3. ANALISIS DAN DESAIN SISTEM

Pada bab ini akan dibahas lebih lanjut mengenai analisis dan desain sistem yang digunakan dalam pembuatan website prediksi genre dan rating penonton berdasarkan sinopsis film. Bab ini akan terdiri dari 4 sub-bab, yaitu Analisis Permasalahan dan Kebutuhan, Analisis Data yang digunakan, Flowchart Desain Sistem, dan Desain User Interface website.

3.1 Analisis Permasalahan dan Kebutuhan

Penelitian ini bertujuan untuk memprediksi vote average dan genre berdasarkan sinopsis film berbahasa Indonesia, sebuah area yang belum banyak dieksplorasi dibandingkan dengan sinopsis berbahasa Inggris. Salah satu permasalahan utama yang dihadapi adalah keterbatasan dataset sinopsis film dalam bahasa Indonesia. Mayoritas penelitian sebelumnya menggunakan dataset berbahasa Inggris, yang menyebabkan adanya kekurangan referensi dan benchmark dalam bahasa Indonesia. Hal ini menjadi tantangan tersendiri dalam memastikan kualitas data yang digunakan untuk melatih model.

Kebutuhan akan model yang akurat dan efisien sangat penting bagi industri film, baik untuk rekomendasi film kepada penonton maupun untuk analisis pasar oleh produser dan distributor film. Oleh karena itu, penelitian ini memerlukan pengumpulan dataset sinopsis film berbahasa Indonesia yang memadai dan berkualitas. Dataset tersebut harus diolah dan dibersihkan untuk memastikan validitas dan reliabilitas data yang digunakan dalam pelatihan model. Penelitian ini akan membandingkan dua model utama, yaitu IndoBERT dan Indonesian RoBERTa, yang keduanya dirancang khusus untuk bahasa Indonesia namun memiliki arsitektur yang sedikit berbeda. Perbandingan ini penting untuk menentukan model mana yang lebih akurat dan efisien dalam konteks prediksi vote average dan genre film.

Selain itu, implementasi dan evaluasi dua model bahasa alami, IndoBERT dan Indonesian RoBERTa, menjadi kebutuhan utama dalam penelitian ini. Perbandingan performa kedua model ini diharapkan dapat memberikan wawasan yang berharga mengenai model mana yang lebih unggul dalam memprediksi vote average dan genre berdasarkan sinopsis film berbahasa Indonesia. Dengan demikian, penelitian ini tidak hanya berkontribusi pada pengembangan teknologi prediktif dalam konteks bahasa Indonesia, tetapi juga membuka jalan bagi penelitian

lebih lanjut yang dapat memanfaatkan model bahasa alami untuk berbagai aplikasi dalam industri kreatif.

3.2 Analisis Data

Data yang digunakan dalam skripsi ini adalah data dari TMDB. TMDb adalah singkatan dari "The Movie Database". Ini adalah platform daring yang menyediakan informasi terkait film dan acara televisi. TMDb adalah database yang dikuratori oleh pengguna, artinya pengguna dari seluruh dunia dapat menyumbangkan dan memperbarui informasi tentang film, termasuk detail seperti judul, genre, sinopsis, pemeran, staf produksi, tanggal rilis, poster, dan banyak lagi.

Selain itu, TMDb juga menyediakan API (Application Programming Interface) yang memungkinkan pengembang perangkat lunak untuk mengakses data TMDb dan mengintegrasikannya ke dalam aplikasi atau situs web mereka. Ini memungkinkan pengembang untuk membuat aplikasi atau layanan yang memanfaatkan informasi film dari TMDb, seperti aplikasi streaming, situs review film, dan banyak lagi. TMDb adalah salah satu sumber data utama yang digunakan oleh banyak aplikasi dan situs web terkait film di seluruh dunia.

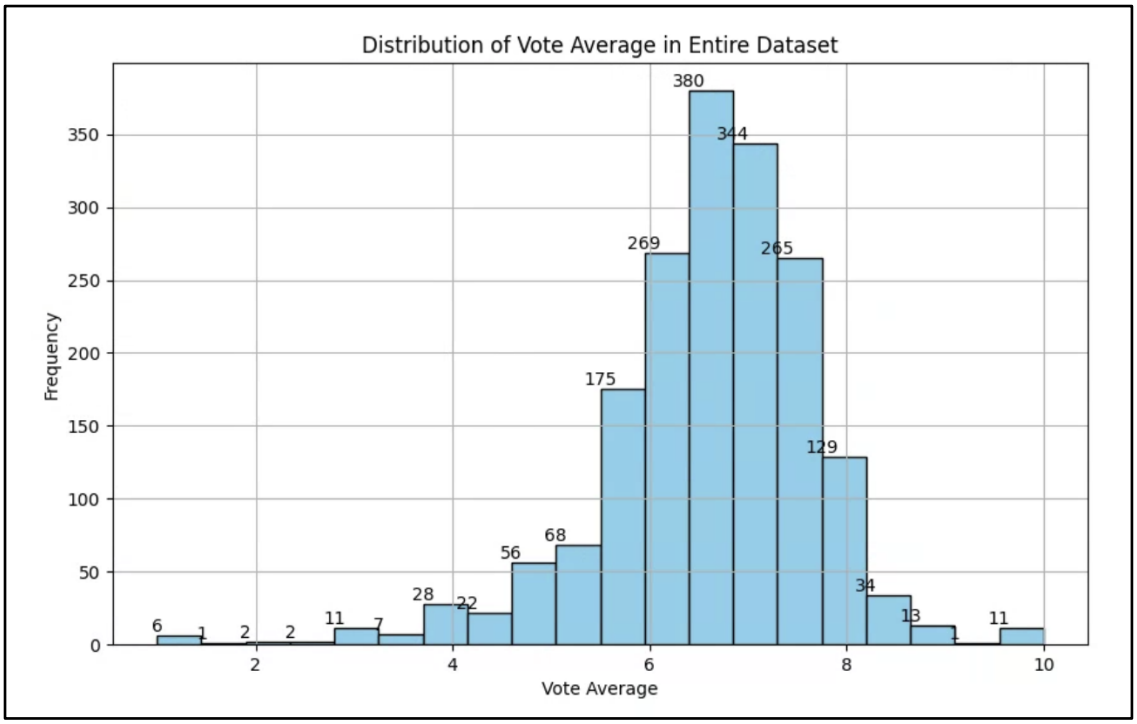
```
                                                  overview  vote_average
26      Menyusul bunuh diri seorang pria, waktu berjal...         7.498
2730    Dua orang penipu asal Spanyol memenangkan peta...         7.300
2736    Taxi 2 merupakan sebuah film asal Prancis yang...         6.244
4454    Setahun setelah membuang jenazah pria yang tid...         6.355
4516    Di dunia di mana mutan (manusia berkekuatan su...         7.002
...                                                   ...           ...
503510  Diadopsi keluarga yang sama setelah masa kecil...         6.502
503512  Ryo Saeba si buaya darat adalah detektif swast...         6.500
503640  Sekar, berusaha menghidupi 3 anaknya sendirian...         7.000
506538  Pada masa susu dan sereal ialah menu wajib sar...         5.279
506545  Film dokumenter ini menyelidiki berbagai miste...         7.667

                      genre_ids
26                         [18]
2730    [10751, 12, 16, 35, 14]
2736                  [28, 35]
4454                      [35]
4516              [12, 28, 878]
...                         ...
503510             [10749, 18]
503512            [28, 18, 35]
503640                [18, 27]
506538                [35, 36]
506545                    [99]

[1824 rows x 3 columns]
```

Gambar 3.1 Bentuk dataset

Dari Gambar 3.1 dapat dilihat bahwa kita hanya menggunakan 3 kolom dari dataset, yaitu id, overview, vote average, dan genre_ids. Setelah data di cleaning dan preprocessing, data yang akan digunakan memiliki 1.824 row. Dataset yang digunakan memiliki distribusi di Gambar 3.2 dan statistik di Gambar 3.3.

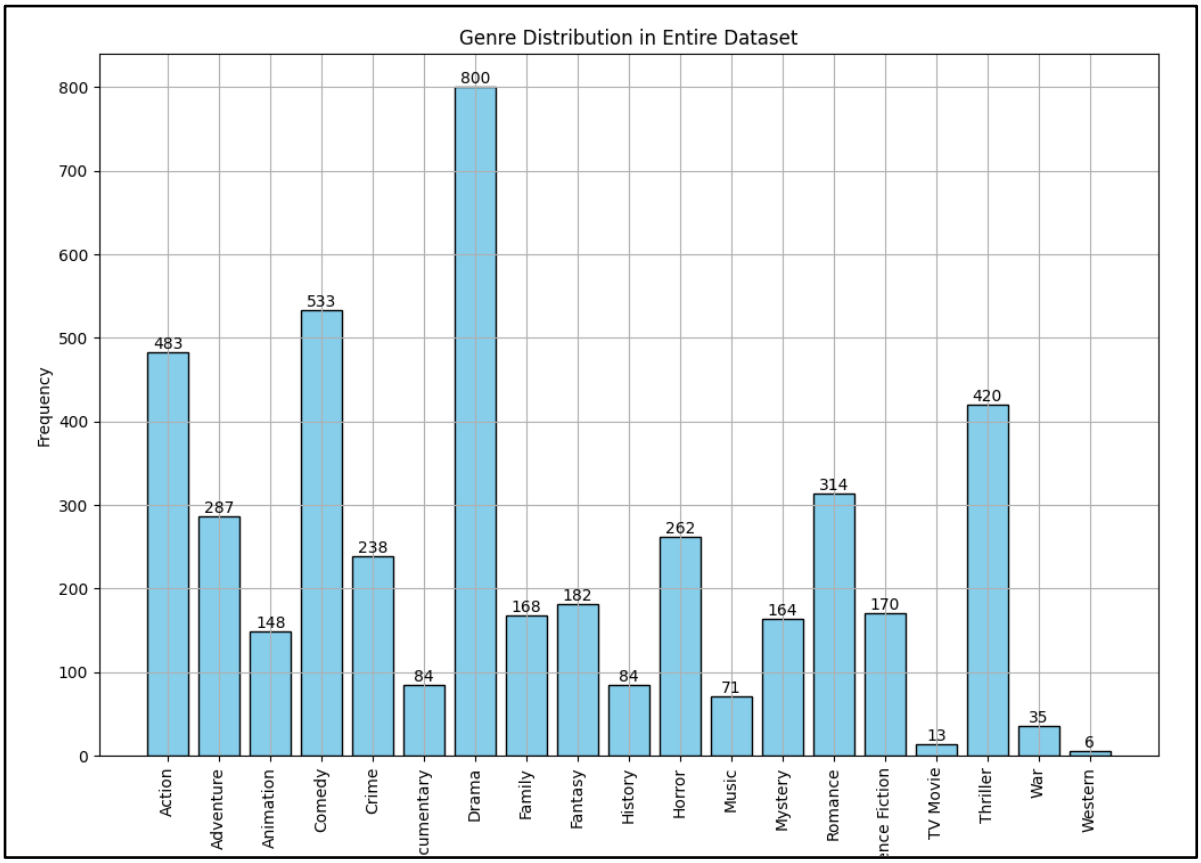


Gambar 3.2 Distribusi Vote Average

```
Statistics for 'vote_average' column in Entire Dataset:
Mean: 6.602064692982457
Standard Deviation: 1.0756575547948237
Minimum Value: 1.0
25th Percentile: 6.05425
50th Percentile (Median): 6.7
75th Percentile: 7.2837499999999995
Maximum Value: 10.0
Width of each bin: 0.45
```

Gambar 3.3 Analisis Vote Average

Untuk genre pada dataset, distribusi data digambarkan pada gambar 3.4 dibawah.



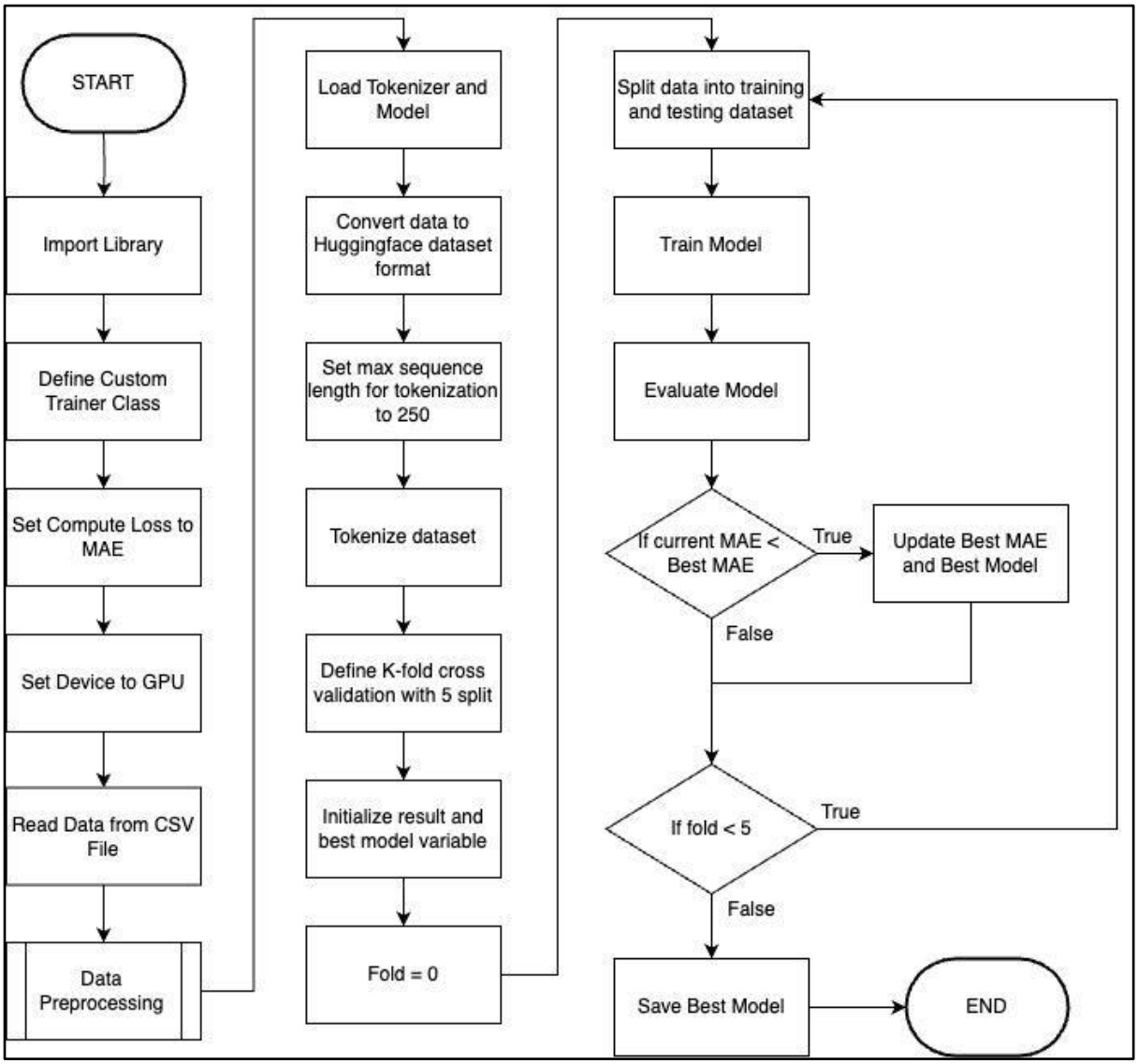
Gambar 3.4 Distribusi Genre

3.3 Desain Sistem

Desain sistem membahas alur terbentuknya input data, proses dan output yang diharapkan. Juga disertakan user interface berbentuk Website untuk memprediksi genre dan rating penonton dari sinopsis yang diinputkan

3.3.1 Model

Model yang digunakan dalam skripsi ini adalah IndoBERT dan Indonesian RoBERTa. Saya menggunakan base model yang sudah di pretrained, kemudian melanjutkan training dengan dataset yang saya miliki. Alur training dapat dilihat di gambar 3.4



Gambar 3.5 Flowchart Training Model

Dari flowchart di gambar 3.5, dapat dilihat bahwa kita pertama mengeset Loss function menjadi MAE, load dataset, melakukan Data Preprocessing yang dijabarkan di gambar 3.6, lalu kita convert ke dalam format dataset hugging face, tokenisasi dengan max length 250, lalu melakukan training dan evaluasi dengan 5 Fold Cross Validation hingga mendapatkan model dengan MAE terendah.

Berikut adalah penjelasan mengenai pertimbangan dalam pemilihan parameter yang digunakan:

- **num_train_epochs=20:**

- Jumlah epoch ditetapkan sebanyak 20 untuk memberikan waktu yang cukup bagi model untuk belajar pola dari data. Ini memastikan model dapat konvergen dengan baik, tetapi juga harus dipantau untuk menghindari overfitting.

- **per_device_train_batch_size=16 dan per_device_eval_batch_size=16:**
  - Batch size untuk pelatihan dan evaluasi diatur sebesar 16 per perangkat (GPU/CPU). Ini adalah nilai yang umum digunakan yang menyeimbangkan antara efisiensi memori dan stabilitas pelatihan. Batch size yang terlalu besar dapat menyebabkan penggunaan memori yang berlebihan, sementara yang terlalu kecil dapat memperlambat pelatihan.
- **warmup_steps=500:**
  - Jumlah warmup steps ditetapkan sebesar 500 untuk menghindari lonjakan awal dalam pembaruan parameter yang dapat mengakibatkan ketidakstabilan dalam pelatihan. Warmup steps membantu dalam memulai pembelajaran secara lebih halus.
- **evaluation_strategy="epoch":**
  - Strategi evaluasi ditetapkan pada setiap epoch, yang berarti model akan dievaluasi setelah setiap epoch selesai. Ini memberikan gambaran performa model secara berkala dan membantu dalam memantau overfitting atau underfitting.
- **save_strategy="no":**
  - Strategi penyimpanan diatur ke "no", yang berarti model tidak akan disimpan secara otomatis pada setiap epoch atau langkah. Ini dapat menghemat ruang penyimpanan, tetapi juga berarti pengguna harus memastikan untuk menyimpan model terbaik secara manual jika diperlukan.
- **max_length=250:**
  - Menetapkan panjang maksimum token yang dihasilkan oleh tokenizer. Rata-rata sinopsis film memiliki panjang sekitar 200 kata, sehingga menetapkan max_length sedikit lebih tinggi pada 250 membantu memastikan semua informasi penting dalam sinopsis terjaga tanpa terpotong.

  - Dengan menetapkan max_length=250, kita memastikan bahwa sinopsis yang lebih panjang tidak dipotong secara berlebihan, sementara sinopsis yang lebih pendek tetap diproses dengan padding untuk mencapai panjang tetap.

Pemilihan parameter ini bertujuan untuk menciptakan proses pelatihan yang efisien dan stabil, dengan evaluasi berkala untuk memastikan model belajar dengan baik dari data tanpa overfitting. Selain itu, pengaturan direktori output dan logging membantu dalam manajemen hasil pelatihan dan pemantauan proses.

3.3.2 Scraping Data

Dalam penelitian ini, data yang digunakan diperoleh melalui proses scraping dari API situs TMDb (The Movie Database) menggunakan akun developer. Proses ini menghasilkan 3.625 data mentah yang kemudian melalui tahap cleaning dan preprocessing untuk memastikan kualitas data yang lebih baik. Setelah proses tersebut, jumlah data yang siap digunakan berkurang menjadi 1.824 data. Scraping dari TMDb ini merupakan langkah penting dalam penelitian ini karena menyediakan informasi film yang komprehensif dan relevan untuk dianalisis lebih lanjut.

Dari data yang diperoleh, penelitian ini fokus pada tiga kolom utama, yaitu overview, genre_ids, dan vote_average. Kolom overview memberikan deskripsi singkat tentang plot atau isi film, sementara genre_ids mencakup informasi mengenai jenis atau kategori film berdasarkan identifikasi numerik (array of integer) yang telah ditetapkan oleh TMDb. Kolom vote_average menunjukkan rata-rata penilaian atau rating yang diberikan oleh pengguna TMDb. Dengan menggunakan ketiga kolom ini, penelitian bertujuan untuk mengeksplorasi hubungan antara genre dan rating film, serta bagaimana deskripsi film dapat mempengaruhi persepsi dan penilaian penonton.

3.3.3 Data Cleaning dan Preprocessing

Data cleaning dan preprocessing merupakan tahapan penting dalam penelitian data yang bertujuan untuk menghasilkan dataset yang berkualitas sebelum dilakukan analisis lebih lanjut. Tujuannya adalah agar model yang dibuat dapat menjadi akurat, dapat diandalkan, dan relevan dengan tujuan penelitian yang ditetapkan. Data cleaning yang diterapkan dalam penelitian ini adalah:

- Menghilangkan data NULL

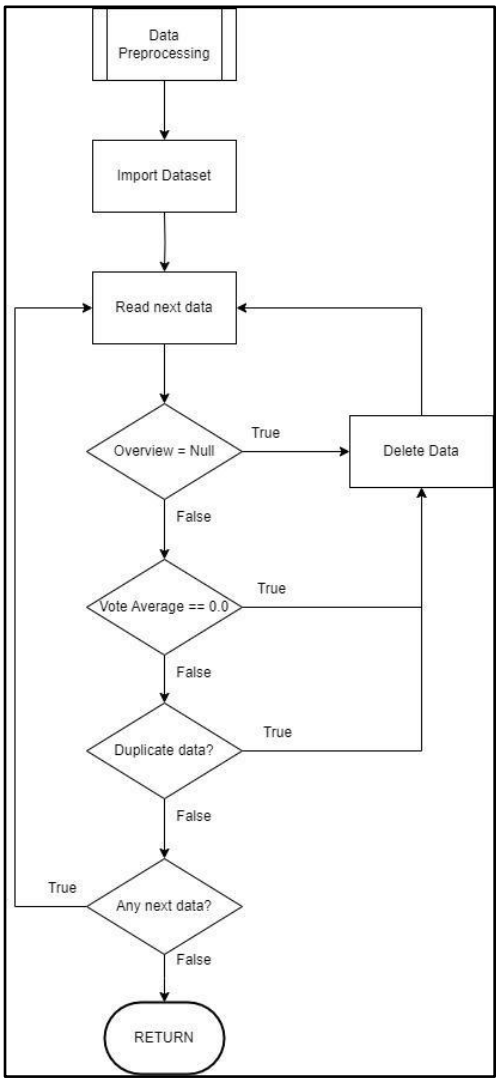
  Data yang memiliki sinopsis null dihilangkan, bisa disebabkan karena kesalahan input atau tidak memiliki sinopsis dalam bahasa Indonesia
- Menghilangkan data duplikat

Untuk memastikan tidak ada data yang duplikat agar model tidak bias

- Drop data dengan Vote Average 0.0 (belum pernah di rating/belum pernah tayang)

  Data dengan Vote Average 0.0 menandakan bahwa film tersebut belum pernah di rating, belum pernah tayang, atau masih dalam proses pembuatan, sehingga data tersebut dihapus.

Flowchart data cleaning dapat dilihat pada Gambar 3.6.



Gambar 3.6 Flowchart Data Preprocessing

Untuk data preprocessing dan tokenization dalam penelitian ini, dibantu dengan library "AutoTokenizer" dimana yang dilakukan oleh library tersebut adalah:

- Lowercase

- Trim Whitespace (di awal dan akhir kalimat)
- Menghapus
- Tokenisasi

  Tokenizer memecah teks input menjadi unit-unit yang lebih kecil, yang disebut token. Pada model IndoBERT, tokenisasi dilakukan dengan menggunakan teknik WordPiece, yang memecah kata menjadi sub-kata atau bahkan karakter jika diperlukan.
- Penambahan Token Khusus

  Tokenizer menambahkan token khusus ke dalam teks, seperti [CLS] di awal setiap teks dan [SEP] di akhir setiap teks atau setiap pasangan teks. Token [CLS] digunakan sebagai representasi keseluruhan dari teks input untuk tugas klasifikasi, sedangkan token [SEP] digunakan untuk memisahkan teks dalam tugas yang melibatkan pasangan teks, seperti tugas pemahaman dua kalimat.
- Konversi ke ID Token sesuai model
- Padding dan Truncation

  Teks yang telah di-tokenisasi kemudian dipadatkan (padding) atau dipotong (truncated) agar semua input memiliki panjang yang sama. Padding dilakukan dengan menambahkan token khusus [PAD] hingga teks mencapai panjang maksimal yang telah ditentukan. Truncation dilakukan dengan memotong teks yang terlalu panjang sehingga sesuai dengan panjang maksimal yang diizinkan oleh model.
- Pembuatan Attention Mask

  Tokenizer juga menghasilkan attention mask, yang merupakan array biner yang menunjukkan token mana yang harus diperhatikan oleh model. Token yang merupakan bagian dari teks asli mendapatkan nilai 1, sementara token yang ditambahkan sebagai padding mendapatkan nilai 0.

3.3.4 Tokenization

Tokenisasi adalah proses dalam pemrosesan bahasa alami yang memecah teks menjadi unit-unit kecil yang disebut token. Dalam konteks data cleaning dan preprocessing, tokenisasi sangat penting karena memungkinkan pemrosesan data yang lebih lanjut dengan lebih efisien.

Misalnya, dalam analisis teks, tokenisasi memungkinkan pembagian kalimat menjadi kata-kata atau frasa-frasa yang dapat diolah secara terpisah. Ini membantu dalam membersihkan data dari karakter-karakter yang tidak diinginkan, seperti tanda baca atau karakter khusus, dan memastikan bahwa data siap digunakan untuk analisis lebih lanjut. Selain itu, tokenisasi juga memungkinkan representasi data yang lebih terstruktur, memudahkan pemodelan data dan ekstraksi fitur-fitur penting untuk tujuan penelitian. Dengan demikian, tokenisasi merupakan langkah awal yang krusial dalam mempersiapkan data untuk analisis data yang berkualitas.

Dalam penelitian ini, metode tokenisasi yang digunakan adalah tokenisasi berbasis model pralatih dari Hugging Face, yaitu "cahya/roberta-base-indonesian-522M" dan "indolem/indobert-base-uncased". Setelah proses tokenisasi, langkah selanjutnya adalah mentransformasikan data teks menjadi representasi numerik yang dapat digunakan dalam model machine learning. Transformasi ini dilakukan dengan mengubah teks menjadi vektor menggunakan model transformator pralatih. Model ini tidak hanya melakukan tokenisasi tetapi juga mengubah setiap token menjadi representasi vektor yang mempertimbangkan konteks dan hubungan antar kata dalam kalimat. Seluruh proses tokenisasi dilakukan dengan library “AutoTokenizer” dari Hugging Face.

Proses transformasi ini melibatkan langkah-langkah berikut:

1. **Tokenisasi**: Menggunakan `AutoTokenizer` dari Hugging Face untuk memecah teks menjadi token dengan mempertahankan informasi konteks.
2. **Penyandian dan Trunkasi**: Token hasil tokenisasi disandikan menjadi ID numerik dan di-trunkasi atau dipadatkan hingga panjang maksimum yang ditentukan (misalnya, 250 token).
3. **Representasi Vektor**: Menggunakan model "cahya/roberta-base-indonesian-522M" dan "indolem/indobert-base-uncased", teks yang sudah ditokenisasi kemudian diproses untuk menghasilkan vektor yang mewakili setiap token dalam konteks kalimat.

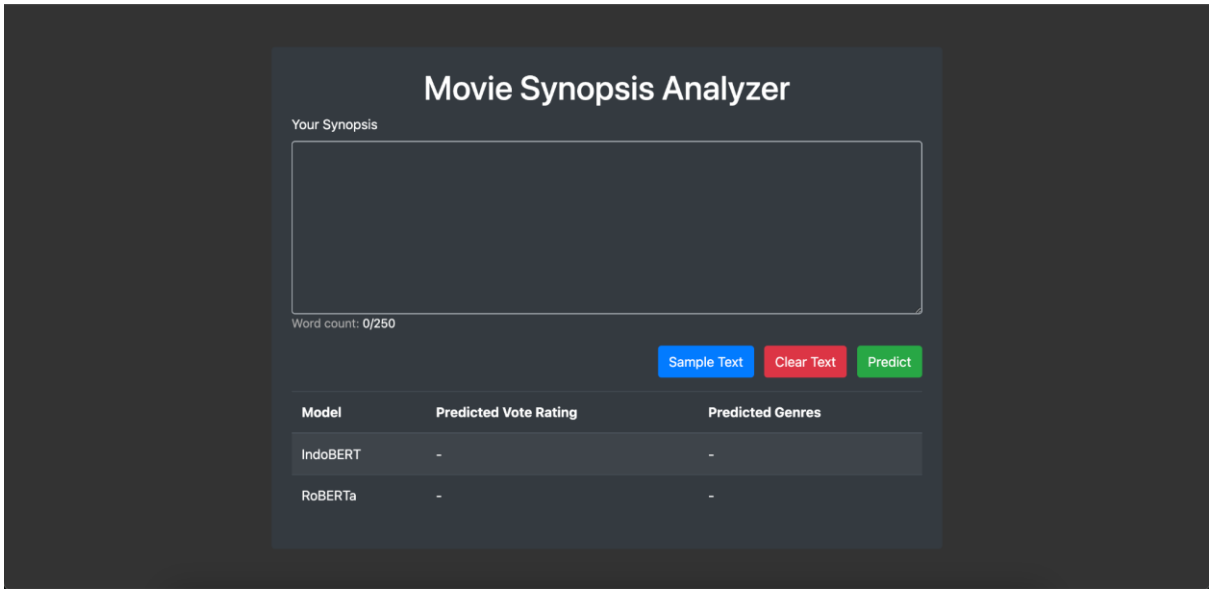
Dalam penelitian ini, dua model pralatih, yaitu RoBERTa Base dan IndoBERT, digunakan dan dibandingkan untuk mengevaluasi performanya dalam tugas regresi. Model pertama yang digunakan adalah "cahya/roberta-base-indonesian-522M" yang dioptimalkan untuk bahasa Indonesia, dan model kedua adalah "indolem/indobert-base-uncased", yang juga merupakan model pralatih untuk bahasa Indonesia. Kedua model ini digunakan untuk mentransformasikan

teks menjadi representasi vektor yang kaya informasi dan mempertimbangkan konteks penggunaan kata-kata dalam kalimat.

Dengan demikian, proses tokenisasi dan transformasi teks menjadi vektor dalam penelitian ini menggunakan dua model pralatih berbasis transformator dari Hugging Face, yaitu RoBERTa Base dan IndoBERT.

3.4 Desain Website

Desain website untuk skripsi ini saya buat minimalis dan sesederhana mungkin agar dapat digunakan oleh semua kalangan. Website ini memiliki 1 buah textbox untuk user menginput sinopsis film serta sebuah tombol untuk memprediksi sinopsis film tersebut. Kemudian di bagian bawah tombol akan muncul tabel hasil dari prediksi sinopsis film tersebut. Desain website digambarkan pada gambar 3.7.



Gambar 3.7 Design UI