1. Pendahuluan

Latar Belakang

Bahasa adalah alat komunikasi yang terorganisasi dalam bentuk satuan-satuan, seperti kata, kelompok kata, klausa, dan kalimat yang diungkapkan baik secara lisan maupun tulis [1]. Indonesia adalah negara kepulauan yang terdiri dari banyak suku, provinsi serta daerah sehingga Indonesia memiliki banyak bahasa daerah. Indonesia merupakan negara yang memiliki Bahasa daerah terbanyak kedua di dunia setelah Papua New Guinea [2]. Penggunaan bahasa daerah di Indonesia masih terbilang cukup banyak terlebih pada daerah-daerah terpencil. Tidak sedikit masyarakat Indonesia yang masih menggunakan bahasa daerah di kota-kota besar Indonesia. Namun pada akhirnya terdapat banyak penduduk Indonesia yang masih bingung untuk mengerti bahasa daerah apa yang digunakan saat mendengar pembicaraan seseorang ataupun saat berkomunikasi dengan orang. Walaupun penggunaan bahasa daerah sudah dilakukan dari dulu, bahasa daerah merupakan bagian dari sebuah kebudayaan masyarakat yang bersifat dinamis yaitu mengalami perubahan-perubahan yang tentunya juga bisa mengarah pada pergeseran bahasa jika tidak diperhatikan dengan seksama [3].

Identifikasi bahasa merupakan langkah pra-pemrosesan penting dalam banyak sistem otomatis yang beroperasi menggunakan teks tertulis seperti *Text Classification* (TC). Sebagai contoh, pada task TC, penerapan identifikasi bahasa merupakan salah satu *preprocessing* yang menunjukkan kinerja baik, dilihat dari nilai *precision* dan *recall* [4]. Identifikasi bahasa dimulai pada tahun 1965 oleh ahli statistik Mustone, yang mengajarkan komputer untuk membedakan pada tingkat kata yang membedakan bahasa Inggris, Swedia, dan Finlandia [5]. Untuk menghasilkan identifikasi Bahasa dibutuhkan metode TC yaitu salah satu cabang dalam *Natural Language Processing* (NLP). Seperti namanya, TC dapat digunakan untuk mengklasifikasikan teks, biasanya dengan indikator atau aturan khusus untuk setiap kelas [6].

Pada penelitian ini penulis akan melakukan identifikasi pada 10 bahasa daerah Indonesia yaitu Bahasa daerah Aceh, Bali, Banjar, Bugis, Madura, Minangkabau, Jawa, Ngaju, Sunda, dan Toba Batak. Dataset tersebut diambil dari penelitian sebelumnya yaitu *NusaX* [7]. Untuk mengimplementasikan identifikasi Bahasa terdapat beberapa metode yang dapat digunakan dalam melakukan identifikasi Bahasa daerah ini. Mengacu kepada penelitian sebelumnya tentang TC yang digunakan dalam identifikasi bahasa [5], penelitian tersebut menggunakan beberapa metode yang juga akan diimplementasikan pada penelitian ini yaitu model *Support vector machine* (SVM), *Naïve Bayes Classifier* (NBC), *Decision Tree* (DT), *Rocchio Classification* (RC), *Logistic Regression* (LR), *Random Forest* (RF). Pengembangan model juga menggunakan dua fitur yang berbeda yaitu n-gram dan TF-IDF.

Topik dan Batasan

Penelitian ini mengangkat topik pengidetifikasian bahasa dan pengukuran kinerja model pada 10 bahasa daerah Indonesia, dengan menggunakan dataset dari penelitian NusaX dengan menggunakan enam model yang akan menghasilkan kinerja yang berbeda-beda.

Tujuan

Tujuan Tugas Akhir ini adalah untuk melakukan identifikasi bahasa menggunakan pendekatan *Text Classification* dengan pengaplikasian 10 bahasa daerah Indonesia. Pendekatan *Text Classification* yang digunakan adalah penggabungan enam model dan dua ekstraksi fitur.x

Organisasi Tulisan

Pada jurnal ini berisi bagian abstrak, pendahuluan, studi terkait, sistem yang dibangun, evaluasi, kesimpulan.