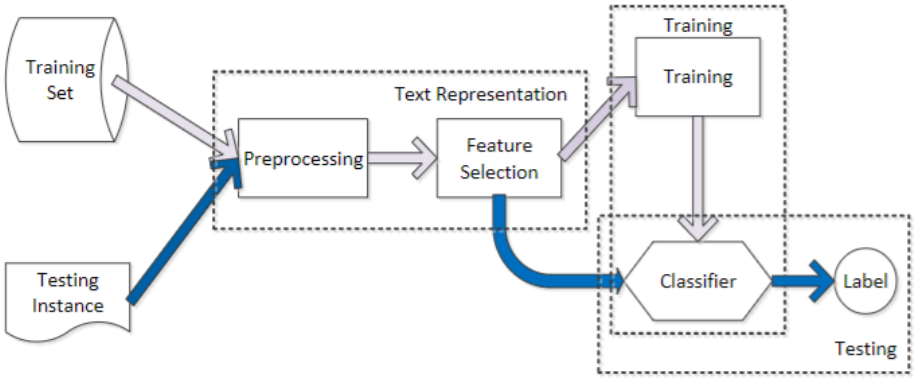
BAB II
TINJAUAN PUSTAKA

2.1. Text Classification

Text classification adalah proses memprediksi kategori dari data teks berdasarkan fitur yang telah diekstraksi dari data mentahnya [9]. Teks yang dikategorikan manusia dapat terpengaruh oleh pendapat pribadi manusia. Hal itu menyebabkan text classification sebaiknya dilakukan oleh mesin sehingga hasilnya lebih reliable. Metode-metode untuk melakukan text classification sangat banyak. Contohnya adalah k nearest neighbors (KNN), Naive Bayes, maximum entropy, and support vector machines (SVM).



Gambar 1. Proses Text Classification

Dalam melakukan text classification, terhadap beberapa langkah yaitu, pertama pada data mentah akan dilakukan preprocess. Preprocess dilakukan agar data bisa diolah oleh mesin. Setelah itu, data yang telah di preprocessing akan dilakukan features extraction. Feature Extraction adalah proses pengubahan format teks dari yang tidak terstruktur menjadi terstruktur sehingga algoritma machine learning akan dapat mengklasifikasi ke dalam kelas yang benar [10]. Selanjutnya, teks akan dilatih untuk menjadi model classifier. Setelah menjadi model classifier, akan dilanjutkan proses testing untuk mengetahui seberapa baik kinerja dari model.

2.2. Algoritma BERT dalam Text Classification

BERT (Bidirectional Encoder Representation from Transformer) adalah model NLP yang dikeluarkan oleh Google pada tahun 2018 yang dibuat dengan

beberapa lapis Transformer. Karena terdiri dari beberapa lapis transformer, BERT dapat mempelajari hubungan kontekstual antara kata-kata pada kalimat inputan. Transformer bertujuan untuk memeriksa kata-kata dalam query yang kompleks untuk mengaitkannya sehingga dapat memahami makna kalimat dan mengerti keseluruhan artinya [11].

Inti dari arsitektur BERT terdiri dari dua, yaitu encoder yang berfungsi untuk membaca teks inputan lalu menghasilkan representasi vektor dari kata-kata inputan dan decoder yang dapat melakukan task yang sudah diperintah [12].

Input dan output pada BERT akan berupa token yang mengandung satu kalimat atau dua kalimat. BERT menggunakan WordPiece Embeddings dengan 30.522 token kata. BERT akan membagi kalimat menjadi kata yang selanjutnya akan dilakukan tokenisasi berdasarkan dari kosa-kata BERT. Jika terdapat kata yang tidak terdapat pada kosakata BERT, maka kata tersebut akan dijadikan sub-bab [13].

2.3. Studi Literatur

Dalam beberapa tahun belakang, sudah dilakukan beberapa penelitian mengenai klasifikasi teks dengan metode BERT. Berikut merupakan penelitian yang menjadi acuan dasar dalam penelitian ini :

1. Penelitian yang dilakukan oleh Raden Mas Rizqi Wahyu Panca Kusuma Atmaja dan Wiyli Yustanti dengan judul "Analisis Sentimen Customer Review Aplikasi Ruang guru dengan metode BERT".

   Penelitian ini bertujuan untuk melakukan analisa sentimen terhadap aplikasi Ruang guru di Google Play sehingga dapat dimanfaatkan untuk memaksimalkan fitur yang dirasa kurang oleh pengguna. Data yang digunakan merupakan hasil dari scraping. Data berjumlah 5437 dengan kategori positif sebesar 5254, negatif sebesar 167, dan netral sebesar 16. Untuk proporsi data latih dan data uji sebesar 70:30. Dari hasil penelitian didapatkan nilai akurasi sebesar 99%, presisi sebesar 64.13% dan recall sebesar 99% dengan 10 kali epoch [14].
2. Penelitian yang dilakukan oleh Antonius Angga Kurniawan, Sarifuddin Madenda, Setia Wirawan, dan Ruddy J. Suhatril dengan judul Multidisciplinary classification for Indonesian scientific articles abstract using pre-trained BERT model.

Penelitian ini bertujuan untuk melakukan klasifikasi multidisiplin untuk abstrak artikel ilmiah bahasa Indonesia sehingga dapat ditunjukkan keterkaitan dari abstrak dan kategori. Data yang digunakan berjumlah 9000 dan kategorinya berjumlah 9. Penyetelan parameter yang dilakukan adalah ukuran batch sebesar 32, learning rate sebesar 1e-5, dan rasio data 9:1. Dari hasil penelitian didapatkan nilai akurasi sebesar 90.8% [15].