PROJEK AKHIR

KLASIFIKASI SENTIMEN KOMENTAR PENGGUNA APLIKASI XYZ MENGGUNAKAN INDOBERT

SULISTIA FAHRI

(PREVIEW PROJECT)

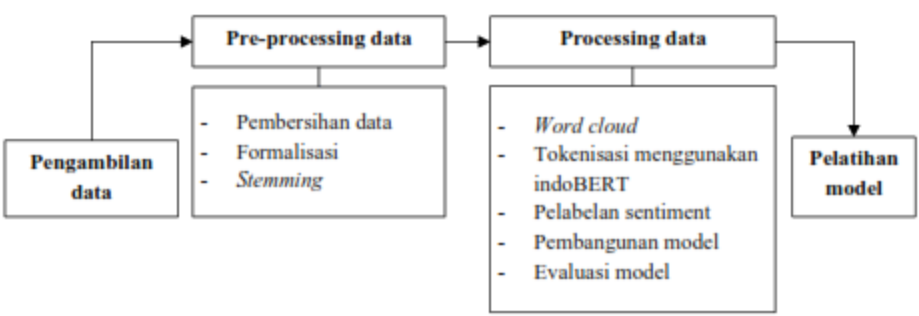
PROGRAM STUDI STATISTIKA JURUSAN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS SYIAH KUALA, BANDA ACEH
DESEMBER, 2023

Teknologi dan informasi yang berkembang menjadi perubahan yang berdampak besar khususnya dalam interaksi antar manusia. Banyak interaksi masyarakat yang saat ini dapat dilakukan secara *online* baik itu dalam bidang pendidikan, kesehatan, maupun bidang lainnya. Hal ini memudahkan akses informasi bagi masyarakat dalam melakukan aktivitas. XYZ adalah aplikasi dengan platform digital layanan kesehatan yang menyediakan informasi serta konsultasi terkait Kesehatan dengan proses daring. XYZ memiliki lebih dari satu juta pengguna sehingga banyak permasalahan yang mungkin terjadi dalam aplikasi ini. Beberapa permasalahan tersebut berupa komentar dari konten aplikasi maupun dari sistem privasi dan keamanan dalam aplikasi.

Data yang digunakan untuk dianalisis adalah respon atau komentar sentimen negatif, netral atau positif dari pengguna aplikasi XYZ. Data ini merupakan *review* pengguna bersumber dari *Google Play Store* yang akses datanya didapatkan melalui: id.codigo.XYZ. Analisis klasifikasi dilakukan dalam penelitian ini bertujuan untuk mengetahui masukan kinerja dan konten dalam aplikasi tergolong ke dalam komentar positif, netral atau bahkan negatif. Hal tersebut juga dapat membuktikan bahwa jika banyak review pengguna yang menunjukkan komentar positif maka aplikasi berfungsi dengan baik, begitu pula sebaliknya.

IMPLEMENTASI METODE DAN STRUKTUR DATA YANG DIGUNAKAN

Impelementasi metode yang digunakan dalam analisis data adalah klasifikasi indoBERT. Metode ini diaplikasikan pada data menggunakan matriks. IndoBERT atau Indonesia *Bidirectional Encoder Representations from Transformers* adalah model BERT Bahasa Indonesia dengan basis *Deep Learning* di mana model *pre-trained* yang dilatih menggunakan Transformer akan melibatkan *encoder* dan *decoder*. Berikut merupakan struktur Langkah-langkah penggunaan metode yang dilakukan dalam analisis data:



ANALISA *SUMMARY DATA*

a) Pengambilan data

Pengambilan data dilakukan berdasarkan komentar sentimen pengguna aplikasi yang diambil dari *Google Play Store*.

```
!pip install google_play_scraper
```

```
Collecting google_play_scraper
  Downloading google_play_scraper-1.2.4-py3-none-any.whl (28 kB)
Installing collected packages: google_play_scraper
Successfully installed google_play_scraper-1.2.4
```

```
from google_play_scraper import Sort, reviews_all
import pandas as pd
```

```
result = reviews_all(
        'id.codigo.xyz',
        sleep_milliseconds=0, # defaults to 0
        lang='id', # defaults to 'en'
        country='id', # defaults to 'us'
        sort=Sort.NEWEST
    )

df = pd.DataFrame.from_records(result)
df = df[['at', 'content', 'score', 'userName']]
```

Pemanggilan data XYZ menggunakan bantuan dari library pandas dan google_play_scraper adalah library Python yang memungkinkan pengambilan data dari *Google Play Store* berupa ulasan aplikasi XYZ dengan nama *package* aplikasi id.codigo.XYZ. dengan lang dan country disesuaikan dengan Bahasa dan negara yang berkode "id" berupa Indonesia.

```
df['at'] = pd.to_datetime(df['at'])

# mulai 2020-01-01
filtered_df = df[df['at'] >= '2020-01-01']
```

```
#download ke csv
filtered_df.to_csv('data_KlikDokter.csv', index = False)
```

```
filtered_df.head()
```

	at	content	score	userName
0	2023-12-18 16:25:03	Biasanya sering konsultasi disini walau pelaya...	1	Raablgs
1	2023-12-18 11:09:34	Udah bayar uang konsultasi udah chat sama dokt...	1	Emmi Arisasmita
2	2023-12-13 11:50:00	Aplikasi gak jelas Lick bgt cari duit Sama sa...	1	Gandi Rafan
3	2023-12-06 10:17:25	Payah ah, dikasih waktu 30 menit, tapi doktern...	2	Sadam Hadiansyah
4	2023-11-29 06:02:07	👍	5	Rizdath Production

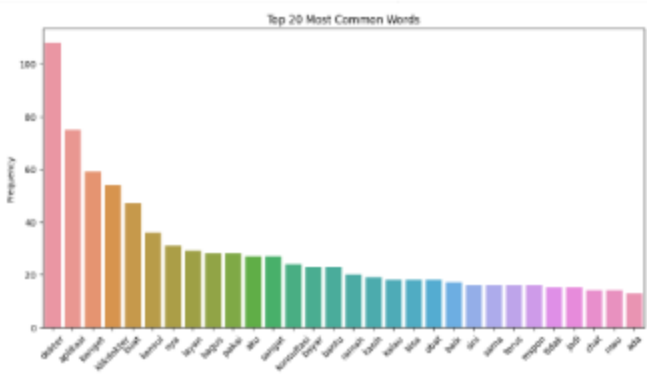
Pemanggilan data dilakukan mulai tanggal 1 Januari 2020 hingga data terbaru pada tanggal 18 Desember 2023. Data ini kemudian disimpan dalam bentuk csv, selanjutnya dilakukan penampilan data ter-atas berupa komentar sentimen pengguna aplikasi XYZ.

b) *Pre-processing data*

Pembersihan data meliputi penghilangan karakter-karakter yang tidak diperlukan atau tidak memberikan makna penting dalam sebuah sentimen. Seperti penjelasan sebelumnya terkait proses-proses yang dilakukan dalam *Pre-processing* ini meliputi *Initial Filtering, case folding, remove stopword,* Formalisasi, dan *Stemming*.

c) *Summary data*

Proses penampilan gambar pada *summary data* ditampilkan dalam chart kata paling banyak digunakan dan gambar *word cloud*.



Berdasarkan gambar di atas menunjukkan bahwa kata paling banyak disebutkan dalam komentar sentimen pengguna aplikasi adalah dokter, aplikasi, banget, dan seterusnya.

Berdasarkan gambar di atas menunjukkan bahwa sentimen relatif positif. Hal tersebut dapat dilihat melalui *reviews* atau kata-kata yang sering muncul cenderung bermakna positif seperti dokter, aplikasi, banget, XYZ, buat, dan seterusnya kata yang cenderung terlihat menunjukkan sifat yang positif adalah ramah, bantu, bagus, dan sebagainya.

ANALISIS DATA

1. Analisis sentiment

	contentp_clean	sentiment
0	biasa sering konsultasi sini layan nya yaaa se...	negative
1	bayar uang konsultasi chat sama dokter ndg bal...	negative
2	aplikasi jelas licik banget cari uang sama tip...	negative

Proseses ini digunakan untuk memetakan skor sentimen ke dalam kelas sentimen yang bersifat negatif (untuk skor 1 atau 2), netral (untuk skor 3), dan postif (untuk skor lainnya)

2. *Encoding* label pada kelas sentiment menjadi angka

```
from sklearn.preprocessing import LabelEncoder
label_encoder = LabelEncoder()
df["label"] = label_encoder.fit_transform(df["sentiment"])
np.save('bert_classes.npy', label_encoder.classes_)
```

```
df
```

	contentp_clean	sentiment	label
0	biasa sering konsultasi sini layan nya yaaa se...	negative	0
1	bayar uang konsultasi chat sama dokter ndg bal...	negative	0
2	aplikasi jelas licik banget cari uang sama tip...	negative	0
3	payah ah kasih waktu menit dokter super lambat...	negative	0
5	aku suka banget sama apk sangat bantu terima k...	positive	2
...	...	...	...
195	dulu pakai aplikasi klikdokter makin kesini ma...	positive	2
196	sangat luar biasa aplikasi kasih saya antusias...	positive	2
197	mantap aplikasi	positive	2
198	aplikasi enak pakai sih tampil oke banget buat...	positive	2
199	asli aplikasi nyaman pakai fiturnya lengkap ja...	positive	2

198 rows × 3 columns

Proses ini perlu dilakukan tujuannya adalah agar mempermudah penggunaan dalam model BERT. Beberapa fungsi yang digunakan adalah LabelEncoder (merupakan fungsi dari scikit-learn) untuk meng-*encode* label kategori menjadi angka. Kemudian hasil yang didapatkan dibuat dalam kolom baru yaitu “Label”, hasil kelas sentimen yang didapatkan disimpan ke dalam bert_classes.npy menggunakan fungsi np.save().

Proses perubahan label menjadi angka tersebut perlu dilakukan karena BERT hanya menerima input angka.

3. Pembuatan model BERT

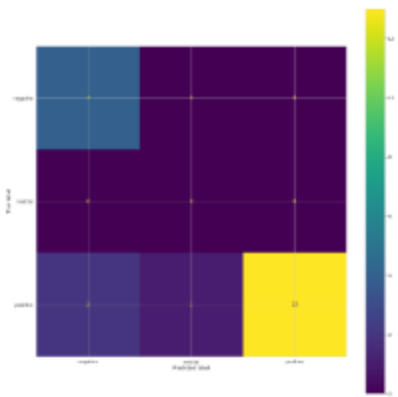
   Pembuatan model akan meliputi banyak proses termasuk melibatkan DataLoader dan pembagian data yang dilakukan sebelumnya. Proses pembuatan model ini dapat dievaluasi menggunakan data test.

```
Truncation was not explicitly activated but `max_length` is provided a specific value, please use `truncation=True` to explicitly truncate
Truncation was not explicitly activated but `max_length` is provided a specific value, please use `truncation=True` to explicitly truncate
/usr/local/lib/python3.10/dist-packages/transformers/tokenization_utils_base.py:2614: FutureWarning: The `pad_to_max_length` argument is
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/transformers/tokenization_utils_base.py:2614: FutureWarning: The `pad_to_max_length` argument is
  warnings.warn(
0.8500000000000001
```

   Hasil evaluasi model menggunakan data test menunjukkan nilai *accuracy* yang didapatkan adalah sebanyak 0.8500000000000001 atau 85%.

4. Evaluasi model

   Evaluasi model dapat ditampilkan dalam bentuk visualisasi menggunkan bantuan library skelarn dan matholip.



   Hasil evalusi menggunakan model indoBERT secara lebih rinci adalah sebagai berikut:

```
print(classification_report(ytest, ypred))
```

	precision	recall	f1-score	support
negative	0.67	1.00	0.80	4
neutral	0.00	0.00	0.00	0
positive	1.00	0.81	0.90	16
accuracy			0.85	20
macro avg	0.56	0.60	0.57	20
weighted avg	0.93	0.85	0.88	20

Berdasarkan hasil evaluasi di atas menunjukkan bahwa nilai *accuracy* yang dihasilkan adalah sebanyak 85% nilai ini tergolong tinggi, Adapun nilai *precision*, *recall*, dan *F1-score* yang dihasilkan masing-masing nilainya adalah:

Label	***Precision***	***Recall***	***F1-score***
Negatif	0,67	1,00	0,80
Netral	0,00	0,00	0,00
Positif	1,00	0,80	0,90

5. Prediksi sentiment baru

```
if __name__ == "__main__":
        df = predict("Aplikasinya sangat membantu banyak pertanyaan kesehatan")
        #df = df[['tanggal', 'bundle', 'label', 'is_sentiment', 'contentp_clean']]
        encoder = LabelEncoder()
        encoder.classes_ = np.load('bert_classes.npy', allow_pickle=True)
        model = SentimentClassifier(3)
        model.load_state_dict(torch.load('bert_symptoms.bin',  map_location=torch.device('cpu')))
        model = model.to(device)
        testing_data_loader = create_data_loader(df, tokenizer, MAX_LEN, BATCH_SIZE)
        y_review_texts, y_pred, y_pred_probs = get_predictions(
          model,
          testing_data_loader
        )
        ypred = encoder.inverse_transform(y_pred)
        df["Topic_category"] = ypred
        print(df)
```

```
/usr/local/lib/python3.10/dist-packages/torch/utils/data/dataloader.py:557: UserWarning: This DataLoader will create 4 worker processes
  warnings.warn(_create_warning_msg(
Truncation was not explicitly activated but `max_length` is provided a specific value, please use `truncation=True` to explicitly trunca
/usr/local/lib/python3.10/dist-packages/transformers/tokenization_utils_base.py:2614: FutureWarning: The `pad_to_max_length` argument is
  warnings.warn(
                                contentp_clean Topic_category
0  Aplikasinya sangat membantu banyak pertanyaan ...       positive
```

```
df
```

	contentp_clean	Topic_category
0	Aplikasinya sangat membantu banyak pertanyaan ...	positive

Kalimat sentimen baru yang diprediksi yaitu:

"Aplikasinya sangat membantu banyak pertanyaan Kesehatan"

Hasil kategori dari kalimat tersebut menunjukkan bahwa sentimen termasuk ke dalam topik **positif** dan hal tersebut sesuai dengan konten dalam teks.