



Hak cipta dan penggunaan kembali:

Lisensi ini mengizinkan setiap orang untuk menggubah, memperbaiki, dan membuat ciptaan turunan bukan untuk kepentingan komersial, selama anda mencantumkan nama penulis dan melisensikan ciptaan turunan dengan syarat yang serupa dengan ciptaan asli.

Copyright and reuse:

This license lets you remix, tweak, and build upon work non-commercially, as long as you credit the origin creator and license it on your new creations under the identical terms.

**DETEKSI KESALAHAN EJA KATA LULUH PADA BERITA
DENGAN ALGORITMA JACCARD SIMILARITY
(STUDI KASUS : TRIBUNNEWS)**



SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh
Gelar Sarjana Komputer (S.Kom.)

Nicholas Evan

0000027900

UMN

UNIVERSITAS

MULTIMEDIA

NUSANTARA

**PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA**

TANGERANG

2023

**DETEKSI KESALAHAN EJA KATA LULUH PADA BERITA
DENGAN ALGORITMA JACCARD SIMILARITY
(STUDI KASUS : TRIBUNNEWS)**



SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh
Gelar Sarjana Komputer (S.Kom.)

**Nicholas Evan
0000027900**

UMN

UNIVERSITAS

MULTIMEDIA

NUSANTARA

**PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA**

TANGERANG

2023

HALAMAN PERNYATAAN TIDAK PLAGIAT

Dengan ini saya,

Nama : Nicholas Evan

Nomor Induk Mahasiswa : 00000027900

Program Studi : Informatika

Skripsi dengan judul:

Deteksi Kesalahan Eja Kata Luluh pada Berita dengan Algoritma Jaccard Similarity (Studi Kasus : Tribunnews)

merupakan hasil karya saya sendiri bukan plagiat dari karya ilmiah yang ditulis oleh orang lain, dan semua sumber baik yang dikutip maupun dirujuk telah saya nyatakan dengan benar serta dicantumkan di Daftar Pustaka.

Jika di kemudian hari terbukti ditemukan kecurangan/ penyimpangan, baik dalam pelaksanaan Skripsi maupun dalam penulisan laporan Skripsi, saya bersedia menerima konsekuensi dinyatakan TIDAK LULUS untuk Tugas akhir yang telah saya tempuh.

Tangerang, 3 Januari 2023



(Nicholas Evan)

UMM
UNIVERSITAS
MULTIMEDIA
NUSANTARA

HALAMAN PENGESAHAN

Skripsi dengan judul

DETEKSI KESALAHAN EJA KATA LULUH PADA BERITA DENGAN ALGORITMA JACCARD SIMILARITY (STUDI KASUS : TRIBUNNEWS)

oleh

Nama : Nicholas Evan
NIM : 00000027900
Program Studi : Informatika
Fakultas : Fakultas Teknik dan Informatika


Telah diujikan pada hari Senin, 9 Januari 2023

Pukul 10.00 s/s 11.30 dan dinyatakan

LULUS

Dengan susunan penguji sebagai berikut


Ketua Sidang


(Moeljono Widjaja, B.Sc., M.Sc., Ph.D.)
NIDN: 0311106903

Penguji


(Angga Aditya Permana, S.Kom.,
M.Kom.)
NIDN: 0407128901

Pembimbing


(Marlinda Vasty Overbeek, S.Kom, M.Kom)
NIDN: 0818038501

Digitally signed
by Marlinda Vasty
Overbeek
Date: 2023.01.18
12:50:45 +07'00'

Ketua Program Studi Informatika,


(Marlinda Vasty Overbeek, S.Kom., M.Kom.)
NIDN: 0818038501

Digitally signed
by Marlinda
Vasty Overbeek
Date: 2023.01.18
12:50:59 +07'00'

**HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK
KEPENTINGAN AKADEMIS**

Sebagai sivitas akademik Universitas Multimedia Nusantara, saya yang bertanda tangan di bawah ini:

Nama : Nicholas Evan
NIM : 00000027900
Program Studi : Informatika
Fakultas : Teknik dan Informatika
Jenis Karya : Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada **Universitas Multimedia Nusantara** hak Bebas Royalti Non-eksklusif (*Non-exclusive Royalty-Free Right*) atas karya ilmiah saya yang berjudul:

**DETEKSI KESALAHAN EJA KATA LULUH PADA BERITA
DENGAN ALGORITMA JACCARD SIMILARITY
(STUDI KASUS : TRIBUNNEWS)**

Beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Non eksklusif ini Universitas Multimedia Nusantara berhak menyimpan, mengalih media / format-kan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mempublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis / pencipta dan sebagai pemilik Hak Cipta. Demikian pernyataan ini saya buat dengan sebenarnya.

Tangerang, 3 Januari 2023

Yang menyatakan

UNIVERSITAS
MULTIMEDIA
NUSANTARA


Nicholas Evan

Halaman Persembahan / Motto

"When you get tired, learn to rest, not to quit."

Banksy

UMMN

UNIVERSITAS
MULTIMEDIA
NUSANTARA

KATA PENGANTAR

Puji Syukur atas berkat dan rahmat kepada Tuhan Yang Maha Esa, atas selesainya penulisan laporan Skripsi ini dengan judul: Deteksi Kesalahan Eja Kata Luluh pada Berita dengan Algoritma Jaccard Similarity (Studi Kasus : Tribunnews) dilakukan untuk memenuhi salah satu syarat untuk mencapai gelar Sarjana Komputer Jurusan Informatika Pada Fakultas Teknik dan Informatika Universitas Multimedia Nusantara. Saya menyadari bahwa, tanpa bantuan dan bimbingan dari berbagai pihak, dari masa perkuliahan sampai pada penyusunan skripsi ini, sangatlah sulit bagi saya untuk menyelesaikan skripsi ini. Oleh karena itu, saya mengucapkan terima kasih kepada:

1. Bapak Dr. Ninok Leksono, selaku Rektor Universitas Multimedia Nusantara.
2. Dr. Eng. Niki Prastomo, S.T., M.Sc., selaku Dekan Fakultas Teknik dan Informatika Universitas Multimedia Nusantara.
3. Ibu Marlinda Vasty Overbeek, S.Kom., M.Kom., selaku Ketua Program Studi Informatika Universitas Multimedia Nusantara dan Pembimbing pertama yang telah banyak meluangkan waktu untuk memberikan bimbingan, arahan dan motivasi atas terselesainya skripsi ini.
4. Keluarga dan teman-teman yang telah memberikan bantuan dukungan material dan moral, sehingga penulis dapat menyelesaikan tugas akhir ini.
5. Jessica Augustine S. yang telah memberikan dukungan dan bantuan moral, selama pengerjaan tugas akhir ini.
6. Jerico Olwen, selaku teman seperjuangan selama pengerjaan tugas akhir.

Semoga skripsi ini bermanfaat, baik sebagai sumber informasi maupun sumber inspirasi, bagi para pembaca.

Tangerang, 3 Januari 2023



Nicholas Evan

**DETEKSI KESALAHAN EJA KATA LULUH PADA BERITA
DENGAN ALGORITMA JACCARD SIMILARITY
(STUDI KASUS : TRIBUNNEWS)**

Nicholas Evan

ABSTRAK

Bahasa Indonesia merupakan bahasa nasional yang digunakan dalam kehidupan sehari-hari, namun kesalahan berbahasa kerap terjadi dalam di sekitar kita, salah satunya pada portal berita *online*. Kesalahan berbahasa merupakan penyimpangan bahasa dari kaidah tata bahasa dan salah satunya adalah peluluhan fonem. Hal ini terjadi akibat penulisan dilakukan secara *manual* sehingga memungkinkan untuk terjadinya kesalahan pengetikan. Dengan terjadinya kesalahan peluluhan ini, dilakukanlah penelitian yaitu pembuatan sistem dengan menggunakan algoritma *Jaccard Similarity* untuk mendeteksi kesalahan eja pada kata terluluh. *Jaccard Similarity* merupakan algoritma yang digunakan untuk membandingkan dokumen untuk menghitung kesamaan nilai dari dua dokumen. Evaluasi dilakukan dengan menggunakan *confusion matrix* yang kemudian diambil *F-1 score*-nya selain itu efisiensi sistem juga diperhitungkan. Hasil deteksinya memiliki *F-1 score* sebesar 66.6% dan efisiensi sistem dipengaruhi oleh jumlah kalimat dan jumlah kata yang terluluh. Sistem yang dibangun dapat mendeteksi kesalahan eja pada kata terluluh saat dihadapkan dengan berita dari portal berita Tribun.

Kata kunci: berita, *Jaccard Similarity*, kesalahan eja, peluluhan fonem, sistem deteksi



Detection of Spelling Errors in Indonesian Language News with Jaccard Similarity Algorithm (Case : Tribun News)

Nicholas Evan

ABSTRACT

Language is an organized communication tool in the form of units such as words, groups of words, clauses and sentences. Indonesian is the national language which should be used in accordance with Enhanced Indonesian Spelling, but language errors often occur accidentally on online news portals. Language errors are language deviations from grammatical rules and one of them is phoneme decay. This may occur due to writing done manually so that it allows typing errors. With the occurrence of this error, research was carried out, namely making a system using the Jaccard Similarity algorithm to detect misspellings in melted words. Jaccard Similarity is an algorithm used to compare documents to calculate the similarity of values from two documents. Evaluation is done by using the confusion matrix which then takes the F-1 score besides that system efficiency is also taken into account. The detection results have an F-1 score of 66.6% and system efficiency is influenced by the number of sentences and the number of words decayed. The system built can detect misspelled words when dealing with news from the Tribun news portal.

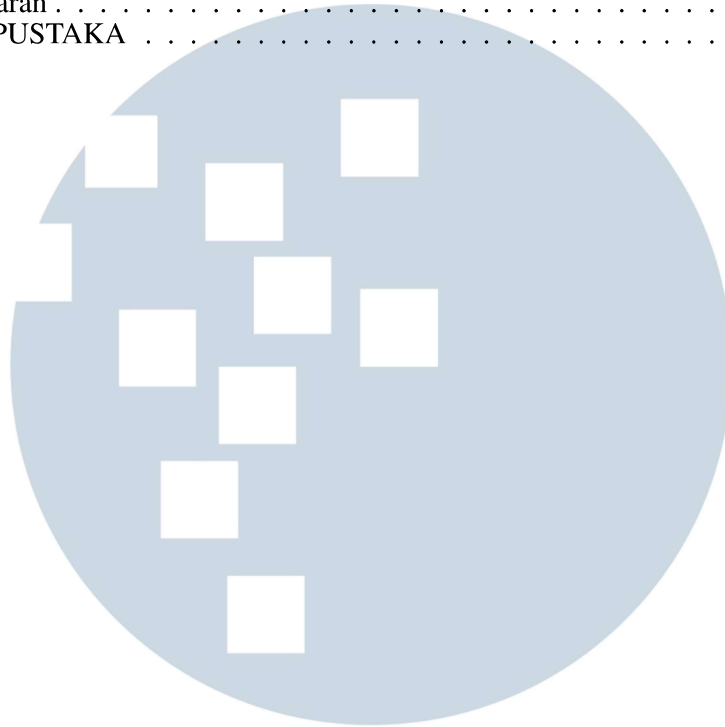
Keywords: *detection system, Jaccard Similarity, misspelling, news, phoneme decay*



DAFTAR ISI

HALAMAN JUDUL	i
PERNYATAAN TIDAK MELAKUKAN PLAGIAT	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERSETUJUAN PUBLIKASI ILMIAH	iv
HALAMAN PERSEMBAHAN/MOTO	v
KATA PENGANTAR	vi
ABSTRAK	vii
ABSTRACT	viii
DAFTAR ISI	ix
DAFTAR GAMBAR	xi
DAFTAR TABEL	xii
DAFTAR KODE	xiii
DAFTAR LAMPIRAN	xiv
BAB 1 PENDAHULUAN	1
1.1 Latar Belakang Masalah	1
1.2 Rumusan Masalah	2
1.3 Batasan Permasalahan	2
1.4 Tujuan Penelitian	3
1.5 Manfaat Penelitian	3
1.6 Sistematika Penulisan	3
BAB 2 LANDASAN TEORI	5
2.1 Tinjauan Teori	5
2.1.1 Portal Berita	5
2.1.2 TRIBUN NEWS	5
2.1.3 Natural Language Processing	6
2.1.4 Text Preprocessing	6
2.1.5 Jaccard Similarity	7
2.1.6 Confusion Matrix	7
BAB 3 METODOLOGI PENELITIAN	10
3.1 Pengumpulan Data	10
3.2 Proses Data menjadi Dataset	10
3.3 Praproses	11
3.4 Jaccard Similarity	12
BAB 4 HASIL DAN DISKUSI	14
4.1 Spesifikasi Sistem	14
4.2 Implementasi	14
4.2.1 Pengumpulan Data	14
4.2.2 <i>Dataset</i>	15
4.2.3 Praproses	17
4.2.4 Jaccard Similarity	19
4.3 Uji Coba	20
4.3.1 Uji Coba Benar	20
4.3.2 Uji Coba Kesalahan	22
4.3.3 Uji Coba Dengan Parameter	23
4.3.4 Perhitungan Confusion Matrix	28
4.4 Evaluasi	28
4.4.1 Evaluasi Confusion Matrix	28
4.4.2 Evaluasi Efisiensi Sistem	29

BAB 5	SIMPULAN DAN SARAN	31
5.1	Simpulan	31
5.2	Saran	31
DAFTAR PUSTAKA		32



UMMN

UNIVERSITAS
MULTIMEDIA
NUSANTARA

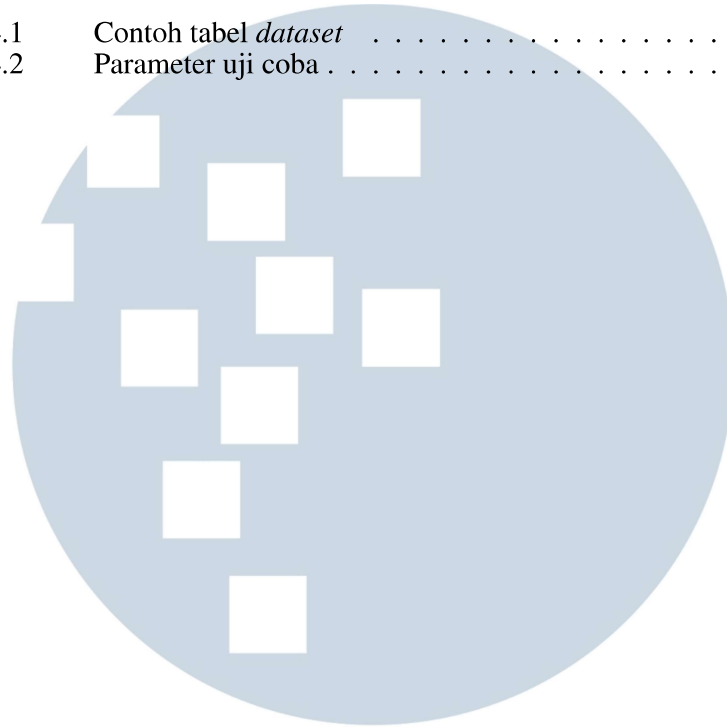
DAFTAR GAMBAR

Gambar 2.1	Logo Tribunnews	5
Gambar 2.2	Confusion matrix	8
Gambar 3.1	Diagram alir metodologi penelitian	10
Gambar 3.2	Diagram alir praproses	11
Gambar 3.3	Diagram alir perhitungan <i>jaccard similarity</i>	13
Gambar 4.1	Contoh data berita	15
Gambar 4.2	Contoh data dalam excel	15
Gambar 4.3	Potongan berita benar	21
Gambar 4.4	Hasil uji coba potongan berita benar	22
Gambar 4.5	Potongan berita salah	22
Gambar 4.6	Hasil uji coba potongan berita salah	23
Gambar 4.7	Hasil uji coba parameter 2 kalimat 25 kata	24
Gambar 4.8	Hasil uji coba parameter 2 kalimat 44 kata	25
Gambar 4.9	Hasil uji coba parameter 4 kalimat 55 kata	26
Gambar 4.10	Hasil uji coba parameter 4 kalimat 95 kata	26
Gambar 4.11	Hasil uji coba parameter 6 kalimat 124 kata	27
Gambar 4.12	Hasil uji coba parameter 6 kalimat 148 kata	27
Gambar 4.13	Confusion matrix hasil uji coba	28
Gambar 4.14	Evaluasi uji coba efisiensi	30
Gambar 4.15	Grafik uji coba efisiensi	30



DAFTAR TABEL

Tabel 4.1	Contoh tabel <i>dataset</i>	16
Tabel 4.2	Parameter uji coba	23



UMMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

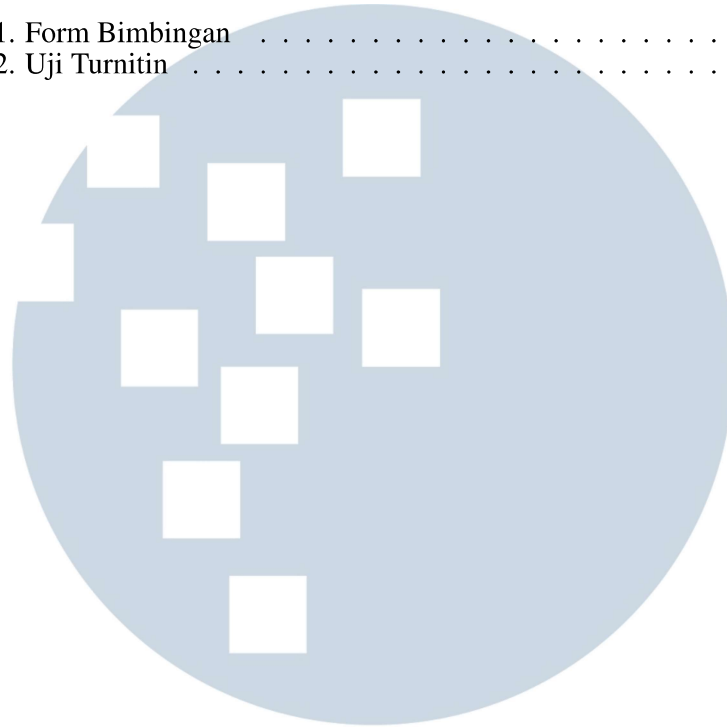
DAFTAR KODE

4.1	Potongan kode membaca file excel	16
4.2	Potongan kode membuat dataset	17
4.3	Potongan kode melakukan filtering	17
4.4	Potongan kode import regular expression	18
4.5	Potongan kode import string	18
4.6	Potongan kode case folding	18
4.7	Potongan kode tokenisasi kata	18
4.8	Potongan kode formula jaccard similarity	19
4.9	Potongan kode pengecekan dan perhitungan jaccard similarity	19



DAFTAR LAMPIRAN

Lampiran 1. Form Bimbingan	33
Lampiran 2. Uji Turnitin	35



UMMN

UNIVERSITAS
MULTIMEDIA
NUSANTARA