

BAB I

PENDAHULUAN

1.1 Latar Belakang

Bahasa merupakan sarana yang digunakan untuk melakukan komunikasi dan interaksi antara individu satu dengan lainnya. Indonesia yang merupakan negara kepulauan dengan letak geografis, kehidupan sosial, suku, dan budaya yang beragam, hal ini menyebabkan Indonesia juga memiliki bahasa yang beragam di tiap daerahnya [1]. Salah satu contoh dari bahasa daerah yang terdapat di Indonesia adalah seperti yang ada pada Jawa Timur, yaitu Bahasa Madura. Mayoritas masyarakat Madura menggunakan Bahasa Madura sebagai bahasa ibu atau bahasa utamanya. Sementara itu untuk kemampuan menggunakan bahasa nasionalnya sendiri yaitu bahasa Indonesia, pada umumnya hanya didominasi oleh beberapa kalangan saja, terutama bagi mereka yang telah memperoleh pendidikan tertentu. Hal ini tentunya dapat menyebabkan sulitnya berkomunikasi dengan mayoritas masyarakat tersebut jika mereka tidak dapat menggunakan Bahasa Madura [2]. Untuk mengatasi permasalahan tersebut, maka diperlukan suatu sarana yang dapat membantu masyarakat dalam memahami makna kata atau kalimat dari suatu bahasa secara spesifik, sarana semacam ini umumnya disebut sebagai mesin penerjemah [3].

Mesin penerjemah sendiri merupakan alat yang sangat penting untuk menerjemahkan suatu bahasa ke bahasa lainnya secara otomatis. Tujuan dari dibuatnya alat ini adalah untuk memudahkan dalam memahami arti kata atau kalimat dari suatu bahasa, tanpa mengharuskan seseorang untuk membuka kamus terlebih dahulu [3]. Dalam ilmu komputer, penerjemahan termasuk dalam kategori pemrosesan bahasa alami atau *Natural Language Processing* (NLP), yang mana dalam bidangnya sendiri penerjemahan disebut dengan *Machine Translation* (MT) [4]. Terdapat beberapa pendekatan yang umum digunakan dalam bidang MT, beberapa diantaranya seperti *Rule-based Machine Translation* (RMT), *Statistical Machine Translation* (SMT), *Phrase-based Machine Translation* (PMT), dan *Phrase-based Statistical Machine Translation* (PSMT), dan juga *Neural Machine Translation* (NMT). Dari beberapa pendekatan yang telah disebutkan tersebut,

NMT menjadi salah satu metode yang dapat menciptakan model dengan hasil yang cukup efektif saat digunakan dalam melakukan penerjemahan, hal ini dikarenakan NMT merupakan metode baru yang dibuat dari hasil pengembangan metode sebelumnya [4].

Untuk modelnya sendiri, NMT dibangun menggunakan framework dengan model *sequence to sequence* (SeqToSeq) yang menggunakan arsitektur *encoder-decoder*. Arsitektur ini bekerja dengan cara melakukan konversi satu per satu dari kata sumber guna mendapatkan hasil yang direpresentasikan dalam bentuk vektor tetap (*encoder*), kemudian hasil yang telah direpresentasikan tersebut diekstrak satu per satu sesuai dengan kata yang menjadi target menggunakan RNN (*decoder*) [5].

Dalam implementasinya di lapangan, saat ini RNN memiliki dua jenis baru dari hasil pengembangannya yaitu LSTM (Long Short Terms Memory) dan GRU (Gated Recurrent Unit). Alasan dari dibuatnya hasil pengembangan tersebut dikarenakan keterbatasan RNN yang dalam menyimpan informasi dengan akurat jika data yang diberikan cukup panjang, oleh karena itu dikembangkan model LSTM dan GRU yang pada masing-masing arsitekturnya memiliki gerbang untuk menyeleksi data pada setiap *sequence* data yang diinputkan [6].

Baik LSTM maupun GRU, keduanya merupakan hasil pengembangan dari RNN yang sangat populer digunakan dalam mengerjakan berbagai tugas dengan model data sequensial. Yang membedakannya adalah GRU merupakan versi lebih simpel dari LSTM karena memiliki lebih sedikit gerbang dan tidak memiliki *memory unit*, sehingga waktu yang dibutuhkan untuk menyelesaikan tugasnya relatif lebih cepat dari LSTM. Namun untuk tingkat akurasi, *output* yang dihasilkan oleh LSTM relatif lebih tinggi daripada GRU dalam berbagai kasus pengolahan bahasa alami [7].

Meskipun penggunaan metode RNN sudah cukup bagus dalam mengolah data dengan bentuk sequensial, namun model ini masih memiliki keterbatasan dalam menjalankan tugas yang menggunakan arsitektur *encoder-decoder* seperti penerjemahan bahasa. Hal ini dikarenakan *decoder* pada arsitektur ini hanya menerima output *hidden state* terakhir dari *encoder*, yang mana jika data yang diproses oleh encoder ini cukup panjang, maka *hidden state* terakhir dari *encoder* akan banyak kehilangan informasi dari data yang diproses sebelumnya. Oleh sebab

itu, maka diperlukan suatu metode yang dapat menyelaraskan setiap *output hidden state encoder* terhadap target *time-step* pada *decoder*, metode ini disebut dengan Attention [8].

Namun dari banyaknya penelitian yang telah dilakukan menggunakan model NMT, umumnya tiap korpus yang digunakan memiliki panjang minimal 4 kata. Oleh sebab itu, diajukannya penelitian ini adalah selain untuk dapat menerapkan metode NMT berbasis *Attention* dengan model LSTM pada mesin penerjemah Bahasa Madura-Bahasa Indonesia karena terbukti memiliki performa yang cukup baik dalam memproses kalimat yang dukup panjang, penelitian ini juga dilakukan untuk mengetahui akurasi yang dihasilkan metode mesin penerjemah tersebut apabila korpus yang digunakan sebagai dataset hanya terdiri dari kata, frasa, ataupun kalimat pendek yang panjangnya kurang dari 4 kata. Penelitian ini dirancang menggunakan dataset berupa kamus digital Bahasa Madura-Bahasa Indonesia, yaitu MadureseSet [9]. Sementara untuk model *Attention*nya, model yang digunakan adalah *Additive Attention* [8].

1.2 Rumusan Masalah

1.2.1 Permasalahan

Permasalahan yang menjadi dasar dari penelitian ini adalah tentang pengembangan model mesin penerjemah Bahasa Madura-Bahasa Indonesia dengan arsitektur *encoder-decoder* berupa pengaruh panjang korpus terhadap akurasi yang dihasilkan oleh NMT berbasis *Attention* dalam penerjemahan Bahasa Madura ke Bahasa Indonesia. Dilakukannya penelitian terhadap permasalahan ini sendiri adalah untuk mengetahui efektivitas penggunaan NMT berbasis *Attention* berdasarkan panjang korpus yang digunakan.

1.2.2 Usulan Solusi

Untuk mengatasi permasalahan tersebut, solusi yang diajukan adalah sebuah mesin penerjemah Bahasa Madura-Bahasa Indonesia menggunakan arsitektur *encoder-decoder* dengan model LSTM berbasis *Attention*, yang mana untuk model datasetnya dibentuk menjadi beberapa macam berdasarkan panjang kalimat pada korpusnya. Sementara itu, untuk dataset yang digunakan dalam merancang mesin penerjemah ini adalah dataset yang didapat dari kamus dan buku terjemahan Bahasa Madura – Bahasa Indonesia. Untuk tahap *preprocessing*nya,

mesin penerjemah ini menggunakan lima tahapan, yaitu *case folding*, *punctuation removal*, *tokenization*, *add padding*, dan *one-hot encoding*. Kemudian untuk tahap evaluasinya, mesin penerjemah ini menggunakan metode *BiLingual Evaluation Understudy* (BLEU) untuk melakukan evaluasi terhadap hasil terjemahannya.

1.2.3 Pertanyaan peneliti

Berdasarkan latar belakang di atas, pertanyaan peneliti pada penelitian ini yaitu:

1. Bagaimana hasil perbandingan antara penggunaan dataset dengan panjang korpus maksimal 2 kata dan dataset dengan panjang korpus minimal 3 kata terhadap performa model LSTM berbasis *Attention*?
2. Berapakah hasil akurasi terbaik yang dapat dihasilkan oleh Mesin Penerjemah Bahasa Madura – Bahasa Indonesia berbasis *Attention* pada Model LSTM?

1.3 Batasan Masalah

Untuk mengantisipasi adanya pelebaran dari permasalahan yang sedang dibahas, tentunya diperlukan adanya batasan dari permasalahan tersebut. Berikut merupakan batasan masalah yang peneliti berikan untuk penelitian ini.

1. Dataset yang digunakan antara lain adalah korpus gabungan hasil seleksi dari buku Tata Bahasa Bahasa Madura [10] dengan data sebanyak 316 pasangan korpus, Pedoman Umum Ejaan Bahasa Madura Yang Disempurnakan [11] dengan data sebanyak 74 pasangan korpus, Morfologi dan Sintaksis Bahasa Madura [12] dengan data sebanyak 618 pasangan korpus, dan juga kamus digital Bahasa Madura-Bahasa Indonesia (MadureseSet [9]) dengan data sebanyak 36.086 pasangan korpus. Yang mana dari keempat sumber dataset tersebut, diperoleh total data sebanyak 37.094 pasangan korpus.
2. Metode penerjemah yang digunakan adalah NMT LSTM dengan arsitektur *encoder-decoder* berbasis *Attention*.
3. Berfokus pada penerjemahan bahasa Madura ke bahasa Indonesia.
4. Tidak adanya perhatian khusus terhadap tingkatan bahasa (seperti Enje' Iyeh; Engghi Enten; dan juga Engghi Bhunten) dan dialek dari bahasa Madura.

1.4 Tujuan dan Manfaat

Adapun tujuan dan manfaat dari dilakukannya penelitian ini, yaitu.

1.4.1 Tujuan

Tujuan dari dilakukannya penelitian ini adalah untuk mengetahui efektif atau tidaknya model mesin penerjemah ini jika digunakan korpus yang hanya memiliki panjang maksimal 2 kata.

Selain itu, penelitian ini juga dilakukan untuk dapat mengetahui tingkat akurasi terbaik yang dapat dihasilkan oleh Mesin Penerjemah Bahasa Madura – Bahasa Indonesia berbasis *Attention* pada Model LSTM.

1.4.2 Manfaat

Dengan diterapkannya Mesin Penerjemah Bahasa Madura – Bahasa Indonesia berbasis *Attention* pada Model LSTM ini, diharapkan dapat digunakan untuk membantu masyarakat Madura dalam belajar Bahasa Indonesia ataupun sebagai sarana untuk membantu berinteraksi dengan masyarakat luar Madura.

1.4.3 Sistematika Penulisan

Dalam skema penulisan laporan, pada penelitian ini dibagi menjadi beberapa bab yang sesuai dengan fungsinya masing-masing, hal ini dilakukan guna mempermudah melakukan pencarian informasi yang dibutuhkan, dan juga untuk menunjukkan mengenai penggunaan penulisan yang sistematis. Untuk pembagiannya, laporan penelitian ini dibagi menjadi beberapa bab berikut:

1 BAB I PENDAHULUAN

Pada bab ini berisi penjelasan mengenai latar belakang, identifikasi masalah, rumusan masalah, tujuan, ruang lingkup serta sistematika penulisan.

2 BAB II. KAJIAN PUSTAKA

Bab ini berisi tentang teori-teori dan beberapa penelitian terkait yang penulis gunakan sebagai dasar pengetahuan dan dijadikan sebagai bahan acuan untuk melakukan pembangunan rancangan dari mesin penerjemah yang dibuat.

3 BAB III. METODOLOGI PENELITIAN

Pada bab ini berisi tentang penjelasan mengenai perancangan dari sistem yang dibuat, yang mana di dalamnya termasuk penjelasan mengenai dataset, metode, serta skenario pengujian yang digunakan dalam merancang sistem dalam penelitian ini.

4 BAB IV. HASIL DAN PEMBAHASAN

Pada bab ini berisi tentang penjelasan mengenai lingkungan kerja yang digunakan, kode program yang telah dibuat, serta hasil uji coba dari setiap skenario yang telah dilakukan.

5 BAB V. PENUTUP

Pada bab ini berisi tentang kesimpulan dan saran dari penelitian yang telah dilakukan dan untuk yang akan dilakukan pada masa yang akan datang.