BAB I
PENDAHULUAN

1.1. Latar Belakang

Dunia saat ini tengah memasuki era revolusi industri 4.0 atau revolusi industri dunia keempat di mana teknologi menjadi basis dalam setiap aspek kehidupan manusia. Aspek pendidikan pun tak luput dari pengaruh revolusi industri 4.0. Dunia pendidikan dituntut mengikuti perkembangan teknologi yang sedang berkembang saat ini dengan cara memanfaatkan teknologi informasi sebagai fasilitas dalam membantu proses belajar dan mengajar. (Riyana, 2018) mengatakan tantangan pendidikan di era revolusi industri 4.0 berupa perubahan dari cara belajar, pola berpikir serta cara bertindak para peserta didik dalam mengembangkan inovasi kreatif berbagai bidang.

Universitas Teknokrat Indonesia sebagai salah satu universitas terbaik di provinsi Lampung dalam menghadapi era revolusi industri 4.0, memanfaatkan teknologi informasi untuk membantu proses belajar mengajar serta meningkatkan kualitas pengaksesan sistem informasi kampus dengan menggunakan *Chatbot* Sistem Informasi Akademik Artifisial yang selanjutnya disebut SIAA. *Chatbot* adalah sebuah program komputer yang dirancang untuk dapat berinteraksi atau berkomunikasi dengan manusia berbasiskan audio atau teks (J. Bang, 2015) (Rebedea, 2013). *Chatbot* diharapkan mampu untuk menyimulasikan dialog layaknya yang dilakukan oleh manusia setidaknya terdiri atas dua bagian utama yaitu bagian pemodelan bahasa dan algoritme komputasi untuk mengolah informasi yang

diberikan oleh pengguna dalam bentuk bahasa sehari-hari manusia sehingga dapat diproses oleh sistem *chatbot*.

Chatbot adalah sebuah sistem, yang memiliki arsitektur terdiri dari beberapa modul yaitu *Messaging Backend, Natural Language Processing, Natural Language Understanding, Decision Engine, Natural Language Generation, Data Connector,* dan *Custom Data Source* (Anbotux, 2017). *Natural Language Processing* adalah bidang ilmu komputer dan linguistik yang berkaitan dengan interaksi antara komputer dan bahasa alami manusia (Kumar, 2013). *Natural Language Processing (NLP)* berkaitan dengan bagaimana teknologi dapat secara bermakna menafsirkan dan bertindak atas input bahasa manusia. *Natural Language Processing* memungkinkan teknologi seperti *Amazon Alexa* dan *Google Assistant* untuk memahami apa yang dikatakan pengguna dan bagaimana bereaksi terhadapnya. Tanpa *Natural Language Processing*, *chatbot* yang membutuhkan input bahasa relatif tidak berguna (Phillips, 2018). *Natural Language Processing (NLP)* akan menerima pesan dari pengguna yang kemudian akan diterjemahkan sebagai perintah. Selanjutnya, sistem *chatbot* akan merespons sesuai *Custom Data Source* yang dimiliki.

Chatbot selayaknya seperti sistem komputer yang lainnya memiliki permasalahan. Permasalahan ini terjadi pada modul *Natural Language Processing* dan *Natural Language Understanding*. Hal ini karena *Natural Language Processing (NLP)* merupakan kunci untuk keberhasilan atau kegagalan *chatbot* (Phillips, 2018), masalah yang terjadi berkaitan tentang bagaimana memahami pesan dan merespons dengan tepat, mengekstraksi

informasi, dan manajemen dialog (Minh, 2017) (Phillips, 2018), Permasalahan ini terjadi karena terdapat kesalahan pengetikan, tata bahasa dan penggunaan bahasa yang buruk dalam pesan pengguna yang mempengaruhi interaksi dengan *chatbot*. Karena itu sebelum pesan pengguna di proses lebih lanjut, perlu dilakukan proses normalisasi teks. Normalisasi teks adalah proses pengubahan singkatan, salah pengetikan, akronim, angka, tanggal, waktu, karakter-karakter khusus, dan simbol-simbol dengan bentuk huruf alfabet lengkap sehingga tidak terjadi ambiguitas (Dutoit, 1997).

Untuk dapat mendeteksi dan memperbaiki kesalahan pengetikan dan kata tidak baku dalam pesan dibutuhkan sebuah algoritme yang dapat mengubah kata tidak baku menjadi kata baku yang paling cocok yaitu algoritme jarak *string* (Saragih, 2017). Algoritme jarak *string* adalah jarak edit minimum antara dua *string* yang terdiri dari jumlah minimum operasi pengeditan, penyisipan, penghapusan, penggantian karakter (Jurafsky, 2017). Ada beberapa algoritme jarak *string* yang bisa digunakan dalam proses normalisasi teks antara lain, Damerau-Levenshtein Distance yang memiliki kelebihan yaitu dapat menghitung empat operasi (penyisipan, penghapusan, penggantian, dan transposisi) cocok dengan lebih dari 80% dari semua kesalahan ejaan manusia (Damerau, 1964). Lalu *Hamming Distance* yang hanya memungkinkan substitusi, karena itu hanya berlaku untuk string dengan panjang yang sama. Namun, untuk membandingkan string dengan panjang berbeda, atau *string* di mana tidak hanya penggantian tetapi juga penyisipan atau penghapusan, metrik yang lebih canggih seperti Levenshtein lebih tepat (Hamming, 1950). *Longest Common Subsequence* (LCS) *distance*

dapat membandingkan dua *file*, dan dengan menemukan kesamaan, menghitung perbedaannya. Ini dapat digunakan untuk memperbaiki kesalahan ejaan dan membandingkan dua urutan genetik untuk homologi (kesamaan) dan lebih cepat untuk *string* yang lebih panjang dari sekitar 200.000 karakter tetapi algoritme ini hanya memungkinkan penyisipan dan penghapusan, bukan substitusi (W.J.Masek & M.S.Paterson, 1980).

Dari tiga algoritme di atas, algoritme jarak *string* yang dipilih untuk normalisasi teks yang dapat memperbaiki kesalahan pengetikan adalah Damerau-Levenshtein distance. Dalam penelitian Raffael Vogler pada tahun 2013 yang melakukan pembandingan dengan sembilan algoritme jarak *string* dan menghasilkan kesimpulan bahwa algoritme yang cocok untuk menangani masalah kesalahan pengetikan adalah variasi Levenshtein adalah yang paling baik. Algoritme ini dinilai yang paling cocok karena dalam algoritme ini, penghitungan operasi yang diperlukan untuk mengubah *string* a menjadi *string* b lebih lengkap yaitu empat operasi (penyisipan, penghapusan, penggantian, dan transposisi) (Vogler, 2013) berbeda dengan *Hamming Distance* yang hanya menghitung operasi substitusi dan LCS hanya operasi penyisipan dan penghapusan. Dalam percakapan *chatbot*, waktu respon *chatbot* juga sangat penting, maka dalam proses normalisasi dibutuhkan algoritme yang memiliki waktu pemrosesan cepat. Permasalahan muncul saat proses perbaikan kesalahan pengetikan, karena terdapat proses membandingkan dengan data korpus kata baku yang besar dan cukup memakan waktu. Untuk mengatasi permasalahan tersebut maka data korpus kata baku akan dibuat dalam bentuk Prefix Tree. Prefix Tree atau *Trie* adalah

struktur data pohon terurut yang menggunakan *string* sebagai kunci. Dalam penelitian Philip Khripkov pada tahun 2019, menghasilkan kesimpulan bahwa penggunaan Prefix Tree dapat mempercepat proses perbandingan *string* dengan 830 ribu kata *data set* sebanyak tiga kali lebih cepat dibandingkan dengan bentuk *data set* biasa. Dengan Prefix Tree maka struktur pengambilan data informasi yang efisien yang dapat kita gunakan untuk mencari kata dalam waktu O(M), di mana M adalah panjang *string* maksimum. Dengan Prefix Tree maka dapat menghindari banyak pekerjaan karena dapat memproses kata-kata secara berurutan, jadi tidak perlu mengulangi baris untuk awalan huruf yang sama. (Griffo, 2017). Untuk meningkatkan hasil perbaikan kata tidak baku, maka akan dilakukan penghitungan jarak kedekatan antara masing-masing huruf yang berbeda dari kata dalam percakapan *chatbot* dan kata dari hasil penghitungan jarak *string* data korpus. Penghitungan dilakukan dengan melihat posisi huruf dalam *layout* papan ketik QWERTY. Perhitungan akan menggunakan rumus *Pythagoras* dan hasil akhir akan dipilih adalah kata yang memiliki nilai terkecil.

Oleh karena itu, untuk mengatasi permasalahan di atas, yaitu kesalahan pengetikan dan bentuk kata tidak baku dalam pesan percakapan pengguna *chatbot*, penulis melakukan penelitian "Normalisasi Teks Pada *Chatbot* Sistem Informasi Akademik Menggunakan Algoritme Damerau–Levenshtein Distance dan Prefix Tree (Studi Kasus: Universitas Teknokrat Indonesia)" di mana nantinya diharapkan dapat menyelesaikan permasalahan dan

meningkatkan *chatbot* dalam memahami pesan dan merespons dengan tepat pesan pengguna.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah penulis uraikan di atas, maka penulis merumuskan masalah sebagai berikut:

1. Bagaimana melakukan normalisasi teks pada pesan pengguna *chatbot*?
2. Bagaimana mendeteksi dan memperbaiki kata tidak baku dan salah pengetikan dengan algoritme jarak *string* Damerau–Levenshtein Distance?
3. Bagaimana mengubah bentuk data korpus dalam bentuk Prefix Tree*?*

1.3. Batasan Masalah

Batasan masalah digunakan supaya pembahasan tidak keluar dari jalur yang dibuat, maka batasan masalah yang terkait dalam penelitian ini sebagai berikut:

1. Penelitian ini hanya berfokus pada proses normalisasi teks, deteksi dan perbaikan kata tidak baku dan salah pengetikan.
2. Penelitian ini menggunakan data pesan berbahasa Indonesia.
3. Bahasa pemrograman yang digunakan adalah *Python* sebagai *backend* dari normalisasi teks dan PHP (*Hypertext Preprocessor*) sebagai *frontend*-nya.

1.4. Tujuan Penelitian

Tujuan yang ingin dicapai dari hasil penelitian ini adalah sebagai berikut:

1. Melakukan normalisasi teks pada pesan pengguna *chatbot*.
2. Mendeteksi dan memperbaiki kata tidak baku dan salah pengetikan dengan algoritme jarak *string* Damerau–Levenshtein Distance.
3. Mengubah bentuk data korpus dalam bentuk Prefix Tree.

1.5. Manfaat Penelitian

Manfaat yang diharapkan dari hasil penelitian ini adalah:

1. Menambah pengetahuan dan meningkatkan keahlian mahasiswa atau pihak-pihak lain yang memiliki antusias dan minat dalam bidang *chatbot* dan pengolahan bahasa alami.
2. Bagi penulis dapat meningkatkan pemahaman berpikir ilmiah dan logis dalam menganalisis masalah khususnya terkait dalam bidang *chatbot* dan pengolahan bahasa alami.
3. Bagi Universitas diharapkan dapat menyelesaikan permasalahan pada *chatbot* sistem informasi akademik dalam pengolahan dan pemahaman pesan sehingga dapat meningkatkan kualitas *chatbot* itu sendiri.

1.6. Sistematika Penulisan

Dalam penulisan laporan tugas akhir, masing-masing bagian terdiri dari sub-sub bagian, yaitu sebagai berikut:

BAB I PENDAHULUAN

Bab ini menguraikan tentang latar belakang masalah, rumusan masalah, batasan masalah, tujuan penulisan penelitian, manfaat penelitian, dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Bab ini menjelaskan teori-teori tentang hal-hal yang berhubungan dengan penelitian yang akan dibahas yaitu normalisasi teks, algoritme jarak *string*, algoritme *Dameau-Lavenshtein distance* dan referensi-referensi tentang normalisasi teks yang akan digunakan.

BAB III METODE PENELITIAN

Bab ini membahas tentang tahapan penelitian, alat penelitian, metode pengumpulan data, analisis masalah, usulan penelitian, dan metode pengujian.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menjelaskan tentang hasil penelitian dan pembahasan normalisasi teks pada *chatbot* sistem informasi akademik menggunakan algoritme Damerau–Levenshtein Distance dan metode Prefix Tree.

BAB V SIMPULAN DAN SARAN

Bab ini berisikan tentang kesimpulan dan saran dari hasil penelitian.