BAB 1

PENDAHULUAN

1.1 Latar Belakang

Dalam melakukan aktivitas sehari-hari, kita akan menerima berbagai informasi. Penyajian informasi dalam berbagai bentuk seperti tulisan, gambar, tabel, diagram, audio, video, dan lain-sebagainya. Seiring dengan perkembangan teknologi, informasi-informasi tersebut dapat dengan mudah kita terima dan bagikan ke orang lain melalui aplikasi-aplikasi di dalam *smartphone* yang kita gunakan sehari-hari.

Dari 272,1 juta populasi di Indonesia, sebanyak 59% atau 160 juta orang merupakan pengguna aktif media sosial, di mana penggunaan rata-rata media sosial di Indonesia pada tahun 2020 adalah 3 jam 26 menit (Kemp, 2020). Dari keseluruhan pengguna sosial media di Indonesia, 99% di antaranya menggunakan aplikasi *messaging* dalam sebulan terakhir (Kemp, 2020). Besarnya penggunaan sosial media terutama dalam aplikasi *messaging* berdampak pada banyaknya informasi berupa teks yang beredar dan diterima oleh kita sehari-hari. Tulisan atau teks tersebut perlu diolah dengan baik agar tidak terjadi disinformasi yang dapat merugikan banyak pihak. Terdapat banyak cara dalam melakukan pengolahan data berupa teks, salah satunya adalah klasifikasi teks atau *text classification*.

Berdasarkan data sensus yang beragam dari seluruh negara, terdapat sekitar 990 ribu orang yang dapat berbahasa Indonesia di seluruh dunia (tidak termasuk negara Indonesia). Selain itu, bahasa Indonesia juga termasuk ke dalam 10 bahasa yang paling sering digunakan di seluruh dunia (Eberhard, et al., 2021). Semakin

berkembangnya ilmu mengenai *Natural Language Processing* (NLP) di Indonesia, mulai banyak penelitian yang menggunakan model *text classification* atau klasifikasi teks untuk bahasa Indonesia.

Dalam membuat sebuah model *text classification*, diperlukan data teks yang mencukupi agar dapat digunakan sebagai data *train*. Kebanyakan dari penelitian NLP saat ini hanya fokus pada 20 bahasa dari total 7000 bahasa yang digunakan oleh manusia, sehingga masih banyak bahasa yang kurang dipelajari atau biasa dirujuk sebagai *low-resource languages* (Magueresse, et al., 2012). Berdasarkan sejarah, bahasa Indonesia sendiri merupakan turunan bahasa dari bahasa Melayu, sehingga juga termasuk ke dalam *low-resource languages* (Murakami, 2019). Banyaknya bahasa yang belum dipelajari ini membutuhkan sebuah *machine learning* yang dapat mengatasi permasalahan data yang tidak mencukupi pada bahasa-bahasa tersebut. Salah satu solusi untuk mengatasi kurangnya data pada *low-resource languages* adalah dengan memanfaatkan *transfer learning* yang memungkinkan sebuah model untuk melakukan *multilingual text classification*.

Dalam menyelesaikan permasalahan tersebut, terdapat beberapa penelitian terdahulu yang dikembangkan oleh Google dan Facebook, di antaranya adalah BERT (Devlin, et al., 2019), *multilingual* BERT (mBERT) yang merupakan pengembangan dari BERT, RoBERTa (Liu, et al., 2019), dan XLM (Lample & Conneau, 2019). Namun masih terdapat beberapa kekurangan dan aspek yang dikorbankan dalam membuat model *multilingual text classification* seperti terbatasnya jumlah data teks, banyak *low-resource languages* yang belum dapat diprediksi oleh model, dan nilai akurasi yang masih belum dapat mengungguli *monolingual model* dengan konsisten pada seluruh bahasa. Keterbatasan model dari

metode-metode di atas membuat sebuah urgensi dalam melakukan pengolahan teks menggunakan *machine learning* pada data teks dengan *low-resource languages*, termasuk data teks berbahasa Indonesia.

Walaupun bahasa Indonesia termasuk ke dalam *low-resource languages* dan memiliki keterbatasan data, terdapat sebuah penelitian yang membuat sebuah model menggunakan IndoBERT untuk melakukan beberapa pekerjaan termasuk klasifikasi teks. Penelitian ini mampu menghasilkan akurasi yang lebih tinggi dibandingkan model-model lain seperti fastText, mBERT, bahkan model *multilingual* terbaru yaitu XLM-R (Willie, et al., 2020). Namun, model klasifikasi teks untuk bahasa Indonesia itu saja tidaklah cukup untuk menyelesaikan permasalahan di Indonesia dikarenakan penggunaan bahasa yang lebih dari satu. Terdapat 2 pilihan bahasa yang marak digunakan oleh orang Indonesia terutama para generasi muda dalam menggunakan salah satu platform media sosial terbesar (Instagram), yaitu bahasa Indonesia dan bahasa Inggris (Abraham, 2017).

Pemilihan penggunaan bahasa dalam melakukan aktivitas di media sosial membuat sebuah permasalahan baru dalam melakukan klasifikasi teks. Selain memerlukan *resource* yang lebih besar, membuat 2 buah model untuk melakukan klasifikasi teks dengan bahasa yang berbeda juga tidaklah efektif. Berdasarkan tingkat kebutuhan dan efektivitas pekerjaan dalam melakukan klasifikasi teks di Indonesia, dibutuhkan sebuah model yang mampu melakukan *multilingual text classification*, terutama untuk data teks bahasa Indonesia dan bahasa Inggris.

Pada tahun 2020, terdapat sebuah penelitian yang mengembangkan metode baru bernama XLM-RoBERTa. *Cross Lingual Model – RoBERTa* atau XLM-R merupakan pengembangan dari XLM dan BERT. Hasil penelitian ini menunjukkan,

untuk pertama kalinya, bahwa XLM-R memungkinkan untuk membuat sebuah model berukuran raksasa yang dapat melakukan *multilingual text classification* dengan tingkat akurasi yang lebih baik daripada kebanyakan *monolingual text classification model* dan dapat digunakan pada 100 bahasa, termasuk *low-resource languages* seperti bahasa Indonesia (Conneau & Khandelwal, 2020).

Dalam penelitian ini, metode XLM-R akan digunakan untuk melakukan *multilingual text classification* terhadap *news dataset* dengan bahasa yang berbeda (bahasa Indonesia dan Inggris) untuk menentukan kategori berita berdasarkan judul berita. *News dataset* yang digunakan adalah kumpulan berita lokal berbahasa Indonesia pada tahun 2017 dan berita Internasional berbahasa Inggris dari tahun 2012 hingga 2018. Penelitian ini memiliki fokus dan tujuan utama untuk membuat sebuah model yang dapat melakukan *multilingual text classification* pada teks berbahasa Indonesia menggunakan *transfer learning* pada model XLM-R yang sudah dilatih sebelumnya pada lebih dari 100 bahasa yang berbeda.

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, dapat dirumuskan beberapa masalah yang hendak diselesaikan dalam penelitian ini, yaitu:

1. Bagaimana cara mengimplementasikan XLM-R sebagai model klasifikasi teks berbahasa Inggris dan Indonesia?
2. Bagaimana performa model XLM-R pada model klasifikasi teks berbahasa Inggris dan Indonesia?

1.3 Batasan Masalah

Batasan masalah dalam penelitian ini dapat dijabarkan sebagai berikut:

1. Lingkup bahasa dan *dataset* yang akan digunakan adalah *news dataset* berbahasa Inggris dan Indonesia
2. *News dataset* berbahasa Indonesia menggunakan kumpulan berita dari situs berita terkenal (detik.com) pada tahun 2017 dan *news dataset* berbahasa Inggris menggunakan kumpulan berita HuffPost dari tahun 2012 hingga 2018
3. Terdapat 7 kategori berita dalam *news dataset* yang akan digunakan dalam penelitian ini, yaitu kategori *Sports*, *Food, World News, Healthy & Living, Travel, Business & Finance,* dan *Tech & Internet*

1.4 Tujuan Penelitian

Tujuan dari penelitian ini dapat dijabarkan sebagai berikut:

1. Mengimplementasikan XLM-R sebagai model klasifikasi teks berbahasa Inggris dan Indonesia
2. Mengukur performa model XLM-R pada model klasifikasi teks berbahasa Inggris dan Indonesia

1.5 Manfaat Penelitian

Manfaat dari penelitian ini dapat dijabarkan sebagai berikut:

1. Mengembangkan penelitian *multilingual Natural Language Processing* (NLP)
2. Membantu *developer* dari situs berita ataupun aplikasi berita dalam menyelesaikan pekerjaan secara otomatis

1.6 Sistematika Penulisan

Dalam penyusunan laporan skripsi ini, digunakan sistematika penulisan sebagai berikut:

BAB 1 PENDAHULUAN

Pada bab ini berisikan latar belakang masalah, batasan masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan laporan skripsi.

BAB 2 LANDASAN TEORI

Pada bab ini menjelaskan tentang teori-teori yang digunakan dalam penelitian ini. Teori tersebut meliputi konsep dasar maupun metode yang digunakan untuk memperoleh hasil yang diharapkan pada penelitian ini. Terdapat beberapa teori yang dijelaskan pada bab ini, diantaranya adalah *Natural Language Processing*, *Multilingual Text Classification*, dan XLM-R atau *Cross Lingual Model – RoBERTa*.

BAB 3 METODOLOGI PENELITIAN

Pada bab ini berisikan metode penelitian secara urut mulai dari perancangan sistem, implementasi, evaluasi, hingga penulisan laporan.

BAB 4 HASIL DAN DISKUSI

Pada bab ini menjelaskan tentang implementasi yang dilakukan oleh penulis berdasarkan perancangan yang telah dibuat sebelumnya. Implementasi dilakukan untuk memperoleh hasil yang diharapkan pada penelitian yang dilakukan. Pada

akhir bab ini juga dituliskan hasil penelitian dan analisa dari implementasi yang telah berhasil dilakukan.

BAB 5 SIMPULAN DAN SARAN

Pada bab ini memberikan kesimpulan berdasarkan hasil implementasi dan analisa yang telah dilakukan. Berdasarkan kesimpulan tersebut, diberikan juga saran untuk pengembangan penelitian lebih lanjut.