BAB 1
PENDAHULUAN

1.1 Latar Belakang

Indonesia tumbuh sebagai bangsa yang memiliki kebudayaan yang tidak sedikit, tersebut pula dengan bahasa yang dimilikinya. (Collins, 2014), menuturkan setidaknya ada 706 variasi bahasa yang terdapat di Indonesia terlepas dari faktor yang mempengaruhinya. Didukung dengan posisi geografis serta demografisnya, menjadikan keberagaman kebudayaan terkait pembahasaan begitu banyak ditemukan. Variasi yang tercipta dari tiap satuan kecil di wilayah tertentu baik yang mendekati maupun yang masih punya kesamaan dengan sebaran bahasa utamanya, menjadikan hal terkait pengumpulan serta pengolahan data terkait menjadi tantangan yang cukup besar.

Pembelajaran bahasa dengan mesin sekarang ini memang sudah mencapai pada tahap dimana jumlah data memegang kunci penting pada hasilnya. Berbagai pengembangan dilakukan, terutama banyak terjadi di era *Neural Machine Translation* (NMT). Namun sampai saat ini pengembangan tersebut hanya mampu bekerja di lingkup yang terorganisir dengan baik, serta dengan jumlah data yang cukup besar juga. Muncul pertanyaan bagaimana jika semua hal diatas berkebalikan dengan kondisi yang diharapkan.

Salah satu ide besar untuk mengembangkan pembelajaran lebih lanjut untuk hal ini adalah dengan membangun sistem secara *low-resource* atau sumber daya terbatas, yang akan memfokuskan media pembelajar untuk bekerja dan mampu membentuk pengetahuan yang dapat berasal dari beberapa metode yang sudah ada pada kondisi diatas. Penekanan pada sistem ini bisa dikaitkan dalam beberapa hal, namun umumnya lebih ke kuantitas sumber data yang dikumpulkan. Dengan alasan masalah *underfitting* dan *overfitting* saat generalisasi, hasil penelitian terdahulu lebih banyak diarahkan ke penggunaan data dalam jumlah besar. Sekarang ini sudah banyak muncul artikel terkait yang membahas kemungkinan sistem dengan spesifikasi diatas (Alencar, 2019; Folkman, 2019; Maheswari, 2018) dan penelitian dari (Sennrich & Zhang, 2019; Wdowiak, 2021), yang mana memberi jalan untuk bahasa dengan kondisi “susah diperoleh” untuk ikut dikembangkan. Dengan harapan untuk dapat menemukan ukuran pasti dalam

pengembangannya, dibutuhkan lebih banyak penelitian dengan topik terkait untuk mewujudkannya.

Arsitektur NMT sudah umum terdengar serta digunakan sebagai bagian arsitektur sistem *machine translation* sejak naik di WMT16, yang mana mempunyai kelebihan dapat melakukan berbagai tugas terkait penerjemahan bahasa secara tepat dan mudah. Dengan adanya pengembangan seperti model *Seq2Seq* dengan *LSTM* (Sutskever et al., 2014), dan mekanisme *attention* (Bahdanau et al., 2015) menjadikan dasar penting bagi NMT untuk muncul kembali, yang mampu melampaui nilai metrik BLEU untuk beberapa riset yang menggunakan sistem dan arsitektur terdahulu.

NMT memiliki berbagai potensi yang layak untuk diuji. Terutama terkait bentukan arsitektur, algoritma kerja, serta komponen tambahan lainnya yang digunakan, yang sudah menjadi isu utama terkait implementasinya secara efektif. Ada beberapa pilihan untuk memudahkan penelitian dalam hal ini untuk menyederahanakan prosesnya, yaitu dengan bantuan *toolkit* yang terfokus untuk menyelesaikan permasalahan terkait basis sistem NMT. Hal tersebut menjanjikan penelitian terkait penerjemahan bahasa dengan basis *neural* untuk mendapat berbagai kemudahan dalam penggunaannya.

Tersebutlah *toolkit* MarianNMT, Sebuah *toolkit* NMT yang dikembangkan oleh tim *Microsoft Translator* dengan harapan menciptakan *toolkit* yang *resource-friendly* dan dapat mencapai kecepatan *training* dan penerjemahan yang tinggi (Junczys-Dowmunt et al., 2018). Kemudahan dan kecepatan yang ditawarkan MarianNMT dapat menjadi pertimbangan bagus untuk memulai proyek terkait tugas mesin penerjemah berbasis *neural* untuk sistem lokal yang mengandalkan sumber daya *hardware* dalam *device,* terutama dengan GPU atau CPU.

Fokus penelitian ini diarahkan untuk menguji penerjemahan NMT untuk pasangan bahasa EN-ID (Inggris ke Indonesia) pada dua kasus berbeda. Akan digunakan MarianNMT sebagai *toolkit* NMT, dan menggunakan model untuk *training* dalam contoh di *repository* Marian sebagai model dasar NMT. Penelitian akan berfokus untuk menemukan perbandingan antar data uji berbeda dengan menggunakan MarianNMT sebagai *toolkit,* untuk membentuk model terjemahan dalam percobaan yang berbeda tersebut.

1.2 Perumusan Masalah

Menemukan perbandingan antar data uji berbeda yaitu pada hasil terjemahannya pada korpus Wikimedia dan QED&TED dengan MarianNMT, yang akan dibagi menjadi beberapa topik seperti, cara setup, training, penerjemahan, serta evaluasi dan pengujian hasilnya.

1.3 Tujuan Penelitian

Tujuan dari penelitian ini untuk mendapat hasil uji dengan *toolkit* MarianNMT. Hasil uji yang didapat akan menjadi pembanding untuk sumber data yang berbeda.

1.4 Pembatasan Masalah

Beberapa hal yang menjadi batasan dalam penelitian ini adalah sebagai berikut.

1. Penerjemahan satu arah dari bahasa Inggris ke bahasa Indonesia.
2. Alur penelitian yang digunakan akan dibuat mendekati contoh dari "Nematus-Style Shallow RNN" sebagai implementasi model dalam MarianNMT.
3. Skor pada pengujian hasil otomatis dengan skor BLEU dan SpBLEU.
4. Penelitian dijalankan secara lokal dan *offline*
5. Semua permasalahan pemrosesan akan dibatasi sesuai kemampuan *toolkit* dan model yang digunakan saja.

1.5 Sistematika Penulisan

Sistematika penulisan penelitian ini disusun untuk memberikan gambaran umum tentang penelitian yang dijalankan. Sistematika laporan tugas akhir ini disusun dalam 5 (lima) bab yang terdiri dari Bab I Pendahuluan, Bab II Tinjauan Pustaka, Bab III Metodologi Penelitian, Bab IV Hasil dan Analisis Sistem, serta Bab V Penutup.

Bab I Pendahuluan adalah bab yang berisi latar belakang, perumusan masalah, tujuan penelitian, pembatasan masalah, dan sistematika penulisan.

Bab II Tinjauan Pustaka adalah bab yang berisi uraian sistematis tentang hasil-hasil penelitian yang didapat oleh peneliti terdahulu dan landasan teori yang

ada hubungannya dengan penelitian yang akan dilakukan.

Bab III Metodologi Penelitian adalah bab yang berisi tentang bahan penelitian, perangkat penelitian yang digunakan, metode yang akan digunakan pada penelitian, dan perancangan pengujian yang akan dilakukan pada penelitian.

Bab IV Hasil dan Analisis adalah bab yang berisi hasil penelitian, penjelasan mengenai implementasi metode yang digunakan, hasil analisis dari setiap pengujian. Bagian yang ditampilkan akan dilakukan analisis terlebih dahulu untuk mengarah kepada suatu kesimpulan.

Bab V Penutup adalah bab yang berisi kesimpulan dari penelitian yang telah dilakukan dan saran atau rekomendasi untuk perbaikan, pengembangan atau kesempurnaan dan kelengkapan penelitian yang telah dilakukan.