



YOUTUBE COMMENT SENTIMENT ANALYSIS AND SUMMARIZATION FROM 'COURSERA REVIEW' VIDEOS

FELIX CRISTIANO BUNGARAN - 2602170612

KAYLA MASAYUNINGTYAS - 2602141871

MEISA KAMILIA - 2602135446

CONTENT

- 01** LATAR BELAKANG
- 02** METODOLOGI
- 03** HASIL DAN EVALUASI
- 04** KESIMPULAN

LATAR BELAKANG

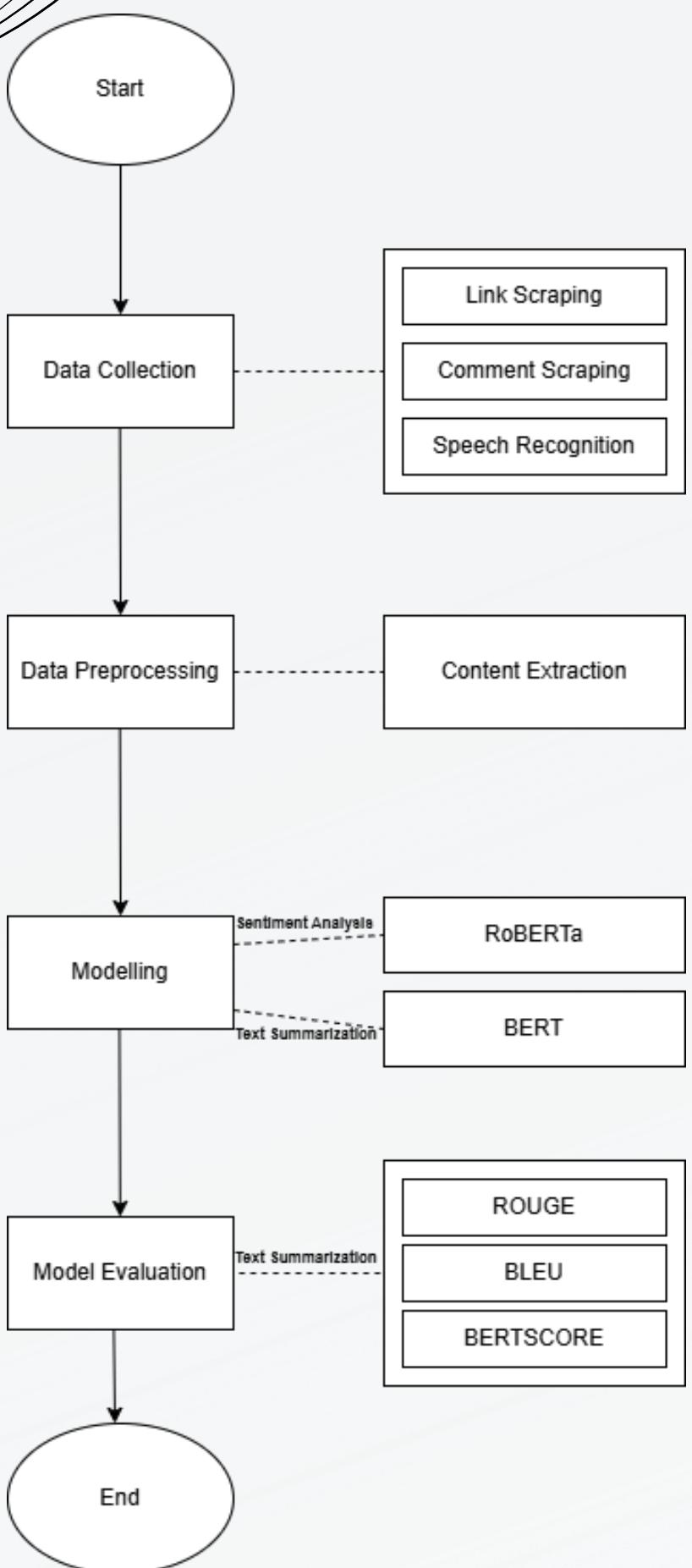


MOOCs seperti Coursera telah berkembang pesat, dari **300.000** pengguna pada 2011 menjadi 220 juta pada 2021 (McKinsey, 2022). Dengan jumlah pengguna yang terus meningkat, platform menghadapi tantangan dalam menganalisis ulasan pengguna secara manual karena volume data yang besar. Ulasan yang tidak terstruktur sulit dianalisis tanpa teknologi otomatis.



Analisis sentimen dan text summarization dapat mengelola volume data besar di platform seperti Coursera. Analisis sentimen mengkategorikan feedback pengguna menjadi positif, negatif, atau netral untuk meningkatkan kualitas. Teknik text summarization, seperti BERT, menyaring informasi kunci dan memberikan ringkasan relevan untuk pengambilan keputusan yang lebih baik.

METODOLOGI



Data Collection

- **Link Scraping**

Link scraping menggunakan **Selenium** memungkinkan pengambilan data seperti **title**, **creator**, dan **link** dari halaman hasil pencarian YouTube berdasarkan kata kunci tertentu, dalam hal ini "coursera review." Dengan **total link 26**

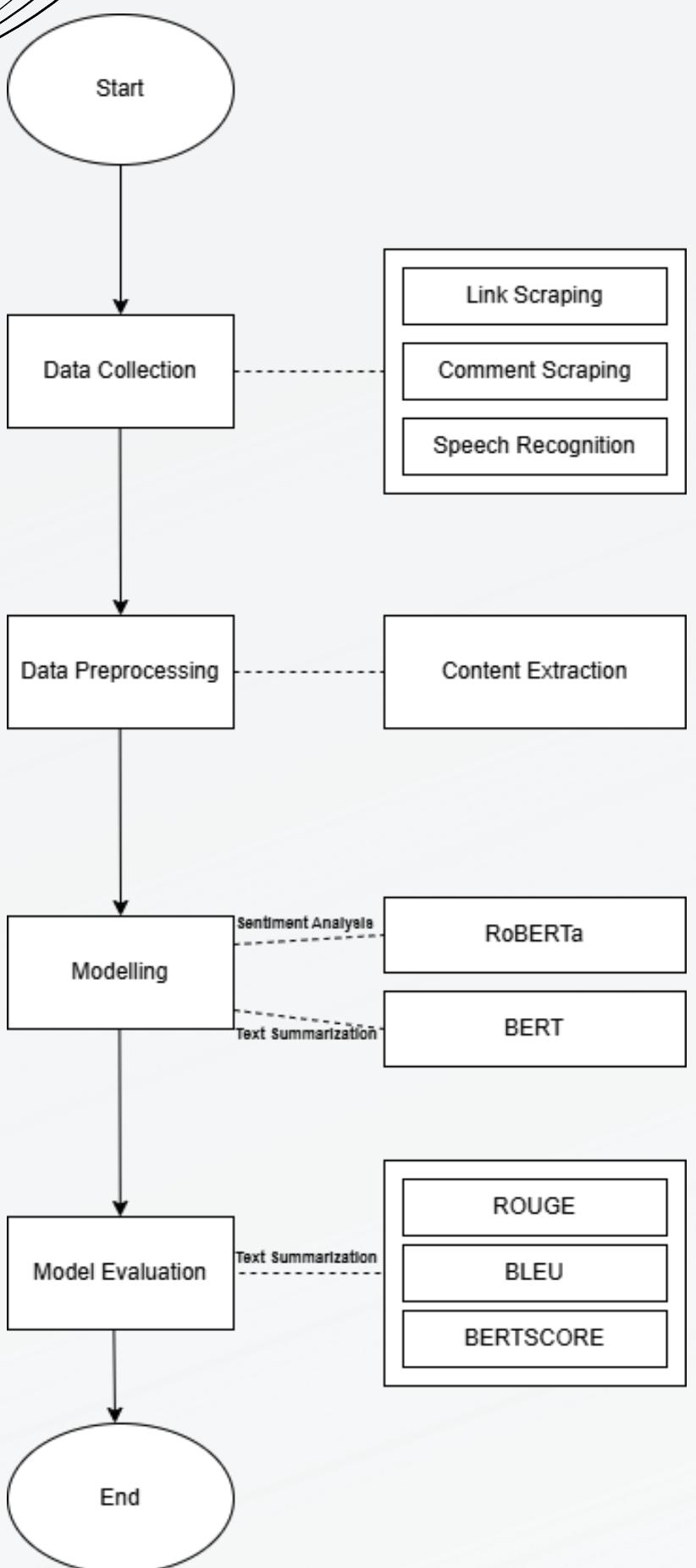
- **Comment Scraping**

Comment scraping pakai **Selenium** dipakai untuk mengumpulkan **comment** dan **creator** dari video YouTube. Dengan cara, daftar link video dibaca dari file, lalu browser otomatis buka satu per satu video. Dengan **total comment 1348**

- **Speech Recognition**

Speech recognition menggunakan PyTubeFix untuk mengunduh audio dari video YouTube berdasarkan daftar **link**, lalu mengonversi file MP3 ke WAV menggunakan FFmpeg. File WAV diproses menggunakan library SpeechRecognition untuk **mengubah audio menjadi teks** dengan **membagi audio menjadi potongan 30 detik**.

METODOLOGI

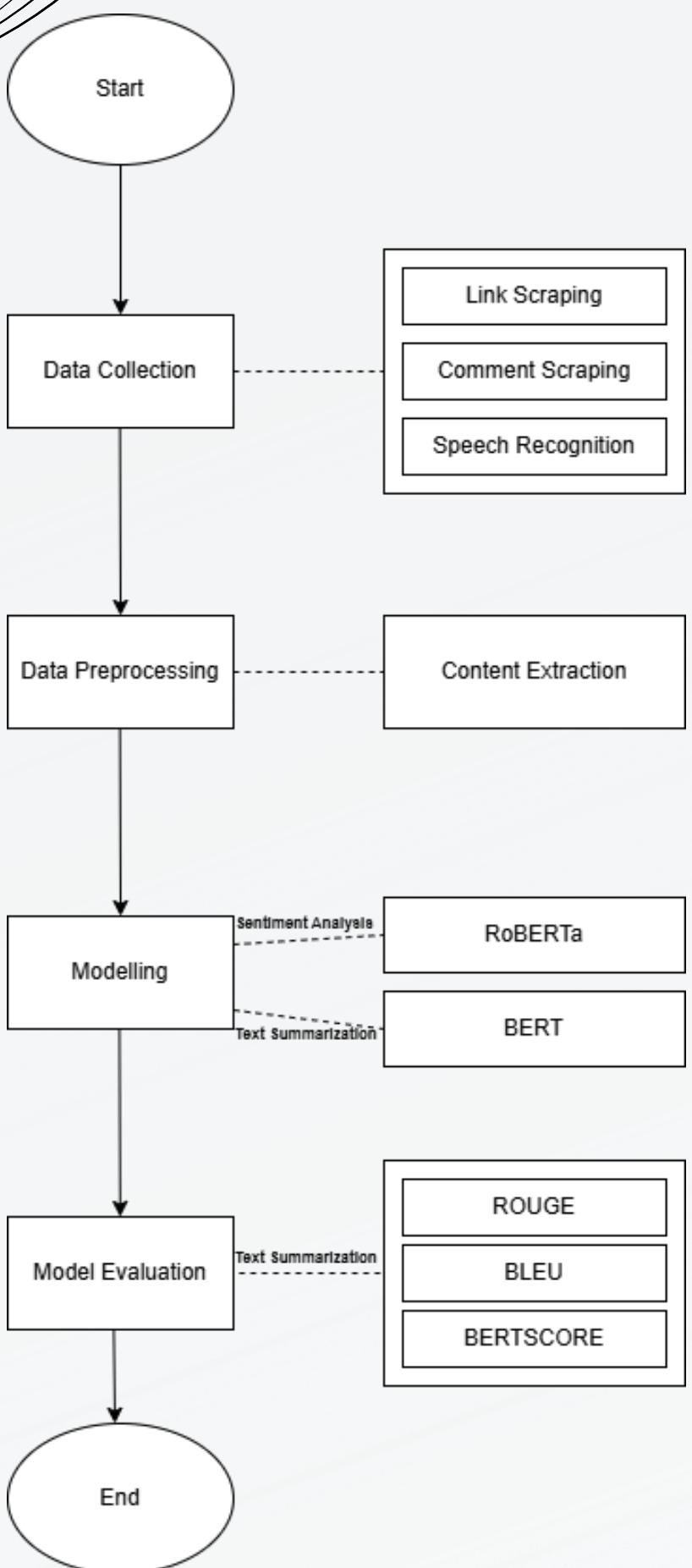


Data Preprocessing

- **Content extraction**

Content extraction dilakukan dengan tujuan mengambil informasi relevan dari data teks yang berhubungan dengan topik tertentu. Dalam hal ini, digunakan keyword seperti **['coursera', 'online course', 'platform', 'course', dan 'learning']** untuk mengekstraksi konten dari **data comment**. Keyword ini dirancang untuk mencakup berbagai aspek yang terkait dengan Coursera sebagai platform pembelajaran online.

METODOLOGI



Modelling

- **Analisis Sentimen**

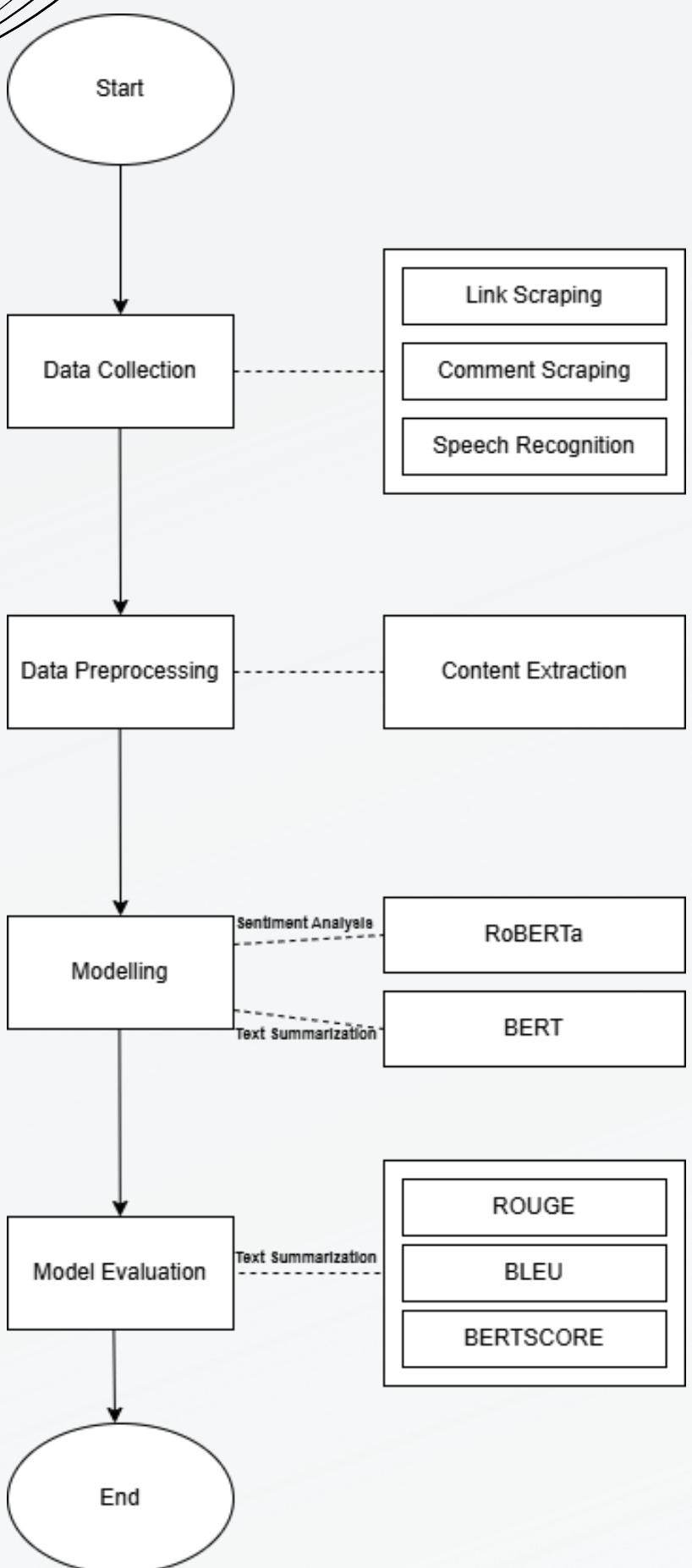
Analisis sentimen menggunakan **RoBERTa** melibatkan penerapan model transformer yang dilatih dengan optimisasi lebih lanjut untuk menangani teks dalam jumlah besar dan dapat memahami konteks secara lebih baik.

- **Text Summarization**

Text summarization menggunakan metode **BERT** telah dilakukan dalam penelitian ini untuk merangkum konten dari video-video di Youtube yang membahas mengenai *review* dari Coursera.

Dalam penelitian ini, *text summarization* menghasilkan ringkasan sebanyak **19 video Youtube** yang membahas mengenai *review* dari Coursera

METODOLOGI



Model Evaluation

- Mengevaluasi model hasil analisis sentimen menggunakan RoBERTa, kita bisa mengukur jumlah prediksi yang masuk ke dalam masing-masing kelas sentimen seperti 'Positive', 'Negative', dan 'Neutral'. Dengan menghitung distribusi hasil klasifikasi dari model, kita dapat mengetahui berapa banyak data yang diprediksi dengan masing-masing label ini, yang akan memberikan gambaran seberapa baik model dalam mengidentifikasi masing-masing kelas.
- Mengevaluasi model hasil text summarization menggunakan BERT, kita dapat menggunakan beberapa metrik evaluasi seperti BLEU , BERTScore, dan ROUGE. Ketiga metrik ini memberikan pandangan yang komprehensif mengenai kualitas ringkasan yang dihasilkan oleh model BERT.

HASIL DAN EVALUASI

SENTIMENT ANALYSIS

RoBERTa		
	labels	counts
0	positive	228
1	negative	180
2	neutral	179

- Dari hasil distribusi, metode RoBERTa menunjukkan distribusi yang seimbang. Dengan hasil yang cenderung seimbang antara label sentimen positif, negatif dan netral dapat mengindikasikan bahwa metode RoBERTa mampu memahami nuansa sentimen yang lebih kompleks. Model ini tidak hanya sensitif terhadap teks positif, tapi juga mampu mengidentifikasi teks dengan sentimen negatif dan netral.

HASIL DAN EVALUASI

TEXT SUMMARIZATION

Rouge

- Rouge 1 memiliki skor 22,6%
- Rouge 2 memiliki skor 13,3%
- Rouge L memiliki skor 22,5%

Blue

- Bleu memiliki skor 1,56% menunjukkan bahwa ada sedikit atau hampir tidak ada kesesuaian kata secara langsung antara ringkasan yang dihasilkan dengan referensi.

BERTscore

- Bertscore memiliki skor 86,5%, lebih tinggi dibandingkan ROUGE dan BLEU. Ini menunjukkan bahwa meskipun ada perbedaan kata atau struktur, makna dan inti dari ringkasan yang dihasilkan sangat relevan dengan referensi.

CONCLUSION

Dengan menghitung jumlah prediksi untuk kelas "Positif," "Negatif," dan "Netral", evaluasi model analisis sentimen RoBERTa memberikan wawasan tentang distribusi sentimen kumpulan data. Ini membantu pemahaman kita tentang akurasi model dalam klasifikasi sentimen.

Evaluasi BERT terhadap model ringkasan teks menggunakan metrik seperti ROUGE, BLEU, dan BERTScore memberikan gambaran menyeluruh tentang kualitas ringkasan yang dihasilkan. Bersama-sama, ketiga ukuran ini memberikan gambaran menyeluruh tentang seberapa efektif model tersebut menghasilkan ringkasan yang akurat dan relevan. BERTScore memberikan penilaian yang lebih menyeluruh berdasarkan kesamaan semantik, sementara metrik BLEU dan ROUGE berkonsentrasi pada kesamaan n-gram.

**THANKS FOR
YOUR ATTENTION!**

