BAB I
PENDAHULUAN

1.1 Latar Belakang

Pemilihan Gubernur merupakan momen penting dalam dinamika politik daerah di Indonesia. Besarnya peran Gubernur sebagai kepala daerah tingkat provinsi membuat pembahasan terkaitnya cenderung lebih aktif dan melibatkan lebih banyak perhatian masyarakat (Jumadin & Wibisono, 2020). Dalam era digital saat ini, media sosial telah menjadi sarana publik untuk bebas mengekspresikan pendapat, termasuk dalam lingkup demokrasi (Kasmani, 2024), seperti yang terlihat pada pemilihan kepala daerah. Tingginya aktivitas masyarakat dalam membahas pemilihan Gubernur melalui media sosial menghasilkan data komentar dalam jumlah yang besar, namun komentar tersebut sering menggunakan bahasa informal, bahasa sehari-hari, singkatan, serta bercampur dengan bahasa daerah dan dialek lokal yang sulit dipahami (Rizkyna, Nisa, & Aulia, 2020). Hal tersebut dapat menghambat pemrosesan data komentar apabila data tersebut hendak digunakan untuk analisis atau tujuan lainnya, sehingga diperlukan proses normalisasi untuk mengubah teks informal dari data komentar tersebut menjadi teks yang sesuai dengan kaidah bahasa Indonesia.

Normalisasi teks mengacu pada proses mengubah kata informal menjadi kata formal dalam bahasa Indonesia. Fokus utama dari normalisasi adalah mengubah kata-kata tidak baku, singkatan, dan bahasa sehari-hari yang umum ditemukan dalam komentar media sosial menjadi kata-kata baku sesuai Kamus Besar Bahasa Indonesia (KBBI). Proses normalisasi manual membutuhkan waktu dan sumber daya yang sangat banyak, terutama jika data yang diolah berjumlah besar. Normalisasi secara manual juga rentan terhadap inkonsistensi dan kesalahan manusia dalam prosesnya. Selain itu, jumlah data komentar di media sosial akan selalu meningkat yang membuat normalisasi secara manual semakin tidak mungkin dan tidak efisien untuk dilakukan. Oleh karena itu, diperlukan solusi otomatisasi untuk mengatasi masalah ini.

Statistical Machine Translation (*SMT*) merupakan salah satu solusi untuk masalah normalisasi teks secara otomatis (Costa-jussà & Banchs, 2013). Pendekatan ini

memungkinkan model untuk mempelajari pola-pola perubahan dari bahasa sumber ke bahasa target (informal ke formal) berdasarkan dari *dataset* korpus paralel yang tersedia (Macken & Lefever, 2008). Penerapan *SMT* dapat mengotomatisasi proses normalisasi sekaligus menjaga konsistensi hasil dengan menggunakan model statistik yang telah dilatih. Salah satu pendekatan dari *SMT* yang sering digunakan adalah *Phrase-Based Statistical Machine Translation* (*PBSMT*) yang memproses teks dalam bentuk frasa.

Dalam penelitian ini, *PBSMT* menjadi pilihan yang sesuai karena jumlah data komentar terkait pemilihan Gubernur memadai untuk kebutuhan metode ini, seperti yang dilakukan oleh (Wibowo dkk., 2020) yang menggunakan 2500 kalimat paralel informal-formal. *PBSMT* juga dapat memberikan performa yang baik untuk data yang memiliki banyak *noise* seperti data komentar sosial media, sebagaimana dibuktikan oleh penelitian (Eleison, Hutahaean, Tampubolon, Panggabean, & Fitriyaningsih, 2022). Selain itu, metode ini juga efektif untuk menangani bahasa daerah seperti yang ditunjukkan dalam penelitian (Lestari, Ardiyanti, & Asror, 2021) yang berhasil menerapkan PBSMT untuk terjemahan bahasa Jawa meski dengan data yang terbatas. Akan tetapi, *PBSMT* dalam penerapannya pada data media sosial memiliki tantangan utama yaitu bagaimana sistem menangani kata-kata di luar kosakata yang telah dilatih atau *out-of-vocabulary* (*OOV*) (Lochter, Silva, & Almeida, 2020). Masalah *OOV* ini semakin krusial mengingat bahasa sehari-hari di media sosial yang terus berkembang mengikuti tren. Dalam hal ini, *framework PBSMT* bernama *moses* memiliki kemampuan menangani *OOV* melalui berbagai opsi yang disediakannya.

Pemilihan judul "Normalisasi Komentar Media Sosial Pasangan Calon Gubernur 2024 dengan *Statistical Machine Translation*" ini didasarkan pada urgensi permasalahan teks informal di media sosial dalam konteks pemilihan gubernur yang telah dipaparkan dan potensi pendekatan *SMT* yaitu *PBSMT* sebagai solusi yang telah terbukti efektif untuk normalisasi bahasa Indonesia. Penggunaan *PBSMT* untuk normalisasi komentar media sosial pasangan calon gubernur 2024 diharapkan dapat memberikan solusi terkait proses normalisasi teks media sosial pada penelitian yang akan dilakukan. Metodologi yang dikembangkan juga diharapkan

dapat menjadi landasan untuk pengembangan sistem serupa dalam lingkup pemilihan kepala daerah lainnya di Indonesia.

1.2 Rumusan Masalah

Adapun rumusan masalah yang menjadi fokus dalam penelitian ini adalah sebagai berikut.

1. Bagaimana implementasi model *PBSMT* untuk normalisasi komentar media sosial pasangan calon Gubernur 2024 berbasis *website*?
2. Bagaimana performa model *PBSMT* dalam melakukan normalisasi komentar media sosial pasangan calon Gubernur 2024 berdasarkan metrik evaluasi *Bilingual Evaluation Understudy Score* (*BLEU*)?

1.3 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah untuk memberikan solusi atas rumusan masalah yang telah disampaikan, yaitu sebagai berikut.

1. Untuk mengetahui implementasi model *PBSMT* untuk normalisasi komentar media sosial pasangan calon Gubernur 2024 berbasis *website*.
2. Untuk mengukur performa model *PBSMT* dalam melakukan normalisasi komentar media sosial pasangan calon Gubernur 2024 berdasarkan metrik evaluasi *BLEU.*

1.4 Batasan dan Asumsi Penelitian

Penelitian ini dirancang dengan menetapkan batasan yang jelas dan asumsi mendasar untuk menjaga fokus penilitian serta mencapai tujuan yang telah direncanakan.

1.4.1. Batasan Penelitian

Adapun batasan yang ditetapkan dalam penelitian ini adalah sebagai berikut.

1. Penelitian ini hanya mengambil data dari komentar di platform media sosial Instagram yang berkaitan dengan pasangan calon Gubernur Jawa Timur 2024.
2. Periode pengambilan data dibatasi selama masa kampanye, yaitu dari 25 September 2024 hingga 23 November 2024, untuk menangkap interaksi dan volume komentar yang tinggi pada waktu tersebut.

3. Normalisasi teks dalam penelitian ini difokuskan pada konversi bahasa tidak baku menjadi bahasa baku sesuai dengan Kamus Besar Bahasa Indonesia.
4. Model hanya dibangun dan dijalankan pada *operating system* Ubuntu 18.04.

1.4.2. Asumsi Penelitian

Berdasarkan penelitian terdahulu, diasumsikan bahwa penerapan *PBSMT* menggunakan *moses* dengan *KenLM* dapat menghasilkan normalisasi teks yang sesuai dengan tata bahasa Indonesia baku. Tantangan dalam menangani *OOV* yang sering muncul pada komentar media sosial diharapkan dapat diatasi dengan metode-metode penanganan *OOV* yang disediakan moses. Jumlah data komentar selama masa kampanye diasumsikan mencukupi untuk membentuk *dataset* pelatihan model *PBSMT* yang dapat menghasilkan performa optimal dalam normalisasi teks komentar media sosial pasangan calon Gubernur Jawa Timur 2024.

1.5 Manfaat Penelitian

Adapun manfaat penelitian ini bagi pengembangan studi dan penelitian di masa mendatang adalah sebagai berikut.

1. Menyediakan landasan metodologi untuk pengembangan model normalisasi teks bahasa Indonesia.
2. Menyediakan *dataset* paralel informal-formal yang dapat digunakan untuk penelitian sejenis.
3. Menyediakan hasil penelitian dan analisis sebagai referensi penelitian normalisasi teks bahasa Indonesia.

1.6 Sistematika Penulisan

Sistematika penulisan dalam penelitian ini terdiri dari enam bab yang disusun dengan tujuan agar pembahasan dapat lebih terfokus pada pokok permasalahan dan tidak melebar ke permasalahan lain atau di luar batasan pokok permasalahan yang telah dijabarkan sebelumnya. Berikut sistematika penulisan proposal tugas akhir.

BAB I PENDAHULUAN

Bab ini berisikan gambaran umum tentang latar belakang penelitian, masalah yang diangkat, batasan penelitian, manfaat.

BAB II LANDASAN TEORI

Bab ini berisikan landasan teori yang menjadi acuan dasar penelitian termasuk konsep *SMT* dan keseluruhan *moses pipeline* untuk proses *preprocessing* hingga evaluasi.

BAB III METODOLOGI PENELITIAN

Bab ini memaparkan langkah-langkah yang dilakukan selama penelitian, meliputi pengumpulan data, pembangunan sistem, pengujian, dan pembangunan *website*.

BAB IV HASIL DAN PEMBAHASAN

Bab ini memaparkan hasil penelitian dan skenario pengujian yang dilakukan. Serta hasil analisis dari sistem yang dibuat.

BAB V KESIMPULAN DAN SARAN

Bab ini berisi rangkuman dari hasil penelitian yang telah dilakukan, termasuk pencapaian tujuan penelitian, evaluasi terhadap implementasi metode *PBSMT*, serta saran untuk pengembangan penelitian lebih lanjut.