**Cyberspace**: Jurnal Pendidikan Teknologi informasi
Volume 6, Nomor 2, Oktober 2022, hal. 130-141
ISSN 2598-2079 (print) | ISSN 2597-9671 (online)

# DETEKSI KATA TAK BAKU DAN KESALAHAN PENULISAN KATA PADA TUGAS AKHIR MAHASISWA MENGGUNAKAN METODE *DICTIONARY LOOKUP*

***Alim Misbullah[1], Viska Mutiawani[1], dan Cut Sri Mulyani[1]**

[1]Jurusan Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Syiah Kuala, Darussalam, Banda Aceh, 23111, Indonesia
E-mail: misbullah@unsyiah.ac.id, viska.mw@unsyiah.ac.id,
cut.sri@s1.informatika.unsyiah.ac.id

**Abstract**

Dictionary lookup is a simple searching method that aimed to find a word in the dictionary. The technique is effectively implemented to find the incorrect spelling based on the lexical resource. This research will create a system to detect non-standard and misspell words by looking into Standard Indonesian Language Dictionary (KBBI) and Kateglo Dictionary. In addition, the lookup process of the system is optimized by utilizing an index in the database. To validate the system, the front-end interface is created by displaying the options for end-user to upload their documents. The system will be evaluated by using 40 documents of the students' final project. The testing result shows that the average document processing speed is 187 words per second. The average of incorrect words in each document is 2.76%, with a precision value is 28.71%. The low precision is caused by the higher value of False Positive, which was determined by non-existence words in the dictionary. However, the system has shown good performance by obtaining higher True Negative values from each document which implies the accuracy is also high.

**Keywords:** *dictionary lookup, word index, database, speed, accuracy*

**Abstrak**

*Dictionary lookup* merupakan sebuah metode pencarian sederhana yang bertujuan untuk menemukan kata di dalam sebuah kamus. Metode ini secara efektif dapat diimplementasikan untuk menemukan kesalahan penulisan kata berdasarkan sumber leksikal. Penelitian ini dilakukan untuk membangun sebuah sistem yang dapat mendeteksi kata yang tidak standar dan kesalahan penulisan kata dengan melakukan pencarian kata tersebut di dalam Kamus Besar Berbahasa Indonesia (KBBI) dan Kamus Kateglo. Selain itu, process pencarian kata di dalam kamus dioptimisasi dengan mengutilisasi *index* yang ada pada *database*. Validasi sistem dilakukan dengan membangun sistem antarmuka yang menampilan opsi kepada pengguna untuk mengunggah dokumennya. Sistem dievaluasi menggunakan 40 dokumen tugas akhir mahasiswa. Hasil uji menunjukkan bahwa rata-rata dokumen dapat diproses dengan kecepatan 187 kata tiap detik. Rata-rata kesalahan kata yang terdapat di setiap dokumen adalah 2,76% dengan nilai presisi adalah 28,71%. Nilai presisi yang rendah disebabkan oleh tingginya nilai *False Positive* yang ditentukan

oleh tidak adanya kata tersebut di dalam kamus. Namun, sistem yang dibangun sudah memperoleh nilai *True Negative* yang tinggi sehingga akurasi juga menjadi tinggi.

**Kata Kunci:** *dictionary lookup, index* kata*, database, kecepatan, akurasi*

## 1. Pendahuluan

Bahasa Indonesia merupakan bahasa resmi yang digunakan di Indonesia. Pada penulisan karya tulis ilmiah seperti skripsi, jurnal, tesis, disertasi, dan beberapa karya sejenis lainnya, semua harus ditulis dengan Bahasa Indonesia ragam baku. Penggunaan bahasa baku pada tulisan-tulisan dimaksud akan membuat para pembaca lebih mudah memahami konteks yang dimaksud oleh penulis dan memperkecil kemungkinan salah tafsir.

Kualitas sebuah karya tulis ilmiah tidak hanya dilihat dari isi dan ide atau pemikiran penulis, tetapi juga penggunaan bahasa dalam penulisan. Salah satu penggunaan Bahasa Indonesia pada karya tulis ilmiah adalah penulisan skripsi tugas akhir yang diwajibkan oleh setiap universitas untuk jenjang sarjana. Setiap universitas menerbitkan buku panduan yang menyebutkan bahwa Bahasa Indonesia yang digunakan dalam menulis proposal atau laporan tugas akhir adalah bahasa tulis yang baku. Penggunaan ejaan juga harus berpedoman pada ejaan Bahasa Indonesia yang telah disempurnakan [1]. Semua kata yang digunakan juga harus merujuk kepada Kamus Besar Bahasa Indonesia (KBBI).

Salah satu cara untuk mengetahui apakah kata yang digunakan dalam sebuah tulisan itu baku atau tidak adalah dengan merujuk ke kata-kata yang terdapat di dalam KBBI. Jika setiap kata dari tulisan yang diperiksa terdapat di dalam kamus, maka kata itu adalah layak disebut kata baku. Sebaliknya, jika kata tersebut tidak terdapat di dalam kamus maka kata tersebut kemungkinan bukan termasuk kata baku, atau merupakan istilah asing (bukan Bahasa Indonesia), atau merupakan kata dengan kesalahan tipografi (salah tik). Salah satu kesulitan yang akan dihadapi saat melakukan pengecekan yaitu adanya kata-kata yang merupakan istilah dari bahasa asing dikarenakan tulisan-tulisan yang berasal dari tema keilmuan matematika dan pengetahuan alam, khususnya bidang informatika atau ilmu komputer.

Berdasarkan hal di atas, maka penelitian ini dilakukan untuk memeriksa kebenaran tulisan dan pengetikan dari naskah tugas akhir mahasiswa Informatika yang saat ini telah menjadi alumni. Tahapan yang akan dilakukan nanti dimulai dengan membentuk dua jenis kamus, yaitu kamus Bahasa Indonesia yang berisi kosakata Bahasa Indonesia dari semua kelas kata dan kamus istilah asing. Kosakata Bahasa Indonesia didapat dari situs KBBI daring, sedangkan kumpulan istilah asing didapat dari situs kateglo.com yang dikembangkan oleh Ivan Lanin, Romi Hardiyanto, dan Arthur Purnama. Kamus inilah yang akan digunakan nantinya dalam memeriksa tulisan tugas akhir mahasiswa.

Salah satu penelitian terkait yang dilakukan oleh Hamzah [2] pada tahun 2016 mengungkapkan bahwa kesalahan penulisan ejaan kata umumnya disebabkan oleh dua hal, yaitu galat tipografi dan ketidaktahuan penulis terhadap kosakata baku Bahasa Indonesia. Galat tipografi biasanya terjadi karena kelalaian penulis yang tidak disengaja seperti fokus yang terganggu, slip tangan atau jari akibat kedekatan tombol huruf saat mengetik, atau sebab lainnya.

Metode yang digunakan dalam pemeriksaan naskah mahasiswa untuk menemukan kesalahan penulisan adalah metode *dictionary lookup*. Metode ini telah dilakukan di beberapa penelitian terkait deteksi kesalahan kata seperti yang dikaji oleh Maghfira [3] dan Hadi [4]. Menurut Maghfira [3] dalam artikel jurnalnya, metode *dictionary lookup* dinilai efektif dalam menentukan suatu ejaan kata bernilai benar atau salah berdasarkan

*Lexical Resource*. Metode ini tergolong sederhana apalagi untuk menentukan non-word error. Namun penelitian tersebut belum mempertimbangkan istilah dari bahasa asing yang digunakan dalam pengujiannya. Oleh karena itu, penelitian ini selain memeriksa kesalahan penulisan dengan metode *dictionary lookup*, juga akan mempertimbangkan istilah bahasa asing dalam pengujiannya.

## 2. Kajian Pustaka

### A. *Kata Baku dalam Bahasa Indonesia*

Dalam KBBI Edisi Keempat disebutkan pengertian baku adalah pokok, utama; tolok ukur yang berlaku untuk kuantitas dan kualitas yang ditetapkan berdasarkan kesepakatan; standar. Sementara menurut Kosasih dan Hermawan **[5]** kata baku adalah kata yang diucapkan atau ditulis oleh seseorang sesuai dengan kaidah atau pedoman yang dibakukan. Kaidah standar yang dimaksud dapat berupa pedoman ejaan yang disempurnakan (EYD), tata bahasa baku, dan kamus. Kata baku umumnya sering dipakai pada kalimat resmi atau ragam bahasa baku, baik itu melalui lisan ataupun tulisan. Kata baku dalam bahasa Indonesia ini juga memiliki beberapa ciri-cir. Pertama, baik secara lisan maupun tulisan, kata baku digunakan dalam situasi resmi, seperti surat menyurat dinas, perundang-undangan, karangan ilmiah, laporan penelitian dan lainnya. Ragam bahasa baku tidak diwarnai atau dicampuri oleh dialek atau logat tertentu. Kedua, baik secara lisan maupun tulisan, kata baku menggunakan ketentuan-ketentuan yang berlaku dalam Pedoman Umum Ejaan Bahasa Indonesia. Ketiga, baik secara lisan maupun tulisan, ragam baku memenuhi fungsi gramatikal seperti subjek, predikat, dan objek secara eksplisit dan lengkap [6].

Setiorini [7] dalam artikel jurnalnya mengatakan bahwa penggunaan bahasa akan berubah sesuai dengan kebutuhan penuturnya. Sebagai contoh, bahasa yang digunakan saat seseorang berpidato atau berceramah dalam sebuah seminar akan berbeda dengan bahasa yang digunakannya saat mengobrol atau bercengkerama dengan keluarganya. Bahasa itu akan berubah lagi saat ia menawar atau membeli sayuran di pasar. Kesesuaian antara bahasa dan pemakaiannya ini disebut ragam bahasa. Dalam penggunaan bahasa (Indonesia) dikenal berbagai macam ragam bahasa dengan pembagiannya masing-masing, seperti ragam formal-semi formal-nonformal; ujaran-tulisan; jurnalistik; iklan; populer dan ilmiah.

Menurut Hasan Alwi [8], ragam bahasa ini memiliki dua ciri, yaitu kemantapan dinamis dan kecendekiaan. Kemantapan dinamis berarti aturan dalam ragam bahasa ini telah berlaku dengan mantap, tetapi bahasa ini tetap terbuka terhadap perubahan (terutama dalam kosakata dan istilah). Ciri kecendekiaan terlihat dalam penataan penggunaan bahasa secara teratur, logis, dan masuk akal. Ragam bahasa ini bersifat kaku dan terikat pada aturan-aturan bahasa yang berlaku. Sebagai bahasa baku, terdapat standar tertentu yang harus dipenuhi dalam penggunaan ragam bahasa ilmiah. Standar tersebut meliputi penggunaan tata bahasa dan ejaan bahasa Indonesia baku. Tata bahasa Indonesia yang baku meliputi penggunaan kata, kalimat, dan paragraf yang sesuai dengan kaidah baku. Kaidah tata bahasa Indonesia yang baku adalah kaidah tata bahasa Indonesia yang sesuai dengan aturan berbahasa yang ditetapkan oleh Pusat Bahasa Indonesia. Sementara itu, kaidah ejaan bahasa Indonesia yang baku adalah kaidah ejaan bahasa Indonesia. Sesuai dengan ragam bahasanya, aturan-aturan ini mengikat penggunaan bahasa dalam karya tulis ilmiah.

Salah satu funsgi kata baku dalam Bahasa Indonesia adalah sebagai kerangka acuan [6]. Kata baku sebagai kerangka acuan artinya kata baku menjadi patokan bagi

benar atau tidaknya pemakaian bahasa seseorang atau kelompok. Namun terkadang banyak kata tak baku yang sangat sering ditemukan pada tulisan ilmiah seperti yang ditunjukkan pada Tabel 1.

Tabel 1. Contoh Kata Tak Baku dan Kata Baku

| No. | Kata Tak Baku | Kata Baku |
|---|---|---|
| 1 | Analisa | Analisis |
| 2 | Praktek | Praktik |
| 3 | Desaign | Desain |
| 4 | Aktip | Aktif |
| 5 | Diagnosa | Diagnosis |
| 6 | Detil | Detail |
| 7 | Efektifitas | Efektivitas |
| 8 | Hirarki | Hierarki |
| 9 | Hipotesa | Hipotesis |
| 10 | Katagori | Kategori |

B. *Typographical Error (Kesalahan Pengetikan)*

Ketika membuat karya tulis ilmiah, sering sekali terjadi kesalahan dalam hal pengetikan. Kesalahan tersebut dapat berupa kurangnya pengetahuan mahasiswa akan ejaan yang benar sesuai dengan kamus besar Bahasa Indonesia, kelalaian mahasiswa yang tidak disengaja, kesalahan pengaturan aplikasi yang digunakan untuk media pengetikan dan beberapa hal lain yang menyebabkan terjadinya kesalahan ejaan kata [2].

Kesalahan ejaan merupakan keadaan di mana terjadi kesalahan penulisan susunan kata. Berdasarkan sejarahnya, awalnya keadaan ini berhubungan dengan kesalahan penulisan kata secara manual, namun saat ini hal tersebut juga dapat terjadi pada proses pengetikan yang dilakukan dengan bantuan mesin ketik dan komputer. Hal tersebut dapat terjadi dikarenakan kesalahan mekanik juga tangan atau jari memeleset saat mengetik, selain itu terkadang juga disebabkan oleh ketidaktahuan seseorang tentang bagaimana pengejaan tulisan yang benar [3].

Berdasarkan jenis katanya *typographical error* dapat dibedakan menjadi 2 tipe yaitu *non-word spelling error* dan *real-word spelling error*. *Non-word spelling error* merupakan kesalahan penulisan kata di mana kata tersebut tidak dapat ditemukan dalam kamus (tidak memiliki makna). Sedangkan *real-word spelling error* merupakan kesalahan penulisan kata di mana kata tersebut dapat ditemukan dalam kamus (memiliki makna) namun bukan kata yang dimaksud dalam dokumen [3].

Tahap awal yang dilakukan untuk menghasilkan tulisan yang bebas dari salah eja adalah pendeteksian. Menurut Maghfira [3], deteksi kesalahan ejaan merupakan proses pengecekan validitas suatu kata dalam bahasa tertentu, suatu kata disebut valid jika kata tersebut dapat ditemukan dalam *lexical resource*. *Lexical resource* merupakan *database* di mana data di dalamnya dapat berupa *corpus*, *lexicon*, *word list* atau bentuk lain. Proses utama dari *error detection* adalah membandingkan kata dalam teks dengan kata yang terdapat pada *lexical resource*. Banyak metode yang dapat digunakan untuk proses deteksi kesalahan penulisan kata, namun yang umumnya digunakan untuk deteksi *non-word error* adalah deteksi menggunakan *Dictionary Lookup* dan *N-Gram Analysis*.

Metode *dictionary lookup* merupakan metode yang sering digunakan dalam menentukan *non-word error*. Proses yang dilakukan pada metode ini yaitu melakukan pengecekan apakah kata yang dimaksud terdaftar dalam kamus atau tidak, jika tidak ada maka kata ini dianggap sebagai *non-word*. Cara ini termasuk cara yang efektif untuk menentukan kata termasuk salah penulisannya atau tidak, namun jumlah kata dalam

kamus yang banyak dapat berakibat pada proses pengecekan menjadi lama, oleh karena itu dibutuhkan teknik optimasi pada teknik pencarian kata. Teknik optimasi dapat dilakukan dengan penggunaan binary search dan hash seperti pada penelitian [9].

C. *Dictionary Lookup*

Dictionary lookup merupakan metode yang melakukan pencarian secara sederhana untuk melihat keberadaan kata di dalam kamus atau daftar kata yang telah dibuat (Putra, Sujaini, & Safriadi, 2018). Metode dictionary lookup merupakan sebuah metode yang sering digunakan dalam menentukan non-word error, yaitu kesalahan penulisan kata yang menjadikan ia tidak dapat ditemukan dalam kamus. Proses yang dilakukan pada metode ini yaitu melakukan pengecekan apakah kata yang dimaksud terdaftar dalam kamus atau tidak, jika tidak ada maka kata ini dianggap sebagai non-word. Menurut Soleh **[9]**, cara ini termasuk cara yang efektif untuk menentukan kata termasuk salah penulisannya atau tidak, namun jumlah kata dalam kamus yang banyak dapat berakibat pada proses pengecekan menjadi lama. Oleh karena itu dibutuhkan teknik optimasi pada teknik pencarian kata. Pada penelitian Soleh et al. (2011) teknik optimasi dilakukan dengan penggunaan binary search dan hash.

D. *PHP Framework Laravel*

PHP adalah singkatan perulangan untuk PHP: Hypertext Preprocessor, yaitu bahasa pemrograman yang digunakan secara luas untuk penanganan pembuatan dan pengembangan sebuah situs web. Ada banyak framework PHP yang popular, salah satunya Laravel. Framework Laravel dibuat oleh Taylor Otwell, Proyek Laravel dimulai pada April 2011. Laravel merupakan Framework PHP yang menekankan pada kesederhanaan dan fleksibilitas pada desainnya. Sama seperti framework lainnya, Laravel dibangun dengan basis MVC (Model, View, Controller). Banyak situs survei yang merilis bahwa framework Laravel ini merupakan yang terpopuler dibandingkan framework PHP lainnya. (Rohman, 2014)

## 3. Metode Penelitian

Penelitian ini terbagi dalam beberapa tahapan seperti yang terlihat pada Gambar 1. Tahap awal yang dilakukan pada penelitian ini yaitu mengidentifikasi masalah. Hal ini bertujuan untuk memastikan ruang lingkup penelitiannya. Pada tahap ini diidentifikasi latar belakang serta rumusan masalah yang akan diselesaikan.

Gambar 1. Alur Tahapan Penelitian

Selanjutnya, langkah yang dilakukan pada tahapan kedua yaitu mencari sumber kajian sebagai pengetahuan dasar dari berbagai macam literatur yang dapat dijadikan landasan dalam pengerjaan penelitian ini. Tujuan studi literatur dilakukan untuk memahami lebih dalam permasalahan dari penelitian ini. Sumber yang digunakan dalam studi literatur ini berasal dari beberapa artikel dan buku yang membahas topik yang relevan dengan penelitian ini.

Tahapan ketiga merupakan tahapan yang sangat penting dalam penelitian ini karena pengumpulan data merupakan komponen pendukung utama untuk melakukan penelitian ini. Data yang dikumpulkan berasal dari beberapa sumber diantaranya data kamus Bahasa Indonesia pada situs KBBI daring dan data kamus istilah asing berasal dari situs *kateglo.com*. Pengambilan data dilakukan dengan menggunakan teknik *scraping* dengan memanfaatkan fitur dari *webscraper.io*. Semua kata diambil dari situs halaman per halaman hingga tuntas, kemudian dilakukan *proprocessing* untuk membersihkan halaman tersebut sebelum disimpan dalam format *plain text*. Adapun data naskah tugas akhir mahasiswa dikumpulkan dari arsip Jurusan Informatika atau alumni secara langsung.

Tahapan implementasi program dikerjakan dengan menggunakan bahasa PHP dan *framework* Laravel. Kode skrip disusun untuk memproses tahapan dari awal hingga akhir, dimulai dari proses *preprocessing*, kemudian proses *Dictionary Lookup* pada saat melengkapi kamus dan saat mencari kata yang salah pada naskah tugas akhir. Tahapan terakhir yang dilakukan adalah analisa hasil yaitu terbentuknya kamus kota kata yang lengkap sehingga dapat digunakan untuk memeriksa dan menemukan kata-kata yang salah dan keliru dalam tulisan ilmiah tugas akhir mahasiswa Jurusan Informatika. Saat melakukan *Dictionary Lookup* pada pengecekan tugas akhir, kata-kata yang tidak terdapat di kamus akan ditandai dan dihitung jumlahnya pada setiap dokumen. Kata-kata salah yang sudah dikumpulkan dalam satu dokumen dianalisis kesalahannya. Lalu akan diuji keakuratan algoritma pada penelitian ini dengan menggunakan *confusion matrix*.

## 4. Hasil dan Pembahasan

Tahap awal dalam pembangunan kamus diawali dengan mengumpulkan kosakata. Kosakata yang dibutuhkan meliputi kosakata bahasa Indonesia, singkatan- singkatan, serta istilah asing. Kosakata diambil dari situs yang terdiri dari banyak halaman, sehingga digunakan *web scraper* untuk mempercepat pengumpulan kata. Setelah dilakukan proses scraping, diperoleh kata bahasa Indonesia dari tujuh kelas kata yaitu adjektiva, adverbia, nomina, numeralia, partikel, pronomina, dan verba dengan jumlah 77.348 kata. Singkatan-singkatan beserta istilah asing dari situs Kateglo didapat sebanyak 191.200 kata. Karena adanya duplikasi dan kata homonim dari seluruh halaman yang telah dilakukan *scraping*, maka kata yang sama dihapus sehingga menghasilkan kata yang unik berjumlah 130.431 kata.

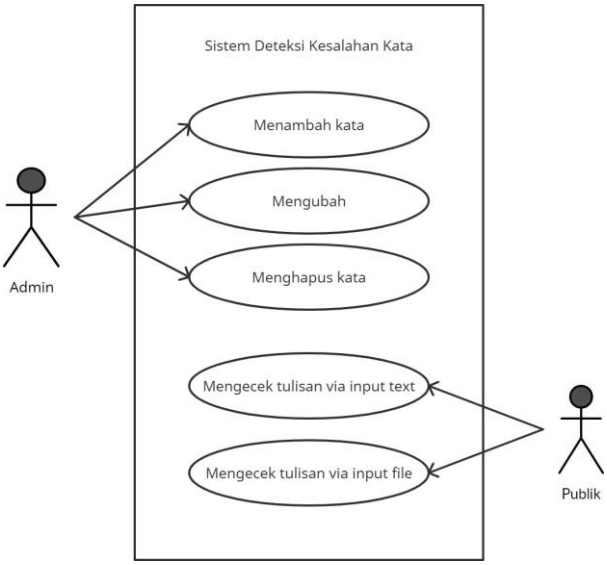
Selanjutnya, berdasarkan tujuan awal dari penelitian ini adalah mendeteksi keberadaan kata tak baku dan kesalahan penulisan kata pada naskah tugas akhir mahasiswa Jurusan Informatika Universitas Syiah Kuala. Objek data yang akan dijadikan sebagai sampel adalah bab satu dari naskah tugas akhir mahasiswa yang mewakili dari setiap tahun kelulusan. Oleh karena itu, berdasarkan data jumlah alumni Jurusan Informatika yang terlihat pada Tabel 2, maka data sampel yang akan digunakan sebagai pengujian berjumlah 40 dokumen tugas akhir. Jumlah ini dianggap memadai untuk dijadikan sampel yang optimal jika dibandingkan dengan jumlah populasi semua lulusan dari setiap tahun sebanyak 263 mahasiswa.

Tujuan dari penelitian ini adalah membentuk kamus kosakata dengan lengkap, memeriksa dan menemukan kata yang salah dan keliru dalam tulisan, serta menganalis seberapa banyak kekeliruan yang terdapat dalam tulisan-tulisan tersebut. Tujuan-tujuan tersebut dicapai dengan implementasi *Dictionary Lookup* yang melakukan pencarian sederhana untuk melihat keberadaan kata di dalam kamus dan untuk menemukan non word error.

Tabel 2. Data jumlah lulusan Jurusan Informatika, Universitas Syiah Kuala

| No. | Tahun Lulus | Jumlah Lulusan |
|---|---|---|
| 1 | 2015 | 41 mahasiswa |
| 2 | 2016 | 59 mahasiswa |
| 3 | 2017 | 27 mahasiswa |
| 4 | 2018 | 35 mahasiswa |
| 5 | 2019 | 43 mahasiswa |
| 6 | 2020 | 44 mahasiswa |
| 7 | 2021 | 14 mahasiswa |

Sistem pendeteksian kesalahan kata ini diimplementasikan dalam bentuk aplikasi berbasis web. Halaman web ini memiliki dua sisi pengguna, yaitu admin dan pengguna publik seperti yang digambarkan dalam diagram use case pada Gambar 2. Admin memiliki fungsi untuk mengelola kosakata dalam kamus seperti menambahkan, mengubah, dan menghapus kata. Pengguna publik hanya dapat mengakses halaman terluar yang berisi kolom input teks dan kolom input fail untuk mendeteksi kata yang salah.




Gambar 2. Diagram use case pengguna halaman deteksi kesalahan kata

Dalam implementasinya, data kamus disimpan dalam *database* dengan menggunakan *index* untuk setiap kata di dalam tabel. *Index* adalah sebuah objek dalam sistem *database* yang dapat mempercepat proses pencarian (*query*) data. *Index* merupakan objek struktur data tersendiri yang tidak bergantung kepada struktur tabel. Setiap *index* terdiri dari nilai kolom dan penunjuk ke baris yang berisi nilai tersebut. Penunjuk tersebut secara langsung menunjuk ke baris yang tepat pada tabel, sehingga menghindari terjadinya *full table-scan*. Pendeteksian kata ini membutuhkan pencarian berulang yang mengharuskan pengecekan ke kamus untuk setiap kata. Oleh sebab itu penggunaan *index* akan mempercepat kinerja sistem dalam melakukan pencarian secara berulang.

Sebanyak 40 dokumen tugas akhir mahasiswa Jurusan Informatika, Universitas Syiah Kuala digunakan sebagai sampel dalam proses pendeteksian terhadap kata-kata yang tidak baku dan salah tik. Hasil pada Tabel 3 menunjukkan jumlah kesalahan yang muncul pada setiap dokumen yang diperiksa.

Tabel 3. Uji pendeteksian kesalahan penulisan kata dalam dokumen

| Dokumen | Jumlah Kata | Kata Yang Salah | Persentase Kata Yang Salah (%) | Waktu Eksekusi (ms) |
|---|---|---|---|---|
| 1 | 517 | 9 | 1,74 | 3,31 |
| 2 | 616 | 6 | 0,97 | 4,04 |
| 3 | 830 | 34 | 4,10 | 4,87 |
| 4 | 425 | 19 | 4,47 | 2,77 |
| 5 | 655 | 15 | 2,29 | 4,26 |
| 6 | 754 | 20 | 2,65 | 4,04 |
| 7 | 1052 | 54 | 5,13 | 5,78 |
| 8 | 758 | 14 | 1,85 | 4,55 |
| 9 | 558 | 18 | 3,23 | 3,23 |
| 10 | 678 | 34 | 5,01 | 3,95 |
| 11 | 975 | 13 | 1,33 | 5,74 |
| 12 | 970 | 17 | 1,75 | 5,91 |
| 13 | 679 | 29 | 4,27 | 4,32 |
| 14 | 708 | 23 | 3,25 | 4,21 |
| 15 | 842 | 4 | 0,48 | 5,06 |
| 16 | 758 | 17 | 2,24 | 4,46 |
| 17 | 752 | 19 | 2,53 | 4,19 |
| 18 | 745 | 4 | 0,54 | 4,47 |
| 19 | 715 | 20 | 2,80 | 4,29 |
| 20 | 849 | 14 | 1,65 | 5,32 |
| 21 | 752 | 3 | 0,40 | 4,29 |
| 22 | 462 | 6 | 1,30 | 2,80 |
| 23 | 374 | 18 | 4,81 | 2,37 |
| 24 | 802 | 18 | 2,24 | 4,83 |
| 25 | 527 | 20 | 3,80 | 3,21 |
| 26 | 843 | 12 | 1,42 | 4,99 |
| 27 | 566 | 1 | 0,18 | 3,40 |
| 28 | 680 | 11 | 1,62 | 4,10 |
| 29 | 575 | 20 | 3,48 | 0,62 |
| 30 | 640 | 11 | 1,72 | 3,99 |
| 31 | 467 | 14 | 3,00 | 2,90 |
| 32 | 695 | 20 | 2,88 | 3,83 |
| 33 | 756 | 45 | 5,95 | 4,58 |
| 34 | 706 | 24 | 3,40 | 4,52 |
| 35 | 603 | 5 | 0,83 | 3,64 |
| 36 | 763 | 42 | 5,50 | 2,75 |
| 37 | 939 | 29 | 3,09 | 5,57 |
| 38 | 723 | 19 | 2,63 | 4,41 |
| 39 | 791 | 44 | 5,56 | 5,17 |
| 40 | 866 | 37 | 4,27 | 4,96 |

Berdasarkan Tabel 3 di atas, kemudian ingin dihitung rata-rata kecepatan sistem dalam proses eksekusi teks. Untuk itu, perlu dianalisis kemampuan jumlah kata per detik dalam setiap dokumen yang diproses. Hasil perhitungan jumlah kata per detik dalam setiap dokumen terlihat pada Tabel 4. Jika dihitung rata-ratanya, maka seluruh nilai pada Tabel 4 menghasilkan nilai rata-rata kecepatan eksekusi sebesar 187 kata per milidetik,

atau 187.414 kata per detik. Penggunaan metode *Dictionary Lookup* dengan memanfaatkan *index* pada DBMS dinilai efektif dari segi waktu eksekusi yang dibutuhkan.

Tabel 4. Perhitungan Kecepatan Eksekusi

| Dokumen | Kecepatan (Kata/milidetik) | Kecepatan (Kata/detik) |
|---|---|---|
| 1 | 156,14 | 156.137 |
| 2 | 152,39 | 152.392 |
| 3 | 170,14 | 170.141 |
| 4 | 153,09 | 153.087 |
| 5 | 153,71 | 153.713 |
| 6 | 186,62 | 186.624 |
| 7 | 181,81 | 181.809 |
| 8 | 166,58 | 166.582 |
| 9 | 172,37 | 172.371 |
| 10 | 171,64 | 171.637 |
| 11 | 169,73 | 169.733 |
| 12 | 164,09 | 164.093 |
| 13 | 157,02 | 157.023 |
| 14 | 167,84 | 167.844 |
| 15 | 166,23 | 166.229 |
| 16 | 169,94 | 169.944 |
| 17 | 179,08 | 179.082 |
| 18 | 166,32 | 166.321 |
| 19 | 166,58 | 166.581 |
| 20 | 159,43 | 159.428 |
| 21 | 175,00 | 174.998 |
| 22 | 164,58 | 164.577 |
| 23 | 157,74 | 157.739 |
| 24 | 165,86 | 165.864 |
| 25 | 164,16 | 164.164 |
| 26 | 168,64 | 168.644 |
| 27 | 166,02 | 166.021 |
| 28 | 165,60 | 165.603 |
| 29 | 913,86 | 913.859 |
| 30 | 160,07 | 160.072 |
| 31 | 160,58 | 160.580 |
| 32 | 181,17 | 181.169 |
| 33 | 164,80 | 164.803 |
| 34 | 155,94 | 155.943 |
| 35 | 165,38 | 165.378 |
| 36 | 277,43 | 277.434 |
| 37 | 168,30 | 168.301 |
| 38 | 163,60 | 163.601 |
| 39 | 152,84 | 152.841 |
| 40 | 174,38 | 174.375 |

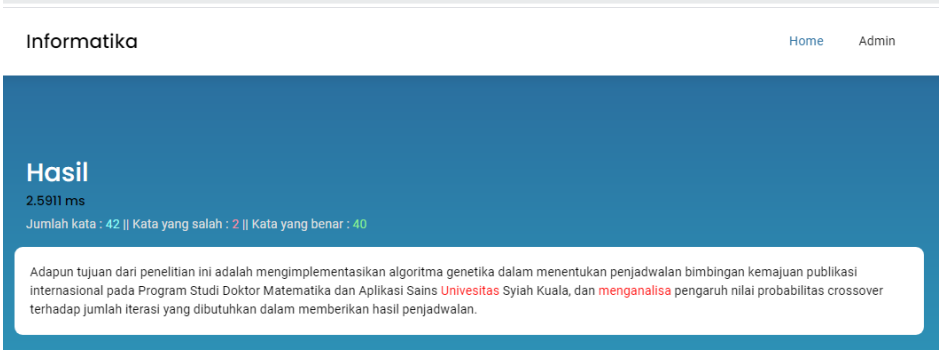
Selanjutnya untuk mengevaluasi kinerja sistem, perlu dilakukan suatu pengujian. Pada penelitian ini kinerja sistem akan diuji dengan confusion matrix.. Penentuan TP (True Positive), FP (False Positive), TN (True Negative), dan FN (False Negative) didasari pada benar salahnya sistem mengelompokkan sebuah kata ke dalam kategori benar atau salah. Secara lebih sederhana, dua kelas yang dimilkii adalah positif dan negatif. Kelas positif adalah kelas milik tulisan berwarna merah atau kata yang terdeteksi salah tik atau tidak baku. Kelas negatif adalah kelas milik tulisan berwarna hitam atau yang dianggap benar. Agar lebih jelas, TP adalah tulisan yang dimerahkan dan ternyata

valid. TN adalah tulisan yang dihitamkan dan ternyata valid. FP adalah tulisan yang dimerahkan dan ternyata tidak valid. FN adalah tulisan yang dihitamkan dan ternyata tidak valid.

Tabel 5. Perhitungan nilai presisi pada setiap dokumen

| **Dokumen** | **Jumlah Kata** | **Kata Terdeteksi Salah** | **TP (*True Positive*)** | **FP (*False Positive*)** | **Presisi** |
|---|---|---|---|---|---|
| 1 | 517 | 9 | 4 | 5 | 44,44 % |
| 2 | 616 | 6 | 2 | 4 | 33,33 % |
| 3 | 830 | 34 | 18 | 16 | 52,94 % |
| 4 | 425 | 19 | 8 | 11 | 42,11 % |
| 5 | 655 | 15 | 0 | 15 | 0,00 % |
| 6 | 754 | 20 | 0 | 20 | 0,00 % |
| 7 | 1052 | 54 | 14 | 40 | 25,93 % |
| 8 | 758 | 14 | 5 | 9 | 35,71 % |
| 9 | 558 | 18 | 7 | 11 | 38,89 % |
| 10 | 678 | 34 | 6 | 28 | 17,65 % |
| 11 | 975 | 13 | 5 | 8 | 38,46 % |
| 12 | 970 | 17 | 5 | 12 | 29,41 % |
| 13 | 679 | 29 | 13 | 16 | 44,83 % |
| 14 | 708 | 23 | 3 | 20 | 13,04 % |
| 15 | 842 | 4 | 3 | 1 | 75,00 % |
| 16 | 758 | 17 | 6 | 11 | 35,29 % |
| 17 | 752 | 19 | 1 | 18 | 5,26 % |
| 18 | 745 | 4 | 3 | 1 | 75,00 % |
| 19 | 715 | 20 | 1 | 19 | 5,00 % |
| 20 | 849 | 14 | 4 | 10 | 28,57 % |
| 21 | 752 | 3 | 0 | 3 | 0,00 % |
| 22 | 462 | 6 | 4 | 2 | 66,67 % |
| 23 | 374 | 18 | 0 | 18 | 0,00 % |
| 24 | 802 | 18 | 4 | 14 | 22,22 % |
| 25 | 527 | 20 | 2 | 18 | 10,00 % |
| 26 | 843 | 12 | 0 | 12 | 0,00 % |
| 27 | 566 | 1 | 0 | 1 | 0,00 % |
| 28 | 680 | 11 | 1 | 10 | 9,09 % |
| 29 | 575 | 20 | 6 | 14 | 30,00 % |
| 30 | 640 | 19 | 12 | 7 | 63,16 % |
| 31 | 467 | 14 | 5 | 9 | 35,71 % |
| 32 | 695 | 20 | 8 | 12 | 40,00 % |
| 33 | 756 | 45 | 21 | 24 | 46,67 % |
| 34 | 706 | 24 | 8 | 16 | 33,33 % |
| 35 | 603 | 5 | 1 | 4 | 20,00 % |
| 36 | 763 | 42 | 22 | 20 | 52,38 % |
| 37 | 939 | 29 | 4 | 25 | 13,79 % |
| 38 | 723 | 19 | 9 | 10 | 47,37 % |
| 39 | 791 | 44 | 4 | 40 | 9,09 % |
| 40 | 866 | 37 | 3 | 34 | 8,11 % |

Nilai presisi dari setiap dokumen data sampel sangat bervariasi seperti yang terlihat pada Tabel 5. Jika dirata- ratakan, maka nilai presisi dari semua sampel adalah 28,71%. Artinya, nilai ini menjawab pertanyaan seberapa besar persentase jumlah kata yang benar-benar keliru dari seluruh kata yang dikategorikan keliru oleh sistem. Nilai presisi yang rendah ini jika dianalisis secara sederhana berdasarkan kata-kata yang dianggap salah oleh sistem, maka ditemukan bahwa penyebab utamanya adalah jumlah pada False Positive. Pada kolom FP, sebagian besar kata merupakan kata yang dianggap keliru oleh sistem karena kata tersebut adalah nama tempat, nama merek, nama orang, kutipan langsung dari bahasa lain, dan lainnya. Sedangkan sistem tetap akan mengategorikannya sebagai kata yang salah karena tidak berada dalam kamus.




Gambar 3. Hasil pendeteksian kata yang salah pada teks.

Hasil nilai presisi yang rendah pada Tabel 5 tidak menjadikan sistem ini buruk. Berdasarkan pengamatan sederhana pada hasil deteksi tulisan, sebagian besar kata yang benar tetap dikategorikan sebagai kata yang benar (dihitamkan) karena sudah tentu kata-kata tersebut berada di dalam kamus seperti terlihat di halaman sistem pada Gambar 3. Pengecualian terjadi pada kata-kata salah tik namun menghasilkan makna lain yang masih berada dalam kamus (word error). Melakukan pemindaian pada seluruh dokumen sampel untuk menemukan jumlah True Negative dan False Negative dianggap kurang efisien karena jumlah data yang sangat besar. Oleh karena itu, cukup dengan mengetahui bahwa nilai porsi TN pada semua sampel adalah sangat besar, maka kita dapat mengetahui bahwa nilai akurasi juga sangat besar.

## 5. Kesimpulan

Penelitian ini menghasilkan sebuah kamus kosakata dan aplikasi halaman web untuk mendeteksi kesalahan penulisan kata akibat salah tik maupun penggunaan kata tak baku. Kosa kata yang diperoleh sebanyak 131.495 yang terdiri dari kata bahasa Indonesia, singkatan, akronim, dan istilah asing. Selain itu, besar nilai rata-rata kecepatan deteksi tulisan pada percobaan sampel adalah 187.414 kata per detik. Metode *Dictionary Lookup* dinilai efektif dari segi waktu yang dibutuhkan. Hasil percobaan pendeteksian kata dari sampel yang digunakan menunjukkan bahwa rata-rata kesalahan yang ada di dalam setiap dokumen adalah sebesar 2,76%, dengan nilai presisi sebesar 28,71%. Nilai akurasi dan presisi dari hasil deteksi sangat bergantung pada tingkat kelengkapan kamus kosakata. Semakin lengkap isi kamus maka hasil yang didapat akan semakin baik.

## References

[1] U. Syiah Kuala, Panduan Tugas Akhir dan Tesis Fakultas Matematika dan Ilmu Pengetahuan Alam, Banda Aceh: FMIPA, Universitas Syiah Kuala, 2019.

[2] H. R.N., "Aplikasi Perbaikan Ejaan Pada Karya Tulis Ilmiah di Program Studi Teknik Informatika dengan Menerapkan Algoritma Levenshtein Distance," Universitas Nusantara PGRI, Kediri, 2016.

[3] T. N. Maghfira, I. Cholissodin and A. Wahyu, "Deteksi Kesalahan Ejaan dan Penentuan Rekomendasi Koreksi Kata yang Tepat Pada Dokumen Jurnal JTIIK Menggunakan Dictionary Lookup dan Damerau-Levenshtein Distance," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer,* vol. 1, pp. 498-506, 2017.

[4] A. Hadi, M. Fachrurrozi and N. Yuslianti, "Analisa Perbandingan Algoritma Damerau-Levensthein Distance dan Soundex Similarity Pada Pengkoreksian Ejaan Kata Otomatis," Fakultas Ilmu Komputer, Universitas Sriwijaya, Palembang, 2019.

[5] E. Kosasih and W. Hermawan, "BAHASA INDONESIA Berbasis Kepenulisan Karya Ilmiah dan Jurnal," CV. Thursina, Bandung, 2012.

[6] S. and S. Saudah, Buku Ajar Bahasa Indonesia Akademik, Yogyakarta: Pustaka Pelajar, 2016.

[7] S. R.A, Analisis Penggunaan Tata Bahasa Indonesia dalam Penulisan Karya Tulis Ilmiah : Studi Kasus Artikel Ilmiah, Visi Pustaka, 2010.

[8] H. Alwi, Tata bahasa baku bahasa Indonesia, Jakarta: Perum Balai Pustaka, 1998.

[9] M. Y. Soleh and A. Purwarianti, A non word error spell checker for Indonesian using morphologically analyzer and HMM, New Jersey, USA: IEEE, 2011.