

Bab I

Pendahuluan

Pada bagian ini akan dijelaskan beberapa hal penting terkait tentang latar belakang penelitian, topik yang diangkat, batasan masalah, tujuan, dan rencana kegiatan yang akan dikerjakan pada tugas akhir ini.

1.1 Latar Belakang

Media sosial merupakan media dalam jaringan yang memungkinkan manusia untuk dapat saling berinteraksi tanpa dibatasi oleh ruang dan waktu.

Data melimpah dari unggahan pengguna media sosial dinilai berpotensi menjadi sumber informasi yang sangat berharga, dengan data dari media sosial kita dapat mengumpulkan informasi seperti preferensi pengguna media sosial, memprediksi tren yang akan datang, bahkan memperkirakan emosi pengguna media sosial saat mengirimkan unggahan di media sosialnya.

Kebanyakan media sosial memberi kebebasan pada penggunanya untuk berekspresi dalam unggahan mereka, termasuk dalam gaya penulisan yang digunakan, namun hal ini sering kali dinilai sebagai sebuah permasalahan untuk penelitian di bidang *Natural Language Processing* (NLP), penulisan yang tidak sesuai standar bahasa, serta istilah asing dan baru yang digunakan dalam suatu kelompok tertentu sering kali menjadi permasalahan dalam penelitian berbasis teks.

Berdasarkan penelitian yang dilakukan oleh Budiasa dan Savitri (2019) didapatkan hasil yang menyatakan bahwa setidaknya 76% generasi muda saat ini menggunakan kata-kata gaul (slang) dalam percakapan sehari-hari mereka.

Menurut Lourentzou dkk (2019) keberadaan noise di dalam teks dapat menimbulkan dampak negatif pada beberapa penelitian NLP, disaat bersamaan normalisasi dapat mengurangi dampak dari noise dan meningkatkan performa dari permodelan NLP [5].

Normalisasi adalah proses standarisasi data yang umumnya dilakukan sebagai salah satu tahap praproses dalam proses penambangan data. Tujuan dilakukannya normalisasi adalah agar proses penambangan data bisa terhidar dari ambiguitas dan bias, termasuk juga diantaranya standarisasi format penulisan dan penyesuaian pada bahasa yang tidak sesuai.[9].

Ada beberapa penelitian yang membahas tentang normalisasi teks pada media sosial, diantaranya adalah penelitian yang dilakukan oleh Riyaddulloh dan Romadhony (2021) dan Sebastian dan Nugraha (2019).

Pada penelitiannya, Sebastian dan Nugraha (2019) mendapatkan kesimpulan bahwa normalisasi kata slang berbasis responden kuisioner hanya mampu menangani 59.67% dari kata slang yang diujikan[3], sedangkan penelitian Riyaddulloh dan Romadhony (2021) membuktikan adanya peningkatan nilai performansi sistem klasifikasi sentimen dari 88% meningkat menjadi 91% setelah menjalani proses normalisasi [2].

Pada penelitian ini akan dibangun permodelan proses normalisasi teks dari kalimat tidak baku yang digunakan sehari-hari dalam media sosial menjadi kalimat baku.

Mengadopsi dari permodelan Mesin Translasi yang dikembangkan dalam penelitian Fadel Razsiah (2023) [8] dan dan I Gede Bintang Arya Budaya (2022) [4] yang membahas pembangunan mesin translasi yang dapat menerjemahkan bahasa daerah ke bahasa indonesia, sistem akan dibangun dalam bentuk mesin translasi dengan arsitektur *Sequence to Sequence* serta pendekatan *Attention* dan *Long-Short Term Memory* (LSTM), model akan dibangun menggunakan bahasa pemrograman Python dengan *library* yang disediakan oleh Tensorflow.

1.2 Topik dan Batasan

Berdasarkan latar belakang yang telah dijelaskan, topik yang terkait dalam penelitian ini diantaranya adalah:

1. Penelitian ini melakukan pemrosesan teks, dimana proses pengubahan bentuk teks yang tidak terstruktur, diubah menjadi data yang lebih terstruktur dan dapat diterima oleh sistem lanjutan dengan baik.
2. Proses normalisasi dilakukan dengan permodelan mesin translasi dengan arsitektur *Sequence to Sequence* yang dirasa paling tepat untuk menangani data dalam bentuk serangkaian kalimat. Target hasil dari tugas akhir ini adalah kalimat dalam format baku.
3. Pada penelitian ini dilakukan beberapa aksi normalisasi, diantaranya penggantian kata seperti kata gaul dan bahasa asing, serta penghilangan karakter spesial dan alamat url.

Berdasarkan latar belakang yang telah dijelaskan, batasan dari permasalahan pada tugas akhir ini adalah:

1. Penelitian ini hanya dilatih pada unggahan dengan bahasa utama bahasa indonesia.

2. Model hanya bekerja searah dari bahasa indonesia tidak baku menjadi bahasa indonesia baku
3. Bahasa tidak baku yang dapat di proses oleh media hanya bahasa tidak baku indonesia dan bahasa inggris.

1.3 Tujuan

Tujuan yang ingin dicapai dalam penulisan proposal tugas akhir ini adalah:

1. Membangun sebuah sistem normalisasi yang mampu merepresentasikan kalimat tidak baku kedalam bentuk yang lebih baku.
2. Sistem yang dibangun kedepannya dapat membantu meningkatkan performansi penelitian di bidang Penambangan Teks dan Pemrosesan Bahasa Alami yang menggunakan data serupa.

1.4 Rencana Kegiatan

Penelitian ini mulai dirumuskan sejak Desember 2022 dan diharapkan dapat diselesaikan pada Juni 2023, rencana kegiatan dibagi menjadi beberapa tahapan, yaitu:

- Identifikasi Masalah
Kegiatan yang dilakukan pada tahap ini adalah pengamatan pada lingkungan sekitar, merumuskan masalah yang akan dibahas dan tujuan penelitian dilakukan.
- Studi literatur
Pada tahap ini dilakukan pembelajaran terhadap topik permasalahan dan solusi yang sudah ada dengan cara mengumpulkan literatur berupa buku, jurnal penelitian dan sumber-sumber lain yang berhubungan dengan topik permasalahan.
- Pengumpulan Data
Pengumpulan data dilakukan dengan beberapa cara yaitu menggunakan dataset dari penelitian mesin translasi lain dengan target bahasa indonesia, penggalan komentar dari media sosial Twitter, dan melalui kuisioner kalimat sehari-hari.
- Perancangan Sistem
Setelah melakukan beberapa tahapan sebelumnya, pada kegiatan ini dilakukan perancangan sistem yang dapat menjadi solusi untuk permasalahan yang diangkat sebagai topik penelitian.

- Implementasi dan Pengujian Sistem

Pada tahap ini dilakukan implementasi sistem berdasarkan rancangan sebelumnya, setelah permodelan sistem selesai dibangun dan dilatih dengan data latih, performa sistem akan diuji menggunakan data uji.

- Analisis Hasil Pengujian

Setelah implementasi dan pengujian, pada tahap ini hasil pengujian akan dianalisis dan menarik kesimpulan hasil penelitian berdasarkan nilai akurasi dari sistem yang dibangun.

- Penulisan Laporan

Kegiatan ini mulai dilakukan bersamaan dengan kegiatan identifikasi masalah, kegiatan ini dinyatakan selesai ketika analisis hasil selesai dilakukan dan tidak ada lagi perbaikan pada model yang dibangun ataupun laporan yang ditulis.