BAB I
PENDAHULUAN

1.1 Latar Belakang

Teks atau bacaan merupakan suatu media yang digunakan sebagai sumber informasi maupun sumber pembelajaran bagi manusia. Dalam bidang pendidikan teks mempunyai peran yang sangat penting, sebagai seorang mahasiswa akan dituntut untuk menulis berbagai karya tulis seperti makalah, jurnal, paper dan skripsi. Teks yang baik dan layak untuk dipublikasikan adalah teks yang mengacu pada kaidah kebahasaan PUEBI (Pedoman Umum Ejaan Bahasa Indonesia) [1]. Dalam penulisannya, biasanya karya tulis dibuat dalam bentuk digital. Di dalam penulisan teks banyak hal yang harus diperhatikan seperti penggunaan Bahasa Indonesia yang harus sesuai dengan EYD (Ejaan Yang Disempurnakan).

Kesalahan pemakaian ejaan sering ditemukan dalam sebuah penulisan, penyebabnya antara lain penulis masih kurang paham mengenai ejaan, kurang terbiasa menggunakan ejaan, maupun faktor lingkungan penulis [2]. Kesalahan yang banyak dilakukan oleh penulis biasanya adalah kesalahan dalam pengetikan/penulisan kata, hal ini bisa menyebabkan suatu kata bermakna ambigu dan tidak terdefinisi di dalam KBBI (Kamus Besar Bahasa Indonesia). Kesalahan dalam pengetikan atau *typographical error* merupakan hal yang sering terjadi dalam penulisan. Maka dari itu untuk membantu pekerjaan penulis dalam melakukan editing pada suatu teks dibutuhkan suatu sistem pengimplementasian dalam mendeteksi dan mengoreksi kesalahan dalam pengetikan.

Dalam proses pengoreksian ejaan kata jika ditemukan ejaan kata yang kurang sesuai maka harus dilakukan koreksi dengan pencarian kemungkinan kata yang sesuai. Dalam proses pencarian kemungkinan kata tersebut, dibutuhkan

suatu pendekatan pencarian *string* yaitu menggunakan algoritma *string matching* dan *phonetic* [3]. Salah satu algoritma *string matching* adalah *Levenshtein Distance* dan salah satu algoritma *phonetic* adalah *Soundex*. *Levenshtein Distance* adalah algoritma *string matching* yang digunakan untuk menghitung jarak antara dua buah *string* [4]. Sedangkan *Soundex* adalah algoritma yang mengoreksi kesalahan ejaan berdasarkan bunyi[3]. Selain dua algoritma tersebut pengoreksian kesalahan pengetikan dapat diimplementasikan menggunakan algoritma *Recurrent Neural Network* (RNN) dan *FastText*. Algoritma RNN dapat mengoreksi kesalahan pengetikan dengan cara memprediksi kata yang salah pengetikannya. Sedangkan *FastText* mengoreksi kesalahan pengetikan berdasarkan kemiripan vektornya.

Beberapa peneliti telah melakukan penelitian mengenai sistem koreksi kesalahan dalam pengetikan dan memperbaiki agar menjadi kata terdefinisi menggunakan berbagai metode. Salah satunya Reza Fauzan, Joni Riadi, Fuad Sholihin [5] yaitu melakukan penelitian dengan membandingkan algoritma *Levenshtein* dan *Jaro Winkler*. Penelitian ini bertujuan untuk menentukan kemiripan kata. Hasil dari pengujian yang dilakukan, algoritma *Levenshtein* mendapatkan akurasi lebih besar dari *Jaro Winkler*. *Levenshtein* mendapat akurasi sebesar 46,15%, sedangkan *Jaro Winkler* mendapatkan akurasi sebesar 38,46%. Dari penelitian ini diketahui bahwa kedua algoritma tersebut hanya melihat dari huruf yang ada di dalam kata tersebut sehingga kata-kata yang dianggap mirip adalah kata yang memiliki tulisan hampir sama. Sedangkan kata yang memiliki tulisan berbeda namun maknanya sama tidak dapat diselesaikan menggunakan algoritma ini.

Berdasarkan penelitian Abson Hadi[3], penelitian ini dilakukan untuk membandingkan algoritma *Damerau-Levenshtein Distance* dan *Soundex Similarity* untuk mengoreksi kesalahan ejaan kata secara otomatis. Algoritma *Damerau-Levenshtein Distance* melakukan koreksi kesalahan ejaan kata berdasarkan kecocokan kata, sedangkan *Soundex Similarity* mengoreksi

kesalahan ejaan berdasarkan bunyi. Pengujian dalam penelitian ini dilakukan dengan memberikan 50 kata dengan kesalahan ejaan untuk dikoreksi secara otomatis menggunakan kedua algoritma tersebut. Hasil dari penelitian ini menunjukkan bahwa *Damerau-Levenshtein Distance* berhasil mendapat akurasi sebesar 72% dengan keberhasilan koreksi kata sebanyak 36 kata dan *Soundex Similarity* berhasil mendapatkan akurasi sebesar 68% dengan keberhasilan koreksi kata sebanyak 34 kata pada pengoreksian ejaan kata secara otomatis.

Dari penelitian Fendy Augusfian, dkk[6] algoritma RNN dapat diimplementasikan untuk melakukan pengoreksian kesalahan ketik. Penelitian ini menggunakan algoritma *Damerau Levenshtein Distance* dan *Recurrent Neural Network*. Algoritma *Damerau Levenshtein Distance* digunakan untuk menghitung perubahan yang terjadi pada kata yang diperiksa dengan kata yang sesungguhnya dan memperbaikinya. Sedangkan algoritma *Recurrent Neural Network* digunakan untuk menghasilkan perbaikan yang sesuai dengan data yang sudah latih disebelumnya. Hasil akurasi dari *Damerau-Levenshtein Distance* dan *Recurrent Neural Network* menghasilkan akurasi kata sebesar 21,3%. Hasil pengujian ulang metode koreksi gabungan *Damerau Levenshtein Distance* dan *Recurrent Neural Network* memiliki akurasi kata rata-rata sebesar 74%.

Penelitian telah dilakukan oleh Ahmad Arif Samudro yaitu melakukan normalisasi teks pada media sosial. Tujuan dari penelitian ini adalah menormalisasikan teks dari bentuk yang tidak baku menjadi bentuk yang baku. Penelitian ini menggunakan algoritma *Levenshtein Distance, Jaro Winkler Distance, Fasttext,* dan *Word2Vec.* Dengan menggunakan algoritma *FastText* akan didapatkan rekomendasi kata yang memiliki probabilitas kata terbesar dengan input yang diberikan. Pada penelitian ini dilakukan perbandingan antara algoritma *FastText,* dan *Word2Vec.* Hasil akurasi model *FastText* tertinggi adalah 45,60%, sedangkan *Word2Vec* adalah 44,60%. Selain itu untuk normalisasi dengan pengecekan kamus metode *Levenshtein Distance* dan *Jaro Winkler Distance* mendapatkan akurasi sebesar 80,76%.

Meskipun penelitian terkait telah dilakukan sebelumnya, namun perbandingan akurasi empat metode dalam mengoreksi kesalahan pengetikan sesuai konteks kalimat belum pernah dilakukan sebelumnya, sehingga belum diketahui jelas algoritma mana yang lebih baik. Selain itu dalam penelitian sebelumnya belum ada yang membandingkan kecepatan pemrosesan algoritma untuk mengetahui kecepatan pemrosesan algoritma terbaik. Maka dari itu, dilakukan penelitian ini dengan tujuan untuk melakukan perbandingan akurasi dan kecepatan pemrosesan antara algoritma *Levenshtein Distance, Soundex, Recurrent Neural Network,* dan *FastText* dalam melakukan koreksi ejaan kata sesuai dengan konteks kalimat.

1.2 Rumusan Masalah

Berdasarkan latar belakang tersebut, maka perumusan masalah dari penelitian ini adalah :

1. Bagaimana hasil perbandingan akurasi dari model *Levenstein Distance, Soundex, FastText, dan Recurrent Neural Network?*
2. Bagaimana tingkat akurasi training dan testing RNN dalam menentukan rekomendasi kata selanjutnya?
3. Bagaimana kecepatan pemrosesan menggunakan *Levenstein Distance, Soundex, FastText, dan Recurrent Neural Network?*

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah :

1. Mengetahui hasil perbandingan akurasi model terbaik dari algoritma *Levenshtein Distance, Soundex, FastText,* dan *Recurrent Neural Network.*
2. Mengetahui tingkat akurasi *training* dan *testing* algoritma *Recurrent Neural Network.*

3. Mengetahui hasil perbandingan kecepatan pemrosesan terbaik dari algoritma *Levenshtein Distance, Soundex, FastText,* dan *Recurrent Neural Network.*

1.4 Batasan Masalah

Batasan masalah yang ada pada penelitian ini adalah :

1. Hanya dapat melakukan koreksi pada kata berbahasa Indonesia.
2. Hanya bahasa selevel novel, kurang lebih 7000 kata.
3. Model dan sistem hanya dapat mengkoreksi satu *typo* dalam satu kalimat.
4. Model dapat memperbaiki kata yang salah dengan konteks kalimat menjadi kata yang tepat dengan konteks kalimat.
5. *Database* kata yang digunakan adalah Kamus Besar Bahasa Indonesia (KBBI).

1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah :

1. Model yang dibuat dapat mengidentifikasi dan memperbaiki kesalahan pengetikan sesuai dengan konteks kalimat.
2. Hasil dari penelitian ini dapat digunakan sebagai acuan atau referensi dalam pengembangan penelitian selanjutnya yang berhubungan dengan koreksi kesalahan penulisan kata.