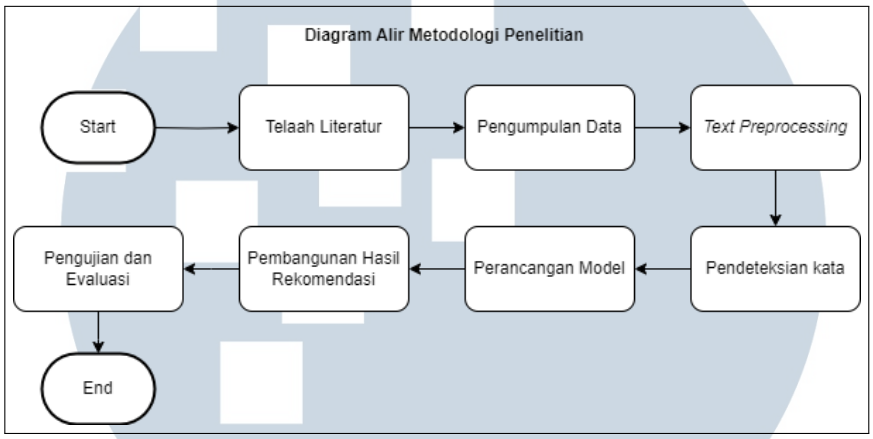
BAB 3
METODOLOGI PENELITIAN

Metodologi Penelitian ini dapat dilihat pada gambar 3.1.



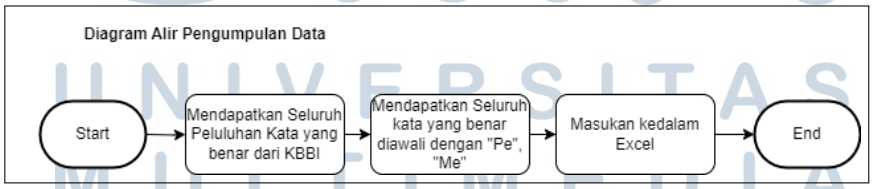
Gambar 3.1 Diagram Alir Metodologi Penelitian

3.1 Telaah Literatur

Pada tahap ini, berbagai penelitian teoritis dan praktis dilakukan untuk mengembangkan model NLP yang bertujuan untuk mendeteksi peluluhan kata beserta dengan rekomendasi untuk peluluhan kata yang salah menggunakan *Damerau-Levenshtein Distance* dan BERT

3.2 Pengumpulan Data

Pada tahap ini yang dapat dilihat pada Diagram Alir 3.2, dilakukan pengumpulan data untuk Peluluhan Kata yang benar sesuai dengan KBBI dan kata-kata yang diawali dengan ”Pe”, ”Me”. Kata-kata tersebut akan dimasukkan ke dalam excel untuk digunakan pada model.

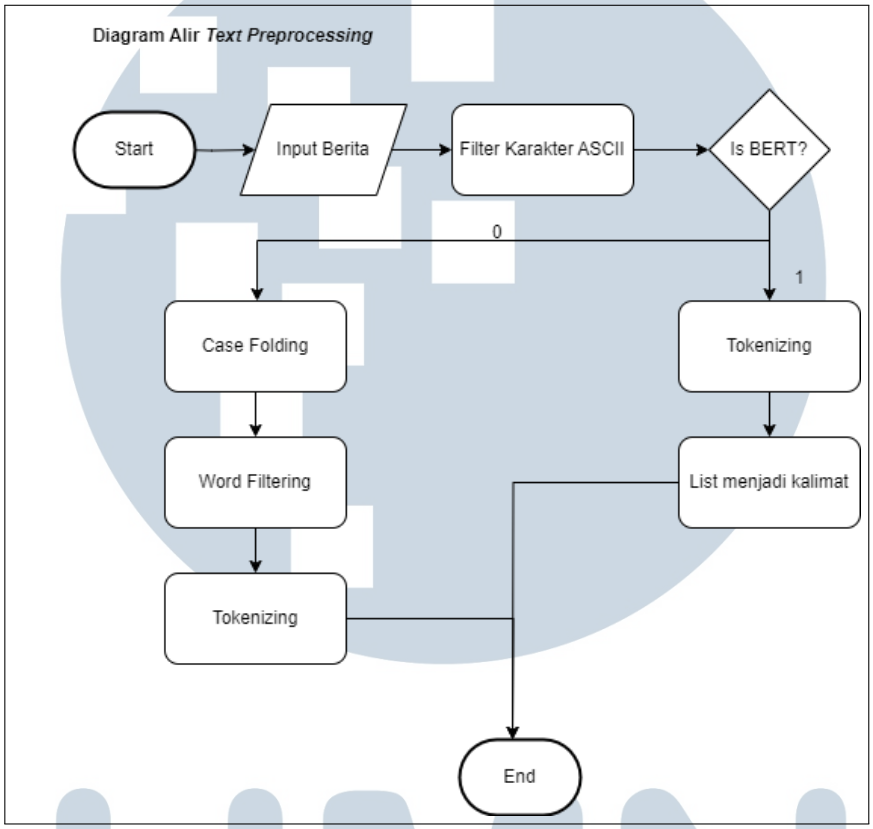


Gambar 3.2 Diagram Alir Pengumpulan Data

3.3 *Text Preprocessing*

Proses dimulai dengan mengumpulkan data-data yaitu mengumpulkan semua kata yang berawal dari input berita Tribunnews. Berita akan dilakukan penyesuaian dengan karakter ASCII

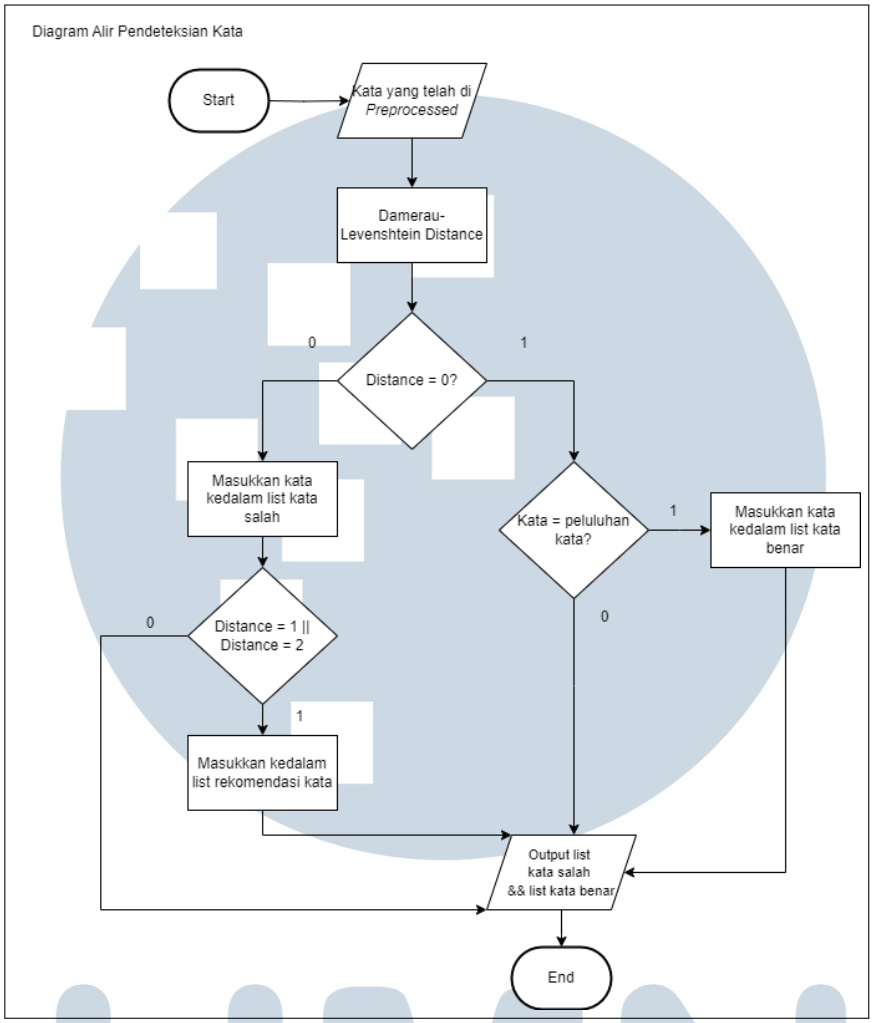
terdekat. *Preprocess* akan dibagi menjadi dua, yaitu *preprocess* untuk BERT, dan *preprocess* untuk *Damerau-Levenshtein Distance*. Untuk *Damerau-Levenshtein Distance*, perlu dilakukan *Case Folding*, *Word Filtering* dan terakhir di*tokenizing*. Sedangkan untuk BERT, hanya perlu dilakukan *tokenizing* dan *tokenized data* akan di*list* menjadi kalimat.



Gambar 3.3 Diagram Alir *Text Preprocessing*

3.4 Pendeteksian Kata

Proses dimulai dengan menggunakan data yang telah di*preprocess* untuk dilakukan pengecekan satu-persatu kata dengan dataset kata yang benar menggunakan algoritma Damerau-Levenshtein Distance. Jika kata memiliki *distance* sama dengan nol, akan dilakukan proses pengecekan kembali, apakah kata tersebut terdapat dalam dataset peluluhan kata benar. Jika ditemukan, akan dimasukkan ke dalam *list* kata yang benar. Jika tidak, akan dilanjutkan ke dalam proses berikutnya. Jika kata memiliki *distance* tidak sama dengan nol, kata tersebut akan dimasukkan ke dalam *list* kata yang salah. Selanjutnya, jika kata tersebut memiliki *distance* satu maupun dua, kata tersebut akan dimasukkan ke dalam *list* rekomendasi kata yang benar. Gambar 3.4 merupakan diagram alir dari pendeteksian kata.



Gambar 3.4 Diagram Alir Pendeteksian Kata

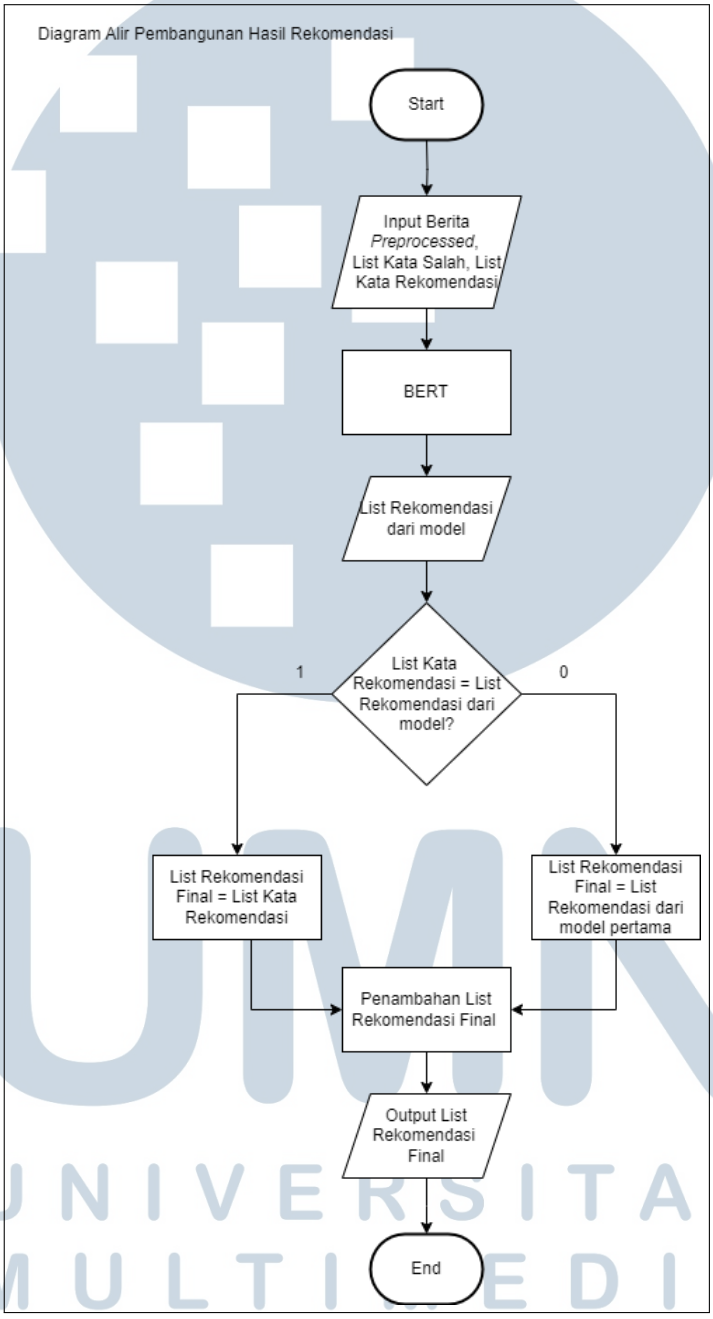
3.5 Perancangan Model

Tahap berikutnya adalah membuat model menggunakan Transformer BERT. model yang digunakan adalah model yang sudah dikembangkan dan diunggah pada Hugging Face [32]. Model sudah dilatih menggunakan dataset berita berbahasa Indonesia pada tahun 2018 yang memiliki ukuran file sebesar 448 mb.

3.6 Pembangunan Hasil Rekomendasi

Model akan diberi masukan data berita yang telah disesuaikan dengan penambahan ”[MASK]” pada setiap kalimat yang memiliki kata yang salah, sesuai dengan *list* kata salah pada tahap sebelumnya. Model akan menghasilkan rekomendasi kata-kata yang sesuai dengan konteks kalimat tersebut. Jika hasil rekomendasi model sama dengan *list* kata rekomendasi pada tahap sebelumnya, kata tersebut akan dimasukkan ke dalam list rekomendasi *final*. Jika tidak, kata rekomendasi pertama yang dihasilkan oleh model akan dimasukkan ke dalam *list* rekomendasi

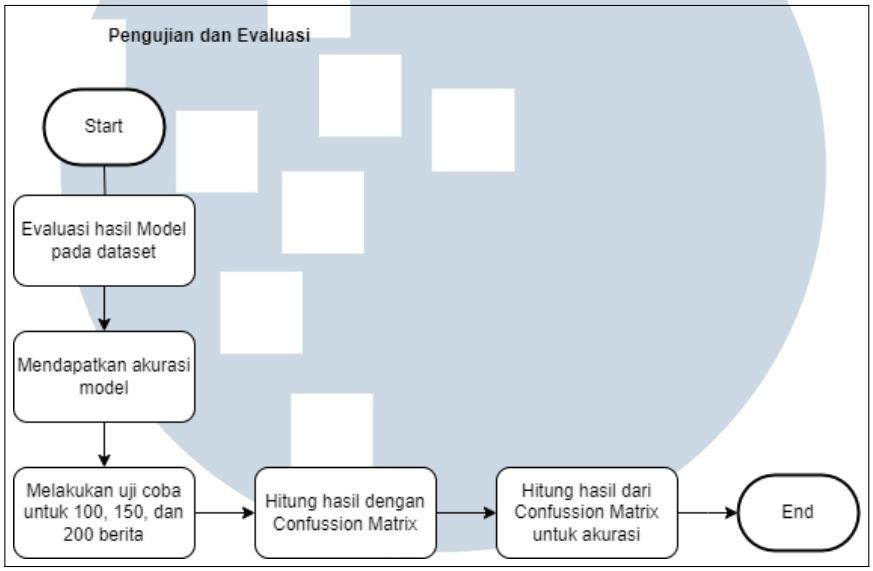
final. *List* rekomendasi *final* juga akan ditambahkan dengan dua kata pertama pada *list* kata rekomendasi yang telah didapatkan pada tahap sebelumnnya. Terakhir, tahap ini akan menghasilkan *list* rekomendasi *final* yang berisikan satu sampai dengan tiga kata.



Gambar 3.5 Diagram Alir Pembangunan Hasil Rekomendasi

3.7 Pengujian dan Evaluasi

Tahap pengujian dan evaluasi dilakukan dengan cara menghitung kata-kata yang berhasil dideteksi menggunakan sistem secara manual. Setelah data didapatkan, digunakan *Confussion Matrix* berdasarkan skenario diambil dari berita Tribunnews sebanyak 100, 150, dan 200 berita. Berdasarkan hasil dari Confusion Matrix akan dilakukan perhitungan akurasi dan F1 Score dari studi kasus artikel portal berita Tribunnews.



Gambar 3.6 Diagram Alir Pengujian dan Evaluasi