



NORMALISASI TEKS PADA TEKS TWITTER BERBAHASA INDONESIA MENGGUNAKAN ALGORITME JARAK STRING PADA R

© Hak cipta milik IPB (Institut Pertanian Bogor)

TRI SONY SARAGIH



**DEPATEMEN ILMU KOMPUTER
MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT PERTANIAN BOGOR
BOGOR
2017**

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



© Hak cipta milik IPB (Institut Pertanian Bogor)

Bogor Agricultural U

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



PERNYATAAN MENGENAI SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa skripsi berjudul Normalisasi Teks pada Twitter Berbahasa Indonesia Menggunakan Algoritme Jarak *String* pada R adalah benar karya saya dengan arahan dari komisi pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir skripsi ini.

 Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Juli 2017

Tri Sony Saragih
NIM G64130020



1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



© Hak cipta milik IPB (Institut Pertanian Bogor)

Bogor Agricultural U

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



ABSTRAK

TRI SONY SARAGIH. Normalisasi Teks pada Twitter Berbahasa Indonesia Menggunakan Algoritme Jarak *String* pada R. Dibimbing oleh MUHAMMAD ABRAR ISTIADI.

Data *tweet* yang ada pada Twitter sering digunakan untuk keperluan *text mining*. Salah satu contohnya adalah untuk klasifikasi. Namun, data *tweet* tersebut sering menggunakan kata yang tidak baku sesuai bahasa Indonesia sehingga sulit digunakan untuk *text mining*. Oleh karena itu perlu dibangun sebuah fungsi yang dapat mengubah setiap kata yang tidak baku tersebut menjadi kata baku. Implementasi dalam pengubahan kata tidak baku menjadi baku pada penelitian ini menggunakan algoritme jarak *string* yang ada dalam pemrograman R. Data yang digunakan berupa yang kamus berisi kata *slang* dan perbaikannya beserta kamus yang berisi kata baku. Algoritme jarak *string* bekerja untuk membandingkan dua *string* dalam menentukan perbedaan jarak sehingga diperoleh jarak kedua *string*. Namun, pengubahan *string* pada penelitian ini tidak hanya berdasarkan jarak antar *string*, tetapi juga melakukan perubahan kata berdasarkan kamus kata *slang*. Penelitian ini melakukan normalisasi *tweet* dalam bahasa Indonesia. Terdapat 200 kata tidak baku dari Twitter yang digunakan untuk pengujian fungsi. Hasil menunjukkan bahwa nilai akurasi tertinggi pada penelitian ini adalah 69% dengan menggunakan metode *longest common substring* (*lcs*) dan kamus korpus Kompas yang sesuai KBBI.

Kata kunci: jarak *string*, kata baku, kata tidak baku, pengubahan kata, Twitter

ABSTRACT

TRI SONY SARAGIH. Normalization of Twitter Text in Indonesian Language Using String Distance Algorithm in R. Supervised by MUHAMMAD ABRAR ISTIADI.

Twitter text data (tweets) are often used for text mining, for example, text classification. However, tweets often contain incorrect words according to the Indonesian language, and thus, it is difficult to be used for text mining. Therefore, a function to transform incorrect words to correct words is needed. The transformation of the incorrect words in this study uses string distance algorithm in R. In addition, a dictionary containing slang words and a dictionary containing correct words are used. String distance algorithms can be used to compare two strings to measure their distance. However, in this study, the strings are normalized not only based on the distance between the strings but also based on slang dictionary. In this study, we only consider tweets in the Indonesian language. Two hundred incorrect words from Twitter were used for testing. Results show that the highest accuracy obtained in this study was 69% using longest common substring (*lcs*) method and Kompas dictionary which corresponds to KBBI dictionary.

Keywords: string distance, correct word, incorrect word, inversion word, Twitter



© Hak cipta milik IPB (Institut Pertanian Bogor)

Bogor Agricultural U

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



NORMALISASI TEKS PADA TWITTER BERBAHASA INDONESIA MENGGUNAKAN ALGORITME JARAK STRING PADA R

© Hak cipta milik IPB (Institut Pertanian Bogor)

Bogor Agricultural U

TRI SONY SARAGIH

Skripsi
sebagai salah satu syarat untuk memperoleh gelar
Sarjana Komputer
pada
Departemen Ilmu Komputer

**DEPARTEMEN ILMU KOMPUTER
MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT PERTANIAN BOGOR
BOGOR
2017**

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



© Hak cipta milik IPB (Institut Pertanian Bogor)

Bogor Agricultural U

Pengaji : 1 Husnul Khotimah, SKomp MKom

2 Dr Medria Kusuma Dewi Hardhienata, SKomp

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



Judul Skripsi: Normalisasi Teks pada Twitter Berbahasa Indonesia Menggunakan Algoritme Jarak String pada R
Nama : Tri Sony Saragih
NIM : G64130020

© Hak cipta milik IPB (Institut Pertanian Bogor)

Disetujui oleh

Muhammad Abrar Istiadi, SKomp MKom
Pembimbing

Disetahui oleh



Drs. Agus Duono, MSi MKom
Ketua Departemen

Tanggal Lulus: 19 JUL 2017

Bogor Agricultural U

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



© Hak cipta milik IPB (Institut Pertanian Bogor)

Bogor Agricultural U

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



PRAKATA

Puji dan syukur penulis panjatkan kepada Tuhan Yang Maha Esa atas segala karunia-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul “Normalisasi Teks pada Twitter Berbahasa Indonesia Menggunakan Algoritme Jarak String pada R”. Skripsi ini disusun sebagai syarat mendapat gelar Sarjana Komputer (SKomp) pada Program Sarjana Ilmu Komputer di Fakultas Matematika dan Ilmu Pengetahuan Alam Institut Pertanian Bogor (IPB).

Terima kasih penulis ucapkan kepada Bapak Muhammad Abrar Istiadi, SKomp MKom selaku pembimbing yang telah memberikan bimbingan, saran, arahan, dan bantuan selama penyelesaian skripsi, serta Ibu Husnul Khotimah, SKomp MKom dan Ibu Dr Medria Kusuma Dewi Hardhienata, SKomp yang telah berkenan menjadi penguji. Ungkapan terima kasih juga disampaikan kepada ayah, ibu,kakak, adik, serta seluruh keluarga yang telah memberikan dukungan, doa, motivasi, dan kasih sayangnya. Terima kasih juga disampaikan kepada Yohannes Bela Kurniawan atas bantuannya, rekan satu bimbingan Fakhri Izzudin, Faldhi Gifari, serta rekan-rekan seperjuangan di Ilmu Komputer IPB angkatan 50 atas segala kebersamaan, bantuan, dukungan, serta kenangan bagi penulis selama menjalani masa studi.

Semoga karya ilmiah ini bermanfaat.

Bogor, Juli 2017

Tri Sony Saragih

Q
Hak Cipta

Hak Cipta Milik IPB (Institut Pertanian Bogor)

Bogor Agricultural U

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



© Hak cipta milik IPB (Institut Pertanian Bogor)

Bogor Agricultural U

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



DAFTAR ISI

DAFTAR TABEL	vi
DAFTAR GAMBAR	vi
DAFTAR LAMPIRAN	vi
PENDAHULUAN	1
Latar Belakang	1
Perumusan Masalah	2
Tujuan Penelitian	2
Manfaat Penelitian	2
Ruang Lingkup Penelitian	2
METODE	2
Pengumpulan Data	3
Praproses Data	3
Normalisasi Teks	3
Evaluasi Normalisasi Teks	7
Lingkungan Pengembangan	7
HASIL DAN PEMBAHASAN	8
Pengumpulan Data	8
Praproses Data	9
Normalisasi Teks	10
Evaluasi Normalisasi Teks	13
SIMPULAN DAN SARAN	16
Simpulan	16
Saran	16
DAFTAR PUSTAKA	16
LAMPIRAN	18
RIWAYAT HIDUP	24

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.

b. Pengutipan tidak merugikan kepentingan yang wajar IPB.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.

b. Pengutipan tidak merugikan kepentingan yang wajar IPB.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.

Hak Cipta Dilindungi Undang-Undang

DAFTAR TABEL

1	Perbandingan 10 metode <i>stringdist</i> menggunakan data kamus KBBI	13
2	Perbandingan 10 metode <i>stringdist</i> menggunakan kamus korpus Kompas	14
3	Perbandingan 10 metode <i>stringdist</i> menggunakan kamus korpus Kompas sesuai KBBI	15

DAFTAR GAMBAR

1	Tahapan penelitian	3
2	Perintah pengambilan data <i>tweet</i> pada RStudio	8
3	Tampilan Twitter <i>apps</i>	9
4	Tahapan algoritme normalisasi gabungan	11
5	Contoh pemanggilan fungsi untuk 1 kata	13
6	Contoh pemanggilan fungsi 1 kalimat	13

DAFTAR LAMPIRAN

1	Contoh kamus KBBI	18
2	Contoh kamus Kompas	19
3	Contoh kamus kata <i>slang</i>	20
4	Kata uji dari <i>tweet</i>	21
5	Contoh data <i>tweet</i> mengandung kata tidak baku	22
6	Contoh kamus Kompas sesuai KBBI	23



PENDAHULUAN

Latar Belakang

Twitter merupakan salah satu media sosial yang digunakan oleh masyarakat untuk berkomunikasi, mengutarakan pendapat, dan mengutarakan perasaan pengguna. Dari berbagai jenis media sosial, Twitter merupakan salah satu yang paling populer. Hal ini dibuktikan dengan adanya 110 juta *tweet* per hari dan jumlah pengguna lebih dari 200 juta (Sarwani dan Mahmudy 2015). Namun, keterbatasan jumlah karakter tulisan dalam Twitter yaitu hanya 140 karakter menyebabkan pengguna Twitter sering melakukan penyingkatan. Singkatan tersebut mengakibatkan kata menjadi tidak baku (Wahyuningtyas 2016). Selain itu, penyebab ketidakbukan kata adalah kesalahan pengetikan atau sering disebut *typo*.

Sebelumnya penelitian terkait telah dilakukan oleh Adriyani *et al.* (2012). Penelitian tersebut membangun sistem berbasis web untuk menampilkan saran perbaikan kesalahan pengetikan dokumen berbahasa Indonesia. Penelitian tersebut menjelaskan adanya kesalahan pengetikan yang disebabkan oleh beberapa faktor seperti letak huruf pada *keyboard* yang berdekatan, kesalahan karena kegagalan mekanis atau slip dari tangan atau jari, kesalahan yang disebabkan oleh ketaksengajaan, dan kesalahan pengetikan karena kurangnya spasi sehingga kata tersebut tidak memiliki arti.

Dalam implementasi penelitian Adriyani *et al.* (2012), pengecekan *string* dilakukan per kata. Sebelumnya kalimat masukan masuk ke dalam tahap *preprocessing* terlebih dahulu. *Preprocessing* merupakan tahap penghilangan tanda baca dan tokenisasi terhadap setiap kata. Selanjutnya setiap token dilakukan pencocokan *string* ke basis data menggunakan algoritme *levenshtein distance* dan metode empiris. Setelah itu ditampilkan saran kata baku yang paling mendekati dengan kata masukan. Cara normalisasi pada penelitian terkait ini adalah untuk mengganti sebuah kata yang tidak baku menjadi kata baku, harus dilakukan pemilihan kata pada saran kata yang ditampilkan pada web tersebut. Selanjutnya harus melakukan pengubahan kata yang telah dipilih sebelumnya sehingga kata yang salah sebelumnya diganti dengan saran kata yang telah dipilih. Dengan demikian, cara kerja yang dilakukan masih melibatkan pengguna dalam pemilihan kata.

Selanjutnya penelitian terkait yang dilakukan oleh Aziz (2013). Tujuan penelitian tersebut untuk membuat sistem yang mampu menyaring adanya suatu entitas tertentu yang ada di dalam *tweet* dan mengklasifikasikan sentimen *tweet* tersebut. Namun, sebelum melakukan pengklasifikasian sentimen *tweet* tersebut, dilakukan normalisasi teks terhadap data *tweet*. Tujuannya adalah mengubah *tweet* yang awalnya sulit dimengerti menjadi *tweet* yang mudah dimengerti. Hal ini disebabkan adanya kesalahan penulisan kata pada *tweet* yang tidak sesuai dengan kata baku dalam bahasa Indonesia.

Untuk melakukan normalisasi teks, pada penelitian Aziz (2013) dibuat sebuah kamus yang berisi kata yang tidak baku dan perbaikan kata baku dari kata yang tidak baku tersebut. Proses penggantian kata tidak baku menjadi kata baku dilakukan dengan mencari kata yang tidak baku pada *tweet* kemudian menggantinya dengan kata baku yang ada di kamus. Kelemahannya yaitu jika kata

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.

b. Pengutipan tidak merugikan kepentingan yang wajar IPB.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.

tidak baku pada *tweet* tidak ditemukan pada kamus, kata tidak baku tersebut tidak diganti menjadi kata baku.

Oleh sebab di atas, dibutuhkan suatu fungsi yang mampu melakukan normalisasi teks menjadi bentuk kata baku pada data *tweet* karena data *tweet* banyak digunakan untuk keperluan *text mining*. Untuk melakukan normalisasi teks pada penelitian ini digunakan algoritme jarak *string* yang ada dalam R. Algoritme jarak *string* ini dibutuhkan untuk menghitung jumlah perbedaan jarak antar *string* yang selanjutnya digunakan untuk pengubahan *string*.

Perumusan Masalah

- Rumusan masalah dalam penelitian ini adalah:
- 1 Bagaimana mengubah kata tidak baku pada teks di Twitter menjadi kata baku dengan menggunakan algoritme jarak *string* yang ada di R.
 - 2 Bagaimana mengimplementasikan pengubahan kata baku pada pemrograman R.

Tujuan Penelitian

- Tujuan penelitian ini adalah:
- 1 Menerapkan algoritme jarak *string* untuk mengoreksi kata yang sesuai untuk kata tidak baku tertentu.
 - 2 Membuat fungsi yang mampu mengubah kata tidak baku pada teks di Twitter menjadi kata baku.

Manfaat Penelitian

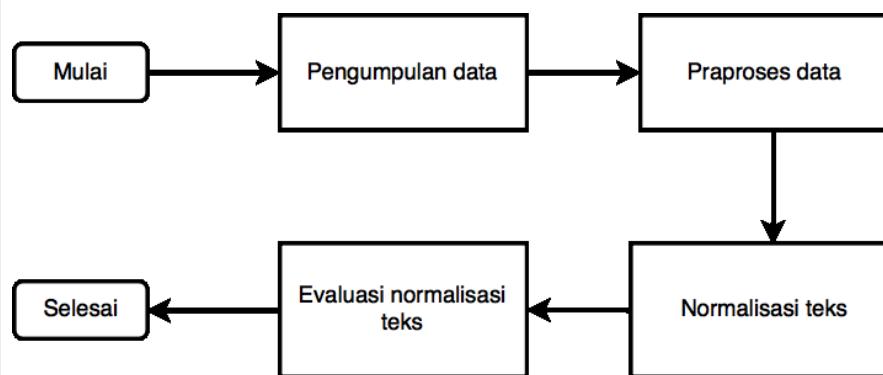
Penelitian ini diharapkan dapat membantu memperbaiki kata tidak baku pada Twitter menjadi kata baku sehingga kata-kata yang dimuat dalam *tweet* lebih mudah untuk diklasifikasikan atau dijadikan keperluan *text mining*.

Ruang Lingkup Penelitian

- Lingkup dari penelitian ini, yaitu:
- 1 Penelitian ini menggunakan data dari Twitter dengan teks dalam bahasa Indonesia sebagai bahan untuk pengujian fungsi apakah berjalan sesuai dengan yang diharapkan atau tidak.
 - 2 Penelitian ini melakukan normalisasi hanya dalam bentuk huruf.
 - 3 Masukan bukan termasuk singkatan dalam bahasa Indonesia.

METODE

Metode yang digunakan dalam penelitian ini terdiri atas 4 tahap yaitu: pengumpulan data, praproses data, normalisasi teks, dan evaluasi normalisasi teks. Alur penelitian dapat dilihat pada Gambar 1.



Gambar 1 Tahapan penelitian

Pengumpulan Data

Tahapan yang dilakukan pertama kali adalah mengumpulkan data kata baku dalam Bahasa Indonesia, kata *slang* dan beberapa *tweet* dari Twitter. Kemudian melakukan perancangan basis data. Perancangan basis data dalam hal ini merupakan pembuatan tabel data untuk menyimpan kata baku yang telah dikumpulkan sebelumnya. Tabel data dibuat dalam format *comma separated value* (csv) sehingga mudah untuk dimasukkan ke dalam pemrograman R. Data tersebut digunakan sebagai kamus data dalam pencocokan *string*.

Praproses Data

Tahapan ini merupakan pengubahan semua isi kamus menjadi huruf kecil atau *lower case*. Hal ini bertujuan menyeragamkan semua huruf agar lebih mudah melakukan pengecekan. Pada tahapan ini dilakukan penghilangan tanda baca secara manual. Selanjutnya dilakukan penyeleksian kata yang ada dalam kamus korpus Kompas yang sesuai dengan KBBI.

Normalisasi Teks

Normalisasi Teks Menggunakan Kamus Kata Tidak Baku

Normalisasi teks pada tahapan ini merupakan pengubahan kata tidak baku menjadi baku berdasarkan kamus kata tidak baku yang dikumpulkan dari penelitian sebelumnya. Masukan yang diberikan pada fungsi dicek keberadaannya pada kamus kata *slang* atau kamus kata tidak baku. Kamus tersebut berisi kata tidak baku dan perbaikan dari kata tidak baku tersebut. Perbaikan kata tidak baku tersebut sudah merupakan kata baku dalam bahasa Indonesia.

Normalisasi Teks Menggunakan Jarak *String* pada R

Algoritme jarak *string* pada R disebut dengan *stringdist*. *Stringdist* telah tersedia pada pemrograman R sehingga membantu proses pencocokan *string* dengan cara memanggil fungsi *stringdist* pada program yang dibangun. Dalam pembangunan *library stringdist* diberikan 10 metode jarak yang berbeda (Loo 2014). Berikut di bawah ini adalah 10 metode yang diimplementasikan pada *stringdist*.

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



- Hak Cipta Dilindungi Undang-Undang**
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.

1 Metode *lv* (*Levenshtein distance*)

Menurut Levenshtein (1966), algoritme *levenshtein distance* melakukan tiga operasi, yaitu penyisipan, penghapusan, dan permutasi. Algoritme ini membandingkan antara dua *string* untuk melakukan perhitungan terhadap jarak kedua *string*. Algoritme *levenshtein distance* bekerja untuk menghitung jumlah minimum pentransformasian suatu *string* menjadi *string* lain. Pentransformasian dapat merupakan penggantian, penghapusan, dan penyisipan. Nilai jarak *levenshtein distance* dihitung menggunakan persamaan:

$$\text{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{jika } \min(i,j) = 0, \\ \min \left\{ \begin{array}{l} \text{lev}_{a,b}(i-1,j) + 1 \\ \text{lev}_{a,b}(i,j-1) + 1 \\ \text{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{array} \right\} & \text{selainnya} \end{cases} \quad (1)$$

dengan nilai $1_{(a_i \neq b_j)}$ sama dengan 0 ketika $a_i = b_j$ dan sama dengan 1, jika sebaliknya. Nilai i merupakan indeks *string* a ke- i dan nilai j merupakan indeks *string* b ke- j .

2 Metode *osa* (*Optimal string alignment*)

Metode *osa* mirip dengan metode *levenshtein distance*, tetapi metode ini memungkinkan melakukan transposisi dengan karakter yang berdekatan. Metode ini bekerja dengan mengubah setiap *substring* hanya sekali saja. Nilai jarak *optimal string alignment* dihitung menggunakan persamaan:

$$\text{osa}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{jika } \min(i,j) = 0, \\ \min \left\{ \begin{array}{l} \text{osa}_{a,b}(i-1,j) + 1 \\ \text{osa}_{a,b}(i,j-1) + 1 \\ \text{osa}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \\ \text{osa}_{a,b}(i-2,j-2) + 1 \text{ jika } a_i = b_{j-1}, a_{i-1} = b_j \end{array} \right\} & \text{selainnya.} \end{cases} \quad (2)$$

3 Metode *dl* (*Full damerau levenshtein distance*)

Metode ini mirip dengan metode *osa*, tetapi dapat melakukan beberapa kali penyuntingan terhadap setiap *substring*. Nilai jarak *damerau levenshtein distance* dihitung menggunakan persamaan:

$$\text{dl}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{jika } \min(i,j) = 0, \\ \min \left\{ \begin{array}{l} \text{dl}_{a,b}(i-1,j) + 1 \\ \text{dl}_{a,b}(i,j-1) + 1 \\ \text{dl}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \\ \min_{(i,j) \in \Delta} \text{dl}(i-1,j-1) + [(|a| - i) + (|b| - j) - 1] \end{array} \right\} & \text{selainnya} \end{cases} \quad (3)$$

dengan $|a|$ merupakan panjang *string* a dan $|b|$ merupakan panjang *string* b .

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.

4 Metode *hamming distance*

Metode ini bekerja dengan mengukur jarak antar dua *string* yang memiliki panjang *string* yang sama dan membandingkan setiap simbol yang ada pada setiap *string* dengan posisi yang sama. *Hamming distance* dari dua *string* adalah jumlah simbol dari kedua *string* yang berbeda. Sebagai contoh *Hamming distance* antara *string* “toned” dan “roses” adalah 3. Metode ini juga digunakan untuk mengukur jarak antar dua *string binary* misalnya jarak antara 10011101 dengan 10001001 adalah 2. Nilai jarak *hamming distance* dihitung menggunakan persamaan:

$$\text{hamming}_{a,b}(i,j) = \sum_{i=1}^{|a|} [1 - \delta(a_i, b_j)] \text{ jika } |a| = |b|, \\ \infty \text{ selainnya} \quad (4)$$

dengan $\delta(a_i, b_j) = 1$ jika $a_i = b_j$ dan 0 selainnya.

5 Metode *lcs* (*Longest common substring distance*)

Metode ini dapat memperoleh *string* terpanjang dengan cara memasangkan karakter *a* dan *b*. Selain itu, metode ini juga dapat mempertahankan urutan dari setiap karakter. Penghitungan jaraknya sama seperti metode *edit distance* dengan bobot sama dengan 1 dan bekerja hanya pada penghapusan dan penyisipan. Nilai jarak *longest common substring distance* dihitung menggunakan persamaan:

$$\text{lcs}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{jika } \min(i,j) = 0, \\ \text{lcs}_{a,b}(i-1, j-1) & \text{jika } a_i = b_j, \\ 1 + \min\{\text{lcs}_{a,b}(i-1, j), \text{lcs}_{a,b}(i, j-1)\} & \text{selainnya.} \end{cases} \quad (5)$$

6 Metode *qgram* (*q-gram distance*)

Metode ini bekerja dengan mengambil *substring* *q* berurutan dari sebuah *string* awal. Perhitungan dibatalkan ketika *q* lebih besar daripada panjang dari *string*. Nilai jarak *q-gram distance* dihitung menggunakan persamaan:

$$\text{qgram}(a, b, q) = \|v(a; q) - v(b; q)\|_1 = \sum_{i=1}^{|\Sigma|^q} |v_i(a; q) - v_i(b; q)| \quad (6)$$

dengan $v(a; q)$ menjadi set unik *q-gram* dalam *string* *a* dan $v(b; q)$ adalah set unik *q-gram* dalam *string* *b*.

7 Metode *cosine* (*Cosine distance*)

Metode ini bekerja berdasarkan metode *q-gram*. Nilai jarak *cosine distance* dihitung menggunakan persamaan:

$$\cos(a, b, q) = 1 - \frac{v(a; q) \cdot v(b; q)}{\|v(a; q)\|_2 \|v(b; q)\|_2} \quad (7)$$



1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.

dengan $v(a; q)$ dan $v(b; q)$ dibulatkan ke atas.

8 Metode *jaccard* (*Jaccard distance*)

Metode ini bekerja berdasarkan metode *q-gram*. Namun perhitungan nilai jarak *jaccard distance* dihitung menggunakan persamaan:

$$\text{jaccard}(a, b, q) = 1 - \frac{|Q(a; q) \cap Q(b; q)|}{|Q(a; q) \cup Q(b; q)|} \quad (8)$$

dengan $Q(a; q)$ merupakan token *string* a setelah dibagi q dan $Q(b; q)$ merupakan token *string* b setelah dibagi q .

9 Metode *jw* (*Jaro or Jaro-Winker distance*)

Metode ini disebut juga dengan istilah *jaro-distance*. Nilai jarak *jaro distance* dihitung menggunakan persamaan:

$$\text{jaro}_{a,b}(i, j) = \begin{cases} \max(i, j) & \text{jika } \min(i, j) = 0, \\ 1 & \text{ketika } m = 0 \text{ dan } |a| + |b| > 0, \\ 1 - \frac{1}{3} \left(\frac{m}{|a|} + \frac{m}{|b|} + \frac{m - T}{m} \right) & \text{selainnya} \end{cases} \quad (9)$$

dengan m adalah jumlah karakter yang cocok antara *string* a dan *string* b . Nilai jarak *jaro-winkler distance* dihitung menggunakan:

$$\text{jw}_{a,b}(i, j) = \text{jaro}_{a,b}(i, j)[1 - p\ell] \quad (10)$$

dengan $\text{jaro}_{a,b}(i, j)$ adalah jarak *Jaro*. Pada formula tersebut, ℓ diperoleh dengan menghitung mulai dari masukan *string*, setelah diperoleh berapa banyak karakter dari karakter pertama yang tidak cocok antara dua string, dengan maksimal adalah 4. Faktor p adalah faktor penalti, dalam *Winkler* sering dipilih adalah 0.1.

10 Metode *soundex* (*distance based on soundex encoding*)

Metode *soundex* merupakan *string* yang diterjemahkan ke dalam sebuah kode *soundex*. Jarak antara *string* diberikan sama dengan 0 jika *string* tersebut memiliki kode *soundex* yang sama. Sebaliknya, jika kode *soundex* antar *string* berbeda diberi nilai 1. *Soundex recoding* hanya berlaku untuk karakter dalam rentang a-z dan A-Z. Kode *soundex* adalah:

- a, e, i, o, u, y, h, w → dihapus atau 0
- b, f, p, v → 1
- c, g, j, k, q, s, x, z → 2
- d, t → 3
- 1 → 4
- m, n → 5
- r → 6

Normalisasi Teks Gabungan

Tahapan ini merupakan penggabungan normalisasi berdasarkan kamus kata tidak baku dan berdasarkan implementasi algoritme jarak *string* yang ada pada R, yaitu *stringdist*. Cara normalisasi dilakukan dengan mengimplementasikan data *tweet* pada pemrograman R dalam bentuk satu kata maupun kalimat. Sebagai contoh *tweet* dengan kalimat yang tidak baku adalah “*ywdh itu udh pilihannya dia. nanti klw dia ada apa2 sama doi km nasehatin dgn baik. itu aja sih saran aku ya.*”. Dari kalimat tersebut didapatkan beberapa kata yang bukan baku seperti *ywdh*, *udh*, *klw*, *apa2*, *doi*, *km*, *nasehatin*, *dgn*, dan *aja*. Dengan adanya fungsi yang telah dibangun, setiap kata tidak baku tersebut diubah menjadi kata baku sesuai dengan keberadaanya pada kamus kata tidak baku ataupun kedekatannya terhadap kata yang ada di kamus. Contoh perbaikan kata yang diharapkan dari kata tidak baku tersebut adalah *ywdh* menjadi *ya sudah*, *udh* menjadi *sudah*, *klw* menjadi *kalau*, *apa2* menjadi *apa-apa*, *doi* menjadi *dia*, *km* menjadi *kamu*, *nasehatin* menjadi *menasihati*, *dgn* menjadi *dengan*, dan *aja* menjadi *saja*. Dengan demikian, teks tersebut lebih mudah digunakan untuk keperluan *text mining*.

Evaluasi Normalisasi Teks

Tahapan evaluasi normalisasi teks adalah tahap mengevaluasi fungsi terhadap *output* yang dihasilkan. Hal ini bertujuan mengetahui apakah *output* yang dihasilkan sudah sesuai dengan yang diharapkan. Cara melakukan evaluasi terhadap *output* yang dihasilkan adalah dengan membandingkan hasil normalisasi menggunakan algoritme jarak *string* dengan hasil normalisasi manual dan dihitung jumlah normalisasi yang benar untuk menghitung akurasi atau persentasenya. Perhitungan persentasenya menggunakan persamaan:

$$\text{Persentase} = \frac{\text{jumlah kata benar}}{\text{jumlah kata uji}} \times 100\%$$

dengan *jumlah kata benar* merupakan banyaknya kata hasil normalisasi yang sesuai dengan yang seharusnya dan *jumlah kata uji* merupakan banyaknya kata yang diujikan untuk dinormalkan.

Akurasi yang diperoleh menggunakan algoritme jarak *string* adalah dengan menguji coba masukan terhadap fungsi yang telah dibangun. Hasil manualnya atau tanpa menggunakan metode diperoleh dengan membuat daftar kata tidak baku dan perbaikan dari kata tidak baku tersebut.

Lingkungan Pengembangan

Lingkungan implementasi yang digunakan sebagai berikut:

Perangkat lunak:

- Microsoft Windows 10
- RStudio 1.0.44
- Excel 2013

Perangkat keras:

- Prosesor Intel® Celeron® CPU B820 @ 1.70GHz
- RAM 2 GB

HASIL DAN PEMBAHASAN

Pengumpulan Data

Data kamus diperoleh dari *database* aplikasi yaitu kamus Stardict¹ yang berisi kata baku berdasarkan Kamus Besar Bahasa Indonesia (KBBI) yang terdiri atas 71116 baris. Contoh isi kamus tersebut dapat dilihat pada Lampiran 1. Data tersebut terurut abjad. Data kamus yang lain diambil dari korpus Kompas yang terdiri atas 10000 kata (Lanin *et al.* 2013). Data tersebut berasal dari arsip berita tahun 2012 situs berita kompas.com. Contoh isi kamus korpus Kompas dapat dilihat pada Lampiran 2. Data korpus Kompas ini terurut berdasarkan tingkat keseringan penggunaan kata tersebut pada arsip berita Kompas yang dikumpulkan. Data kamus kata baku berguna sebagai bahan acuan menentukan apakah masukan yang diberikan ada pada kamus atau tidak. Selain itu, kamus kata baku juga sebagai acuan untuk menampilkan saran kata terdekat dari perbaikan kata tidak baku. Data kamus kata tidak baku yang diperoleh dari penelitian Aziz (2013) terdiri atas 3718 baris. Sebagian isi kamus kata *slang* atau kamus kata tidak baku dapat dilihat pada Lampiran 3. Data tersebut menjadi acuan jika kata masukan ada pada kamus kata tidak baku, keluaran yang ditampilkan adalah perbaikan dari kata tidak baku tersebut.

Pengujian dengan metode pengukuran jarak didapatkan melalui *tweet* yang dikumpulkan berdasarkan kata kunci tertentu. Pemilihan kata kunci tidak memilliki kriteria tertentu. Hasil pengumpulan mendapatkan 200 kata uji yang dapat dilihat pada Lampiran 4. Kata-kata tersebut tidak ada pada kamus kata baku dan juga tidak ada pada kamus *slang* atau kamus kata tidak baku. Data dari Twitter diambil dengan mengaktifkan/install terlebih dahulu paket twitteR pada RStudio. Sebelumnya harus dilakukan pendaftaran terlebih dahulu pada Twitter *apps* menggunakan akun pribadi. Setelah terdaftar ditampilkan *API key*, *API secret*, *access tokens* dan *access tokens secret* yang digunakan untuk menghubungkannya pada pemrograman RStudio. Setelah mengaktifkan paket twitteR pada RStudio, kemudian menjalankan perintah seperti Gambar 2. Kotak berwarna biru diisi dengan kode yang ditampilkan pada akun pribadi pada Twitter *apps*. Tampilan Twitter *apps* dapat dilihat pada Gambar 3.

```
> api_key = "REDACTED"
> api_secret = "REDACTED"
> access_token = "REDACTED"
> access_token_secret = "REDACTED"
> setup_twitter_oauth(api_key, api_secret, access_token, access_token_secret)
[1] "Using direct authentication"
> tweet<-searchTwitter("bogor")
```

Gambar 2 Perintah pengambilan data *tweet* pada RStudio

¹ Aplikasi Stardict pada softonic: <https://stardict.en.softonic.com/post-download?sl=1>

- Hak Cipta Dilindungi Undang-Undang © Hak Cipta Ilmiah IPB (Institut Pertanian Bogor)
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.

Application Management

Tri Sony S

Test OAuth

Details Settings **Keys and Access Tokens** Permissions

Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key)	[REDACTED]
Consumer Secret (API Secret)	[REDACTED]
Access Level	Read and write (modify app permissions)
Owner	trisonystrg
Owner ID	283033796

Gambar 3 Tampilan Twitter apps

Contoh kata acak yang dijadikan kata kunci untuk mengambil data *tweet* yaitu “baik”, “bogor”, “saya”, “kamu”, “tidak”, dan lain-lain. Berikut adalah salah satu contoh *tweet* yang diperoleh berdasarkan kata acak “baik” yaitu “*Prtumbuhn eknmi saat ini ddominasi olh nontradeable, tp sektor yg tdk baik trutama di sktor prtanian dan industri manufaktr*”. Dari *tweet* tersebut diambil kata tidak baku yaitu *prtumbuhn, eknmi, ddominasi, olh, nontradeable, tp, yg, tdk, trutama, sktor, prtanian, dan manufaktr*. Kata-kata tersebut berguna sebagai bahan uji fungsi dan pembanding setiap algoritme jarak *string* yang ada dalam R. Sebagian contoh data *tweet* dapat dilihat pada Lampiran 5.

Praproses Data

Semua kata yang merupakan isi kamus kata baku dan kamus kata tidak baku diubah menjadi *lower case*. Selain itu, setiap masukan diubah menjadi *lower case* yang diimplementasikan pada program. Hal ini bertujuan untuk menghindari ketidakseragaman saran kata yang ditampilkan. Di sisi lain masukan yang diberikan terhadap pemrograman R merupakan *case sensitive* sehingga memungkinkan terjadi kesalahan kata yang ditampilkan hanya karena perbedaan bentuk tulisan. Misal masukan yang diberikan salah satu pengguna adalah kata “SAya”. Misal sebagian pengguna menuliskan masukan dengan kata “saya”. Kedua kata tersebut memiliki makna yang sama. Namun, jika tidak diubah menjadi *lower case*, salah satu kata tersebut tidak terdapat dalam kamus dan menampilkan kata yang lain. Hal ini berpengaruh terhadap perhitungan jarak kedua kata tersebut sehingga perbaikan kata menjadi tidak sesuai dengan yang seharusnya. Kata perbaikan dari kata “SAya” yang ditampilkan oleh fungsi adalah kata “iya”.

Selain pengubahan huruf kecil, tanda baca yang ada pada *tweet* dihapus secara manual. Hal ini dilakukan karena di dalam kamus tidak terdapat tanda baca. Dengan demikian, tidak perlu dilakukan pengecekan maupun pengubahan tanda baca. Hal



1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.

ini bertujuan agar waktu yang diperlukan untuk pengecekan semakin cepat.

Tahapan selanjutnya dilakukan penyeleksian kata pada korpus Kompas yang sesuai dengan KBBI. Hal ini dilakukan karena kamus korpus Kompas yang diperoleh masih berisi kata-kata yang tidak baku dalam bahasa Indonesia, seperti kata “*and*”, “*you*”, dan lain-lain. Selain itu kamus tersebut masih berisi angka, simbol, nama orang, dan singkatan yang mengakibatkan saran kata yang ditampilkan berdasarkan kamus tidak sesuai dengan yang diharapkan. Jumlah kata pada kamus korpus Kompas yang diperoleh setelah dilakukan seleksi terhadap KBBI terdiri atas 5860 kata. Sebagian contohnya dapat dilihat pada Lampiran 6. Namun, setelah dilakukan penyeleksian kata tersebut, ada beberapa kata hilang yang merupakan kata baku. Kata yang hilang tersebut merupakan kata baku yang memiliki imbuhan. Hal ini terjadi karena kata yang berimbuhan tidak semua berada pada kamus KBBI.

Normalisasi Teks

Normalisasi Teks Menggunakan Kamus Kata Tidak Baku

Normalisasi teks dilakukan dengan mengecek masukan terhadap kamus kata *slang* atau kamus kata tidak baku dari penelitian Aziz (2013). Jika masukan ada pada kamus kata tidak baku, masukan tersebut tidak merupakan kata baku sesuai KBBI sehingga perlu dinormalkan atau dibakukan. Hasil normalisasi yang ditampilkan dari masukan kata tidak baku tersebut adalah perbaikan yang sudah merupakan kata baku. Misal, masukan yang diberikan adalah kata “*yg*”. Kata tersebut merupakan kata tidak baku dan berada pada kamus kata tidak baku. Dengan demikian perbaikan yang ditampilkan adalah perbaikan kata yang sesuai dengan kata “*yg*” yaitu “*yang*”.

Normalisasi Teks Menggunakan Jarak *String* pada R

Normalisasi teks untuk mengubah kata tidak baku menjadi kata baku dilakukan dengan mengimplementasikan algoritme jarak *string* yang ada pada pemrograman R. Setiap algoritme menampilkan jarak antara kata masukan dengan setiap kata yang ada pada kamus kata baku. Dari setiap jarak tersebut diambil jarak minimalnya dan menampilkan kata yang memiliki jarak minimal tersebut. Pengubahan kata tidak baku menjadi baku dilakukan dengan mencari *string* terdekat antara kata tidak baku dengan kata baku yang ada di kamus kata baku. Dengan demikian, perbaikan yang ditampilkan merupakan kata baku karena normalisasi yang dilakukan berdasarkan kamus kata baku.

Normalisasi Teks Gabungan

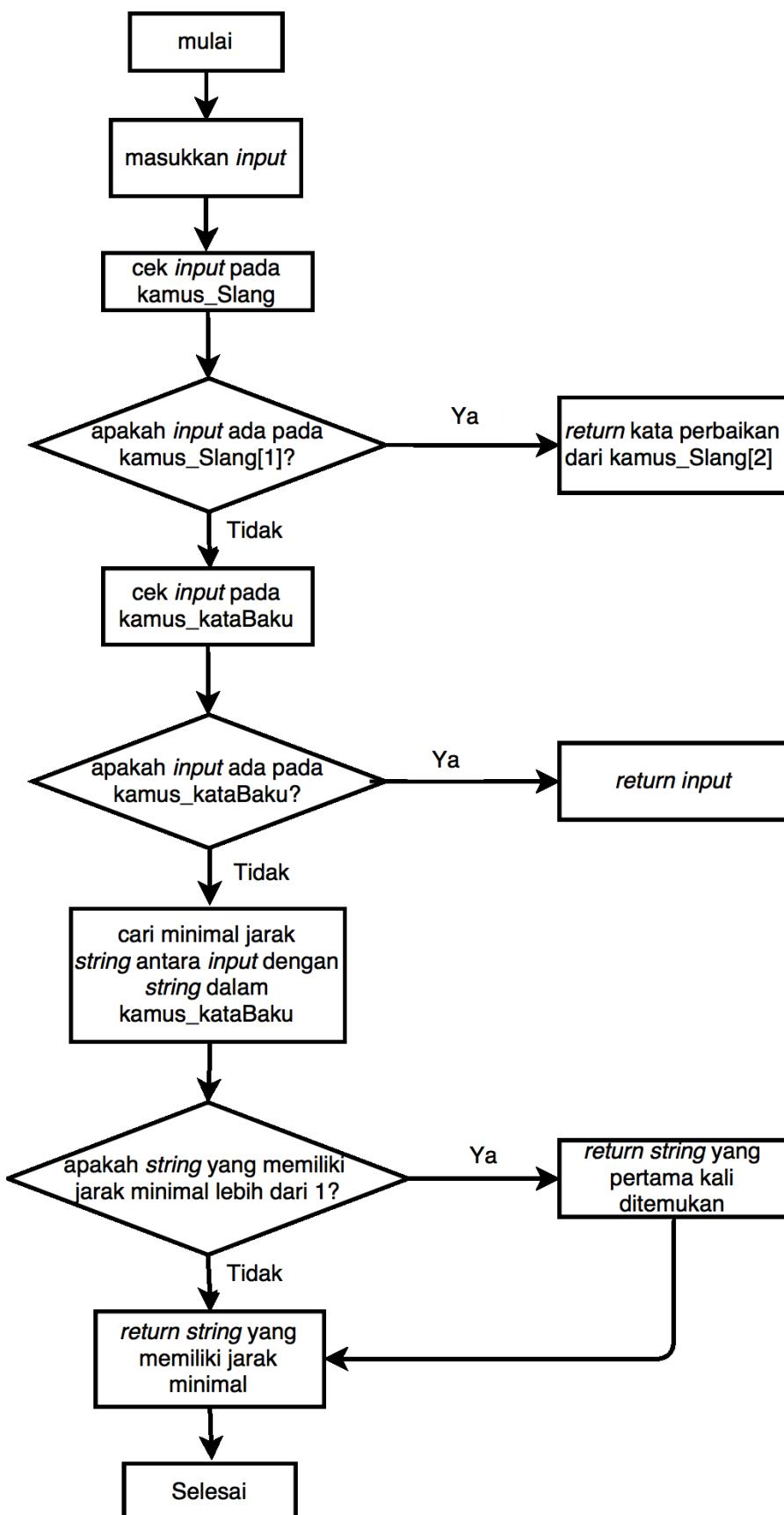
Adapun proses normalisasi teks yang dilakukan terhadap setiap masukan ditunjukkan pada Gambar 4. Pada Gambar 4, kamus_Slang[1] merupakan tabel kamus kata *slang* kolom pertama yang berisi kata tidak baku dan kamus_Slang[2] merupakan tabel kamus kata *slang* kolom kedua yang berisi perbaikan kata tidak baku. Sebelum dilakukan pengubahan kata tidak baku menjadi baku pada penelitian ini, dilakukan pengecekan terhadap masukan yang merupakan angka. Jika masukan merupakan angka, angka tersebut tidak diganti dengan kata ataupun angka lain, melainkan tetap seperti semula. Dengan demikian masukan angka tidak dilakukan normalisasi. Selanjutnya, tahapan normalisasi dimulai dengan melakukan

© Hak cipta milik IPB (Institut Pertanian Bogor)

Bogor Agricultural U

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



Gambar 4 Tahapan algoritme normalisasi gabungan



1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

- a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
- b. Pengutipan tidak merugikan kepentingan yang wajar IPB.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.

pengecekan pada kamus kata *slang* atau kamus kata tidak baku. Jika masukan ada pada kamus kata *slang*, fungsi ini menampilkan perbaikan dari kata tidak baku tersebut. Contoh masukan kata tidak baku “*elu*”. Kata “*elu*” berada pada kamus kata *slang*. Dengan demikian, perbaikan yang ditampilkan adalah kata baku dari kata “*elu*” yaitu “*kamu*”. Selanjutnya, jika masukan tidak ada pada kamus kata *slang*, masukan dicek pada kamus kata baku. Jika masukan ada pada kamus kata baku, fungsi ini menampilkan kata semula. Contoh masukan kata “*saya*”. Kata tersebut sudah merupakan kata baku dan berada pada kamus kata baku sehingga tidak perlu dilakukan perubahan.

Jika masukan tidak berada pada kamus kata *slang* dan tidak berada pada kamus kata baku, langkah selanjutnya adalah mencari minimal jarak antara masukan dengan kata yang ada pada kamus kata baku. Kemudian dilakukan pengecekan kembali untuk melihat apakah jarak minimal masukan dengan kata yang ada dalam kamus hanya satu kata atau lebih. Jika jarak minimal terdiri atas satu kata, perbaikan kata yang ditampilkan adalah kata dengan jarak minimal tersebut. Namun, jika jarak minimal ada dua kata atau lebih, perbaikan kata yang ditampilkan adalah kata yang pertama kali ditemukan. Contoh masukan dengan kata tidak baku “*sambl*”. Kata tersebut memiliki jarak minimal lebih dari satu kata yang berada dalam kamus kata baku, yaitu kata “*sambil*” dan kata “*sambal*”. Kata pertama yang ditemukan adalah kata “*sambil*”, maka perbaikan yang ditampilkan adalah kata “*sambil*”. Kata yang memiliki jarak minimal yang ditampilkan tidak terurut sesuai abjad karena kamus yang dipakai merupakan kamus korpus Kompas, yaitu sesuai tingkat keseringan pemakaian kata.

Pada penelitian ini pengecekan terhadap kamus kata tidak baku dilakukan terlebih dahulu dari pengecekan terhadap kamus kata baku. Hal tersebut disebabkan adanya beberapa kata yang memiliki arti yang ambigu. Misalnya, kata “*doi*” dapat termasuk kata tidak baku dan juga termasuk kata baku. Jika merupakan kata tidak baku, kata “*doi*” memiliki arti “*dia*”. Jika merupakan kata baku, kata “*doi*” memiliki arti “*uang*” menurut arti KBBI daring. Begitu juga dengan kata “*km*”. Kata “*km*” jika dicek terhadap kamus kata tidak baku, perbaikan yang ditampilkan adalah kata “*kamu*”. Namun, jika kata “*km*” dicek terhadap kamus kata baku KBBI, kata tersebut akan tetap “*km*” yang artinya adalah kilometer. Namun, contoh kasus kata tersebut lebih sering digunakan masyarakat pada umumnya sebagai kata tidak baku. Dengan demikian, pengecekan dilakukan terhadap kamus kata *slang* atau kamus kata tidak baku terlebih dahulu. Sehingga perbaikan yang ditampilkan lebih sesuai dengan yang diharapkan.

Implementasi Normalisasi Teks Sebagai Fungsi R

Berdasarkan urutan pengecekan yang ditunjukkan pada Gambar 4, dibangun sebuah fungsi dengan pemrograman R yang mampu melakukan normalisasi setiap masukan. Masukan dapat berupa huruf kapital atau *upper case*, tetapi keluarannya tetap *lower case*. Hal ini telah diatur dalam program sehingga bagaimanapun jenis karakter masukannya yaitu *lower case* atau *upper case*, keluarannya tetap *lower case*. Selain itu, masukan harus merupakan bahasa Indonesia karena data yang digunakan untuk normalisasi merupakan bahasa Indonesia. Pemanggilan fungsi dapat dilakukan dengan perintah seperti Gambar 5. Masukan berupa *string* sehingga

```
> katabaku("DGN")
[1] "dengan"
> katabaku("Dgn")
[1] "dengan"
> katabaku("dgn")
[1] "dengan"
```

Gambar 5 Contoh pemanggilan fungsi untuk 1 kata

diapit oleh tanda kutip (“ ”). Masukan dapat berupa huruf maupun angka, tetapi masukan yang dinormalkan adalah huruf. Katabaku merupakan nama fungsi yang dibangun untuk menampung masukan yang diberikan. Selain itu, fungsi telah mampu menerima masukan berupa kalimat yang ditunjukkan pada Gambar 6.

```
  
katabaku("Prtumbuhn eknmi saat ini ddominasi olh nontradeable tp
sektor yg tdk baik trutama di sktor prtanian dan industri manufak
")
] "pertumbuhan ekonomi saat ini dominasi oleh real tetapi sektor
yang tidak baik terutama di sektor pertanian dan industri manufak
"
```

Gambar 6 Contoh pemanggilan fungsi 1 kalimat

Jika dibandingkan dengan penelitian Aziz (2013) dengan masukan sama seperti pada Gambar 6, kata tidak baku yang tidak diubah yaitu “*prtumbuhan*”, “*eknmi*”, “*ddominasi*”, “*olh*”, “*nontradeable*”, “*sktor*”, “*prtanian*”, “*manufaktr*”. Hal ini terjadi karena kata tidak baku tersebut tidak ada dalam kamus kata tidak baku, sehingga tidak diubah menjadi kata baku.

Evaluasi Normalisasi Teks

Setelah membuat program yang dapat mengeksekusi setiap masukan, diperoleh hasil perbandingan dengan 10 metode yang ada pada *library stringdist* dalam R. Evaluasi normalisasi dilakukan hanya pada 200 kata yang dikumpulkan pada tahap sebelumnya. Berikut Tabel 1 adalah perbandingan setiap metode pengukuran jarak dengan menggunakan data KBBI sebagai kamus acuan.

Tabel 1 Perbandingan 10 metode *stringdist* menggunakan data kamus KBBI

No	Metode	Jumlah kata benar	Persentase
1	<i>Osa</i>	74	37%
2	<i>Lv</i>	75	37.5%
3	<i>Dl</i>	75	37.5%
4	<i>Hamming</i>	17	8.5%
5	<i>Lcs</i>	100	50%
6	<i>Qgram</i>	35	17.5%
7	<i>Cosine</i>	46	23%
8	<i>Jaccard</i>	36	18%
9	<i>Jw</i>	111	55.5%
10	<i>Soundex</i>	9	4.5%

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.

Persentase jumlah kata benar diperoleh dengan mengujikan 200 kata tidak baku dari Twitter. Jumlah kata benar tersebut merupakan hasil dari pendekatan setiap metode yang digunakan. Berdasarkan tabel di atas nilai akurasi tertinggi yang diperoleh adalah 55.5% dengan menggunakan metode *jw* dengan jumlah kata yang benar adalah 111 kata dari 200 kata yang diuji. Akurasi tertinggi diperoleh berdasarkan perhitungan berikut:

$$\text{Persentase} = \frac{111}{200} \times 100\% = 55.5\%$$

Setelah melihat saran kata yang ditampilkan dengan menggunakan KBBI, banyak kata yang tidak sering atau lazim digunakan pada bahasa sehari-hari. Hal ini membuat perbaikan saran kata yang ditampilkan sangat berbeda dengan yang seharusnya. Contohnya kata tidak baku “*asem*” memiliki perbaikan yang seharusnya adalah “*asam*”. Namun, perbaikan yang ditampilkan oleh fungsi ini adalah “*rasem*”. Kata “*rasem*” merupakan kata baku yang artinya adalah “tandan” menurut KBBI daring. Dengan demikian, makna dari kata “*rasem*” sangat berbeda dengan makna yang seharusnya yaitu asam. Selain itu kata “*rasem*” tidak lazim atau jarang digunakan oleh masyarakat sehingga menimbulkan ketidaktahuan makna dari kata tersebut. Oleh karena itu dibutuhkan kamus yang berisi kata-kata yang sering digunakan oleh masyarakat.

Kamus yang diperoleh dari korpus Kompas merupakan kamus yang telah diurutkan berdasarkan frekuensi penggunaan kata tersebut pada korpus yang dikumpulkan. Dalam hal ini kamus tersebut tidak terurut sesuai abjad, tetapi berdasarkan frekuensi penggunaan kata. Namun, kesalahan pengubahan tidak hanya ada pada proses menentukan kedekatan *string*. Kesalahan juga terjadi karena ada beberapa masukan yang memiliki kata baku yang sama tetapi berbeda makna atau memiliki kata baku ambigu. Contoh kata masukan “*km*”, jika dinormalkan berdasarkan urutan pengecekan, kata “*km*” dicek pada kamus kata tidak baku dan diubah menjadi kata “*kamu*”. Namun, kata “*km*” tersebut sudah ada dalam kamus korpus Kompas yang merupakan kata singkatan yang bermakna kilometer. Setelah dilakukan pengujian 200 kata tidak baku berdasarkan fungsi jarak, diperoleh akurasi yang ditunjukkan pada Tabel 2.

Tabel 2 Perbandingan 10 metode *stringdist* menggunakan kamus korpus Kompas

No	Metode	Jumlah kata benar	Persentase
1	<i>Osa</i>	116	58%
2	<i>Lv</i>	116	58%
3	<i>Dl</i>	116	58%
4	<i>Hamming</i>	24	12%
5	<i>Lcs</i>	133	66.5%
6	<i>Qgram</i>	82	41%
7	<i>Cosine</i>	87	43.5%
8	<i>Jaccard</i>	92	46%
9	<i>Jw</i>	124	62%
10	<i>Soundex</i>	67	33.5%

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.

b. Pengutipan tidak merugikan kepentingan yang wajar IPB.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.

Berdasarkan Tabel 2, nilai akurasi tertinggi meningkat menjadi 66.5% dengan jumlah kata yang benar adalah 133 kata menggunakan metode *lcs*. Akurasi tertinggi diperoleh menggunakan metode *lcs* karena metode ini bekerja membandingkan *string* dengan tetap mempertahankan urutannya. Dengan demikian, setiap *string* masukan yang dibandingkan memilih *string* yang memiliki kemiripan terbesar, baik dari segi urutan karakter maupun panjang *substring* yang sama.

Namun, setelah melihat saran kata yang ditampilkan, masih ada perbaikan *string* yang merupakan kata tidak baku dalam bahasa Indonesia ataupun merupakan singkatan, seperti “*entr*” menjadi “*center*”, “*and*” tetap menjadi “*and*”, “*ckp*” menjadi “*kpk*”, dan lain-lain. Dengan demikian, kamus yang digunakan merupakan kamus korpus Kompas yang telah diseleksi kata-katanya sesuai KBBI untuk menghilangkan kata yang tidak baku tersebut. Caranya adalah setelah fungsi *mapu* menerima masukan lebih dari 1 kata tanpa menjadikannya dalam bentuk vektor, maka masukan yang diberikan adalah semua kata yang ada pada kamus korpus Kompas. Masukan tersebut dicek keberadaannya pada kamus KBBI. Jika masukan setiap kata terdapat pada kamus KBBI, *return* kata tersebut. Sebaliknya, jika kata tersebut tidak terdapat pada kamus KBBI, *return null*. Selanjutnya hasil *return* kata disimpan dalam format csv. Dengan demikian, terbentuklah tabel kamus korpus Kompas yang berisi kata baku sesuai KBBI. Tabel tersebut dimasukkan ke dalam RStudio untuk dijadikan kamus baru. Kamus tersebut dijadikan acuan dalam penggantian kata tidak baku menjadi kata baku. Tabel 3 merupakan hasil akurasi menggunakan kamus acuan dari korpus Kompas yang telah diseleksi sesuai KBBI.

Tabel 3 Perbandingan 10 metode *stringdist* menggunakan kamus korpus Kompas sesuai KBBI

No	Metode	Jumlah kata benar	Percentase
1	<i>Osa</i>	119	59.5%
2	<i>Lv</i>	119	59.5%
3	<i>Dl</i>	119	59.5%
4	<i>Hamming</i>	24	12%
5	<i>Lcs</i>	138	69%
6	<i>Qgram</i>	89	44.5%
7	<i>Cosine</i>	89	44.5%
8	<i>Jaccard</i>	89	44.5%
9	<i>Jw</i>	133	66.5%
10	<i>Soundex</i>	71	35.5%

Berdasarkan hasil yang diperoleh dengan menggunakan kamus korpus Kompas sesuai KBBI, diperoleh nilai akurasi meningkat menjadi 69% dengan jumlah kata yang benar adalah 138 dari 200 kata yang diuji. Namun, ada beberapa kata yang ada pada kamus korpus Kompas merupakan kata baku menjadi hilang atau dalam proses sebelumnya *return null*. Kata tersebut merupakan kata baku dalam bahasa Indonesia yang memiliki imbuhan. Hal ini terjadi karena tidak semua kata yang memiliki imbuhan ada pada kamus KBBI. Berdasarkan akurasi tertinggi, fungsi ini dirancang menggunakan metode *stringdist* yaitu *lcs*.



SIMPULAN DAN SARAN

Simpulan

Berdasarkan hasil perbandingan 10 metode, akurasi yang paling tinggi adalah menggunakan metode *lcs* yaitu 69% dan menggunakan data korpus Kompas yang diseleksi sesuai KBBI sebagai kamus. Dengan demikian, dalam implementasi fungsi ini dibuat *default* menggunakan metode *lcs*. Fungsi ini telah mampu melakukan perubahan kata tidak baku menjadi baku, meskipun perubahan yang dilakukan tidak semuanya benar sesuai dengan yang seharusnya. Hal ini disebabkan oleh adanya saran kata lain yang memiliki jarak minimal yang sama dan tidak adanya kata dengan perbaikan seharusnya pada kamus.

Saran

Berdasarkan perbaikan yang ditampilkan pada pemrograman R, tidak semua kata memiliki perbaikan yang benar. Dengan demikian, ada baiknya melakukan normalisasi dengan melihat konteks dari masukan yang diberikan. Selain itu, masih terdapat kesalahan kata yang ditampilkan jika kata tidak baku tersebut mengandung imbuhan karena tidak semua kata yang mengandung imbuhan ada pada kamus. Dengan demikian, ada baiknya menambah isi kamus dengan kata yang berimbuhan ataupun melakukan penyisipan huruf terhadap masukan yang memungkinkan mengandung imbuhan. Selain itu, untuk melakukan pengujian terhadap fungsi yang telah dibangun, ada baiknya dengan menggunakan kata acak yang lebih beragam agar diperoleh kesimpulan yang lebih signifikan terhadap setiap metode yang ada pada *stringdist*.

DAFTAR PUSTAKA

- Adriyani NMM, Santiyasa IW, Muliantara A. 2012. Implementasi algoritma levenshtein distance dan metode empiris untuk menampilkan saran perbaikan kesalahan pengetikan dokumen berbahasa indonesia. 1(1).
- Aziz ATA. 2013. Sistem pengklasifikasian entitas pada pesan Twitter menggunakan ekspresi regular dan *naïve bayes* [skripsi]. Bogor (ID). Institut Pertanian Bogor.
- De Jonge Edwin, Van der loo Mark PJ. 2013. *An Introduction to Data Cleaning with R*. Netherlands (NL). Grafimedia.
- Lanin I, Geovedi J, Soegijoko W. 2013. Perbandingan distribusi frekuensi kata bahasa Indonesia di Kompas, Wikipedia, Twitter, dan Kaskus. *Konferensi Linguistik Tahunan Atma Jaya Kesebelas (KOLITA11)*. 2013 Mei 1-2; Jakarta, Indonesia. Jakarta(ID). Tersedia pada: <https://github.com/ardwort/freq-dist-id/tree/master/data>. [diakses pada 7 April 2017].
- Levenshtein VI. 1966. Binary codes capable of correcting deletion, insertion, and reversals. *Soviet physics docklady*. 10(8):707-710.



- Lhoussain AS, Hicham G, Abdellah Y.2015. Adapting the levenshtein distance to contextual spelling correction. *Technomathematics research foundation*. 12(1): 127-133.
- Sarwani MZ, Mahmudy WY. 2015. Analisis Twitter untuk mengetahui karakter seseorang menggunakan algoritma *naïve bayess classifier*. *Seminar Nasional Sistem Informasi Indonesia (SESINDO)*. 2015 Nov 2-3; Surabaya, Indonesia. Malang (ID).
- Wahyuningtyas A. 2016. Deteksi spam pada Twitter menggunakan algoritme naïve bayes [skripsi]. Bogor (ID). Institut Pertanian Bogor.
- Loo Mark PJ van der. 2014. The stringdist package for approximate string matching. *Contributed research articles*. 6(1):111-122.



Hak cipta milik IPB (Institut Pertanian Bogor)

Bogor Agricultural University

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.

b. Pengutipan tidak merugikan kepentingan yang wajar IPB.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.

Lampiran 1

Contoh kamus KBBI

© Hak cipta milik IPB (Institut Pertanian Bogor)

Bogor Agricultural U

LAMPIRAN

Kata	
abduktor	aborsi legal
abdul	abortif
abece	abortiva
aben	abortus
aberasi	abortus
aberasi cahaya	habitualis
aberasi	abortus
kromosom	inkomplet
abet	abortus komplet
abian	abortus
abid	provokatus
abidin	abrak
abilah	abrakadabra
abilah peringgi	abrar
abimana	abras
abing	abrasi
abiogenesis	abreaksi
abiosfer	abrek
abiotik	abreviasi
abis	abrikos
abisal	abritis
abiseka	abrosfer
abiturien	abrupsi
abjad	absah
abjad fonemis	absen
abjad fonetis	absensi
abjadiah	absente
abiasi	absenteisme
ablaut	abses
ablepsia	absis
abblur	absolusi
abnormal	absolut
abnormalitas	absolutisme
abnus	absonan
aboi	absorb
abolisi	absorben
abon	absorpsi
abonemen	absorpsi aktif

Lampiran 2 Contoh kamus Kompas

Kata			
calon	sekarang	inggris	khusus
wilayah	akibat	maupun	bawah
meski	akhirnya	cara	pihaknya
sempat	petugas	'	sedangkan
kepolisian	musim	laga	kelompok
mengalami	nomor	segera	eropa
sabtu	gedung	kecil	suara
melihat	menyatakan	dinas	berdasarkan
lima	akhir	kegiatan	turun
meminta	rasa	nanti	terlihat
acara	mungkin	ekonomi	2
perempuan	yaitu	tanah	bidang
proses	utama	guru	amerika
sol	sakit	pelatih	merasa
kecamatan	mantan	diduga	ketiga
aksi	bekerja	siswa	melawan
mendapatkan	bermain	surat	sebagian
badan	dollar	informasi	anggaran
seluruh	nama	provinsi	film
demikian	begitu	adanya	memberi
tanpa	utara	posisi	dewan
iana	ibu	pesawat	membantu
mendapat	lama	china	pagi
terhasil	jam	pekan	biasa
program	mengenai	rakyat	hotel
2011	proyek	ruang	unit
sebesar	selalu	bbm	20
terakhir	motor	membawa	milik
asal	awal	berharap	jauh
sebanyak	pembangunan	penting	kendaraan
datang	the	menjelaskan	makanan
1	punya	wib	poin
ujarnya	semakin	menerima	3
aku	naik	hidup	tampil
jelas	kesehatan	direktur	api
tahu	sesuai	data	anak-anak
kementerian	soal	menunjukkan	jaya
paling	polri	kkan	jangan
maka	berlangsung	sistem	media
pasangan	mahasiswa	khusus	jenis
bank			
10			
masa			
mau			

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.

b. Pengutipan tidak merugikan kepentingan yang wajar IPB.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



Lampiran 3 Contoh kamus kata *slang*

Kata <i>slang</i>	Perbaikan kata <i>slang</i>
abis	habis
accent	tekanan
accept	terima
accident	kecelakaan
achievement	prestasi
acra	acara
acrany	acaranya
acrnya	acaranya
action	aksi
active	aktif
activity	aktivitas
actually	sebenarnya
actualy	sebenarnya
ad	ada
ade	ada
adult	dewasa
adventure	petualangan
adventurer	petualang
advice	nasehat
after	setelah
afternum	sore
again	lagi
agency	perwakilan
agent	agen
agk	agak
agkttn	angkatan
agree	setuju
agreement	persetujuan
aing	saya
aj	saja
aja	saja
ajah	saja
aje	saja
ajeh	saja
ajk	ajak
ak	saya
akeh	banyak
akhire	Akhirnya
aktifkn	aktifkan
aku	saya
alhamdlh	alhamdulillah

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.

b. Pengutipan tidak merugikan kepentingan yang wajar IPB.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



Hak cipta milik IPB (Institut Pertanian Bogor)

Bogor Agricultural U

Lampiran 4 Kata uji dari tweet

Kata				
adalah	brapa	kite	mntak	senantias
akhirny	brarti	kkak	mntari	seorg
akn	brganti	klau	mrangkul	sgla
aktf	brjalan	kmudian	msukan	sihat
adlh	brjln	knangan	msyarakat	ska
agr	brsinar	konvensionil	nasehatin	skrang
ambl	brsma	kpda	nda	sktor
and	brsyukur	ksendirian	nebak	slamat
alalg	brtmbah	kurng	nescaya	slamatlah
ankh	bsalah	kwn	ngantok	slah
asem	ckp	lame	ngerti	sllu
bæk	cma	lmbat	olh	smakin
baikk	cuba	lngkh	page	smangat
ban	dapet	logik	pastu	smbungan
ban	dbntuk	lwti	pendem	sn드리
bedmpin	ddominasi	makacih	pernh	snyuman
gan	denger	maklumn	perpindhan	ssuatu
bener	dengn	manaaaa	pd	supay
erenti	die	manufaktr	pjbat	tertanya
bgaimana	diem	member	pling	tggikan
bgaimna	dnia	membri	pmimpin	tidk
bigun	eknmi	mencuba	pncernaan	tk
bhsa	emank	mengantikan	pngalam	tntng
bw	enaq	menghub	pnilaian	trbuka
bnwa	entr	mkanan	pnjara	trbukti
brase	garem	mkcih	prmata	trcapai
bkan	gawl	mk	prmpuan	trjdi
bkerja	grget	mkin	pernh	trnyta
bkin	gregetan	mlalui	prtama	ush
blajar	hbngn	mlindungi	prtanian	ttep
blakang	hngga	mmang	prtumbuhn	walopun
bljar	hkum	mmndang	psal	whai
bleh	hrp	mmperjuangkan	pulak	ykin
bliau	jngan	mmutuskn	rasany	zmn
blikan	jumpe	mndapatkan	sangt	
bnget	kalean	mngadu	satukn	
ngsa	kamuh	mngbh	sbagian	
bntuk	kdng	mnghiraukan	sudh	
bnyak	kemaren	mninggalkan	sebsar	
bprestasi	kesepyan	mnjalnkn	sekale	
brantas	kewajipan	mnolak	sempet	
			semuaaa	

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.

b. Pengutipan tidak merugikan kepentingan yang wajar IPB.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.

Lampiran 5 Contoh data *tweet* mengandung kata tidak baku*Tweet*

Prtumbuhn eknmi saat ini ddominasi olh nontradable,tp sektor yg tdk baik trutama di sktor prtanian&industri manufaktr.

ywdh itu udh pilihannya dia. nanti klw dia ada apa2 sama doi km nasehatin dgn baik. itu aja sih saran aku ya.

Alhamdulillah udh sampe sekolah. ditanya guru kenapa telat pdhl udh jujur tp ga percaya gurunya

baru kali ini ngumpul semua di rumah. senang sekali

uda nyampe ga dpt ucapan slmt pagi malah dpt ucapan slmt datang

sedih yang nggak enak adalah sedih yang nggak ngerti alasannya apa. pengen marah, nggak ngerti juga marah ke siapa.

Klau dulu aq klua awal awal kan senang

Ktika dhia prgi mninggalkan anda & semua sanak sodara anda mnghiraukan anda mka kasih & syg seorang ibulah yg kan mrangkul anda dlam ksendirian.

jadi downloader lagu2 lawas kalo ud dirumah, trs cara nebak judulnya dinya iin dulu sepenggal, ikutan berpacu dlm melodi aja

Gue rasa, baiknya orang sama lo itu tergantung pekara dia butuh sesuatu dari elo atau nggak. Saat dia udah gak butuh lg sama lo, babay.

Jngan pernah kamu bilang smua lelaki itu sama.! Karna raja fir'aun tdak setara dengn nabi musa begitu jga aku tidak setara dengn masa lalumu

Selalu ada masa dimana kita tiba-tiba kesel, pingin ngamuk, gregetan banget, tapi engga ngerti penyebabnya apa.

mang nya kl burung biru pesan taksi lewat telepon jd bukan nya beda dgn konvensionil cegat dipinggir jln ? Kl tarif, se smb

Bukan kelas gue ribut ama manusia jadi2an kyk loe , gue gak mw abisin energi bt balasin mantion loe . Sorry ya

Lampiran 6 Contoh kamus Kompas sesuai KBBI

Kata			
yang	harus	sangat	as
di	kepada	kepala	jadi
dan	lalu	negara	pihak
ini	telah	saja	tetap
itu	setelah	membuat	hasil
dengan	hari	korban	partai
untuk	rumah	tiga	presiden
dari	warga	lagi	memang
dalam	bahwa	baik	barat
aran	baru	kedua	kamis
pada	banyak	waktu	senin
tidak	hal	selain	tempat
juga	melakukan	pemain	tapi
ke	hingga	kalau	bersama
tersebut	anda	tetapi	mulai
ada	menurut	anak	selasa
bisa	belum	ujar	kabupaten
saat	dapat	daerah	kemudian
tahun	lain	sampai	depan
karena	beberapa	agar	timur
sudah	kota	kali	langsung
menjadi	besar	sama	terkait
mereka	persen	tengah	bulan
kata	sekitar	kembali	memberikan
indonesia	tim	sejumlah	bukan
saya	kita	ketika	melalui
lebih	pemerintah	anggota	minggu
atau	pun	serta	gubernur
kami	jika	pertama	jumat
oleh	jalan	seorang	tinggi
sebagai	salah	sejak	selatan
adalah	sementara	dunia	negeri
tak	bagi	polisi	air
satu	memiliki	antara	terus
masih	secara	sehingga	nasional
orang	merupakan	juta	menggunakan
seperti	masyarakat	terhadap	mengaku
dia	atas	rabu	ingin
namun	kasus	harga	setiap
para	terjadi	semua	berada
hanya	selama	ketua	masuk
mengatakan			
dua			

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.

b. Pengutipan tidak merugikan kepentingan yang wajar IPB.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



RIWAYAT HIDUP

Penulis dilahirkan di Sinaman Pematang pada tanggal 31 Maret 1995 dari Ayah bernama Jal Benson Saragih dan Ibu bernama Rosni Girsang. Penulis adalah anak ketiga dari empat bersaudara. Tahun 2013 penulis lulus SMA Swasta Teladan Pematang Siantar dan pada tahun yang sama penulis lulus seleksi masuk Institut Pertanian Bogor melalui jalur SNMPTN dan diterima sebagai mahasiswa di Departemen Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam.

Selama menjadi mahasiswa aktif, penulis pernah menjadi pengurus dalam organisasi KEMAKI (Keluarga Mahasiswa Katolik IPB) pada tahun 2014/2015. Penulis melaksanakan kegiatan Praktik Kerja Lapangan di Pusat Studi Biofarmaka pada tahun 2016. Selain itu, Penulis juga berprestasi dengan meraih juara 1 dalam PEKSIMDA (Pekan Seni Mahasiswa Daerah) tingkat DKI Jakarta mewakili IPB pada tahun 2016 dan meraih juara harapan 2 pada tingkat PEKSIMINAS (Pekan Seni Mahasiswa Nasional) ke-13 pada tahun 2016.

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.