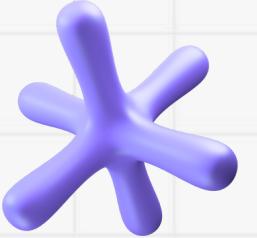


Kelompok 2

Ekstraksi Fitur Kalimat Menggunakan Algoritma Bert



NAMA KELOMPOK



AISYAH NUR KHOIROFIQ



AGUS TUSILAWATI



EKA WIRA YUDHA



TASYA KHADIJAH

Agenda

- 1 DESKRIPSI
- 2 PENGAMBILAN DATA
- 3 TAHAPAN PERANGKAT LUNAK

- 4 HASIL PERANGKAT LUNAK
- 5 KESIMPULAN



Deskripsi



1. Deskripsi Perangkat Lunak

Dalam konteks ekstraksi fitur kalimat, perangkat lunak ini berfungsi untuk mengekstrak representasi numerik dari setiap kalimat dalam teks. Representasi ini mencerminkan pemahaman terhadap konten kalimat tersebut, termasuk hubungan antar kata, sintaksis, dan makna secara keseluruhan.

Algoritma BERT bekerja secara bertahap dalam proses ekstraksi fitur kalimat. Pertama, kalimat-kalimat dalam teks dipecah menjadi token-token kata. Kemudian, setiap token kata tersebut diolah secara bersamaan dengan konteks sebelum dan sesudahnya oleh model BERT. Proses ini memungkinkan BERT untuk menghasilkan representasi vektor yang menggambarkan makna kalimat secara komprehensif.

Dengan menggunakan representasi numerik ini, perangkat lunak tersebut dapat melakukan berbagai tugas, seperti klasifikasi teks, pengelompokan kalimat, atau analisis sentimen. Fitur-fitur ini dapat digunakan sebagai masukan untuk model pembelajaran lebih lanjut dalam aplikasi yang memerlukan pemahaman yang dalam terhadap teks.



2. Deskripsi Jurnal Yang Digunakan



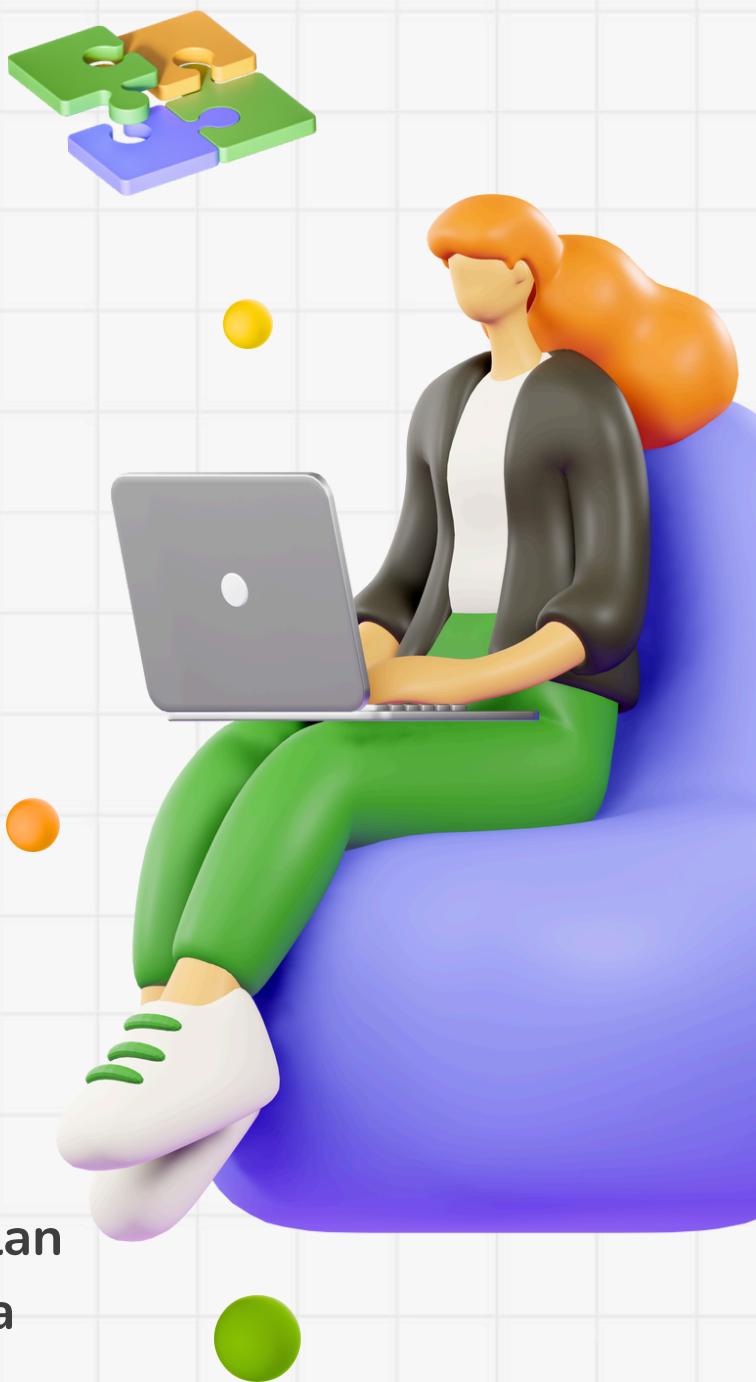
Jurnal "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" adalah sebuah makalah penelitian yang dipublikasikan pada tahun 2018 oleh Jacob Devlin, Ming-Wei Chang, Kenton Lee, dan Kristina Toutanova. Makalah ini menjadi terkenal karena memperkenalkan algoritma BERT (Bidirectional Encoder Representations from Transformers), yang telah menghasilkan kemajuan signifikan dalam bidang pemrosesan bahasa alami.

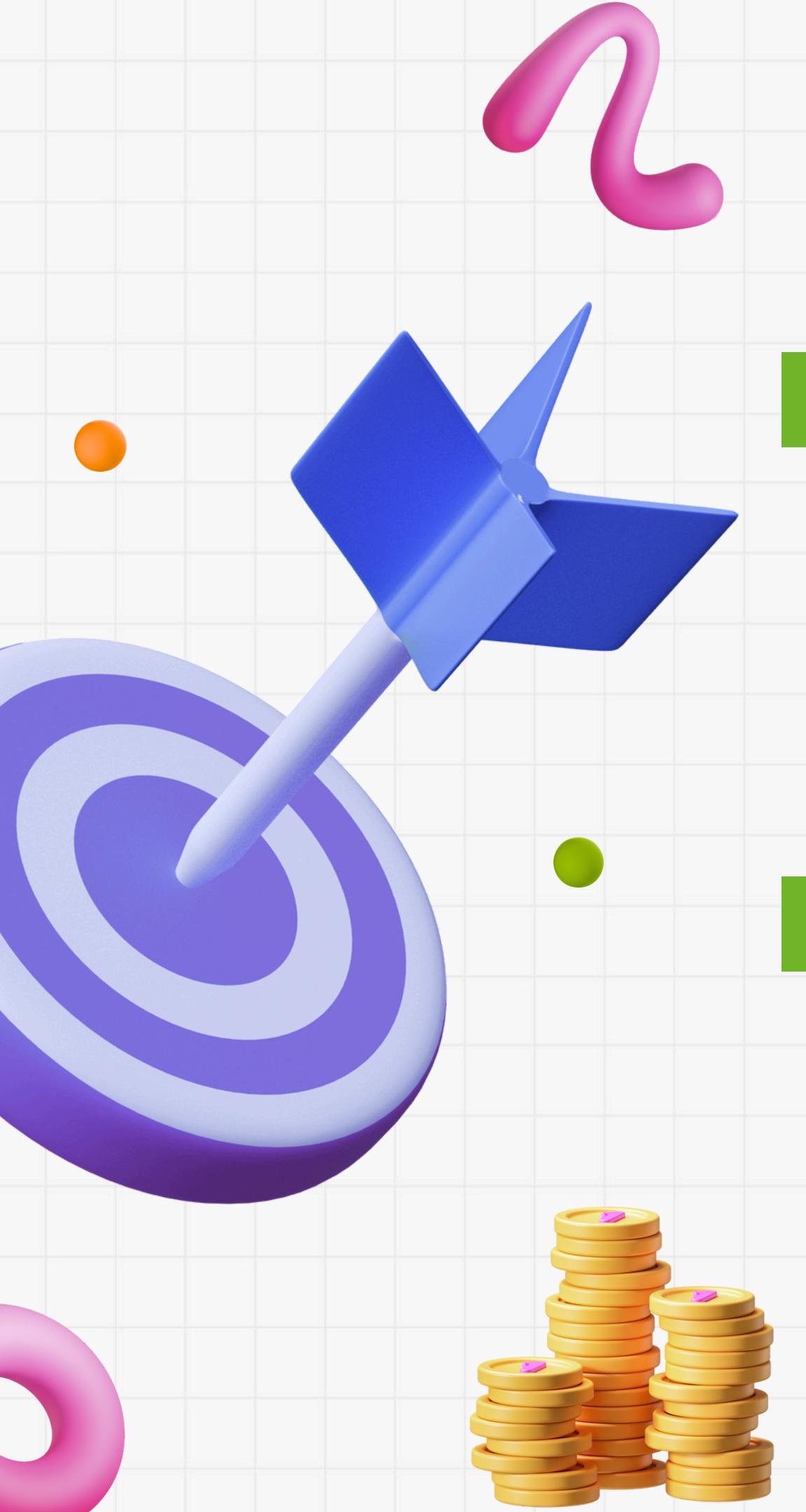


Dalam jurnal ini, para penulis mempresentasikan pendekatan baru untuk pre-training model bahasa yang mendalam. Mereka mengusulkan penggunaan arsitektur transformer dalam kombinasi dengan pre-training berbasis "masked language model" dan "next sentence prediction". Pendekatan ini memungkinkan model untuk memahami konteks kalimat secara global dan menghasilkan representasi yang kaya dan mendalam dari teks.



Hasil eksperimen dalam jurnal ini menunjukkan bahwa BERT berhasil mencapai kinerja yang mengesankan dalam berbagai tugas pemrosesan bahasa alami, termasuk pemahaman teks, pengenalan entitas, dan analisis sentimen. Kesuksesan BERT telah membuatnya menjadi salah satu model bahasa yang paling banyak digunakan dan dipelajari dalam penelitian pemrosesan bahasa alami.





3. Perbedaan Perangkat Lunak dan Jurnal

1

Tujuan Utama

Perangkat Lunak: Fokusnya adalah implementasi konkret dari konsep ekstraksi fitur kalimat menggunakan algoritma BERT. Tujuan utamanya adalah untuk memberikan alat yang dapat digunakan untuk mengekstrak fitur dari kalimat-kalimat dalam teks.

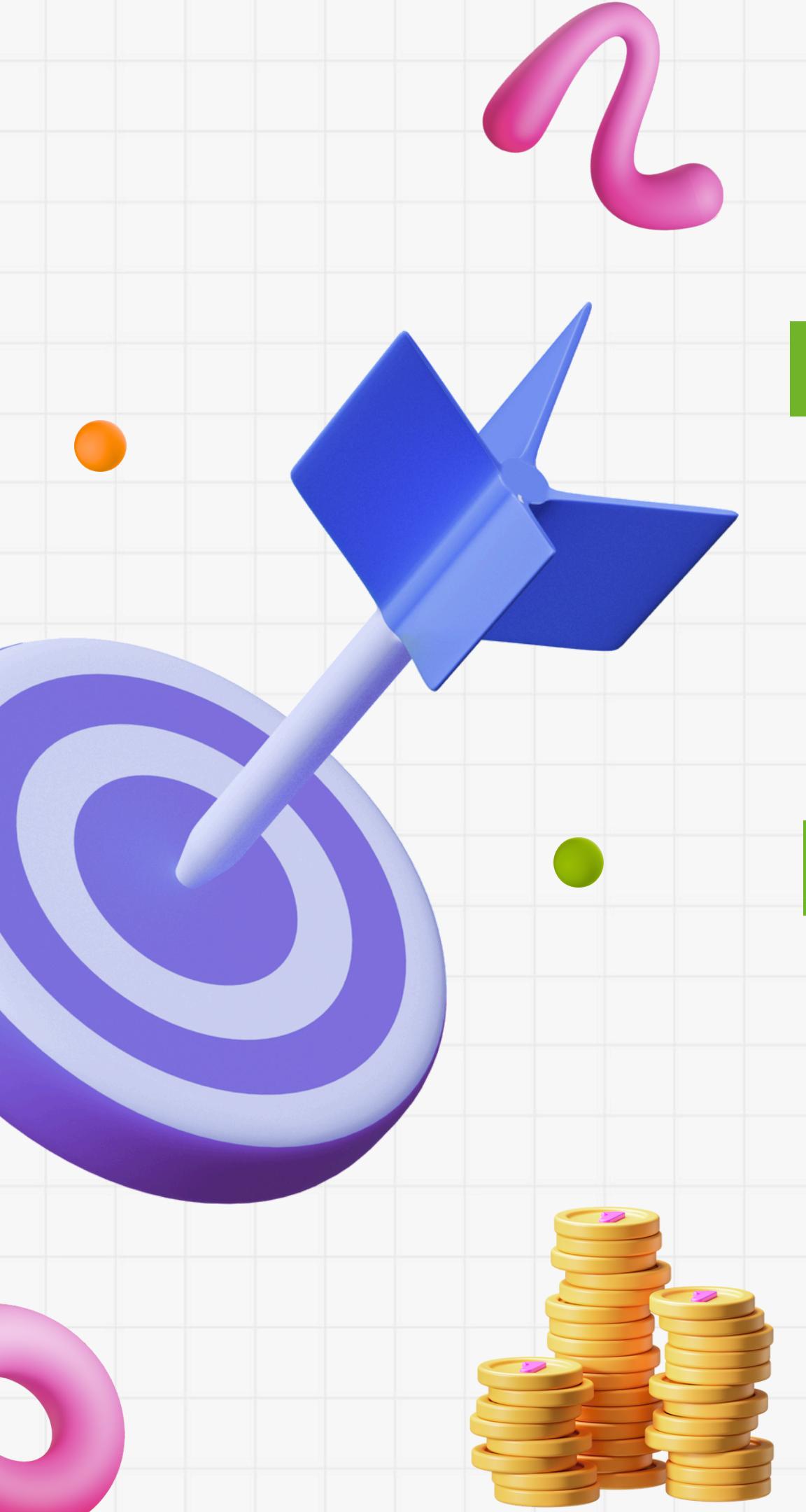
Jurnal: Fokusnya adalah pada penyelidikan teoretis dan eksperimental tentang algoritma BERT itu sendiri, termasuk metodologi pre-training dan evaluasi kinerja. Tujuan utamanya adalah untuk memperkenalkan algoritma BERT dan mengevaluasi kinerjanya dalam pre-training model bahasa yang mendalam

2

Konteks Publikasi

Perangkat Lunak: Dapat berupa aplikasi konkret yang dikembangkan untuk tujuan tertentu, oleh pengembang atau peneliti yang ingin menerapkan teknik ekstraksi fitur kalimat dengan menggunakan algoritma BERT

Jurnal: Merupakan sebuah publikasi akademis yang berupa makalah penelitian yang dipublikasikan dalam jurnal ilmiah atau konferensi yang berisi temuan dan analisis dari penelitian yang dilakukan oleh para penulis



3. Perbedaan Perangkat Lunak dan Jurnal

3

Tingkat Kedalaman Analisis

Perangkat Lunak: Lebih berkaitan dengan implementasi teknis dan operasional dari konsep ekstraksi fitur kalimat menggunakan algoritma BERT. Fokusnya pada bagaimana menggunakan algoritma BERT dalam konteks aplikasi yang spesifik

Jurnal: Lebih dalam secara teoritis, menjelaskan konsep dan metodologi di balik algoritma BERT serta eksperimen yang dilakukan untuk menguji dan mengevaluasi kinerjanya. Ini mencakup pemahaman yang lebih luas tentang bagaimana BERT bekerja dan apa yang membuatnya efektif dalam pre-training model bahasa

4

Output dan Kontribusi

Perangkat Lunak: Menghasilkan alat atau sistem yang dapat digunakan untuk tujuan tertentu, seperti ekstraksi fitur untuk analisis teks atau tugas-tugas pemrosesan bahasa alami lainnya

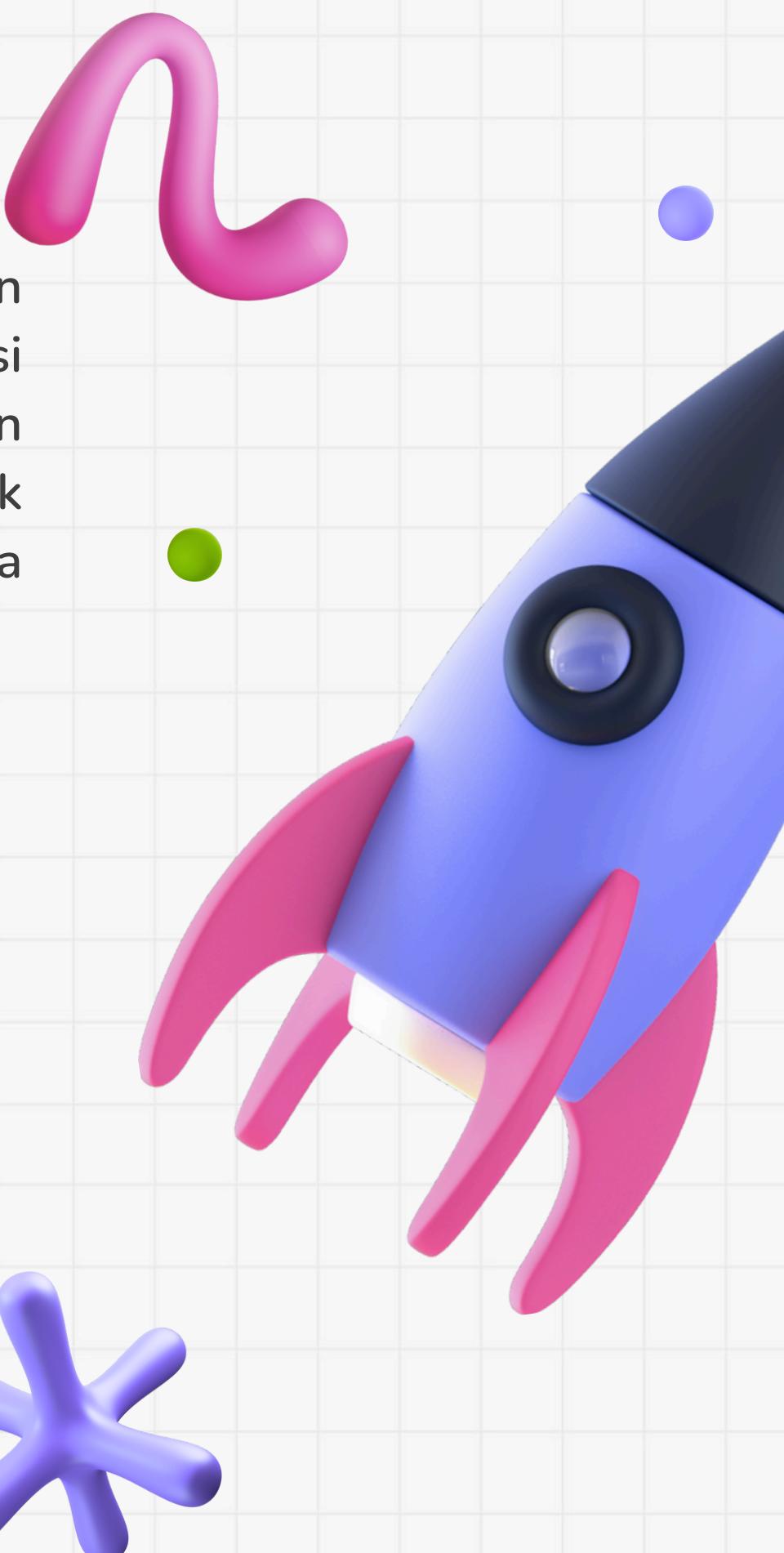
Jurnal: Menghasilkan pengetahuan dan pemahaman yang lebih mendalam tentang algoritma BERT itu sendiri, yang kemudian dapat digunakan oleh peneliti lain untuk memperbaiki atau mengembangkan lebih lanjut teknik pemrosesan bahasa alami

Pengambilan Data



Pengambilan Data

Pada tahap pengambilan data kelompok kami mengambil dataset yang berhubungan dengan ekstraksi fitur kalimat dengan BERT, kami menggunakan data dummy dengan data yang berisi kalimat yang dilabeli 2 nilai yaitu “truthful” dan “deceptive” yang berjumlah 70 baris data train dan 30 baris data test (uji), data dummy berupa teks mentah atau data terstruktur dalam bentuk tabel. dan setelah itu kami Menyimpan data yang diekstrak ke dalam format yang dapat dibaca oleh program berupa file CSV.



| A | B |
|---|-----------|
| 1 text | deceptive |
| 2 My husband and I were in the Fairmont Chicago recently for a conference. We stayed in a spacious suite. All amenities appeared recently updated and in excellent shape. The bed was very comfortable and the room was quiet. Staff were friendly, room was well kept like a typical Hyatt location. Hotel is huge, with two buildings and over 2000 rooms. Good location, close to Michigan Ave and the subway. Parking is expensive but there are several options available. | truthful |
| 3 The Palmer House Hilton, while it looks good in pictures, and the outside, is actually a disaster of a hotel. When I went through, the lobby was dirty, my room hadn't been cleaned, and smell bad. | deceptive |
| 4 Overall, the hotel was okay. Though I did have a problem with the price. The accommodations mentioned on the website were accurate (other than the supposedly luxury mattress, it was uncomfortable). | deceptive |
| 5 We stayed for a one night getaway with family on a thursday. Triple AAA rate of 173 was a steal. 7th floor room complete with 44in plasma TV, Bose stereo, voss and evian water, and gorg | truthful |
| 6 I stayed at the Fairmont during a professional conference. While I have to say that for a conference, some of the meeting rooms were too small, in terms of traveler's comfort the hotel is excellent. | truthful |
| 7 I'll keep it short. My wife, daughter and I stayed at this hotel for 3 nights. We go on several vacations a year and this was a great hotel. Location: It is located in the center of the Magnificent Mile. | truthful |
| 8 Simply a nice place to stay... I had a great deal for the room, and was impressed with the quality. Hotel staff was very helpful; they were able to let me check out and helped me print boarding passes. | truthful |
| 9 Great hotel in a wonderful location. I stayed in this hotel last weekend for one night before coming home and I loved it. I couldn't recommend it highly enough. Evidently popular and busy. | truthful |
| 10 I recently stayed at the Hotel Allegro Chicago with my wife. I was on business but I thought I'd take her so we could see the city as well. Chicago was great, The Allegro...not so much. I'm not sure if it's because we stayed in a different building or what. | deceptive |
| 11 Oh My! More Like A Dank Museum Than A Four Star Hotel! I read so many reviews on this hotel that I was sure my husband and I wouldn't be disappointed on spending our 10th Wedding Anniversary here. | deceptive |
| 12 The Conrad Hotel is a beautiful hotel. Our room was perfect and beautiful in every way. I was very impressed (and I am hard to please and have high standards). It was so clean, aesthetic, and comfortable. | truthful |
| 13 The Amalfi Hotel in Chicago advertises itself as family friendly, yet when I arrived with my children it was anything BUT family friendly. The staff seemed miffed and stuck up, and I had to wait a long time for my room to be ready. | deceptive |

File dataset yang digunakan dapat diakses melalui link berikut: [punctuation-small-train - Google Spreadsheet](#)

Tahapan Perangkat Lunak



Tahapan Perangkat Lunak

Instalasi dan Import Library

- Instalasi Library Menggunakan pip untuk menginstal library bert-for-tf2 yang diperlukan untuk menggunakan BERT dengan TensorFlow 2.
- Import Library Mengimpor library pandas untuk pengolahan data, tensorflow untuk menggunakan model machine learning, tensorflow_hub untuk mengambil model pra-terlatih dari TensorFlow Hub, dan bert untuk utilitas BERT.

```
!pip install bert-for-tf2

import pandas as pd
import tensorflow as tf
import tensorflow_hub as hub
import bert
```

Inisialisasi Tokenizer dan Model BERT

- FullTokenizer Menginisialisasi tokenizer BERT untuk memecah teks menjadi token yang dimengerti oleh BERT.
- max_seq_length Menentukan panjang maksimal urutan token (512 token).
- bert_layer Mengambil layer BERT pra-terlatih dari TensorFlow Hub. Model yang digunakan adalah bert_en_uncased_L-12_H-768_A-12/1, yaitu BERT bahasa Inggris yang tidak menghiraukan huruf besar/kecil dengan 12 lapisan.

```
FullTokenizer = bert.bert_tokenization.FullTokenizer  
max_seq_length = 512  
  
bert_layer = hub.KerasLayer(  
    "https://tfhub.dev/tensorflow/bert_en_uncased_L-12_H-768_A-12/1",  
    trainable=True  
)  
  
vocab_file = bert_layer.resolved_object.vocab_file.asset_path.numpy()  
do_lower_case = bert_layer.resolved_object.do_lower_case.numpy()  
tokenizer = FullTokenizer(vocab_file, do_lower_case)
```



Fungsi Untuk mendapatkan Id Token

get_ids Fungsi untuk mengkonversi token menjadi ID token yang dipahami oleh model BERT dan memastikan panjang urutan sesuai dengan max_seq_length dengan menambahkan padding (0) jika diperlukan.

```
def get_ids(tokens, tokenizer, max_seq_length):
    token_ids = tokenizer.convert_tokens_to_ids(tokens)
    input_ids = token_ids + [0] * (max_seq_length - len(token_ids))
    return input_ids
```

Proses Ekstraksi Fitur dari Dataset Training

- Load Data Membaca dataset punctuation-small-train.csv menggunakan pandas.
- Tokenisasi dan Ekstraksi Fitur Untuk setiap baris dalam dataset:
- Tokenisasi teks menggunakan tokenizer BERT.
- Menambahkan token spesial [CLS] di awal dan [SEP] di akhir token.
- Mengkonversi token menjadi ID token dan menambahkan padding jika perlu menggunakan get_ids.
- Menyimpan ID token bersama label (deceptive) ke dalam list extracted.
- Save Data Mengkonversi list extracted menjadi DataFrame dan menyimpannya ke file punctuation-small-train-extracted.csv.

```
df = pd.read_csv('punctuation-small-train.csv')

extracted = []

for index, row in df.iterrows():
    stokens = tokenizer.tokenize(row['text'])
    stokens = ["[CLS]"] + stokens + ["[SEP]"]

    input_ids = get_ids(stokens, tokenizer, max_seq_length)

    extracted.append([input_ids, row['deceptive']])

df = pd.DataFrame(extracted, columns=['text', 'deceptive'])

df.to_csv('punctuation-small-train-extracted.csv', index=False)
```

Proses Ekstraksi Fitur dari Dataset Uji

- Load Data Membaca dataset punctuation-small-test.csv menggunakan pandas.
- Tokenisasi dan Ekstraksi Fitur Prosesnya sama seperti pada dataset latih:
- Tokenisasi teks, menambahkan token spesial, mengkonversi token menjadi ID token, dan menyimpan hasilnya ke list extracted.
- Save Data Mengkonversi list extracted menjadi DataFrame dan menyimpannya ke file punctuation-small-test-extracted.csv.

```
df = pd.read_csv('punctuation-small-test.csv')

extracted = []

for index, row in df.iterrows():
    stokens = tokenizer.tokenize(row['text'])
    stokens = ["[CLS]"] + stokens + ["[SEP]"]

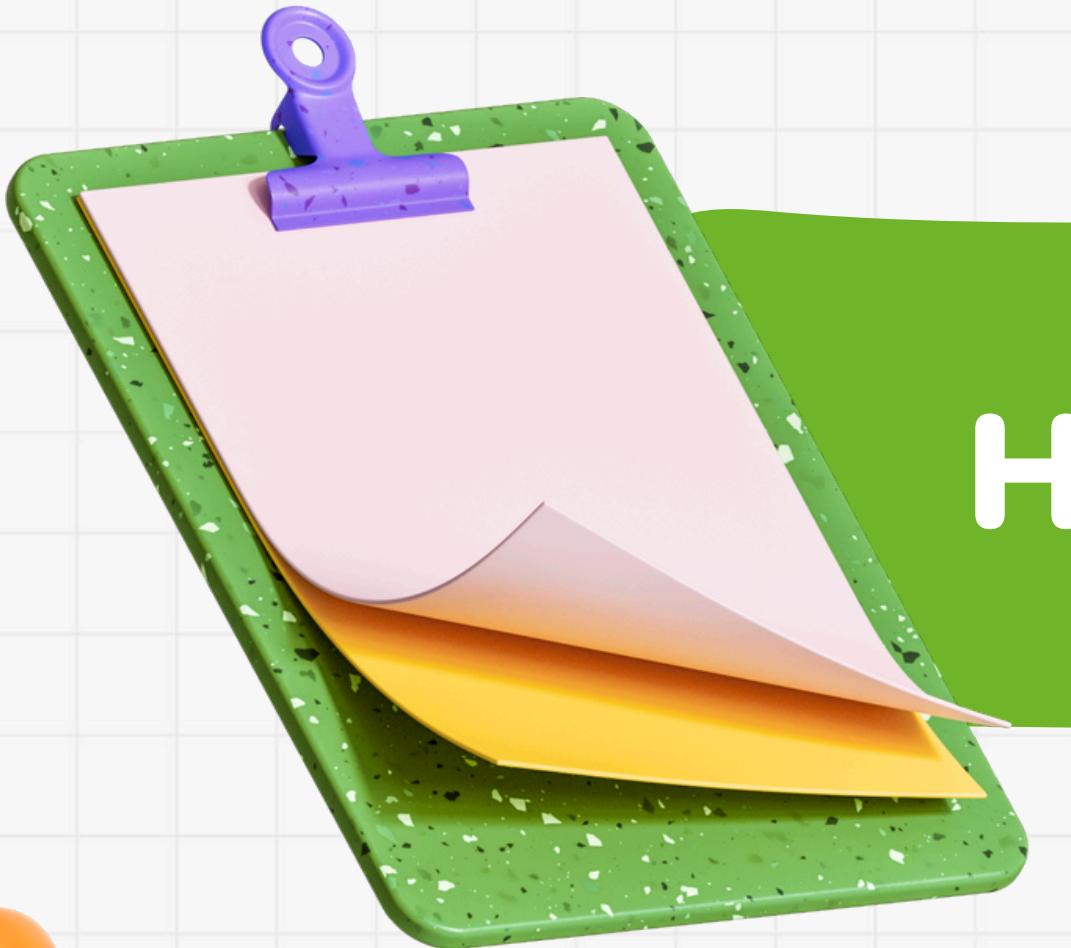
    input_ids = get_ids(stokens, tokenizer, max_seq_length)

    extracted.append([input_ids, row['deceptive']])

df = pd.DataFrame(extracted, columns=['text', 'deceptive'])

df.to_csv('punctuation-small-test-extracted.csv', index=False)
```

Hasil Perangkat Lunak



HASIL PERANGKAT LUNAK

Proses ekstraksi fitur kalimat menggunakan BERT menghasilkan konversi kalimat menjadi representasi numerik berdimensi tinggi yang dapat digunakan dalam berbagai aplikasi NLP. Representasi ini memungkinkan pemahaman yang lebih dalam terhadap struktur dan makna kalimat, serta penerapan pada berbagai model machine learning dan analisis lebih lanjut.

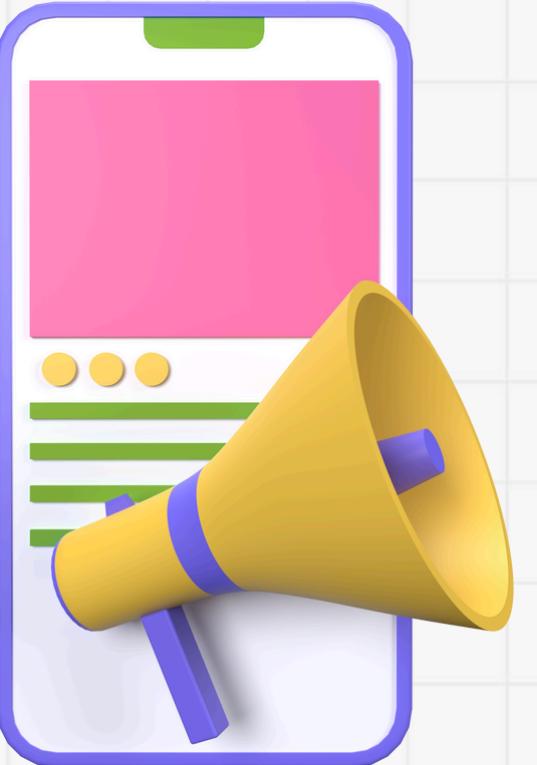


Kesimpulan



Kesimpulan

Ekstraksi fitur kalimat dengan menggunakan BERT bertujuan untuk mengubah kalimat atau teks menjadi representasi numerik yang dapat digunakan oleh model machine learning atau deep learning dalam berbagai aplikasi Natural Language Processing (NLP). BERT (Bidirectional Encoder Representations from Transformers) memberikan representasi kontekstual dari kata-kata dalam sebuah kalimat, yang lebih akurat dalam menangkap makna dan hubungan antar kata dibandingkan dengan metode tradisional. Fungsi dari ekstraksi fitur ini meliputi pemrosesan data NLP, peningkatan akurasi, dan transfer learning.



Adapun tahapan dalam pembuatan ekstraksi fitur ini antara lain instalasi dan import, inisialisasi model dan tokenizer, tokenisasi dan konversi ID, dan ekstraksi fitur dari dataset. Dari tahapan tersebut menghasilkan dua file CSV: punctuation-small-train-extracted.csv dan punctuation-small-test-extracted.csv. Setiap baris dalam file ini berisi representasi numerik teks dalam bentuk ID token sepanjang 512 token, serta label asli untuk pelatihan atau evaluasi model. Representasi ini memungkinkan pemodelan teks yang lebih akurat dan mendalam, memanfaatkan kekuatan BERT dalam memahami konteks dan hubungan antar kata.