BAB 2
LANDASAN TEORI

2.1 Text Classification

Text Classification adalah salah satu bidang dalam NLP yang memiliki tujuan untuk menetapkan label pada sebuah unit tekstual seperti kalimat, kueri, paragraf, dan dokumen. *text classification* dapat diaplikasikan ke dalam beberapa hal diantaranya menjawab pertanyaan, deteksi spam, analisis sentimen, kategorisasi berita, klasifikasi maksud pengguna, dan sebagainya. Data teks yang akan diproses dalam *text classification* bisa berasal dari berbagai sumber, antara lain data web, email, media sosial, ulasan pengguna dan sebagainya [5].

2.2 Text Preprocessing

Dataset yang akan digunakan menggunakan komentar di Twitter, komentar-komentar tersebut masih mengandung beberapa karakter yang tidak dibutuhkan dalam proses *text classification* ataupun ada beberapa kata yang perlu dihilangkan dan diperbaiki agar mempermudah proses *text classification*. Oleh karena itu ada beberapa tahapan dalam melakukan *text preprocessing* diantaranya adalah *case folding*, *tokenizing*, *stop-word removal*, dan *stemming*.

2.2.1 BLOOM Tokenizer

Tokenizing adalah proses memecah sebuah kalimat menjadi kata, frasa, simbol atau elemen bermakna lainnya yang disebut token [14]. Tujuan *tokenizing* adalah eksplorasi kata-kata dalam sebuah kalimat. Contoh proses *tokenizing* dari kalimat ”hari ini aku masak nasi goreng” menjadi pecahan kata ”aku”, ”hari”, ”ini”, ”masak”, ”nasi”, ”goreng”.

2.3 BLOOM

BLOOM merupakan singkatan dari *BigScience Large Open-science Open-access Multilingual Language Model*, BLOOM memiliki 176 miliar model bahasa parameter yang dilatih menggunakan 46 bahasa natural dan 13 bahasa pemrograman yang dibuat oleh kolaborasi dari gabungan ratusan peneliti [6].

BLOOM dilatih dengan 498 *Hugging Face* dataset dengan kapasitas 1.61 terabyte dengan bahasa paling banyak di bahasa Inggris, untuk bahasa Indonesia sendiri memiliki 0,0199 terabyte.

2.4 Metrik Evaluasi

Metrik evaluasi digunakan untuk mengevaluasi hasil performa model yang telah dibuat, metrik evaluasi yang digunakan untuk mengevaluasi penelitian ini diantaranya *accuracy*, *precision*, *F1-Score*, dan *recall* [15].

A Accuracy

Akurasi adalah probabilitas bahwa prediksi model itu benar. Rumus akurasi didapatkan dari jumlah prediksi benar dibagi dengan seluruh prediksi yang dibuat, rumus dari akurasi dapat dilihat dari Persamaan 1

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

B Precision

Presisi adalah metrik evaluasi yang memberikan informasi seberapa besar model dapat dipercaya ketika melakukan prediksi sebagai positif. formula dari presisi dapat dilihat dari persamaan 2

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

C Recall

Recall adalah metrik evaluasi yang memberikan informasi seberapa besar model dapat dipercaya ketika melakukan prediksi individu sebagai positif. Formula dari *recall* dapat dilihat dari persamaan 3

$$Precision = \frac{TP}{TP + FN} \tag{2.3}$$

D F1-Score

F1-Score adalah metrik evaluasi yang memberikan informasi rata-rata dari *precision* dan *recall*, *f1-score* memiliki nilai terbaik 1 dan nilai terburuk 0. evaluasi ini berguna untuk menemukan *trade-off* terbaik antara kedua kuantitas. Formula dari *f1-score* dapat dilihat dari persamaan 4.

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{2.4}$$