MIND Journal | ISSN (p): 2528-0015| ISSN (e): 2528-0902 | Vol. 3 | No. 2 | Halaman 47-56
DOI : https://doi.org/10.26760/mindjournal Desember 2018

Penerapan Algoritma Lemmatization pada Dokumen Bahasa Indonesia

YUSUP MIFTAHUDDIN, JASMAN PARDEDE, RENITA DEWI

[1]Institut Teknologi Nasional Bandung

Email : yusufm@itenas.ac.id

ABSTRAK

Sebuah kata, kalimat maupun tulisan dapat dikatakan layak apabila telah memenuhi PUEBI (Pedoman Umum Ejaan Bahasa Indonesia) dan KBBI (Kamus Besar Bahasa Indonesia). Akan tetapi, sangat banyak ditemukan kesalahan penulisan dalam suatu dokumen seperti karya ilmiah mapun skripsi diantaranya yaitu suatu katayang tidak memenuhi kebakuan dan tidak sesuai dengan PUEBI dan kesalahan penulisan (typographical error) yaitu salah dalam pengetikkan karena kecepatan perpindahan jari yang tidak seimbang dari satu tombol ke tombol lain untuk merangkai kata yang akhirnyamembuat orang salah paham dengan maksud kalimat yang dihasilkan dari susunan kata tersebut karena tidak ada dalam KBBI. Algoritma lemmatization adalah suatu algoritma yang digunakan untuk menemukan bentuk dasar dari suatu kata sehingga dapat dimanfaatkan untukmemeriksa kebenaran dari penggunaan ejaan pada suatu kata. Penelitian ini bertujuan untuk mengukur keakurasian dari penggunaan algoritma lemmatization dalam melakukan seleksi terhadap kata yang salah atau tidak tepat berdasarkan PUEBI dan KBBI sebagai acuan

***Kata kunci**: Kesalahan penulisan, lemmatization, PUEBI, KBBI*

ABSTRACT

A word, sentence or writing can be said worthy if it meets PUEBI and KBBI. However, very many errors found in a document such as scientific papers mapun thesis among which is a word that does not meet the behavior and not in accordance with PUEBI and typographical errors is wrong in typing because the speed of finger movement is not balanced from a single button to another button for stringing words that ultimately make people misunderstand the meaning of the sentence that results from the wording because it is not in KBBI. Lemmatization algorithm is a useful process to find the basic form of a word so it can be used to check the truth of the use of spelling on a word. This study aims to find out the accuracy of the use of lemmatization algorithm in the selection of wrong or incorrect word based on PUEBI and KBBI as a reference

***Keywords**: Typographical error, lemmatization, PUEBI, KBBI*

1. PENDAHULUAN

Bahasa merupakan alat komunikasi dalam bentuk percakapan dengan berbagai rangkaian kata. Bahasa Indonesia digunakan untuk mempersatukan keanekaragaman bahasa di Indonesia **(Saepudin, Apep, 2016).** Suatu tulisan dikatakan layak untuk dipublikasikan atau dibaca oleh masyarakat umum apabila mematuhi aturan atau kaidah penulisan PUEBI **(Gusnita, Nova, 2016)**.Namun, dalam penulisan suatu dokumen seperti karya ilmiah maupun skripsi, tidak jarang ditemui banyak kesalahan penulisan yang tidak memenuhi kebakuan dan tidak sesuai dengan aturan tata bahasa yang berlaku di Indonesia dan juga terdapat kesalahan dalam penulisan (typographical error) yang menghasilkan suatu kata tidak terdefinisi di dalam KBBI (Kamus Besar Bahasa Indonesia).Lemmatization adalah algoritma yang memanfaatkan analisis morfologi dan aturan penulisan pemisahan dan penggabungan kata yang berguna sebagai pemeriksaan kebenaran ejaan berdasarkan 2aturan PUEBI dengan menerapkan algoritma lemmatization, suatu kata dapat diketahui kata dasarnya serta ketepatan penggunaan imbuhannya. Apabila kata yang diperiksa tidak ditemukan kata dasarnya, maka kata tersebut akan masuk kedalam kategori typographical / morphology error.

Dalam proses pembuatan sebuah tulisan karya ilmiah diperlukan penggunaan ejaan dan penggunaan kalimat dan kata yang sesuai dengan Pedoman Umum Ejaan Bahasa Indonesia (PUEBI). Dalam hal ini, masih banyak penulis yang kurang memperhatikan ejaanpenggunaan tanda baca, pemakaian huruf, penulisan kata, penulisan kata ganti dan lain sebagainya (**Jasmienti, 2017**). Selain kesalahan penggunaan ejaan, terdapat satu kesalahan lagi yang sering terjadi yaitu *typographical error* dimana kata yang diketik tidak lengkap atau tidak terdapat pada Kamus Besar Bahasa Indonesia (KBBI) sehingga mengakibatkan kata tersebut sulit dimengerti.

Tujuan pada penelitian ini adalah melakukan pengecekan pada suatu file untuk mencari kata yang tidak terdapat pada KBBI atau ejaan yang tidak terdapat pada PUEBI dengan menerapkan algorima *Lemmatizaion* dalam melakukan proses pencariannya.

Adapun ruang lingkup penelitian yang dibuat adalah sebagai berikut :

1. Literature data menggunakan PUEBI Edisi keempat yang dijadikan sebagai landasan acuan ketepatan penggunaan ejaan
2. Literature acuan kata baku atau kata dasar menggunakan Kamus Besar Bahasa Indonesia (KBBI) Edisi keempat
3. File yang diperiksa berupa masukkan dari suatu berkas yang dibuat dengan minimal Microsoft Word 200 3 yang memiliki ekstensi file .doc (*document)* dan ekstensi file .docx.
4. Melakukan pencarian kesalahan ejaan dengan menggunakan algoritma *lemmatization*.

2. LANDASAN TEORI

2.1 Penggunaan Bahasa Indonesia

Morfologi adalah ilmu yang mempelajari penyusunan kata secara struktural terhadap morfem pembentuknya. Morfem adalah bentuk bahasa terkecil yang tidak dapat dibagi lagi menjadi bagian-bagian yang lebih kecil. Dalam bahasa indoneisa terdapat dua jenis morfem yaitu morfem bebas atau morfem yang dapat berdiri sendiri dan morfem terikat misalnya imbuhan (afiks). Berikut ini merupakan macam-macam afiks dan contohnya :

- Awalan (prefiks) adalah imbuhan yang dibutuhkan di awal kata (‘me-‘, ‘se-‘, ‘ke-‘, ‘di-‘,‘ter-‘ (te-), ‘ber-‘ (be-), ‘per-‘ (pe-), ‘ku-‘, dan ‘kau-‘).
- Akhiran (sufiks) adalah imbuhan yang dibubuhkan di akhir kata (‘-i’, ‘-kan’, dan ‘-an’).
- Sisipan (infiks) adalah imbuhan yang dibutuhkan di tengah kata (‘-el-’, ‘-em-‘ dan ‘-er‘).
- Gabungan awalan dan akhiran (konfiks) merupakan kombinasi dari prefiks dan sufiks. Untuk kombinasi prefiks-sufiks dijelaskan pada tabel 1.

Tabel 1. Kombinasi imbuhan prefiks-sufiks

Prefiks	Sufiks	Contoh
‘me-‘, ‘per-‘, ‘ber-‘, ‘ter-‘, dan ‘di-‘	‘-kan’	Perkenalan – kenal
‘me-‘, ‘per-‘, ‘ter-‘, dan ‘di-‘	‘-i’	Memiliki – milik
‘ber-‘ dan ‘ke-‘	‘-an’	Kepercayaan - percaya
‘ter-‘ + ‘per-‘	-	
‘se-‘ + ‘per-‘	-	
‘ke-‘ + ‘se-‘ + ‘per-‘	-	
‘mem-‘ + ‘per-’	-	
‘di-‘ + ‘per-‘	-	

2.2 Pedoman Umum Ejaan Bahasa Indonesia (PUEBI)

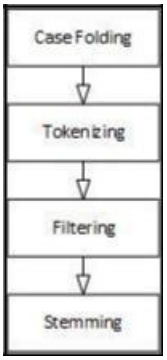
Pada tahun 2015, PUEBI (Pedoman Umum Ejaan Bahasa Indonesia) ditetapkan sebagai pedoman dalam sistem ejaan yang berlaku di Indonesia, dimana ruang lingkupnya terdiri dari penulisan huruf, kata, singkatan dan akronim, angka bilangan, serta tanda baca **(Rangga Nugraha, 2016).**

2.3 Kamus Besar Bahasa Indonesia (KBBI)

KBBI adalah kamus yang dijadikan acuan dalam tata bahasa Indonesia yang lengkap dan dipatenkan oleh Pemerintah Republik Indonesia yang dinaungi oleh Kementrian Pendidikan dan Kebudayaan Indonesia. **(http://www.kamus-kbbi.com/kbbi/)**

2.4 *Text Preprocessing*

Preprocessing adalah proses mempersiapkan teksmenjadi data yang dapat diolah pada tahapan selanjutnya dengan inputan awal berupa dokumen, dengan kata lain bertujuan untuk menghilangkan *noise* yang terdapat pada dokumen teks dan mengambil fitur atau parameter penting yang terdapat pada dokumen teks. Tahapan *text proprecessing* dijelaskan pada Gambar 1.



Gambar 1. Tahap Preprocessing

1. *Case Folding* merupakan proses konversi keseluruhan teks dalam dokumen menjadi suatu bentuk standar (biasanya huruf kecil atau *lowercase*).
2. *Tokenizing* merupakan tahap pemotongan *string input* setiap kata
3. *Filtering* merupakan tahap pemilihan kata-kata penting dari hasil token.
4. *Stemming* adalah tahapan untuk mengelompokkan kata yang memiliki kata dasar dan arti yang serupa tetapi kata tersebut memiliki bentuk atau form yang berbeda

2.5 Algoritma Lemmatization

A. K. Ingason (2008) mengemukakan bahwa "*lemmatization* adalah proses untuk menemukan bentuk dasar dari sebuah kata". **Nirenburg (2009)** menjelaskan bahwa "*lemmatization* adalah proses yang bertujuan untuk melakukan normalisasi pada teks dengan berdasarkan pada bentuk dasar yang merupakan bentuk lemmanya".

Normalisasi adalah mengidentifikasi dan menghapus imbuhan prefiks serta sufiks sebuah kata. Dimana, Lema merupakan bentuk dasar sebuah kata yang memiliki arti tertentu yang berdasar pada kamus **(Suhartono, 2014).** Beberapa proses yang perlu dilakukan dalam algoritma *lemmatization* yaitu sebagai berikut :

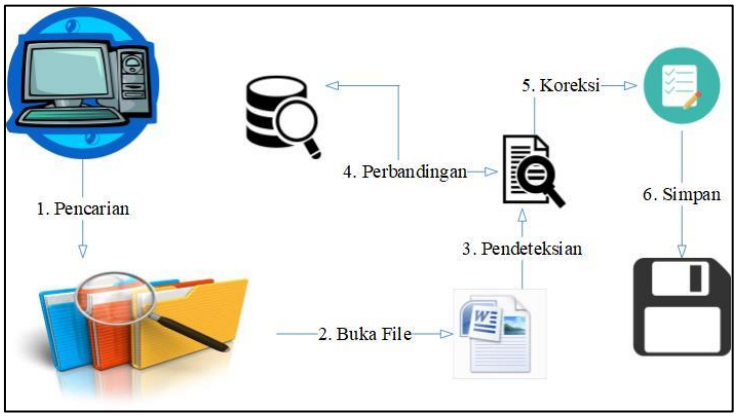
a. *Dictionary Lookup*
b. *Rule Precedence Check*
c. *Inflectional Suffix Removal*
d. *Derivational Suffix Removal*
e. *Derivational Prefix Removal*
f. *Recoding*
g. *Suffix Backtracking*

3. METODE PENELITIAN

3.1 Analisis Proses Sistem

Terdapat 6 (enam) langkah dalam melakukan perbaikan kesalahan ejaan maupun penulisan (*typographical error)*.

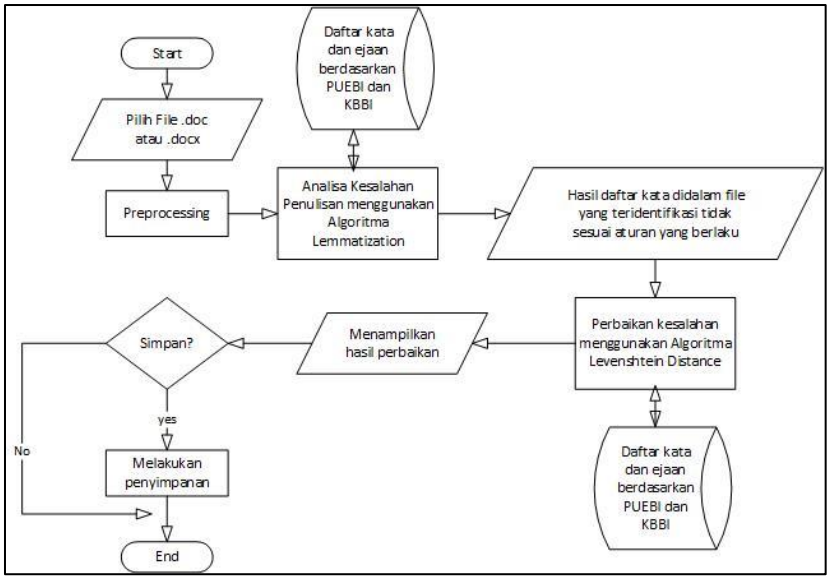
1. Langkah pertama melakukan pencarian terhadap file mana yang akan dilakukan perbaikan.
2. Langkah kedua dengan membuka file yang telah ditemukan untuk dilakukan perbaikan dimana file tersebut dalam format ekstensi .docx untuk versi Microsoft Word terbaru dan ekstensi .doc untuk versi sebelumnya.
3. Langkah ketiga melakukan pendeteksian atau pembacaan terhadap seluruh isi teks yang ada pada file yang sedang dibuka.
4. Langkah keempat melakukan perbandingan dari isi teks yang sedang di uji dengan *database* yang menyimpan ejaan dan kata-kata yang tepat berdasarkan PUEBI dan KBBI. Apabila ada kata-kata atau ejaan yang tidak sesuai maka akan langsung diberikan tanda harus diperbaiki. Algoritma *lemmatization* dengan memanfaatkan analisis morfologiakan digunakan untuk pendeteksian penggunaan ejaan yang tidak sesuai dengan PUEBI. Setelah dilakukan pemeriksaan terhadap kesesuaian ejaan, selanjutnya dilakukan pemeriksaan kesesuai penggunaan kata dengan menggunakan algoritma *Levenshtein Distance* untuk mendeteksi *typographical error* dengan memanfaatkan KBBI sebagai acuan.
5. Langkah kelima melakukan perbaikan terhadap berbagai kata-kata dan ejaan yang telah dideteksi tidak sesuai, dimana perbaikan ejaan dengan menggunakan algoritma *levenshtein distance* yaitu dengan menghitung jarak terdekat dari *string* awal dengan *string* baru sebagai saran kata perbaikannya. Perbaikan yang dilakukan memiliki beberapa jenis operasi yang dapat dilakukan yaitu dengan melakukan penambahan (*insert*) sebuah karakter kedalam sebuah kata yang kurang, penghapusan (*delete*) sebuah karakter didalam suatu kata, serta penggantian karakter (*substitute*) pada suatu karakter pada posisi tertentu dengan karakter lain.
6. Langkah keenam yaitu melakukan penyimpanan terhadap file yang telah dilakukan perbaikan.



Gambar 2. Prinsip kerja keseluruhan sistem

3.2 Flowchart Sistem

Pada Gambar 3 berikut ini merupakan flowchart dari proses *preprocessing.* Dimana sistem akan membaca file yang dimasukkan. Kemudian akan masuk kedalam tahap *tokenizing* yaitu data kalimat yang terbaca akan dilakukan pengecekkan terhadap spasi. Apabila terdapat spasi dalam kalimat yang diuji maka akan dilakukan penghapusan karakter spasi. Apabila telah dihapus karakter spasinya maka akan di *split* menjadi kata. Tahap *tokenizing* selesai apabila semuakalimat didalam file yang diuji telah menghasilkan daftar kata. Setelah *tokenizing,* langkah selanjutnyaadalah *filtering* dimana data kata yang telah didapatkan dari proses *tokenizing* akan diambil *stop word listnya* dengan melakukan pencocokandaftar kata nya menggunakan *stopword list* yang telah disiapkan. Apabila kata yang ada di dalam file yang diuji cocok dengan daftar *stopword list* maka kata tersebut akan dihapus. Selanjutnya langkah terakhir, *case folding* dimana data kata dari hasil filtering akang diubah semua hurufnya menjadi huruf kecil. Kemudian dilakukan penghapusan selain karakter *alphabet.* Apabila semua kata telah menjadi huruf kecil atau *lowercase*. Tahap terakhir dari *preprocessing* adalah tahap *stemming* yang menerapkan algoritma *lemmatization* yang telah dijelaskan pada gambar 2



Gambar 3. Flowchart deteksi kesalahan penulisan

4. HASIL DAN PEMBAHASAN

4.1 Implementasi Graphic User Interface (GUI)

Aplikasi ini memiliki beberapa panel, tombol dan menu. Menu pilih file digunakan untuk mengupload file yang akan diperiksa, tombol tokenize, filtering, stemming dan weight digunakan untuk menampilkan tahapan-tahapan pada lematisasi kata. Tampilan GUI dijelaskan pada Gambar 4.

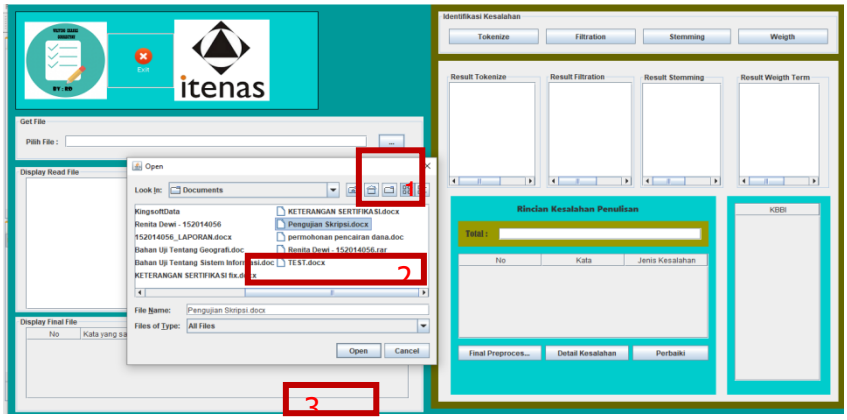
Gambar 4. GUI Sistem Lematisasi Dokumen

4.2 Pengujian Fungsional Sistem

Pengujian yang dilakukan adalah *black box testing* yaitu menguji fungsionalitas pada aplikasi yang telah dibuat, serta untuk menguji apakah fungsionalitas pada aplikasi sesuai dengan perancangan yang telah dibuat.

1. Pengujian tombol pilih file

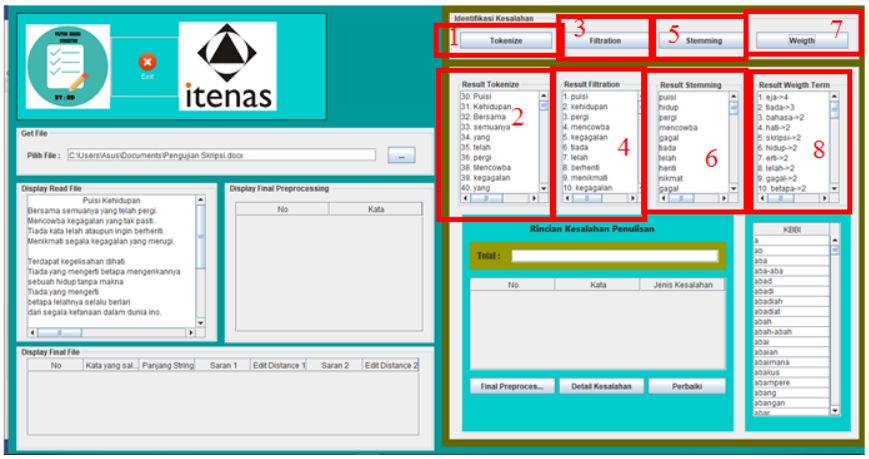
Pada proses buka file, *user* dapat memasukkanfile dalam format *.doc maupun *.docx kemudian tulisan yang ada didalam file tersebut akan ditampilkan pada *text area* yang dijelaskan pada gambar 5 dan 6.



Gambar 5. Pengujian Fungsional Buka File

2. *Text preprocessing*

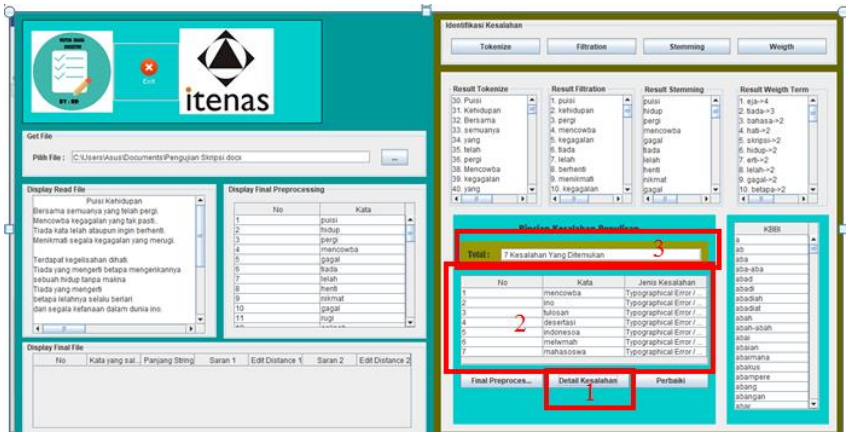
Pada tahapan ini, user dapat mengambil kata-kata yang sudah tepat penulisannya berdasarkan kata dasar yang diperoleh dan juga kata-kata yang tidak tepat yang dijelaskan pada gambar 6.



Gambar 6. Pengujian Fungsional Text Processing

3. Deteksi Kesalahan Penulisan

Pada proses selanjutnya adalah melakukan deteksi kesalahan penulisan dengan menggunakan data dari hasil *preprocessing* yang dibandingkan dengan KBBI. Apabila kata tersebut tidak ada di dalam KBBI maka akan diberikan ciri sebagai *typographical/ morphology error* seperti yang terlihat pada Gambar 7 berikut:



Gambar 7. Pengujian Fungsional Deteksi Kesalahan Penulisan

4.2 Pengujian Sistem

Dalam pengujian dilakukan menggunakan 10 file yang berbeda dengan total kata yang diuji adalah 19682 dan kata yang teridentifikasi salah adalah 3057 seperti yang terlihat pada Tabel 2.Terdapat tiga kategori kesalahan yaitu kesalahan penulisan (*typo* atau *morphology*), bahasa asing yang merupakan kumpulan kata-kata dari bahasa asing masuk ke dalam kategori kata yang salah, dan juga lain-lain merupakan untuk kategori penulisan nama, singkatan, dan lain-lain yang tidak tercantum dalam KBBI Denganmenggunakan rumus × 100% diterapkan pada ketiga jenis kesalahan yang terdeteksi yaitu :

- *Typo / morphology error* =$\frac{498}{3057}$× × 100% =16.291 %
- Bahasa Asing = $\frac{1388}{3057}$× 100% = 45.404 %
- Lain-lain = $\frac{1171}{3057}$× 100% = 38.306 %

Dapat disimpulkan bahwa tingkat kesalahan yang terdeteksi yang paling banyak adalah pada kata bertipe bahasa asing dengan tingkat kesalahan yaitu 45.306%.

Tabel 2. Pengujian Deteksi Kesalahan

No	Materi	Total Kata	Wrong Word	Typo / morph error	Bahasa asing	Lain-Lain
1	Sastra	4504	504	50	160	294
2	Geografi	1102	162	20	75	67
3	Sistem Informasi	1451	608	95	250	263
4	Teknologi	2500	430	50	221	159
5	Sejarah	1500	500	102	300	98
6	Ekonomi	1765	201	41	63	97
7	Biologi	1200	375	52	203	120
8	Kimia	1250	141	52	65	24
9	Matematika	2109	58	12	21	25
10	Bahasa	2301	78	24	30	24
Total		**19682**	**3057**	**498**	**1388**	**1171**

5. KESIMPULAN

Dalam menyusun suatu kata untuk menjadi kalimat yang sesuai sangat diharapkan untuk memperhatikan pemisahan maupun penggabungan kata dan juga penggunaan imbuhan awalan (prefiks), sisipan (infiks), akhiran (sufiks), maupun penggunaan imbuhan awalan dan akhiran yang pemakaiannya sekaligus. Untuk menentukan letak kesalahan dalam penggunaan ejaan yang tidak sesuai dengan KBBI terdapat beberapa proses diantaranya yaitu dengan melakukan *textpreprocessing* dan *lemmatization* dimana dalamprosesnya terdapat pemotongan kata dengan imbuhan yang digunakan, apabila hasil dari pemotongan imbuhan yaitu kata dasar yang terdapat di dalam KBBI maka kata tersebut benar, namun apabila dari pemotongan imbuhan terhadap kata yang diuji menghasilkan suatu kata yang tidak ada didalam KBBI maka akan tergolong kedalam *typographical /morphology error.*|

DAFTAR PUSTAKA

Adikara, Putra Pandu., 2017, Kesalahan Umum dan Perbaikannya dalam Penulisan Karya Ilmiah (Skripsi, Tesis, dll), Department of Informatics, Faculty of Computer Science, Brawijaya University (online)

Andri., dkk., 2014. Aplikasi Koreksi Kesalahan Berbasis Pada Tulisan Berbahasa Indonesia Untuk Meningkatkan Kualitas Penulisan Karya Ilmiah.

Anonim, 2017, Penggunaan Bahasa Indonesia Yang Baik dan Benar (dengan contoh), Mari Kita Belajar (MARKIJAR) (online)

Fahma, Arina Indana., dkk., 2017. Identifikasi Kesalahan Penulisan Kata (Typographical Error)pada Dokumen Berbahasa Indonesia Menggunakan Metode N-Gram dan Levenshtein Distance.

Fayyadh, 2017, Penulisan Artikel Sesuai EYD &Contohnya, Writer Artikel (online), (https://contentwritter.com/penulisan-artikel-sesuai-eyd-contohnya/, diakses 14 Oktober 2017)

Levenshtein, Vladimir., 1965, Binary Codes Capable of Correcting Deletions, Insertions, andReversal. Russia: Soviet Physics Doklady.

Peggy., Hansun, Seng., 2015, Optimasi Pencarian Kata Pada Aplikasi Penerjemah Bahasa Mandarin– Indonesia Berbasis Android Dengan Algoritma Levenshtein Distance

Suhartono, Derwin., 2014. Lemmatization Technique in Bahasa: Indonesian Language (JOURNAL OF SOFTWARE, VOL.9, NO.5).

Susanti, Ratna., 2015., Kesalahan Penggunaan EYD Dalam Karya Ilmiah Mahasiswa Politeknik Indonusa Surakarta.

Wijayanti, Atrianing Yessi., 2016., Analisis Kesalahan Penggunaan Ejaan Pada Skripsi Mahasiswa Program Studi di Pendidikan Guru Sekolah Dasar Fakultas Keguruan dan Ilmu Pendidikan Universitas Darul Ulum Islamic Cente Sudirman GUPPI UNDARIS.

Wirastuti, Intan., 2013. Analisis Kesalahan Berbahasa Pada Penulisan Latar Belakang Skripsi Mahasiswa Non Bahasa dan Sastra Indonesia Universitas Muhammadiyah Surakarta