BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Aksara, sastra, dan Bahasa Bali menjadi sumber imajinasi, kreativitas, dan daya cipta serta merupakan tenaga dalam kebudayaan Bali. Hal tersebut mulai mengalami penurunan khususnya dari segi penggunaan Aksara Bali yang semakin berkurang digunakan pada kehidupan sehari-hari masyarakat Bali (Mudiarta et al., 2020a). Salah satu faktor utama yang menyebabkan penurunan tersebut adalah masyarakat Bali kurang tertarik untuk membaca Aksara Bali karena keengganan dalam mempelajari Aksara Bali yang relatif rumit dalam proses pengenalannya. Dalam Peraturan Gubernur Bali Nomor 80 Tahun 2018 tentang Pelindungan dan Penggunaan Bahasa, Aksara dan Sastra Bali Serta Penyelenggaraan Bulan Bahasa Bali mengatur penggunaan bahasa Bali, yaitu sebagai sarana komunikasi dalam kehidupan keluarga Bali, komunikasi dalam segala kegiatan agama Hindu, adat dan budaya Bali, dan pemberian informasi pada layanan masyarakat baik pada lembaga pemerintahan maupun lembaga swasta sebagai pendamping Bahasa Indonesia (Peraturan Gubernur Bali Nomor 80, 2018).

Aksara Bali berdasarkan fungsinya terbagi menjadi empat yaitu Aksara Wreastra, Aksara Swalelita, Aksara Wijaksara, dan Aksara Modre. Berdasarkan penggunaannya Aksara Bali terbagi menjadi dua jenis Aksara yaitu Aksara Bali Biasa dan Aksara Bali Suci. Aksara Bali Biasa adalah Aksara Bali yang digunakan untuk menulis bahasa sehari-hari yang umum digunakan oleh masyarakat. Aksara Bali Suci adalah Aksara Bali Aksara yang digunakan untuk menulis hal-hal yang berhubungan dengan agama seperti japa mantra, weda, dan rerajahan (Suasta &

Mayun, 1996). Aksara Bali dapat ditemukan dengan mudah di seluruh wilayah pulau Bali, penggunaan Aksara Bali sering digunakan untuk nama jalan, plang kantor, dan naskah lontar (Sutramiani et al., 2021).

Perkembangan teknologi komputer saat ini telah banyak dimanfaatkan untuk melakukan pengenalan karakter optik atau diistilahkan dengan *Optical Character Recognition* (OCR). OCR adalah teknik mengubah teks dan gambar yang dicetak menjadi bentuk karakter digital, yang dapat dimanipulasi oleh mesin. Implementasi OCR telah digunakan pada banyak sektor aplikasi seperti pendidikan, perbankan, keuangan, hukum, dan sebagainya. Sebagian besar pengembangan OCR masih berfokus pada skrip latin bahasa inggris karena didukung oleh standar *Encoding American Standard Code for Information Interchange* atau disingkat ASCII. Keterbatasan kemampuan OCR dalam mengenali skrip non latin menjadi tantangan tersendiri bagi peneliti untuk bisa melakukan improvisasi. Seiring perkembangan teknologi OCR banyak penelitian yang memanfaatkan OCR untuk melakukan pengenalan karakter untuk skrip non latin (Qaroush et al., 2020).

OCR sebagai sebuah teknologi *converter text images* umumnya terdiri dari beberapa sub-proses yaitu pra-pemrosesan, pengenalan karakter, dan pasca-pemrosesan (Smith, 2007). Tahapan pertama adalah pra-pemrosesan gambar dengan meminimalkan efek kendala data umum seperti buram, miring, bintik, dan warna pada gambar karakter bertujuan untuk meningkatkan kemungkinan mengenali data secara akurat. Selanjutnya proses pengenalan karakter melibatkan berbagai pendekatan (pencocokan matriks & ekstraksi fitur) untuk memecah gambar menjadi bagian atau zona yang dapat dikelola dan mengenali karakter yang terkandung di dalam gambar. Terakhir pasca-pemrosesan melibatkan teknik &

algoritma untuk meningkatkan akurasi data yang diekstraksi dengan terlebih dahulu mendeteksi dan kemudian memperbaiki kesalahan dengan membandingkan teks/data yang diekstraksi dengan model *trained* data yang dimasukkan kedalam *engine* OCR (Smith et al., 2009).

Teknologi OCR berkembang pesat dengan terciptanya beberapa *engine* OCR yang bersifat *open source* dan berbayar. Berdasarkan penelitian yang dilakukan oleh (Ramdhani et al., 2021) melakukan komparasi tingkat kinerja dari tiga *engine* OCR yang memiliki tingkat kinerja yang tinggi. Penelitian ini mencoba menguji *engine* OCR mana yang memiliki kinerja tertinggi untuk *Information Extraction* menggunakan *Named Entity Recognition* dengan membandingkan tiga *engine* OCR yaitu Foxit, PDF2GO dan Tesseract. Pengujian tersebut dilakukan dengan 8.562 dokumen sumber daya manusia pemerintah dalam enam kategori dokumen, dua struktur dokumen, dan empat pengukuran. Hasil pengujian dari penelitian tersebut mendapatkan bahwa Tesseract adalah solusi yang paling cocok dan mendapatkan nilai tertinggi dari sisi kinerja dalam melakukan *Information Extraction*. Rincian hasil pengujian tersebut rata-rata PDF2GO mendapatkan kinerja sebesar 86,27% selanjutnya Foxit mendapatkan nilai kinerja sebesar 84,01% dan terakhir Tesseract mendapatkan nilai kinerja 92,46%.

Pada penelitian yang dilakukan oleh (Abdul Robby et al., 2019) memanfaatkan *engine* Tesseract OCR untuk diimplementasikan sebagai mesin pengenalan karakter Aksara Jawa. Penelitian ini bertujuan untuk mempermudah proses pengenalan karakter Aksara Jawa secara otomatis menggunakan aplikasi *mobile*. *Dataset* yang digunakan sebagai sumber data untuk membangun data *training* (*trained data*) *engine* Tesseract OCR berjumlah 5.880 karakter Aksara

Jawa. Untuk membangun *dataset* Aksara Jawa tersebut dikumpulkan dari karakter digital dengan spesifikasi (3 set x 120 karakter) dan tulisan tangan (46 set x 120 karakter). *Tools* training dataset yang digunakan pada penelitian ini adalah *Neural-Network* API dari *engine* Tesseract OCR. Sebelum dilakukan proses pelatihan *dataset* Aksara Jawa diseleksi dengan cara melakukan segmentasi untuk masing-masing karakter dan mengatur variabel untuk *cluster* dari karakter menggunakan JTessBoxEditor. Akurasi tertinggi yang dicapai oleh model yang dihasilkan dari *traineddata* tersebut adalah sebesar 97,50%.

Penelitian berikutnya yang sejenis dengan kasus pengenalan karakter optik non latin adalah penelitian yang dilakukan oleh (Mudiarta et al., 2020). Penelitian ini berfokus pada pelestarian pengetahuan membaca Aksara Bali dalam gambar dengan menggabungkan teknologi informasi dengan disiplin Aksara Bali. Dalam penelitian ini aplikasi OCR dikembangkan pada perangkat berbasis *mobile* dengan fasilitas kamera. Masukan pada aplikasi ini berupa gambar dan diproses dengan teknologi *engine* Tesseract OCR. Untuk melakukan proses *training dataset* Aksara Bali dibuat berdasarkan delapan belas suku kata dasar Aksara Bali dan hanya angka-angka. Alat yang digunakan untuk melakukan proses training adalah jTessBoxEditor, alat ini memiliki sepenuhnya fasilitas otomatis untuk *training dataset*. Hasil pengujian untuk 50 kata, pengenalan 62% diperoleh dengan baik *font* Bali-Simbar berbasis gambar berkualitas.

Dari pemaparan dua penelitian diatas terdapat kemiripan dari sisi *engine Optical Character Recognition* yang digunakan dan proses training data yang dilakukan. Proses *training* data yang dilakukan untuk membuat model *traineddata* memanfaatkan *tools* jTessBoxEditor dengan cara melakukan segmentasi karakter

dari gambar karakter non latin. Proses segmentasi tersebut dilakukan secara bergantian untuk masing-masing *dataset* yang dimiliki. Ada beberapa kelemahan yang terjadi pada dua penelitian tersebut khususnya pada proses *training* data yang dilakukan. Penggunaan *tools* jTessBoxEditor harus dilakukan dengan manual dengan melakukan segmentasi untuk masing-masing *dataset* membuat proses *training* relatif menjadi lebih membutuhkan waktu yang lama. Pada bagian bab saran dari dua penelitian tersebut berfokus kepada peningkatan jumlah dataset yang digunakan.

Berdasarkan kelemahan dan saran dari dua penelitian tersebut dapat disolusikan dengan menggunakan metode *training* data yang berbeda selain menggunakan *tools* jTessBoxEditor ada metode *training* terbaru untuk membuat *traineddata* yaitu dengan menggunakan metode *training* Tesseract OCR terbaru. Metode *training* Tesseract OCR terbaru ini dapat melakukan training dataset secara simultan untuk seluruh *dataset*. Menurut (Idrees & Hassani, 2021) sejak versi 4.0, Tesseract OCR menghadirkan mesin baru berbasis *Long Short-Term Memory* (LSTM). LSTM, sebagai bentuk khusus dari Jaringan Syaraf Tiruan (RNN), memberikan akurasi yang jauh lebih tinggi pada pengenalan gambar dari pada versi Tesseract OCR sebelumnya. Tesseract bisa dilatih dari awal atau disempurnakan berdasarkan bahasa yang sudah terlatih.

Berdasarkan pemaparan latar belakang tersebut maka dilakukan penelitian dengan judul "*Optical Recognition Character* Aksara Bali Menggunakan *Mobile Framework Tesseract OCR*". Tujuan umum penelitian ini adalah untuk membantu masyarakat dalam membaca dan mempelajari Aksara Bali dengan memanfaatkan teknologi khususnya teknologi *smartphone*, memicu tumbuhnya penelitian untuk

implementasi teknologi dalam pelestarian budaya adiluhung bangsa serta mendukung visi pemerintah provinsi Bali dalam melindungi dan melestarikan penggunaan Aksara Bali dalam kehidupan sehari-hari masyarakat Bali. Tujuan khusus penelitian ini adalah untuk mengembangkan sebuah aplikasi *Optical Character Recognition* (OCR) yang mampu mengenali dan malakukan transformasi dari gambar yang memuat tulisan Aksara Bali menjadi *plain text* atau teks digital sehingga dapat memberikan kemudahan dalam mengenali, membaca dan mempelajari Aksara Bali.

1.2 Identifikasi Masalah

Berdasarkan uraian latar belakang yang dipaparkan diatas, maka dapat diidentifikasi beberapa masalah yaitu sebagai berikut:

1. Dalam membantu masyarakat dalam mengenali dan mempelajari Aksara Bali diperlukan pemanfaatan teknologi untuk meningkatkan tingkat pengenalan Aksara Bali.
2. Gawai *smartphone* adalah menjadi kebutuhan untuk generasi *digital native* sehingga teknologi *Optical Character Recognition* dapat menjadi pilihan dalam melakukan pengenalan karakter khususnya untuk karakter Aksara Bali.
3. Teknologi *Optical Character Recognition* dapat menjadi pilihan utama untuk dimanfaatkan dalam proses pengenalan karakter Aksara Bali bagi masyarakat yang awam dengan Aksara Bali.
4. Belum tersedia dataset *image* dan *ground truth* terkait Aksara Bali sebagai data training untuk *Tesseract OCR*.
5. Penelitian tentang penerapan model *training* Tesseract OCR terbaru untuk karakter digital non latin khususnya bahasa yang belum disupport Tesseract OCR belum ditemukan penelitian terkait hal tesebut.

1.3 Batasan Masalah

Berdasarkan ruang lingkup masalah penelitian, maka dapat dirumuskan beberapa Batasan yaitu sebagai berikut:

1. Dataset yang digunakan pada penelitian ini terdiri dari dua jenis dataset yaitu image Aksara Bali dan ground truth berupa *file text*.
2. Dataset Aksara Bali menggunakan *type script* dengan *font style* Noto Sans Balinese.
3. Teknologi *Optical Character Recognition* yang digunakan adalah Tesseract OCR.
4. Input berupa image karakter digital Aksara Bali yang berasal dari gambar yang sudah tersedia atau didapat dari hasil akuisisi gambar dari device aplikasi OCR terpasang.

1.4 Rumusan Masalah

Adapun tujuan yang diharapkan dapat dicapai dari penelitian ini adalah sebagai berikut:

1. Bagaimana implementasi hierarki dataset dalam membangun model language Aksara Bali menggunakan framework Tesseract OCR?
2. Bagaimana kinerja teknologi generate dataset web scrapping dalam mengumpulkan dataset aksara bali dan ground truth text?
3. Bagaimana perbandingan kinerja antara setiap hierarki dataset yang digunakan dalam membangun model language?

1.5 Tujuan Penelitian

Adapun pembatasan masalah terhadap penelitian yang dilakukan adalah sebagai berikut:

1. Untuk mengimplementasi hierarki dataset dalam membangun model language Aksara Bali menggunakan framework Tesseract OCR.

2. Untuk mengetahui kinerja teknologi generate dataset web scrapping dalam mengumpulkan dataset aksara bali dan ground truth text.
3. Untuk mengetahui perbandingan kinerja antara setiap hierarki dataset yang digunakan dalam membangun model language.

1.6 Manfaat Penelitian

Adapun beberapa manfaat penelitian terhadap penelitian yang dilakukan adalah sebagai berikut:

1. Manfaat Teoritis

   Hasil penelitian ini secara teoritis diharapkan mampu menjadi bahan acuan dalam memilih model training data Tesseract OCR yang tepat untuk pengenalan karakter non latin khususnya karakter Aksara Bali.
2. Manfaat Praktis

   Manfaat umum penelitian ini adalah untuk membantu masyarakat dalam membaca dan mempelajari Aksara Bali dengan memanfaatkan teknologi khususnya teknologi *smartphone*, memicu tumbuhnya penelitian untuk implementasi teknologi dalam pelestarian budaya adiluhung bangsa serta mendukung visi pemerintah provinsi Bali dalam melindungi dan melestarikan penggunaan Aksara Bali dalam kehidupan sehari-hari masyarakat Bali. Manfaat khusus penelitian ini adalah untuk mengembangkan sebuah sistem *Optical Character Recognition* (OCR) yang mampu mengenali dan malakukan transformasi dari gambar yang memuat tulisan Aksara Bali menjadi *plain text* dan sehingga dapat memberikan kemudahan dalam membaca dan mempelajari Aksara Bali.

1.7 Rencana Publikasi

Luaran jurnal akan dipublikasikan di jurnal RESISTOR (Rekayasa Sistem Komputer). Jurnal RESISTOR adalah jurnal yang dikelola dan diterbitkan oleh Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM) STMIK STIKOM Indonesia, dengan e-ISSN 2598-9650 dan p-ISSN: 2598-7542. Jurnal RESISTOR

terbit pertama kali pada bulan April 2018 dan memiliki masa terbit dua kali dalam setahun yaitu pada bulan April dan Oktober. Fokus dan ruang lingkup Jurnal RESISTOR meliputi *Biomedical Engineering*, *Cloud Infrastructure*, *Computer Network and Architecture*, *Computer Security*, *Computer Vision*, *Cultural Tourism Application*, *Digital Forensics*, *Embedded System*, *Internet of Things*, *Machine Learning*, *Power*, *Energy*, *and Industry Applications*, *Remote Sensing*, *Robotics and Automation*, *Signal Processing and Analysis*, *Soft Computing*, *Wireless Sensor Network*.