BAB 1

PENDAHULUAN

1.1. Latar Belakang Masalah

Kebutuhan akan internet bagi masyarakat pada masa kini sungguh penting. Dengan internet semua orang dapat mengakses berbagai informasi dan juga hiburan secara daring yang memiliki konten yang beragam. Menurut data "Digital 2020" oleh *we are social,* pengguna internet Indonesia mencapai 175,4 juta dari 272,1 juta rakyat Indonesia. Dari jumlah pengguna internet yang masif tersebut terdapat 35,8% yang mengakses berita secara daring menurut data dari badan pusat statistik yang diolah oleh Beritagar.id pada tahun 2017 dan terus meningkat hingga tahun 2020. Berita daring makin diminati oleh masyarakat karena informasi yang diberikan secara *real-time.* Selain itu, dalam mengakses berita daring juga terbilang mudah karena hanya membutuhkan gawai yang terkoneksi internet. Jumlah pembaca berita daring yang meningkat juga diikuti dengan jumlah berita daring yang makin banyak di internet. Berita-berita tersebut memiliki berbagai kategori yang sudah disusun oleh setiap *platform* penyedia berita daring. Kategori-kategori tersebut seperti bisnis, teknologi, olahraga dan masih banyak lainnya. Dengan jumlah berita daring yang makin banyak, proses pengkategorian berita diharapkan dapat mengimbangi agar tidak terjadi berita tanpa memiliki kategori ataupun diberikan kategori yang tidak tepat. *Editor* yang bertugas untuk mengkategori berita-berita daring tersebut, bisa melakukan kesalahan ataupun memiliki

keterbatasan kecepatan dalam mengkategori berita-berita tersebut. Apalagi berita tersebut harus secepatnya diunggah agar dapat dinikmati oleh pembaca secara tepat waktu.

Dengan permasalahan tersebut, maka terbentuklah penelitian ini untuk mengatasi masalah dalam pengkategorian berita yang banyak dan membutuhkan metode yang cepat dalam mengelompokkan berita-berita dengan kesesuaian konten isi berita. Maka dari itu dibuat sistem pengkategorian berita daring secara otomatis menggunakan pembelajaran mesin. Algoritma yang diimplementasikan adalah *Recurrent Neural Network* (RNN) dengan arsitektur LSTM (*Long Short Term Memory*). Menurut Serban dkk(2019) pada *paper* berjudul "Real-time processing of social media with Sentinel : A syndromic surveillance system incorporating deep learning for health classification" mengatakan bahwa LSTM-RNN adalah algoritma *machine learning* yang memiliki tingkat akurasi yang tinggi untuk melakukan klasifikasi berita menggunakan data teks. Worsham dkk(2018) sebuah model *machine learning* tidak dapat menerima data berupa teks ataupun karakter. Agar dapat diproses sebuah model dalam melakukan klasifikasi maka ditransformasi menjadi barisan angka. Dalam mempersiapkan data masukan tersebut, maka pada penelitian ini menerapkan cabang ilmu pada *artificial intelligence* yaitu *Natural Language Processing.* Lindén dkk(2018) LSTM-RNN memiliki kombinasi terbaik dalam memprediksi suatu kumpulan kata masuk dalam kategori tertentu. Fuks(2018) proses klasifikasi menggunakan beberapa algoritma Jaringan Saraf Tiruan (JST) hasil LSTM-RNN memiliki tingkat akurasi paling tinggi dibanding algoritma JST lainnya. *Natural Language Processing* (NLP)

memberikan kemampuan sebuah mesin untuk membaca, mengerti dan menginterpretasi makna dari bahasa manusia agar dapat melakukan proses klasifikasi dengan kategori sesuai makna sesungguhnya dari kumpulan teks tersebut. Proses-proses tersebut meliputi tokenisasi, *stemming* dan *pad sequencing* .

Berdasarkan penelitian sebelumnya oleh Sari dkk(2020) menunjukkan akurasi lebih dari 90%. Penelitian ini menggunakan metode yang sama yaitu LSTM-RNN dengan menggunakan *word embedding* dari Word2Vec. Dataset yang digunakan merupakan kumpulan artikel berbahasa inggris. Pada penelitian ini, dataset yang digunakan merupakan artikel-artikel bahasa Indonesia yang diambil dari Jakartaresearch dan *web scraping* pada *platform* Kompas.com . Selain itu proses *stemming* menggunakan *library* Sastrawi yang memiliki kumpulan kata bahasa Indonesia yang lengkap. *Pre trained model Word embedding* pada penelitian ini menggunakan FastText. Young dkk(2019) FastText merupakan implementasi pada Word2Vec yang sudah dikembangkan untuk memproduksi vektor walaupun kata tersebut memiliki salah pengetikan. Penelitian ini diharapkan dapat membuat sistem yang secara otomatis dapat mengklasifikasi berita dengan menerapakan algoritma LSTM-RNN.

1.2. Rumusan Masalah

Masalah yang dapat dirumuskan dalam penelitian ini adalah sebagai berikut

1. Bagaimana cara mengklasifikasi kategori berita secara otomatis menggunakan metode LSTM-RNN?

2. Bagaimana tingkat akurasi metode LSTM-RNN dalam mengklasifikasi kategori berita?

1.3. Batasan Masalah

Penelitian ini memiliki batasan-batasan masalah, yaitu :

1. Aplikasi untuk *Natural Processing Language* (NLP) dan implementasi algoritma menggunakan *jupytr notebook* dengan bahasa pemrograman python
2. Pembuatan model dan NLP menggunakan *library* Keras Tensorflow
3. Proses *stemming* menggunakan *library* Sastrawi
4. Model *machine learning* diakses melalui *localhost* menggunakan *framework* Flask
5. Model *machine learning* hanya mengklasifikasi berita dalam 5 kategori yaitu bola, *news,* bisnis, teknologi dan otomotif
6. Dataset berita Indonesia berasal dari Jakartresearch dan *web scraping* dari tahun 2019-2020 yang bersumber dari Liputan6 dan Kompas

1.4. Tujuan Penelitian

Berdasarkan rumusan masalah tersebut, tujuan dari penelitian ini adalah,

1. Mengimplementasi metode LSTM-RNN untuk melakukan klasifikasi kategori berita Indonesia kedalam beberapa kategori yaitu bola, *news*, bisnis, teknologi dan otomotif
2. Mengukur performa dari metode LSTM-RNN dalam mengklasifikasi kategori dengan menggunakan metrik *f1-score, accuracy, recall* dan *precision*

1.5. Manfaat Penelitian

Penelitian ini memiliki berbagai manfaat yaitu sistem yang akan dibuat dapat membuat pengkategoriaan berita-berita Indonesia secara otomatis dan mempermudah para editor dalam mengklasifikasi kategori berita Indonesia dengan menerapkan LSTM-RNN.

1.6. Sistematika Penulisan

Agar penelitian ini dapat mencapai tujuan dan dapat dimengerti oleh para pembaca dan peneliti lainnya, maka materi-materi yang ada pada laporan skripsi ini dikelompokkan menjadi beberapa bab dengan sistematika sebagai berikut.

BAB 1 PENDAHULUAN

Bab ini memiliki substansi berupa latar belakang, rumusan masalah, batasan masalah, tujuan penelitan, manfaat penelitian, dan sistematika penulisan

BAB II LANDASAN TEORI

Bab ini memiliki isi berupa teori-teori yang berkaitan dengan penelitian yang diambil dari buku, jurnal, internet. Teori-teori pada penelitian ini meliputi penjelasan mengenai *NLP*, RNN, LSTM, *softmax activation* dan fasttext

BAB III METODOLOGI DAN PERANCANGAN SISTEM

Bab ini membahasa mengenai tahapan penelitian yang dilakukan berupa *pre processing* hingga pembuatan sistem. Proses *pre processing* dan pembuatan

model *machine learning* dibuat dalam bentuk *flowchart* dan sistem web akan menampilkan antarmuka program.

BAB IV IMPLEMENTASI DAN UJI COBA

Bab ini berisi implementasi dari beberapa komponen NLP untuk memberikan kemampuan pada mesin agar mengerti masukan teks dan alogritma LSTM RNN untuk klasifikasi berita serta hasil dari uji akurasi dengan menggunakan *confusion matrix.*

BAB V SIMPULAN DAN SARAN

Bab ini terdiri dari simpulan pada hasil pengujian sistem dan saran agar penelitian ini dapat dikembangkan lebih baik.