the country...

Recomendation 7: (IDX: 97), score: 0.04143247655591322 | And we're back! Today was part two of Y Combinator...

Recomendation 8: (IDX: 178), score: 0.03992847727515998 | Mumbai-based Eruditus, which works with top univer...

Recomendation 9: (IDX: 564), score: 0.03961769704900882 | Poland is becoming an important European tech ecos...

Recomendation 10: (IDX: 720), score: 0.03781986356587531 | Here's how you can get a second shot at Startup Ba...
```

HASIL EKSPERIMEN (NEURAL NETWORK)

Word2Vec

```
Article Read: Even as more than 150 million people are using dig...
 ------------------------------------------------------------
Recomendation 1: (IDX: 264), score: 0.6808363199234009 | Aug 26 - The following are the details of Indian S...

Recomendation 2: (IDX: 113), score: 0.6790512204170227 | Aug 19 - The following are the details of Indian S...

Recomendation 3: (IDX: 62), score: 0.5325736403465271 | (Reuters) - A look at the key facts and records of...

Recomendation 4: (IDX: 1343), score: 0.5170466899871826 | FILE PHOTO: A view is seen from the Amazon Tall To...

Recomendation 5: (IDX: 214), score: 0.46485865116119385 | Sep 1 (OPTA) - Scoreboard at close of play of 3rd ...

Recomendation 6: (IDX: 67), score: 0.4537274241447449 | Sep 13 (OPTA) - Scores from the LPGA Tour ANA Insp...

Recomendation 7: (IDX: 274), score: 0.45062702894210815 | Aug 30 (OPTA) - Scoreboard at close of play of 2nd...

Recomendation 8: (IDX: 832), score: 0.4493864178657532 | KIGALI (Reuters) - Rwandan President Paul Kagame d...

Recomendation 9: (IDX: 130), score: 0.44323742389678955 | Aug 25 (OPTA) - Scoreboard at close of play on the...

Recomendation 10: (IDX: 84), score: 0.4368095099925995 | Sep 12 (OPTA) - Scores from the European Tour Port...
```

HASIL EKSPERIMEN (NEURAL NETWORK)

Transformer - sentence-transformers/paraphrase-mpnet-base-v2 (768 dimensional dense vector space)

```
Article Read: Even as more than 150 million people are using dig...
 ----------------------------------------------------------
Recomendation 1: (IDX: 6), score: 0.2762608230113983 | CashKaro, one of the leading cashback and coupon s...

Recomendation 2: (IDX: 9), score: 0.25234079360961914 | Byju's has raised $500 million in a new financing ...

Recomendation 3: (IDX: 176), score: 0.2384922057390213 | 3one4 Capital, a venture capital firm in India, to...

Recomendation 4: (IDX: 741), score: 0.29171955585479736 | As a business model, SaaS has expanded to epic siz...

Recomendation 5: (IDX: 8), score: 0.3042358160018921 | Since India enforced a lockdown across the country...

Recomendation 6: (IDX: 1086), score: 0.28184065222740173 | Point72 Ventures, the early-stage investment firm ...

Recomendation 7: (IDX: 1124), score: 0.29140838980674744 | Dawn Capital, the London-based VC that focuses on ...

Recomendation 8: (IDX: 890), score: 0.26587602496147156 | Apple is well known for picking up smaller startup...

Recomendation 9: (IDX: 740), score: 0.23904013633728027 | Once upon a time, fintech founders could pitch 10 ...

Recomendation 10: (IDX: 91), score: 0.24016138911247253 | More than 150 e-commerce and delivery companies gl...
```

HASIL EKSPERIMEN (NEURAL NETWORK)

Transformer - sentence-transformers/all-MiniLM-L6-v2 (384 dimensional dense vector space)

```
Article Read: Even as more than 150 million people are using dig...
 ----------------------------------------------------------
Recomendation 1: (IDX: 566), score: 0.7232925431037607 | Investor interest in no-code, low-code apps and se...

Recomendation 2: (IDX: 176), score: 0.7168260677463922 | 3one4 Capital, a venture capital firm in India, to...

Recomendation 3: (IDX: 179), score: 0.7060372750709747 | Your startup is special and different, and you nee...

Recomendation 4: (IDX: 9), score: 0.6904699084050727 | Byju's has raised $500 million in a new financing ...

Recomendation 5: (IDX: 5), score: 0.683123557261046 | More than a third of small and medium-sized busine...

Recomendation 6: (IDX: 8), score: 0.6814406174305236 | Since India enforced a lockdown across the country...

Recomendation 7: (IDX: 498), score: 0.6803235459609738 | DNX Ventures, an investment firm that focuses on e...

Recomendation 8: (IDX: 733), score: 0.6703664729903799 | Nerdwallet, which provides resources for people lo...

Recomendation 9: (IDX: 499), score: 0.6695382917091338 | The CEO of Pan-African fintech unicorn Interswitch...

Recomendation 10: (IDX: 7), score: 0.6638006522966278 | Airmeet, a startup that offers a platform to host ...
```

ANALISIS HASIL EKSPERIMEN

1. Model dengan neural network (word2vec, transformers) memiliki nilai *cosine_similarity* lebih baik daripada model TF-IDF.
2. Pada model dengan TF-IDF, semakin kecil nilai *min_df*, nilai *cosine_similarity* semakin kecil namun jumlah kolom semakin banyak. Pada *min_df*=0.25, skor mencapai 0.879. Namun hanya terdapat 15 kolom/kata yang digunakan dalam prediksi.
3. Pada model transformer, model sentence-transformers/all-MiniLM-L6-v2 dengan 384 *dimensional dense vector space* menghasilkan nilai *cosine_similarity* lebih baik daripada sentence-transformers/paraphrase-mpnet-base-v2 dengan 768 *dimensional dense vector space*

04 DEMO

THANKS!

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon** and infographics & images by **Freepik**

Please keep this slide for attribution