Versi Online tersedia di : https://jurnal.buddhidharma.ac.id/index.php/algor/index

JURNAL ALGOR

|2715-0577 (Online)| 2715-0569 (Print)

Artikel

KLASIFIKASI BERITA HOAKS TOPIK *COVID-19* DENGAN KLASIFIKASI *ROCCHIO* DAN *COSINE SIMILARITY*

Indrawan Gotama[1], Susanto Hariyanto[2], Hartana Wijaya[2]

[1,2]Universitas Buddhi Dharma, Teknik Informatika, Banten, Indonesia

SUBMISSION TRACK

Recieved: March 28, 2017
Final Revision: May 03, 2017
Available Online: May 15, 2017

KEYWORD

Klasifikasi, Hoaks, COVID-19, Rocchio

KORESPONDENSI

E-mail: indrawangotama@gmail.com

E-mail: sanada000@gmail.com

A B S T R A K

Wabah COVID-19 menyerang dunia serta Indonesia dimulai sejak akhir tahun 2019. Kurangnya pengetahuan tentang pandemi baru ini, menyebabkan banyak informasi yang diragukan kebenarannya tersebar melalui media sosial dan portal berita online. Beberapa pihak tidak bertanggung jawab juga memanfaatkan kondisi ini untuk memperoleh keuntungan pribadi. Fenomena penyebaran berita hoaks dan ujaran kebencian di Indonesia yang merupakan dampak dari perkembangan teknologi dan infromasi sehingga menyebabkan penyebaran berita yang sangat cepat dan tidak diimbangi dengan kesadaran akan penyampaian berita yang benar dan sesuai fakta. Oleh karena itu diperlukan sebuah aplikasi untuk menggolongkan informasi mengenai COVID-19 yang beredar di masyarakat apakah berita tersebut memiliki indikasi hoaks atau fakta. Model yang dibentuk menggunakan klasifikasi Rocchio dengan sumber data berasal dari kompas.com dan turnbackhoax.id. Model yang dibentuk juga dapat memberi informasi mengenai berita yang relevan dengan berita masukan dengan menggunakan metode Cosine Similarity, sehingga dapat memberikan hasil yang dapat dipertanggung jawabkan kepada masyarakat. Dari model yang diusulkan, diharapkan dapat menghasilkan penggolongan indikasi hoaks atau fakta terhadap suatu berita serta menampilkan sumber yang relevan dengan berita yang diuji.

PENDAHULUAN

Perkembangan teknologi informasi yang pesat ditandai dengan munculnya beragam media termasuk media online. Kemudahan penggunaan yang ditawarkan media online, menyebabkan media online menjadi wadah distribusi informasi yang sangat menarik dan sering digunakan pada masyarakat. Media online tidak hanya mengubah bagaimana informasi disampaikan, tetapi juga mengubah bagaimana masyarakat menerima informasi

tersebut. Kasus pertama virus COVID-19 di Indonesia yang berawal dari adanya Warga Negara Asing (WNA) dari Jepang yang positif mengunjungi Indonesia.Minimnya informasi mengenai virus ini menjadi tantangan tersendiri bagi media untuk menyampaikan informasi serta perlunya kewaspadaan dari masyarakat untuk menerima informasi terkait COVID-19.

I. METODE

1.1. Pengumpulan Data

Untuk mendapatkan data yang nantinya akan digunakan dalam penelitian ini, penulis mengumpulkan data awal berupa berita dengan tagar COVID-19 melalui proses web scraping website kompas.com dan turnbackhoax.id dari tanggal 16 Mei 2020 sampai 23 Mei 2020. Data yang telah dikumpulkan disimpan dalam bentuk CSV. Terkumpul 880 data dari periode tersebut.

Pengumpulan data dilakukan melalui metode Web Scraping. Data fakta dikumpulkan melalui teknik scraping pada https://www.kompas.com/COVID-19. Judul berita diambil dari text yang berada pada tag<a> dengan class:'article__link'. Link berita juga diambil dari sumber yang sama yang berada pada tag dengan class : 'article__list clearfix'. Data yang telah diambil disimpan kedalam file csv dengan nama dokumen FaktaKompas diikuti dengan tanggal proses scraping dilakukan, dengan judul kolom berupa judul, link berita serta label berupa fakta. Sedangkan, Data hoaks dikumpulkan dari halaman web https://turnbackhoax.id/tag/COVID-19.
Berita diambil dari tag<h3> dengan class: 'entry-title mh-loop-title'. Sedangkan link berita diambil melalui pencarian tag<a> dalam atribut article dengan class : ’mh-loop-item mh-clearfix post’. Data yang telah diambil disave kedalam file csv dengan judul dokumen Hoaks diikuti dengan tanggal proses scraping dilakukan , dengan judul kolom berupa judul,link berita serta label berupa hoaks.

1.2. Praproses Teks

1.Tokenisasi
Tokenisasi adalah proses memisahkan string mentah menjadi token yang memiliki arti. Kesulitan proses tokenization tergantung pada bahasa yang digunakan. Sebagai pembanding tokenization dalam Bahasa Inggris lebih sederhana dibandingkan dengan Bahasa Jepang [1].

2.*Data Cleansing*
Setelah proses parsing teks dari berbagai sumber data, tantangannya adalah memahami data mentah ini. Pembersihan teks digunakan misalnya pada html dengan menggunakan html_clean, dapat katakan sebagai pembersihan teks. Dalam kasus lain, di mana kita melakukan parsing PDF, mungkin ada karakter noise yang tidak diinginkan, karakter non ASCII yang akan dihapus, dan sebagainya. Sebelum melanjutkan ke langkah berikutnya penghapusan ini bertujuan untuk mendapatkan teks bersih untuk diproses lebih lanjut. Dengan sumber data seperti xml, kita mungkin hanya tertarik pada beberapa elemen pohon tertentu, dengan basis data kita mungkin harus memanipulasi splitter, dan terkadang kita hanya tertarik pada kolom tertentu [1].

3.*Stop Word Removal*
Stop Word Removal adalah salah satu langkah Preprocessing yang paling umum digunakan di berbagai aplikasi NLP. Idenya adalah hanya menghapus kata-kata yang umum terjadi di semua dokumen dalam korpus. Biasanya, awalan dan kata ganti umumnya diklasifikasikan sebagai stop words.. Kata-

kata ini tidak memiliki signifikansi dalam beberapa tugas NLP seperti pencarian informasi dan klasifikasi, yang berarti kata-kata ini tidak terlalu diskriminatif [1].

Tabel 1. Daftar Stopword pada Library Sastrawi

yang	di	kenapa	seterusnya
untuk	dari	yaitu	tanpa
pada	telah	yakni	agak
ke	sebagai	daripada	boleh
para	masih	itulah	dapat
namun	hal	lagi	dsb
menurut	ketika	maka	dst
antara	adalah	tentang	dll
dia	itu	demi	dahulu
dua	dalam	dimana	dulunya
ia	bisa	kemana	anu
seperti	bahwa	pula	demikian
jika	atau	sambil	tapi
jika	hanya	sebelum	ingin
sehingga	kita	sesudah	juga
kembali	dengan	supaya	nggak
dan	akan	guna	mari
tidak	juga	kah	nanti
ini	ada	pun	melainkan
karena	mereka	sampai	oh
kepada	sudah	sedangkan	ok
oleh	saya	selagi	seharusnya
saat	terhadap	sementara	sebetulnya
harus	secara	tetapi	setiap
sementara	agar	apakah	setidaknya
setelah	lain	kecuali	sesuatu
belum	anda	sebab	pasti
kami	begitu	selain	saja
sekitar	mengapa	seolah	toh
bagi	walau	seraya	ya
serta	tolong	apalagi	
amat	tentu	bagaimanapun	

4.*Stemming*

Stemming, secara harfiah, adalah proses menebang cabang-cabang pohon ke batangnya. Jadi secara efektif, dengan menggunakan beberapa aturan dasar, token apa pun dapat ditebang sampai ke dasarnya. *Stemming* lebih merupakan proses berbasis aturan yang kasar di mana kita ingin menggabungkan variasi token yang berbeda. Misalnya, kata makan akan memiliki variasi seperti makan, memakan, dimakan, dan sebagainya. Dalam beberapa aplikasi, karena tidak masuk akal untuk membedakan antara makan dan memakan, biasanya digunakan Stemming untuk memadukan kedua varian kata tersebut menjadi bentuk dasarnya [1].

1.3. Klasifikasi Rocchio

TRAINROCCHIO($\mathbb{C}, \mathbb{D}$)
1 **for each** $c_j \in \mathbb{C}$
2 **do** $D_j \leftarrow \{d : \langle d, c_j \rangle \in \mathbb{D}\}$
3 $\quad \vec{\mu}_j \leftarrow \frac{1}{|D_j|} \sum_{d \in D_j} \vec{v}(d)$
4 **return** $\{\vec{\mu}_1, \ldots, \vec{\mu}_J\}$

APPLYROCCHIO($\{\vec{\mu}_1, \ldots, \vec{\mu}_J\}, d$)
1 **return** $\arg\min_j |\vec{\mu}_j - \vec{v}(d)|$

Klasifikasi *Rocchio* merupakan klasifikasi dengan bentuk linear, dimana klasifikasi linear menerapkan prinsip *contiguity hypothesis*, bahwa dokumen dalam suatu kelas yang sama tidak akan *overlap* dengan kelas yang berbeda Nilai centroid diperoleh dengan menghitung rata-rata vektor pada semua dokumen data latih untuk setiap kelasnya.

Centroid kelas c dihitung dengan persamaan:

$$\vec{u}(c) = \frac{1}{Dc} \Sigma_{d \in Dc} \vec{v}(\mathrm{d})$$

dengan Dc adalah gugus dokumen di kelas c, $\vec{v}$(d) adalah vektor kata-kata dalam kelas c, dan $\vec{u}(c)$ adalah centroid masing-masing kelas [2].

1.4. COSINE SIMILARITY

Cosine Similarity adalah ukuran kesamaan yang dapat digunakan untuk membandingkan dokumen atau, memberikan peringkat dokumen sehubungan dengan vektor kata permintaan yang diberikan. Misalkan x dan y menjadi dua vektor untuk perbandingan. Menggunakan ukuran cosinus sebagai fungsi kesamaan:

$$sim(x,y) = \frac{x.y}{||x||||y||'}$$

Dimana || x || adalah norma Euclidean dari vektor x = (x1, x2, ..., xp), didefinisikan sebagai x12 + x22 + ⋯ + xp2. Secara konseptual, itu adalah panjang vektor. Demikian pula, || y || adalah norma Euclidean dari vektor y. Pengukuran tersebut menghitung kosinus sudut antara vektor x dan y. Nilai cosinus dari 0 berarti bahwa kedua vektor berada pada 90 derajat satu sama lain (ortogonal) dan tidak memiliki kecocokan. Semakin dekat nilai cosinus ke 1, semakin kecil sudut dan semakin besar kecocokan antara vektor [3].

Hasil masukan pengguna diperiksa dengan metode Cosine Similarity untuk menguji apakah data masukan relevan dengan topik COVID-19. Data masukan diuji kemiripan nya dengan data sumber sehingga dapat memberikan feedback kepada user agar kevalidan hasil klasifikasi yang dihasilkan dapat meningkat. Pada proses ini , digunakan TF-IDF dengan n-gram (3,3), dengan maksud menghindari pengguna dari menginput data yang dapat dikategorikan sebagai noise. Karena suatu kalimat berita minimal memiliki subjek , predikat , dan objek.

II. HASIL

Berikut hasil dari sampel Preproses Teks dibandingkan dengan teks asli :

Tabel 2. Tabel Judul berita asli

Judul
Jokowi: Pemerintah Akan Mengatur agar Kehidupan Berangsur Normal
Peneliti LIPI Sarankan 5 Hal Ini dalam Penanganan Pandemi COVID-19
Yurianto: Presiden Minta Gugus Tugas COVID-19 Pastikan PSBB Efektif Tekan Kasus Kematian
Anies: Hanya Izin dari Pemprov DKI yang Diterima Petugas untuk Keluar Jabodetabek
Anies: Masyarakat yang Ingin ke Jakarta Harus Urus Izin Masuk
[SALAH] Manager Giant Pal 6 Banjarmasin Meninggal Kena COVID-19
[SALAH] Rapid Test Di Bandara Soekarno-Hatta Bayar Rp550 Ribu
[SALAH] Peringatan Badai Panas Equinox
[SALAH] Pesan Berantai "Ganjar Pranowo Bolehkan Warga Jawa Tengah Salat Idul Fitri"
[SALAH] Imbauan Walikota Solo, Kita Beli Tiket Kebun Binatang Jurug Karena Pengelola Sudah Tidak Sanggup Memberi Makan Binatang

Tabel 3. Tabel judul berita setelah pra proses

Judul setelah pra proses
jokowi perintah atur hidup angsur normal
teliti lipi saran 5 ini tangan pandemi corona
yurianto presiden minta gugus tugas corona pasti psbb efektif tekan kasus mati
anies izin pemprov dki terima tugas keluar jabodetabek
anies masyarakat ingin jakarta urus izin masuk
manager giant pal 6 banjarmasin tinggal kena corona
rapid test bandara soekarnohatta bayar rp550 ribu
ingat badai panas equinox
pesan beranta ganjar pranowo boleh warga jawa tengah salat idul fitri
imbau walikota solo beli tiket kebun binatang jurug kelola tidak sanggup beri makan binatang

Tabel 4. Hasil klasifikasi pengujian data input

Token	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7	Doc 8	Doc 9	Doc 10	Doc Uji
angsur	2.704748	0	0	0	0	0	0	0	0	0	0
anies	0	0	0	2.299283	2.299283	0	0	0	0	0	0
atur	2.704748	0	0	0	0	0	0	0	0	0	0
badai	0	0	0	0	0	0	0	2.704748	0	0	0
bandara	0	0	0	0	0	0	2.704748	0	0	0	0
banjarmasin	0	0	0	0	0	2.704748	0	0	0	0	0
bayar	0	0	0	0	0	0	2.704748	0	0	0	0
beli	0	0	0	0	0	0	0	0	0	2.704748	0
beranta	0	0	0	0	0	0	0	0	2.704748	0	0
beri	0	0	0	0	0	0	0	0	0	2.704748	0
binatang	0	0	0	0	0	0	0	0	0	0.707107	0.707107
boleh	0	0	0	0	0	0	0	0	2.704748	0	0
corona	0	2.011601	2.011601	0	0	2.011601	0	0	0	0	0
dki	0	0	0	2.704748	0	0	0	0	0	0	0
efektif	0	0	2.704748	0	0	0	0	0	0	0	0
equinox	0	0	0	0	0	0	0	2.704748	0	0	0
fitri	0	0	0	0	0	0	0	0	2.704748	0	0
ganjar	0	0	0	0	0	0	0	0	2.704748	0	0
giant	0	0	0	0	0	2.704748	0	0	0	0	0
gugus	0	0	2.704748	0	0	0	0	0	0	0	0
hidup	2.704748	0	0	0	0	0	0	0	0	0	0
idul	0	0	0	0	0	0	0	0	2.704748	0	0
imbau	0	0	0	0	0	0	0	0	0	2.704748	0
ingat	0	0	0	0	0	0	0	2.704748	0	0	0
ingin	0	0	0	0	2.704748	0	0	0	0	0	0
ini	0	2.704748	0	0	0	0	0	0	0	0	0
izin	0	0	0	2.299283	2.299283	0	0	0	0	0	0
jabodetabek	0	0	0	2.704748	0	0	0	0	0	0	0
jakarta	0	0	0	0	2.704748	0	0	0	0	0	0
jawa	0	0	0	0	0	0	0	0	2.704748	0	0
jokowi	2.704748	0	0	0	0	0	0	0	0	0	0
jurug	0	0	0	0	0	0	0	0	0	0.353553	0.353553
kasus	0	0	2.704748	0	0	0	0	0	0	0	0
kebun	0	0	0	0	0	0	0	0	0	0.353553	0.353553
kelola	0	0	0	0	0	0	0	0	0	2.704748	0
keluar	0	0	0	2.704748	0	0	0	0	0	0	0
kena	0	0	0	0	0	2.704748	0	0	0	0	0
lipi	0	2.704748	0	0	0	0	0	0	0	0	0
makan	0	0	0	0	0	0	0	0	0	0.353553	0.353553
manager	0	0	0	0	0	2.704748	0	0	0	0	0
masuk	0	0	0	0	2.704748	0	0	0	0	0	0
masyarakat	0	0	0	0	2.704748	0	0	0	0	0	0
mati	0	0	2.704748	0	0	0	0	0	0	0	0
minta	0	0	2.704748	0	0	0	0	0	0	0	0
normal	2.704748	0	0	0	0	0	0	0	0	0	0
pal	0	0	0	0	0	2.704748	0	0	0	0	0
panas	0	0	0	0	0	0	0	2.704748	0	0	0
pandemi	0	2.704748	0	0	0	0	0	0	0	0	0
pasti	0	0	2.704748	0	0	0	0	0	0	0	0
pemprov	0	0	0	2.704748	0	0	0	0	0	0	0
perintah	2.704748	0	0	0	0	0	0	0	0	0	0
pesan	0	0	0	0	0	0	0	0	2.704748	0	0
pranowo	0	0	0	0	0	0	0	0	2.704748	0	0
presiden	0	0	2.704748	0	0	0	0	0	0	0	0
psbb	0	0	2.704748	0	0	0	0	0	0	0	0
rapid	0	0	0	0	0	0	2.704748	0	0	0	0
ribu	0	0	0	0	0	0	2.704748	0	0	0	0
rp550	0	0	0	0	0	0	2.704748	0	0	0	0
salat	0	0	0	0	0	0	0	0	2.704748	0	0
sanggup	0	0	0	0	0	0	0	0	0	0.353553	0.353553
saran	0	2.704748	0	0	0	0	0	0	0	0	0
soekarnohatta	0	0	0	0	0	0	2.704748	0	0	0	0
solo	0	0	0	0	0	0	0	0	0	2.704748	0
tangan	0	2.704748	0	0	0	0	0	0	0	0	0
tekan	0	0	2.704748	0	0	0	0	0	0	0	0
teliti	0	2.704748	0	0	0	0	0	0	0	0	0
tengah	0	0	0	0	0	0	0	0	2.704748	0	0
terima	0	0	0	2.704748	0	0	0	0	0	0	0
test	0	0	0	0	0	0	2.704748	0	0	0	0
tidak	0	0	0	0	0	0	0	0	0	2.704748	0
tiket	0	0	0	0	0	0	0	0	0	2.704748	0
tinggal	0	0	0	0	0	2.704748	0	0	0	0	0
tugas	0	0	2.299283	2.299283	0	0	0	0	0	0	0
urus	0	0	0	0	2.704748	0	0	0	0	0	0
walikota	0	0	0	0	0	0	0	0	0	2.704748	0
warga	0	0	0	0	0	0	0	0	2.704748	0	0
yurianto	0	0	2.704748	0	0	0	0	0	0	0	0
Label	Fakta	Fakta	Fakta	Fakta	Fakta	Hoaks	Hoaks	Hoaks	Hoaks	Hoaks	?

Untuk mengklasifikasi data uji , diperlukan rata – rata centroid dari masing-masing kelas yaitu kelas fakta dan hoaks, kedua kelas dihitung dengan rumus :

$$\vec{u}(c) = \frac{1}{Dc} \Sigma_{d \epsilon Dc} \vec{v}(d)$$

$$C_{Fakta} = \frac{1}{5} (\sum X1, \sum X2, \ldots, \sum Xn)$$

$$C_{Hoaks} = \frac{1}{5} (\sum X1, \sum X2, \ldots, \sum Xn)$$

$$C_{Fakta} = \frac{1}{5} (\sum X1, \sum X2, \ldots, \sum Xn)$$

$$= (0.540949618 , 0.919713194 , \ldots ,0)$$

$$C_{Hoaks} = \frac{1}{5} (\sum X1, \sum X2, \ldots, \sum Xn)$$

$$= (0, 0, \ldots ,0)$$

$$C_{Data\ uji} = \frac{1}{1} (\sum X1, \sum X2, \ldots, \sum Xn)$$

$$= (0, 0, \ldots ,0)$$

Selanjutnya mencari jarak antara centroid data uji dengan centroid data kelas dengan Euclidean Distance :

Misal : Data uji = p , fakta = q , hoaks = r

$$d(p, q) = | p - q |$$

$$= ((p1 - q1)^2 + (p2 - q2)^2 + (pn - qn)^2)^{\frac{1}{2}}$$

$$= 17.74056852$$

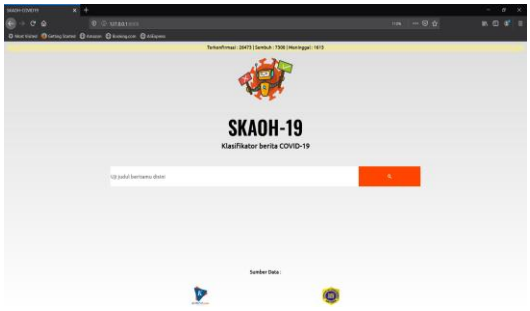
$$d(p, r) = | p - r |$$

$$= ((p1 - r1)^2 + (p2 - r2)^2 + (pn - rn)^2)^{\frac{1}{2}}$$

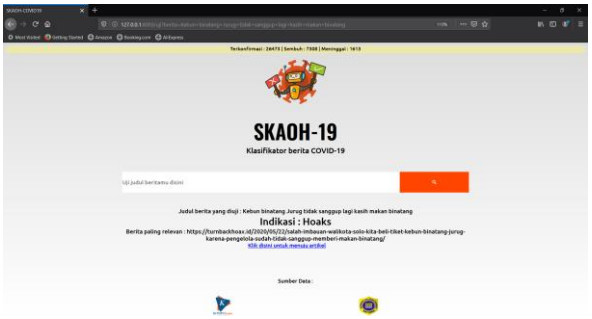
$$= 16.35268721$$

Karena jarak centroid data uji dan hoaks lebih kecil dibanding dengan fakta , maka data uji dklasifikasi sebagai hoaks karena dianggap jarak antara keduanya lebih dekat.

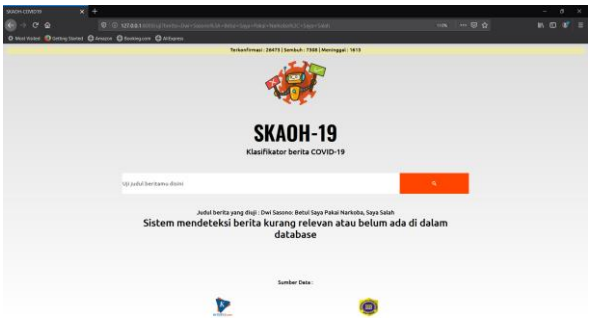
Tampilan Program



Gambar 1. Halaman awal aplikasi



Gambar 2. Halaman hasil klasifikasi



Gambar 3. Halaman data uji tidak relevan

Berikut hasil pengujian model klasifikai dengan cross validation yang dilakukan 5 kali dengan bobot persentase data yang berbeda :

Tabel 5. Tabel akurasi hasil pengujian dengan 5 skema pengujian

Skema	Data Latih (%)	Data Uji (%)	Akurasi (%)
1	90	10	89
2	85	15	86
3	80	20	87
4	75	25	88
5	70	30	88

III. KESIMPULAN DAN SARAN

3.1. Kesimpulan

Setelah melakukan perancangan, pembuatan, dan pengujian terhadap aplikasi ini, maka dapat diambil kesimpulan yaitu :

1. Aplikasi ini membantu pengguna dalam memberikan klasifikasi dan rujukan berita uji terkait COVID-19.
2. Aplikasi ini memiliki rata – rata tingkat akurasi sebesar 87.6% setelah melakukan pengujian dengan 5 skema pengujian yang berbeda.

3.2. Saran

Berdasarkan analisa yang telah dilakukan dan diharapkan penggunaan aplikasi ini berjalan dengan lancar, maka penulis memberikan beberapa saran agar aplikasi ini menjadi lebih baik lagi, antara lain :

1. Menambahkan proses berupa sinonim, agar klasifikasi data menjadi lebih mudah dan akurat.
2. Menambahkan proses untuk mengendalikan pernyataan negasi untuk menghasilkan klasifikasi berita yang lebih baik.

REFERENSI

[1]. Hardeniya, Nitin, Jacob Perkins, Deepti Chopra, Nisheeth Joshi, dan Iti Mathur. Natural Language Processing: Python and NLTK. Birmingham: Packt Publishing Ltd, 2016.

[2]. Afriza, Aulia, dan Julio Adisantoso. "Metode Klasifikasi Rocchio untuk Analisis Hoax ." Jurnal Ilmu Komputer Agri Informatika, 2018: 1-10.

[3]. Lok, , Leon. Finding Similar Names Using Cosine Similarity. 18 Maret 2020. https://towardsdatascience.com/finding-similar-names-using-cosine-similarity-a99eb943e1ab (diakses Juli 4, 2020).

BIOGRAPHY

Indrawan Gotama, lahir di Jakarta pada 19 September 1998. Menyelesaikan kuliah Strata I (S1) pada tahun 2020 pada Program Studi Teknik Informatika di Universitas Buddhi Dharma. Saat ini bekerja sebagai Programmer di PT Gaya Makmur Mobil.

Susanto Hariyanto, lahir di Pontianak pada tahun 1986. Menyelesaikan Magister Komputer di STMIK Eresha tahun 2012. Saat ini mengajar pada Program Studi Teknik Informatika di Universitas Buddhi Dharma sejak tahun 2019. Bidang penelitian dan publikasi ilmiah yang diminati adalah data mining dan Internet of Things.

Hartana Wijaya, Saat ini bekerja sebagai dosen Tetap pada Program Studi Teknik Informatika di Universitas Buddhi Dharma.