BAB 1

PENDAHULUAN

1.1 Latar Belakang Masalah

Bahasa merupakan alat sekaligus kemampuan manusia untuk berkomunikasi antar satu sama lain. Informasi yang disampaikan melalui artikel, buku, koran, surat elektronik dan media tulis lainnya memerlukan penulisan bahasa yang baik, benar dan mudah dipahami. Penulisan kata yang salah atau keliru dapat menyebabkan penyampaian informasi yang salah dan ketidaknyamanan pembaca saat membaca suatu tulisan. Mayoritas media tulis yang ada di Indonesia merupakan tulisan tangan manusia yang tidak luput dari kesalahan penulisan. Kesalahan kata tidak hanya meliputi penulisan kata yang salah (*typography error*), tetapi juga penggunaan kata yang tidak sesuai dengan konteks kalimat (Fahma et al., 2018).

Penelitian serupa pernah dilakukan pada tahun 2017 dengan menggunakan komplain pasien rumah sakit sebagai *dataset* (Rosita et al., 2017). Teks yang dimasukkan akan diproses terlebih dahulu dengan mendeteksi kalimat, tokenisasi dan mengonversikan kata singkatan menjadi kata baku. Penulisan setiap kata yang diproses pada tahap sebelumnya akan dicek dengan metode *dictionary lookup* di mana setiap kata yang tidak terdapat pada kamus akan dianggap sebagai kesalahan eja. Kosa kata dalam kamus didapatkan dari para dokter. Tahap selanjutnya adalah mencari kandidat kata yang berpotensi menjadi solusi untuk menggantikan kata dengan kesalahan pengejaan. Metode yang digunakan adalah dengan mengukur *levenshtein distance* dari kata yang salah dengan setiap kata yang terdapat pada kamus. Kata pertama dengan nilai *levenshtein distance* terkecil akan diusulkan

sebagai pengganti kata tersebut. Perhitungan akurasi dilakukan dengan mengambil rata-rata akurasi perbaikan kata dari penambahan, pengurangan dan substitusi huruf. Berdasarkan 55 komplain pasien yang mengandung kesalahan ejaan, metode *dictionary lookup* dan *levenshtein distance* yang digunakan pada penelitian ini menghasilkan rata-rata akurasi sebesar 97.59% untuk akurasi dalam mendeteksi kesalahan kata dengan *dictionary lookup* dan 94.03% untuk akurasi dalam mengoreksi kesalahan kata dengan *levenshtein distance*.

Penelitian serupa juga dilakukan pada tahun 2019 pada sistem *spell checker* dan *correction* namun dengan metode yang berbeda (Wang dan Zhao, 2019). Metode yang digunakan pada penelitian ini antara lain *word embedding* berdasarkan korpus dari berbagai sumber dan *multi-candidate ranking model* dengan berbagai tingkatan dari *word corrector, word breaking, word concatenation* hingga *unit word corrector*. Penelitian tersebut melakukan perbandingan akurasi *spell correction* menggunakan frekuensi kata unigram dan bigram sebagai *ranking model* dan *word embedding based*. Berdasarkan *dataset* berjumlah kurang lebih 2 juta kata, *word embedding based ranking model* menghasilkan akurasi yang lebih tinggi yaitu pada angka 80.32% dibandingkan hanya dengan menggunakan frekuensi kata unigram dan bigram dalam mengoreksi berbagai kategori kesalahan ejaan yang disebut.

Penelitian lain mengenai *spell correction* juga dilakukan pada tahun 2020 menggunakan metode *levenshtein distance* dan *double metaphone* untuk menghasilkan daftar rekomendasi koreksi kata serta model *Global Vectors for Word Representation* (GloVe) untuk menyortir dan memilih rekomendasi kata yang benar dari daftar tersebut (Huang et al., 2020). Penelitian ini juga melakukan

perbandingan metodenya dengan metode JaSpell4 dan HunspellJNA5 yang merupakan implementasi *spell correction* berbasis bahasa pemrograman Java. Metode yang digunakan pada penelitian tersebut menghasilkan tingkat akurasi dan evaluasi F1-*score* yang lebih tinggi dibandingkan dua metode lainnya yaitu pada angka 86% untuk tingkat akurasi dan 84% untuk nilai F1-*score*.

Adapun penelitian lainnya yang berfokus pada reduksi lingkup pencarian pada sistem *spell correction* menggunakan metode *character neural embeddings* (Pande, 2017). Dalam penelitian tersebut seluruh kata pada korpus direpresentasikan ke dalam *n* dimensi vektor di mana setiap kata yang memiliki jarak *edit* yang kecil akan berada pada *cluster* yang sama. Untuk membentuk *cluster*, digunakan algoritma *Ball Tree* yang mampu mendapatkan *k-nearest-neighbors* dalam kompleksitas waktu logaritmik terhadap jumlah kata pada korpus. Dilakukan juga uji coba dengan parameter *n* sebesar 100 dan *k* sebesar 5000 pada korpus yang memiliki 100000 kata. Rata-rata akurasi *spell correction* berada pada angka 88.2% dengan rata-rata waktu sebesar 52 ms untuk setiap kasus uji. Hasil ini membuktikan bahwa metode reduksi lingkup pencarian ini memungkinkan sistem *spell correction* untuk digunakan secara *real time*.

Berdasarkan penelitian-penelitian serupa yang pernah dilakukan, diperlukan korpus berisi daftar kata benar atau sah beserta bobot setiap kata tersebut dan algoritma yang mampu mendeteksi dan mengoreksi kata *typo* agar dapat mencapai tingkat akurasi deteksi dan koreksi yang tinggi. Bobot setiap kata yang digunakan pada penelitian ini berupa frekuensi kemunculan kata pada proses pengumpulan data. Algoritma *levenshtein* merupakan metode yang sangat baik dalam menghitung jarak *edit* antara dua buah kata. Namun waktu yang diperlukan

oleh algoritma ini berbanding lurus dengan jumlah kata pada korpus dan kata yang perlu dikoreksi. Oleh karena itu, metode *Symspell* digunakan sebagai pengganti algoritma *levenshtein* pada penelitian ini karena *Symspell* dapat menghasilkan daftar kandidat kata pengganti berdasarkan jarak *edit* mencapai 1000 kali lebih cepat dibandingkan *levenshtein* (Garbe, 2012). *Symspell* bekerja lebih cepat karena hanya melakukan operasi penghapusan huruf saja untuk membatasi jarak *edit* pada kandidat kata koreksi. Namun daftar rekomendasi koreksi kata yang dihasilkan belum tentu benar dan tepat karena *threshold* jarak *edit* pada metode *Symspell* sangat rendah. Metode *Symspell* juga digunakan pada penelitian serupa dalam mengoreksi kesalahan ejaan pada hasil *speech recognition* (Ljunglöf dan Kjellberg, 2018). Penelitian tersebut menggunakan metode *Symspell* sebagai salah satu metode perbandingan dalam menyeleksi daftar kandidat kata koreksi yang tepat. Pada penelitian tersebut, metode *Symspell* memiliki tingkat akurasi sebesar 84% dalam mengoreksi kata berbahasa Inggris.

Untuk memilih kata yang tepat untuk mengoreksi kata yang salah diperlukan model yang mampu menyortir dan menyeleksi daftar rekomendasi kata dari berbagai aspek. Model *Bayesian Network* digunakan pada penelitian ini sebagai *ranking model* untuk menentukan kata pengganti yang paling tepat untuk menggantikan setiap kata yang salah pada dokumen. *Bayesian network* digunakan untuk melakukan komputasi probabilitas gabungan munculnya setiap kata pada suatu kalimat, termasuk kata kandidat koreksi. Relasi antar kata pada kalimat akan dipertimbangkan menggunakan frekuensi bigram dari setiap kata dengan kata yang berada di depan dan di belakangnya. *Bayesian Network* juga dapat diaplikasikan untuk mengukur estimasi tingkat kenyamanan pengguna saat menggunakan sebuah

website atau aplikasi (Ron dan Kenett, 2012). Dengan mengumpulkan cukup data mengenai tingkat kepuasan pengguna terhadap waktu yang mereka habiskan selama menggunakan aplikasi, waktu yang diperlukan aplikasi untuk melakukan navigasi halaman serta frekuensi aplikasi digunakan selama satu hari, model *Bayesian network* dapat membangun relasi antar variabel yang ada untuk mengestimasi kepuasan pengguna dalam menggunakan aplikasi tersebut. Penelitian ini dilakukan dengan tujuan membangun aplikasi yang mampu mendeteksi sekaligus mengoreksi pengejaan kata bahasa Indonesia yang salah.

1.2 Rumusan Masalah

Berdasarkan penjelasan latar belakang masalah, didapatkan beberapa rumusan masalah di bawah ini:

a. Bagaimana cara mengimplementasikan algoritma *Symspell* dan *Bayesian Network* dalam mendeteksi dan memperbaiki kesalahan penulisan kata pada dokumen?
b. Bagaimana tingkat akurasi sistem *spell checker* tanpa atau dengan metode *Bayesian Network* dalam mendeteksi dan memperbaiki kesalahan penulisan kata pada dokumen?

1.3 Batasan Masalah

Berikut batasan-batasan penelitian yang dilakukan:

a. Bahasa kata dan bacaan yang dijadikan uji coba pada penelitian ini merupakan bahasa Indonesia. Seluruh kata pada korpus merupakan kata yang sah menurut Kamus Besar Bahasa Indonesia (KBBI). Daftar kata pada

korpus dikumpulkan dari artikel berita daring pada situs berita Kompas.com dan Republika.co.id serta beberapa buku elektronik. Jumlah kata yang terkumpulkan adalah sebanyak 33.283 kata pada korpus unigram dan 445.328 pasangan kata pada korpus bigram.

b. Sistem *spell checker* tidak dapat memperbaiki kesalahan tanda baca pada kalimat. Sistem akan melewati pengecekan kata yang mengandung karakter di luar karakter alfabet arabik. Sistem juga akan melewati pengecekan kata yang merupakan kata entitas atau singkatan. Setiap kandidat kata yang diusulkan oleh sistem terbatas oleh korpus yang telah dibentuk sebelumnya. Teks bacaan yang dijadikan masukan oleh sistem akan diubah menjadi huruf kecil untuk mempermudah proses pengecekan dan pengoreksian penulisan kata yang salah.
c. Data uji coba sistem *spell checker* diambil dari artikel daring pada situs berita Kompas.com yang tidak digunakan pada proses perancangan korpus.
d. Untuk kebutuhan demonstrasi akan dibuat sebuah sistem *spell checker* berbasis aplikasi *desktop* sebagai salah satu hasil penelitian. Sistem dibuat dalam *platform desktop* dikarenakan dapat menyimpan korpus secara local sehingga mempercepat proses impor korpus ke dalam aplikasi dan dapat digunakan tanpa memerlukan akses internet. Pengukuran aspek usabilitas dan tampilan antarmuka sistem tidak akan dilakukan dikarenakan hal tersebut tidak menjadi fokus utama dalam penelitian.

1.4 Tujuan Penelitian

Tujuan penelitian ini adalah:

a. Mengimplementasi algoritma *Symspell* dan *Bayesian Network* dalam mendeteksi dan memperbaiki kesalahan penulisan kata pada dokumen.
b. Mengukur tingkat akurasi sistem *spell checker* tanpa atau dengan metode *Bayesian Network* dalam mendeteksi dan memperbaiki penulisan kata pada dokumen.

1.5 Manfaat Penelitian

Manfaat penelitian ini adalah:

a. Bagi peneliti

   Menambah wawasan pengetahuan mengenai implementasi algoritma *Symspell* dan *Bayesian Network* dalam penentuan koreksi kata yang sesuai.

b. Bagi pengguna

   Dapat menggunakan hasil penelitian ini untuk meminimalisir kesalahan eja yang terdapat pada suatu bacaan.

c. Bagi ilmu pengetahuan

   Memperkaya hasil penelitian sejenis dengan menambah satu lagi penelitian dalam bidang pengambilan keputusan yang optimal dengan data dan spesifikasi yang berbeda.

1.6 Sistematika Penulisan

Sistematika penulisan yang digunakan dalam laporan skripsi ini terdiri dari BAB 1, BAB 2, BAB 3, BAB 4, dan BAB 5 di mana penjelasan setiap bab akan diuraikan sebagai berikut.

BAB 1 PENDAHULUAN

Bab pendahuluan terdiri dari enam bagian, yaitu latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, menfaat penelitian, dan sistematika penulisan laporan.

BAB 2 LANDASAN TEORI

Bab landasan teori terdiri dari empat teori yang digunakan dalam penelitian, yaitu *typography error*, algoritma *Symspell*, model *Bayesian Network* (*N*-gram), dan optimisasi *dynamic programming*.

BAB 3 METODOLOGI PENELITIAN DAN PERANCANGAN SISTEM

Bab metodologi penelitian dan perancangan sistem menjelaskan tentang metode penelitian yang digunakan, serta perancangan sistem yang terdiri dari perancangan korpus, model aplikasi, dan uji coba aplikasi menggunakan *flowchart*.

BAB 4 HASIL DAN DISKUSI

Bab hasil dan diskusi menjelaskan tentang implementasi algoritma *Symspell* dan model *Bayesian Network* dalam mendeteksi dan mengoreksi kesalahan penulisan kata pada teks bacaan, serta memaparkan perbandingan beberapa hasil koreksi menggunakan parameter model yang bervariasi.

BAB 5 SIMPULAN DAN SARAN

Bab simpulan dan saran berisi simpulan dari hasil penelitian dan eksperimen yang dilakukan dalam penelitian, serta saran yang dapat dilakukan untuk mengembangkan aplikasi maupun penelitian lebih lanjut.