

BAB I

PENDAHULUAN

1.1 Latar Belakang

Indonesia terdiri dari berbagai macam pulau, budaya serta suku bangsa sehingga bahasa daerah yang dimiliki juga bervariasi. Bahasa Madura merupakan salah satu bahasa daerah yang ada di Indonesia. Bahasa Madura menduduki posisi keempat sebagai bahasa daerah terbesar yang ada di Indonesia [1]. Bahasa Madura banyak digunakan sebagai sarana komunikasi dalam kehidupan sehari-hari oleh masyarakat Madura, baik yang bertempat tinggal di dalam pulau Madura maupun di luar pulau Madura. Selain itu, bahasa Madura juga digunakan sebagai sarana pendukung sastra dan budaya serta bahasa penghubung intradaerah atau intramasyarakat.

Dengan beranekaragamnya bahasa yang ada di negara Indonesia, dapat menyebabkan terjadinya interferensi bahasa. Interferensi bahasa adalah peristiwa terjadinya penggunaan dua bahasa ketika melakukan komunikasi [2]. Dialek adalah ragam tutur bahasa yang berbeda berdasarkan faktor geografis [3]. Adanya interferensi bahasa bisa berdampak pada sulitnya dalam mempelajari dialek dan cara penulisan ejaan kata yang baik dan benar suatu bahasa. Hal ini khususnya berlaku bagi bahasa daerah dengan dialek dan cara penulisan ejaan yang unik seperti bahasa Madura [4].

Pada umumnya, bahasa Madura dan bahasa Indonesia memiliki perbedaan yang sangat besar dari segi kosakata dan pelafalan [5]. Salah satu contohnya seperti kata “dari” yang berasal dari bahasa Indonesia, dibandingkan dengan kata “*dâri*” yang berasal dari bahasa Madura. Dua kata tersebut memiliki makna yang sama akan tetapi penulisan ejaan serta pelafalannya berbeda. Perbedaannya terletak pada huruf “a” dan “â” dari masing-masing kata. Kejadian seperti inilah yang dapat menyebabkan terjadinya kesalahan penulisan pada ejaan kata bahasa Madura. Selain itu, kesalahan penulisan ejaan kata juga dapat diakibatkan dari tombol huruf yang berdekatan pada perangkat digital sehingga pengguna perangkat digital sering kali salah dalam memencet tombol huruf ketika sedang mengetik.

Pada tahun 2019, penelitian perbaikan ejaan kata dilakukan oleh Santoso dkk. Algoritma yang digunakan yaitu *Damerau Levenshtein Distance* yang dibandingkan dengan algoritma *Levenshtein Distance*. Kedua algoritma tersebut nantinya diujikan terhadap ejaan kata bahasa Indonesia yang salah. Kelebihan dari kedua algoritma ini adalah nilai akurasi perbaikan ejaan kata yang diperoleh algoritma *Damerau Levenshtein Distance* sebesar 75%, sedangkan algoritma *Levenshtein Distance* sebesar 73%. Kekurangan dari kedua algoritma ini yaitu waktu proses yang dibutuhkan lama, saran kata yang ditampilkan terlalu banyak dan tidak dapat mengoreksi kesalahan kata yang berdempetan [6].

Masih pada tahun 2019, dilakukan penelitian lain terkait perbaikan ejaan kata oleh Augustian dkk. Penelitian ini berfokus pada kombinasi algoritma *Damerau Levenshtein Distance* dan *Recurrent Neural Network*. Sistem diujikan terhadap ejaan kata bahasa Indonesia yang salah. Kekurangan dari sistem pada penelitian ini yaitu saran kata yang dihasilkan memiliki nilai rata-rata akurasi awal sebesar 21,3% dengan rata-rata waktu proses sebesar 29,2 s. Sedangkan kelebihan yang dimiliki sistem adalah setelah dilakukan pengujian ulang dapat menghasilkan saran kata dengan nilai rata-rata akurasi sebesar 74% [7].

Kemudian pada tahun 2020, penelitian perbaikan ejaan kata dilakukan dengan menggunakan algoritma *Damerau Levenshtein Distance* dan metode Empiris oleh Wibawa dkk. Kombinasi kedua algoritma tersebut nantinya diujikan terhadap ejaan kata bahasa Indonesia yang salah. Kelebihan yang diperoleh dari kombinasi algoritma ini berupa saran perbaikan kata dengan nilai akurasi sebesar 97%. Kekurangan dari kombinasi algoritma ini yaitu saran kata yang ditampilkan terlalu sedikit [8].

Kemudian Penelitian lain terkait perbaikan ejaan kata dilanjutkan pada tahun 2021 oleh Erwina dkk. Metode yang digunakan yaitu kombinasi *Bigram Vector* dan *Minimum Edit Distance Based Probabilities* yang nantinya diujikan terhadap ejaan kata bahasa Indonesia yang salah. Kelebihan yang diperoleh dari kombinasi metode ini berupa saran perbaikan kata dengan nilai akurasi sebesar 92%. Kekurangan yang dimiliki oleh kombinasi metode ini yaitu tidak dapat mengoreksi 2 kata yang berdempetan akibat tidak adanya spasi [9].

Penelitian mengenai perbaikan ejaan kata selanjutnya dilakukan pada tahun 2022 oleh Sujaini dkk. Metode yang digunakan dalam penelitian yaitu kombinasi antara metode *N-Gram* dan *Jaro-Winkler*. Kombinasi kedua metode tersebut nantinya diujikan terhadap ejaan kata bahasa Indonesia yang salah. Kombinasi metode ini memiliki kelebihan berupa saran kata yang dihasilkan memiliki nilai *precision* sebesar 71,348% dan nilai *recall* sebesar 98,449%. Sedangkan kekurangan yang dimiliki oleh kombinasi metode ini yaitu tidak dapat mengkoreksi 2 kata yang berdempetan akibat tidak adanya spasi [10].

Dari beberapa penelitian yang telah dilakukan sebelumnya maka pada penelitian ini mengimplementasikan program perbaikan ejaan kata bahasa Madura dengan menggunakan algoritma *Damerau Levenshtein Distance* dan metode Empiris. Dengan penggunaan kombinasi kedua algoritma tersebut maka diharapkan program yang dibuat mampu memperbaiki berbagai kesalahan ejaan kata dengan tingkat akurasi yang tinggi seperti penelitian sebelumnya [8]. *Dataset* yang digunakan yaitu kata dasar dalam kamus bahasa Madura sebagai acuan dalam memperbaiki ejaan kata bahasa Madura yang salah. Sedangkan *input* dari *user* berupa teks bahasa Madura dengan ejaan yang salah.

1.2 Perumusan Masalah

1.2.1 Permasalahan

Kesalahan penulisan ejaan kata bahasa Madura sering kali terjadi akibat adanya interferensi bahasa yang mengakibatkan pemahaman dalam menuliskan ejaan kata menjadi berkurang. Contohnya seperti kata “dari” yang berasal dari bahasa Indonesia, dibandingkan dengan kata “*dâri*” yang berasal dari bahasa Madura dimana dua kata tersebut memiliki makna yang sama akan tetapi penulisan ejaan serta pelafalannya berbeda.

Selain itu, kesalahan penulisan ejaan kata bahasa Madura juga dapat diakibatkan dari tombol huruf yang berdekatan pada perangkat digital sehingga pengguna perangkat digital sering kali salah dalam memencet tombol huruf. Kesalahan penulisan ejaan kata bahasa Madura dapat berupa hilangnya spasi, kekurangan karakter huruf, kelebihan karakter huruf, kesalahan urutan huruf, dan kesalahan memasukkan huruf [8].

1.2.2 Solusi Permasalahan

Untuk mengatasi kesalahan penulisan ejaan kata bahasa Madura maka algoritma yang digunakan adalah algoritma *Damerau Levenshtein Distance* yang dikombinasikan dengan metode Empiris sehingga dapat memperbaiki berbagai macam kesalahan ejaan.

1.2.3 Pertanyaan Penelitian

1. Berapa besar akurasi algoritma *Damerau Levenshtein Distance* dalam mendeteksi saran perbaikan ejaan kata bahasa Madura yang baik dan benar?
2. Bagaimana pengaruh metode Empiris terhadap nilai akurasi *Damerau Levenshtein Distance* dalam mendeteksi saran perbaikan ejaan kata bahasa Madura yang baik dan benar?

1.3 Tujuan dan Manfaat

1.3.1 Tujuan Penelitian

1. Mengetahui besar akurasi algoritma *Damerau Levenshtein Distance* dalam mendeteksi saran perbaikan ejaan kata bahasa Madura yang baik dan benar.
2. Mengetahui pengaruh metode Empiris terhadap nilai akurasi program perbaikan ejaan kata bahasa Madura.

1.3.2 Manfaat

Penerapan algoritma *Damerau Levenshtein Distance* dan metode Empiris dapat digunakan untuk melakukan perbaikan ejaan kata bahasa Madura sehingga pada proses selanjutnya mampu membantu pengembangan NLP untuk mendeteksi dan memberikan saran terhadap kesalahan ejaan kata pada sistem pencarian.

1.4 Batasan-Batasan

1. *Dataset* kamus bahasa Madura dibuat oleh peneliti dari Universitas Trunojoyo Madura yang dapat diakses melalui Mendeley [11].
2. *Dataset* yang digunakan berfokus pada table *lemmata*.

3. *Input* tidak mengandung kata berimbuhan, nama orang, nama metode dan istilah nama lainnya.
4. Batas maksimal *edit distance* dari saran perbaikan kata yang ditampilkan adalah 50% dari jumlah huruf pada *input user*.

1.5 Sistematika Laporan

Berikut ini sistematika penulisan laporan Tugas Akhir:

- BAB I PENDAHULUAN

Pada bab ini dibahas mengenai latar belakang penelitian, perumusan masalah, tujuan dan manfaat dilakukannya penelitian, batasan-batasan permasalahan dalam penelitian serta sistematika penulisan laporan.

- BAB II KAJIAN PUSTAKA

Pada bab ini dibahas mengenai teori terkait sistem pengecek ejaan (*spelling checker*) yang dibuat, algoritma yang digunakan serta penelitian terkait yang telah dilakukan sebelumnya sehingga dapat dijadikan sebagai landasan dalam melakukan penelitian.

- BAB III METODE USULAN

Pada bab ini dibahas mengenai *dataset* yang digunakan, arsitektur sistem dan alur dari algoritma yang diterapkan, serta skenario uji coba pada program yang diimplementasikan.

- BAB IV HASIL DAN PEMBAHASAN

Pada bab ini dibahas mengenai implementasi sistem perbaikan ejaan kata bahasa Madura serta hasil pengujian system dengan menggunakan algoritma *Damerau Levenshtein Distance* dan metode Empiris

- BAB V PENUTUP

Pada bab ini dibahas mengenai kesimpulan dari sistem yang telah diimplementasikan serta saran agar sistem dapat lebih dikembangkan pada penelitian berikutnya.