

LAPORAN AKHIR
PEMROSESAN BAHASA ALAMI

EKSTRAKSI FITUR KALIMAT
MENGGUNAKAN ALGORITMA BERT

Kelas Reg L1 – TA Genap 23/24

Kelompok 2:

1. 09021182126004 : Aisyah Nur Khoirofiq
2. 09021182126024 : Agus Tusilawati
3. 09021182126033 : Eka Wira Yudha
4. 09021282126036 : Tasya Khadijah

TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS SRIWIJAYA

2024

Daftar Isi

Daftar Isi.....2

Deskripsi..... 2

Pengambilan Data..... 2

Tahapan Perangkat Lunak..... 2

Hasil Perangkat Lunak..... 2

Kesimpulan..... 2

BAB I DESKRIPSI

1. Deskripsi Perangkat Lunak

Ekstraksi Fitur Kalimat Menggunakan Algoritma Bert adalah sebuah sistem yang menggunakan algoritma BERT (Bidirectional Encoder Representations from Transformers) untuk mengekstrak fitur dari kalimat-kalimat dalam teks. BERT adalah sebuah model memahami konteks dalam teks.

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.

Proses ekstraksi fitur kalimat dengan algoritma BERT biasanya melibatkan langkah-langkah berikut:

- 1) Tokenisasi Kalimat: Kalimat-kalimat dalam teks dipisahkan menjadi token-token kata. Setiap token ini kemudian diubah menjadi representasi vektor.
- 2) Pengolahan Konteks: Setiap token kata diproses oleh model BERT dengan mempertimbangkan konteks sebelum dan sesudahnya. Hal ini memungkinkan BERT untuk memahami hubungan antar kata, sintaksis, dan makna secara menyeluruh.
- 3) Ekstraksi Fitur: Model BERT menghasilkan representasi vektor yang menggambarkan makna kalimat secara komprehensif. Representasi ini sering kali disebut sebagai *embedding* dan dapat digunakan sebagai fitur dalam berbagai tugas pemrosesan bahasa alami.
- 4) Tugas Selanjutnya: Fitur-fitur yang diekstraksi dapat digunakan dalam berbagai tugas pemrosesan bahasa alami seperti klasifikasi teks, pengelompokan kalimat, atau analisis sentimen. Fitur-fitur ini juga dapat menjadi masukan untuk model pembelajaran mesin lainnya atau untuk analisis lebih lanjut dalam aplikasi yang memerlukan pemahaman yang dalam terhadap teks.

Ekstraksi fitur kalimat menggunakan algoritma BERT sangat berguna dalam mengatasi tantangan pemahaman bahasa alami, karena BERT telah terbukti sangat efektif dalam memahami konteks teks dengan baik.

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.

Metode pre-training BERT melibatkan pelatihan model pada dua tugas: pertama, memprediksi kata yang di-masker (masked language model), dan kedua, memprediksi apakah dua kalimat dalam teks adalah kalimat yang berurutan (next sentence prediction). Dengan melakukan pre-training pada data teks yang sangat besar, BERT dapat belajar pemahaman bahasa secara mendalam.

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.

Secara keseluruhan, jurnal ini memberikan kontribusi yang signifikan bagi perkembangan dalam pemahaman bahasa alami dan telah memicu banyak penelitian dan inovasi di bidang tersebut.

3. Perbedaan Perangkat Lunak dan Jurnal

Perbedaan antara "Ekstraksi Fitur Kalimat Menggunakan Algoritma BERT" dan jurnal "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" dapat dijelaskan sebagai berikut:

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) 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

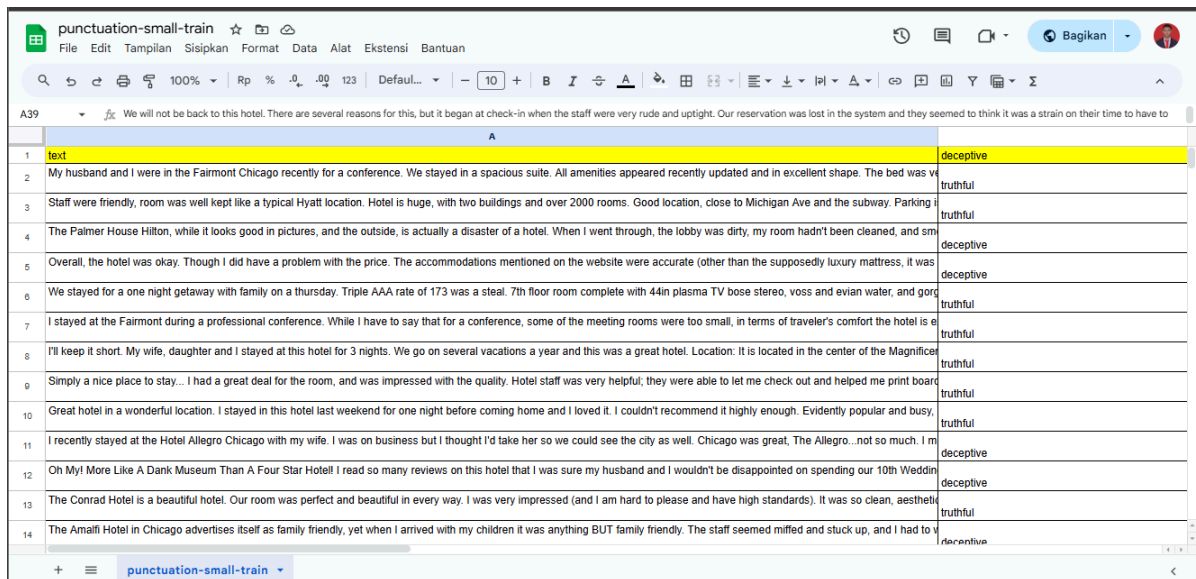
- Perangkat Lunak: Menghasilkan alat atau sistem yang dapat digunakan untuk tujuan tertentu, seperti ekstraksi fitur untuk analisis teks atau tugas-tugas pemrosesan bahasa alami lainnya.
- Jurnal: Menghasilkan pengetahuan dan pemahaman yang lebih mendalam tentang algoritma BERT itu sendiri, yang kemudian dapat digunakan oleh peneliti lain untuk memperbaiki atau mengembangkan lebih lanjut teknik pemrosesan bahasa alami.

Dengan demikian, sementara keduanya terkait dengan algoritma BERT, perbedaan utamanya terletak pada tujuan utama, konteks publikasinya, dan tingkat kedalaman analisis yang disajikan.

BAB II PENGAMBILAN DATA

Tahap 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.



	text	
A39	We will not be back to this hotel. There are several reasons for this, but it began at check-in when the staff were very rude and uptight. Our reservation was lost in the system and they seemed to think it was a strain on their time to have to	
1		A
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 ve	truthful
3	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 i	truthful
4	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 sm	deceptive
5	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	deceptive
6	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
7	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 a	truthful
8	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 Magnific	truthful
9	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 board	truthful
10	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
11	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	deceptive
12	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 Weddin	deceptive
13	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, aesthet	truthful
14	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 v	decentua

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

BAB III TAHAPAN PERANGKAT LUNAK

Dalam pembuatan program ekstraksi fitur ini ada beberapa tahapan yang harus dilalui untuk mendapatkan hasil yang terbaik, antara lain:

1. Instalasi dan Import Library

```
!pip install bert-for-tf2

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

- **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.

2. Inisialisasi Tokenizer dan Model BERT

```
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)
```

- **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.
- **vocab_file & do_lower_case** Mengambil file kosakata dan pengaturan do_lower_case dari model BERT yang diunduh.
- **tokenizer** Menginisialisasi tokenizer dengan file kosakata dan pengaturan yang telah diambil.

3. Fungsi untuk Mendapatkan ID Token

```
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
```

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.

4. Proses Ekstraksi Fitur dari Dataset Training

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

extracted = []

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

    input_ids = get_ids(tokens, 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)
```


- **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.

5. Proses Ekstraksi Fitur dari Dataset Uji

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

extracted = []

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

    input_ids = get_ids(tokens, 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)
```


- **Load Data** Membaca dataset punctuation-small-test.csv menggunakan pandas.
- **Tokenisasi dan Ekstraksi Fitur** Prosesnya sama seperti pada dataset latihan:
 - 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.

BAB IV 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.

Berikut adalah data kalimat yang belum diekstaksi menjadi representasi numerik berdimensi tinggi (data train):

punctuation-small-train.csv ×

text	1 to 10 of 70 entries	Filter 
		deceptive
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. Views were great. I love their products in the bathroom and used them in the pristine tub two out of three nights. The room was so quiet, it was very relaxing. We ate in the restaurant downstairs (Aria) and although pricey, it was excellent. The staff were consistently attentive and responsive. It is evident that everyone is very well-trained. I would love to stay here again.		truthful
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 \$48/day. Ouch! I bid on it on Priceline for \$59.		truthful
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 smelled thoroughly of smoke. When I requested more pillows, the lady on the phone scoffed at me and said she'd send them up. It took over an hour for 2 pillows. This hotel is a good example that what you pay for isn't always what you get. I will not be returning.		deceptive
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 not more comfortable than my own mattress!), but this hotel offers nothing more than a much more affordable hotel! There is a difference between the Palmer House Hilton lobby, and your normal, affordable lobby. But, I do not go to a hotel because it has a fancy lobby; I go to get a good night's sleep. When I want to sight see, I go elsewhere. Like I said, overall, the hotel was okay; but it was not worth the price.		deceptive
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 gorgeous bathroom(no tub but was fine for us) Concierge was very helpful. You cannot beat this location... Only flaw was breakfast was pricey and service was very very slow(2hours for four kids and four adults on a friday morning) even though there were only two other tables in the restaurant. Food was very good so it was worth the wait. I would return in a heartbeat. A gem in Chicago...		truthful
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 exceptional! I can't think of a single thing about my room that wasn't perfect. The accommodations were luxurious and the hotel is conveniently situated within walking distance of many Chicago attractions.		truthful

✓ Connected to Python 3 Google Compute Engine backend


Dan ini adalah data kalimat yang sudah diekstraksi menjadi representasi numerik berdimensi tinggi (menghasilkan file “punctuation-small-train-extracted.csv”):

[illegible]

Berikut adalah data kalimat yang belum diekstaksi menjadi representasi numerik berdimensi tinggi (data uji):

[illegible]

Dan ini adalah data kalimat yang sudah diekstraksi menjadi representasi numerik berdimensi tinggi (menghasilkan file “punctuation-small-test-extracted.csv”):

punctuation-small-test-extracted.csv	punctuation-small-test.csv	✕
		1 to 10 of 30 entries <input type="text" value="Filter"/> 
text		deceptive
This comes a little late as I'm finally catching up on my reviews from the past several months:) A dear friend and I stayed at the Hyatt Regency in late October 2007 for one night while visiting a friend and her husband from out of town. This hotel is perfect, IMO. Easy check in and check out. Lovely, clean, comfortable rooms with great views of the city. I know this area pretty well and it's very convenient to many downtown Chicago attractions. We had dinner and went clubbing with our friends around Division St.. We had no problems getting cabs back and forth to the Hyatt and there's even public transportation right near by but we didn't bother since we only needed cabs from and to the hotel. Parking, as is usual for Chicago, was expensive but we were able to get our car out quickly (however, we left on a Sunday morning, not exactly a high traffic time although it was a Bears homegame day, so a bit busier than usual I would think). No problems at all and the best part is that we got a rate of \$100 through Hotwire, a downright steal for this area of Chicago and the quality of the hotel.		truthful
We arrived at the Omni on 2nd September for a 6 day stay. I took ill when I left the plane after travelling from Manchester so I saw more of the room than I anticipated. I couldn't go out for 4 days. The room was spacious and clean. The bed was extremely comfortable. The bathroom was large and very clean. What more could you ask. We had coffee and juice left outside our door every morning at the time we requested. I managed to go to the 5th floor to see the sun terrace (outside the gym) Having a sun terrace in a city hotel is a great idea but the terrace was a bit grim. A few sunbeds on a concrete floor. There was also noise from an air conditioning unit. It could do with a bit of cheering up. What I saw of Chicago was very pleasing, it has something for everyone. Would I go back to the Omni? yes I would.		truthful
On our visit to Chicago, we chose the Hyatt due to its location in downtown, within walking distance to most major attractions, such as Sears Tower, Magnificent Mile, Grant/Millennium Parks, etc. Subway & bus stops very close by to travel to other locations in the city. Overall the hotel was very nice, clean, and at a great location. Was in a safe area. Went to beautiful Wrigley Field and Soldier Field and saw both teams play. I love Chicago and would definitely stay here again!		truthful
I stayed at the Fairmont Chicago for one night - I'm a frequent business traveler and am very familiar with travel rituals. I checked in late (almost 10pm) due to flight cancellations from my home airport - Atlanta. I took the shuttle to the airport (best option in lieu of the over priced taxis), and it was my first time staying at the hotel. Upon arrival, I immediately noticed the entrance, which appeared very welcoming and warm to me. I checked in with no problems -- and the desk person even asked if I'd prefer a king or double (reconfirming my reservation preference). I always take the king when I can, and the bed was fantastic! It was suited with great linen and these incredible down & feather pillows named Encompass made by The Pillow Factory -- I checked the tags, since I plan to buy some! I ate dinner at the hotel restaurant -		truthful

BAB V 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.