Department of Digital Business
Journal of Artificial Intelligence and Digital Business (RIGGS)
Homepage: https://journal.ilmudata.co.id/index.php/RIGGS
Vol. 4 No. 3 (2025) pp: 3655-3665
P-ISSN: 2963-9298, e-ISSN: 2963-914X

Identifikasi dan Analisis Individual User pada Natural Language dengan N-Gram Analisis

Imamulhakim Syahid Putra[1], Aminullah Imal Alfresi[2]
[1]Fakultas Sains dan Teknologi, Prodi Sistem Informasi, Universitas Islam Negeri Raden Fatah Palembang, Palembang, Indonesia
[1*]imamulhakim_uin@radenfatah.ac.id, [2]aminullah@radenfatah.ac.id

Abstrak

User profil merupakan salah satu pendekatan untuk mengindentifikasi intruder pada sebuah sistem komputer. Profil pengguna sangat penting pada semua aplikasi terutama untuk mengidentifikasi informasi spesifik tentang pengguna itu sendiri. Pada dasarnya profiling membangun informasi atau pengalaman tentang pengguna atau user. Dengan kata lain, profil pengguna digunakan untuk mengumpulkan informasi terkait aktivitas pengguna. Penelitian ini fokus pada sisi psychometric user yaitu mengidentifikasi gaya penulisan user berdasarkan natural language. Tujuan dari penelitian ini adalah untuk memudahkan identifikasi user berdasarkan gaya penulisannya. Sehingga dapat mendeteksi orang yang ingin menyerang sistem keamanan komputer. Analisis yang digunakan dalam penelitian ini adalah N-Gram. N-Gram analisis dapat mengidentifikasi user secara akurat terlihat dari hasil yang dicapai dari setiap jenis N-gram yang digunakan. Pada penelitian ini N-Gram akan membandingkan setiap aktivitas penulisan user sehingga dapat menentukan jenis authorization user.

Kata kunci: N-Gram; User Profil; Identifikasi; Informasi; Penulisan

1. Latar Belakang

Lahirnya situs jejaring sosial telah merubah cara *user* berkomunikasi. Penggunaan dari website tidak hanya fokus pada penyaluran informasi, membangun bisnis online tetapi saat ini website telah digunakan untuk interaksi antar pengguna. Saat ini situs jejaring sosial telah populer digunakan sebagai interaksi antar *user* dan *web interactive* yang menyebabkan penguna lebih aktif berkembang dari website misalnya Facebook dan MSN merupakan revolusi dari web 2.0 (Vosecky et al., 2009). Jaringan sosial dibutuhkan untuk beberapa aktivitas seperti profil dan identifikasi kebiasaan user (Raad et al., 2014). Lebih jauh lagi, keamanan komputer adalah penerapan keamanan informasi, subbidang teknologi pada komputer. Melindungi informasi dari pencurian, kesalahan data, pemeliharaan, dan ketersediaan merupakan salah satu tujuan keamanan komputer. Karena keamanan komputer sering kali berbentuk pembatasan terhadap apa yang tidak boleh dilakukan komputer, hal itu juga memberlakukan batasan pada komputer yang berbeda dari sebagian besar sistem lainnya (Abou-Assaleh et al., 2004). Karena sudah cukup sulit untuk membuat program komputer melakukan apa yang seharusnya mereka lakukan, hal ini membuat komputer dan sistem informasi menjadi jauh lebih rumit. Keamanan komputer sekarang ini tidak hanya terbatas pada ruang lingkup hardware, software dan pengguna tetapi telah berkembang pada sisi interaksi *user* antar ketiga komponen tersebut. Seperti keamanan dari sisi pengumpulan informasi tentang pengguna atau *user* pada situs jejaring sosial sehingga nantinya diharapkan dapat mendapatkan informasi yang akurat tentang pengguna tersebut atau singkatnya disebut *User Profiling*.

User profiling adalah suatu cara untuk mengumpulkan informasi personal atau individu ke katagori-katagori tertentu berdasarkan karakteristik seperti situasi, penampilan dan sifat (VNP Dao, R Vemuri, 2000). *Profiling* di dunia komputer digunakan untuk mendapatkan informasi tentang aktifitas *user*. Selain itu *user profiling* juga dapat berfungsi untuk mendeteksi kelainan atau kecurigaan pada akses sistem komputer. *Profiling* dan identifikasi pengguna digunakan untuk mengidentifikasi karakteristik psikologi pengguna yang bertujuan untuk mengidentifikasi pengguna di jejaring sosial, blog untuk pengguna sebenarnya.

Profiling digunakan diberbagai area teknologi informasi dan komputer, misalnya untuk mengetahui cara penggunaan komputer. Sehingga pengguna dapat mendistribusikan sistem *resource* lebih efisien dan memberikan

layanan terbaik. Selain itu *profiling* digunakan untuk mengidentifikasi *user* pada perdagangan elektronik atau *e-commerce* (G Pannell, 2010).

Analisis N-gram adalah salah satu analisis yang akan digunakan dalam penelitian ini. Menurut (F Luo, Q OU, 2010), "Model bahasa berdasarkan hubungan kolinear" adalah analisis N-gram. Rangkaian dengan n karakter dikenal sebagai hubungan kolinear. Kata-kata yang terus-menerus dibaca dari teks sumber hingga akhir dokumen dipecah menjadi potongan-potongan n-karakter dan dikenakan analisis N-gram (SA Sugianto, L Liliana, 2013).

Sebaliknya, analisis N-gram digunakan dalam pembuatan kata untuk mengekstrak n fragmen kata dari serangkaian kata (kalimat, paragraf, bacaan) yang terus-menerus dibaca dari teks sumber hingga kesimpulan dokumen.

Salah satu manfaat memanfaatkan N-gram daripada kata-kata utuh adalah N-gram lebih kecil kemungkinannya mengalami kesalahan ejaan dalam suatu dokumen (Mustafa, 2005).

Beberapa analisis N-gram digunakan dalam berbagai tujuan salah satu contohnya adalah penelitian yang dilakukan (Zhang et al., 2006) menguraikan penggunaan algoritma *malicious code* ekstraksi fitur berbasis N-gram dengan model bahasa statistik. Dengan menggunakan trigram (3-gram) model yang dapat mendeteksi fitur *malicious code* dan mendeteksi virus. Penggunaan analisis N-gram menawarkan efisiensi dan ketepatan dalam analisis *malicious code*.

Penelitian ini akan menganalisis dua bentuk yang berbeda dari data dalam dua cara, pertama untuk memeriksa apakah dapat mendeteksi ketika pengguna saat tidak cocok dengan profil pengguna dan karenanya penyusup, ini adalah identifikasi negatif. Cara kedua adalah untuk mendeteksi apakah pengguna saat ini dapat diragukan lagi diverifikasi sebagai pengguna benar, ini adalah identifikasi positif. Kebanyakan sistem deteksi intrusi menganggap bahwa pengguna asli sampai anomali atau melanggar peraturan menunjukkan sebaliknya, yaitu, mereka hanya memanfaatkan identifikasi negatif. Namun hal ini mungkin berguna untuk membatasi aktivitas pengguna sampai mereka positif mengidentifikasi diri mereka, mungkin tidak memungkinkan pengguna untuk membuat perubahan signifikan sampai sesi login mereka saat ini telah diidentifikasi positif.

Pemetaan tulisan dari *user* akan menggunakan dua metode yaitu grafik dan t-test. Dari grafik tersebut bisa di lihat apakah *user* yang sama atau *user* yang berbeda mempunyai kesamaaan atau perbedaan dalam gaya penulisan mereka. Kemudian untuk mendapatkan hasil yang terukur maka penulis membandingkan gaya penulisan dari setiap *user* dengan menggunakan alat bantu t-test.

2. Metode Penelitian

Metodologi penelitian berisi tentang pelaksanaan penelitian seperti jenis dan sumber data, pengaturan eksperimental, konsep penelitian dan metode penelitian.

2.1. Panjang Naskah (subtitle tidak tebal)

2.1.1. Jenis dan Sumber Data

1.Jenis Data

Jenis data yang digunakan dalam penelitian ini adalah data kuantitatif, dimana data tersebut diolah dan dianalisis menggunakan teknik perhitungan matematika atau statistika.

2.Sumber Data

Sumber data yang digunakan pada penelitian ini yakni data sekunder. Adapun data sekunder adalah data yang diperoleh atau dikumpulkan peneliti dari berbagai sumber yang telah ada (peneliti sebagai tangan kedua). Data sekunder diperoleh dari berbagai sumber seperti buku, novel, situs jejaring sosial, blog, dan lain-lain.

Adapun sumber data yang diambil untuk penelitian ini yaitu sebagai berikut:

A. Sample Analisis nGram Oprah Winfrey

- Sample dari Blog, referensi sample dari blog yang dijadikan objek penelitian yaitu dengan alamat http://blog.gaiam.com/quotes/authors/oprah-winfrey
- Sample dari Facebook**,** referensi yang dijadikan objek penelitian yaitu: https://www.facebook.com/oprahwinfrey/posts/153682314705526

DOI: https://doi.org/10.31004/riggs.v4i3.2532

- Sample dari Twitter, referensi sample dari twitter yang dijadikan objek penelitian yaitu https://twitter.com/oprah

B. Sample Analisis nGram William Shakespeare

Merupakan 3 judul novel karya dari William Shakespeare yaitu :

- Julius Caesar,
- Othello,
- Tempest

C. Sample Analisis nGram Blog Raditya Dika

- Sample dari Blog, referensi sample dari blog yang dijadikan objek penelitian yaitu dengan alamat http://radityadika.com/category/blog

2.1.2.Pengaturan Eksperimental

Dalam penelitian ini, penulis menggunakan bantuan software aplikasi yang menggunakan bahasa pemrograman java. Ada dua kelas pada aplikasi ini yaitu “Ngram.java” dan “Data.java”. Program ini dapat berjalan dengan perintah “java Ngram [n]” dimana n adalah nilai atau banyaknya N-Gram. Software ini akan menghasilkan N-Gram frekuensi yang akan diletakkan pada “csv” folder pada microsoft excel dan “csv” folder tersebut terdapat data history yang berupa txt.

2.1.3.Konsep dan Metode Penelitian

2.1.3.1. Profiling Menggunakan Natural Language (Bahasa Alami)

Tahap bahasa alami ini akan mengevaluasi penggunaan analisis n-gram untuk profil pengguna sesuai dengan penggunaan bahasa alami.

pertanyaan penelitian 1: Apakah penggunaan analisis n-gram ke profil gaya penulisan pengguna 'dalam situasi jaringan sosial memungkinkan identifikasi pengguna akurat?

Mempelajari analisis studi penelitian pertama, penggunaan analisis n-gram untuk profil pengguna untuk identifikasi pengguna yang akurat. Akibatnya, hal itu memungkinkan identifikasi pengguna positif dan negatif (Hubballi et al., 2011).

Kita perlu membuat n-gram spektrum diketahui pengguna untuk mengisi profil mereka. Jadi untuk setiap pengguna, kami akan menciptakan banyak n-gram spektrum, masing-masing untuk nilai yang berbeda dari n. Setiap spektrum n-gram akan terdiri dari daftar diindeks dari nilai-nilai, di mana indeks berjalan dari 0 sampai 26n-1 - yaitu, ada entri dalam daftar untuk setiap kemungkinan kombinasi huruf alfabet. Sebagai contoh, A = 0 B = 1 C = 2 . . . Z = 25

3-gram, 3-huruf pada waktu memungkinkan nilai antara

AAA BAA CAA ..... ZAA AAB . . . AAZ

Menetapkan setiap huruf nilai 0 ..... 25, sehingga setiap 3 gram di nomor dalam basis 26 menghitung.

Misalnya. ABC = 0 x 26 ^ 2 + 1 x 26 ^ 1 + 2 x 26 ^ 0

0 1 2 = 0 + 26 + 2 = 28

Indeks dari n-gram nilai yang dihitung ini

misalnya. AbC = indeks 28 dalam daftar untuk 3-gram

Perlu diketahui bahwa kami hanya akan melakukan tugas ini untuk surat a..z awalnya (mengabaikan kasus) tetapi jika metode ini ditemukan menjanjikan, itu akan diperluas untuk mencakup semua karakter ASCII.

Untuk setiap sampel, kita menghitung spektrum n-gram teks. Kami kemudian membandingkan n-gram spektrum yang dihasilkan oleh penulis yang sama tetapi menjadi sampel yang berbeda (dalam hal ini, buku yang berbeda), serta membandingkan spektrum n-gram yang dihasilkan oleh penulis yang berbeda. Sebagai penulis yang 'dikenal', perbandingan akan menentukan apakah metode ini cukup akurat untuk mengidentifikasi ketika penulis adalah sama atau berbeda. Perbandingan yang akan dilakukan melalui t-test. Untuk nilai yang diberikan dari n, kami menghitung spektrum n-gram input teks pengguna saat ini, membandingkannya dengan spektrum dengan nilai

DOI: https://doi.org/10.31004/riggs.v4i3.2532

yang sama dari n di profil pengguna, dan jika t-test menunjukkan mereka memiliki sedikit perbedaan, maka pengguna diidentifikasi positif (yaitu penulis yang sama), tetapi jika perbedaan yang signifikan, pengguna akan diidentifikasi secara negatif (yaitu penulis yang berbeda).

Bagian dari penelitian ini adalah untuk menentukan nilai yang paling berguna dari n dalam analisis n-gram, misalnya apakah lebih pendek atau lebih n-gram akan memberikan identifikasi yang paling akurat.

2.1.3.2. N-gram Analisis

Analisis N-gram adalah salah satu analisis yang akan digunakan dalam penelitian ini. Menurut (F Luo, Q OU, 2010), analisis n-gram adalah 'model bahasa berdasarkan hubungan collinear. Hubungan coliniear adalah potongan sejumlah n karakter dari sebuah string. Analisis n-gram ini digunakan untuk mengambil potongan potongan karakter huruf sejumlah n dari sebuah kata yang secara kontinuitas dibaca dari teks sumber hingga akhir dari dokumen (SA Sugianto, L Liliana, 2013).

Sebagai contoh : kata ”TEXT” dapat diuraikan ke dalam beberapa

n-gram berikut:

uni-gram : T, E, X, T

bi-gram : TE, EX, XT

tri-gram :, TEX, EXT

quad-gram : TEXT, EXT_

dan seterusnya. Sedangkan pada pembangkitan kata, analisis n-gram ini digunakan untuk mengambil potongan kata sejumlah n dari sebuah rangkaian kata (kalimat, paragraf, bacaan) yang secara kontinuitas dibaca dari teks sumber hingga akhir dari dokumen.

Sebagai contoh : kalimat ”saya dapat melihat cahaya itu.” Dapat diuraikan ke dalam beberapa n-gram berikut :

uni-gram : saya, dapat, melihat, cahaya, itu

bi-gram : saya dapat, dapat melihat, itu ada

tri-gram : saya dapat melihat, dapat melihat dia

dan seterusnya. Salah-satu keunggulan menggunakan n-gram dan bukan suatu kata utuh secara keseluruhan adalah bahwa n-gram tidak terlalu sensitif terhadap kesalahan penulisan yang terdapat pada suatu dokumen.

2.1.3.3. T-Test

T-test adalah uji komparasi yang dapat dilakukan untuk memutuskan apakah dua set data (sampel) yang sama atau berbeda dan menyimpulkan apakah mereka bisa datang dari populasi yang sama. Hal ini menilai apakah sarana dua kelompok secara statistik berbeda satu sama lain. Analisis ini berguna ketika kita ingin menentukan apakah sarana dua kelompok yang sama atau berbeda. Kami akan menggunakan t-tes untuk menilai gaya penulisan bahasa alami (natural language), antara dua sampel dari pengguna yang sama (untuk tujuan identifikasi positif). Kami selanjutnya mempertimbangkan mana bentuk t-tes yang tepat untuk penelitian ini.

- One sample t-test

Uji satu sampel digunakan untuk memutuskan apakah sampel tertentu berasal dari populasi tertentu. Misalnya, ketika kita ingin tahu tentang sampel spesifik mahasiswa mirip dengan atau berbeda dari mahasiswa pada umumnya. Dalam penelitian saat kita membandingkan serangkaian kata-kata atau perintah dan sementara nanti mungkin layak untuk mengidentifikasi pengguna dari nilai n-gram tunggal, pada tahap awal itu lebih tepat untuk memutuskan apakah individu dapat diidentifikasi dari jumlah yang lebih besar dari tulisan-tulisan mereka.

- Independent t-test

T-test, atau dua sample t-test independent, digunakan untuk menentukan apakah dua sampel secara statistik sama atau berbeda satu sama lain antara sarana dalam dua kelompok yang tidak terkait. Misalnya, ketika kita ingin tahu antara mahasiswa pria dan wanita berbeda atau serupa pada beberapa karakteristik psikologis. Dalam penelitian ini, sampel mungkin tidak berhubungan, terutama ketika membandingkan dua sampel dari pengguna yang sama.

- Dependent t-test

Dependent t-test, disebut juga uji t-test berpasangan-kelompok, uji t-test berkorelasi-kelompok, kelompok matched- uji t atau tergantung kelompok t-test. Uji t-test ini digunakan untuk membandingkan dua sampel yang

DOI: https://doi.org/10.31004/riggs.v4i3.2532

berhubungan (cocok atau terkait dengan cara yang sama) yang baik diukur sekali atau sampel yang sama diukur pada dua kesempatan terpisah. Misalnya, ketika kita ingin tahu bagaimana pengaruh penggunaan obat tertentu untuk insomnia, untuk pasien yang sama atau berbeda setelah mengkonsumsi obat. Dalam hal ini kita akan melihat bagaimana efek dari obat untuk pasien sebelum dan setelah mengkonsumsi obat. Ini sangat cocok untuk penelitian ini seperti yang kita perlu positif mengidentifikasi pengguna dengan membandingkan sampel saat tulisan pengguna untuk sampel yang lebih tua dari tulisan mereka.

3. Hasil dan Diskusi

3.1. Tahapan-Tahapan Analisis N-gram

Tahapan-tahapan analisis nGram menggunakan beberapa tahapan baik untuk Oprah Winfrey, William Shakespeare dan Blog Raditya Dika (2ngram dan 4ngram). Tahapan yang digunakan yaitu :

3.1.1. Compile data

Data sample tersebut yang akan dilakukan analisis disimpan ke folder history berformat txt. (notepad) yang disinkronkan ke program java.

Selanjutnya data yang telah di simpan ke notepad, kemudian tulisan tersebut di compile dengan program java, sehingga menampilkan hasil data yang sudah dicompile ke dalam excel.

3.1.2. Fungsi Match

Fungsi match mencocokkan untuk beberapa kata yang muncul pada parameter.

3.1.3. Fungsi Index

Fungsi index untuk mentedeksi berapa kali dari kata yang sudah terdeteksi.

3.1.4. Fungsi ISNA

Fungsi ISNA digunakan untuk menghasilkan nilai TRUE (Benar) jika data pada cell merupakan Error atau kesalahan #NA.

3.1.5 Total Rata-Rata Standardeviasi

Yaitu menghitung total seluruh nilai n-gram (SUM), menghitung nilai rata-rata dari nilai n-gram (AVERAGE) dan menghitung nilai standar deviasi dari nilai n-gram (STDEV).

3.1.6. Percentage Normalisasi

Fungsi Percentage untuk menghitung setiap nilai dari n-gram dan dibagi dengan nilai total semua n-gram dan waktu untuk seratus.

Rumus untuk percentage yaitu (angka/ total) *100

3.1.7. Zscore Normalisasi

Fungsi Zscore untuk mengetahui lebih detail dimana posisi suatu skor dalam suatu distribusi dan juga memberi tahu berapa jarak skor itu sendiri dengan mean.

Rumus Zscore yaitu (angka-rata-rata)/standardeviasi

3.1.8. T-Test

Adalah uji komparasi yang dapat dilakukan untuk memutuskan apakah set data (sampel) yang sama atau berbeda dan menyimpulkan apakah mereka bisa datang dari populasi yang sama.

Nilai probabilitas p adalah output dengan t-test. Hasil nilai probabilitas adalah dibandingkan dengan tingkat signifikansi yang dipilih untuk menyimpulkan hasil tes.

Sebuah standar umum adalah $\alpha = 0.05$:

- Jika nilai probabilitas atau t hitung sama atau kurang maka tingkat signifikansi, kita dapat menolak hipotesis nol dan menyimpulkan bahwa sampel penulisan yang berbeda dengan user sebenarnya.
- Jika nilai probabilitas atau t hitung lebih dari tingkat signifikansi, kita dapat menerima hipotesis dan menyimpulkan bahwa sampel gaya penulisan yang sama atau mirip dengan user sebenarnya.

DOI: https://doi.org/10.31004/riggs.v4i3.2532

3.2 Hasil

- Sample Oprah Winfrey
  - Facebook dan Twitter

	Facebook	*Twitter*
Mean	0.26455	0.26455
Variance	0.103982	0.139458
Observations	378	378
Pearson Correlation	0.593419	

Korelasi antara dua sampel (Pearson Correlation) adalah suatu ukuran hubungan linier antar variabel. Dimana nilai r hubungan korelasi oprah winfrey facebook dan twitter sebesar = 0,5934 maka dari itu penulis membuat interval kategorisasi kekuatan hubungan korelasi sebagai berikut :

0	:	Tidak ada korelasi
0,00 – 0,25	:	Korelasi sangat lemah
0,25 – 0,50	:	Korelasi cukup
0,50 – 0,75	:	Korelasi kuat
0,75 – 0,99	:	Korelasi sangat kuat
1	:	Korelasi sempurna

Berdasarkan nilai korelasi antara dua sampel sebesar 0,5934 setelah dicocokkan dengan interval kategorisasi kekuatan hubungan korelasi menyatakan bahwa hubungan korelasi oprah winfrey facebook dan twitter termasuk kategori Korelasi Kuat.

- Facebook dan Blog

0	:	Tidak ada korelasi
0,00 – 0,25	:	Korelasi sangat lemah
0,25 – 0,50	:	Korelasi cukup
0,50 – 0,75	:	Korelasi kuat
0,75 – 0,99	:	Korelasi sangat kuat
1	:	Korelasi sempurna

Berdasarkan nilai korelasi antara dua sampel sebesar 0,4998 setelah dicocokkan dengan interval kategorisasi kekuatan hubungan korelasi menyatakan bahwa hubungan korelasi oprah winfrey facebook dan blog termasuk kategori Korelasi Cukup.

- Twitter dan Blog

	Twitter	*Blog*
Mean	0.26455	0.26455
Variance	0.139458	0.136959
Observations	378	378
Pearson Correlation	0.823646	

Korelasi antara dua sampel (Pearson Correlation) adalah suatu ukuran hubungan linier antar variabel. Dimana nilai r hubungan korelasi oprah winfrey twitter dan blog sebesar = 0,8236 maka dari itu penulis membuat interval kategorisasi kekuatan hubungan korelasi sebagai berikut :

DOI: https://doi.org/10.31004/riggs.v4i3.2532

0	:	Tidak ada korelasi
0,00 – 0,25	:	Korelasi sangat lemah
0,25 – 0,50	:	Korelasi cukup
0,50 – 0,75	:	Korelasi kuat
0,75 – 0,99	:	Korelasi sangat kuat
1	:	Korelasi sempurna

Berdasarkan nilai korelasi antara dua sampel sebesar 0,8236 setelah dicocokkan dengan interval kategorisasi kekuatan hubungan korelasi menyatakan bahwa hubungan korelasi oprah winfrey twitter dan blog termasuk kategori Korelasi Sangat Kuat.

- **Sample William Shakespeare**
  - Othello dan Tempest

	Othello	*Tempest*
Mean	0.214592	0.214592275
Variance	0.105052	0.12587893
Observations	466	466
Pearson Correlation	0.655204	

Korelasi antara dua sampel (Pearson Correlation) adalah suatu ukuran hubungan linier antar variabel. Dimana nilai r hubungan korelasi William Shakespeare othello dan tempest sebesar = 0,655 maka dari itu penulis membuat interval kategorisasi kekuatan hubungan korelasi sebagai berikut :

0	:	Tidak ada korelasi
0,00 – 0,25	:	Korelasi sangat lemah
0,25 – 0,50	:	Korelasi cukup
0,50 – 0,75	:	Korelasi kuat
0,75 – 0,99	:	Korelasi sangat kuat
1	:	Korelasi sempurna

Berdasarkan nilai korelasi antara dua sampel sebesar 0,655 setelah dicocokkan dengan interval kategorisasi kekuatan hubungan korelasi menyatakan bahwa hubungan korelasi William Shakespeare othello dan tempest termasuk kategori Korelasi Kuat.

- Othello dan Julius Caesar

	Othello	*Julius Caesar*
Mean	0.214592	0.214592275
Variance	0.105052	0.109514074
Observations	466	466
Pearson Correlation	0.690868	

DOI: https://doi.org/10.31004/riggs.v4i3.2532

Korelasi antara dua sampel (Pearson Correlation) adalah suatu ukuran hubungan linier antar variabel. Dimana nilai r hubungan korelasi William Shakespeare othello dan julius caesar sebesar = 0,6908 maka dari itu penulis membuat interval kategorisasi kekuatan hubungan korelasi sebagai berikut :

0	:	Tidak ada korelasi
0,00 – 0,25	:	Korelasi sangat lemah
0,25 – 0,50	:	Korelasi cukup
0,50 – 0,75	:	Korelasi kuat
0,75 – 0,99	:	Korelasi sangat kuat
1	:	Korelasi sempurna

Berdasarkan nilai korelasi antara dua sampel sebesar 0,6908 setelah dicocokkan dengan interval kategorisasi kekuatan hubungan korelasi menyatakan bahwa hubungan korelasi William Shakespeare othello dan julius caesar termasuk kategori Korelasi Kuat.

- Tempest dan Julius Caesar

	Tempest	*Julius Caesar*
Mean	0.214592	0.214592275
Variance	0.125879	0.109514074
Observations	466	466
Pearson Correlation	0.877582	

Korelasi antara dua sampel (Pearson Correlation) adalah suatu ukuran hubungan linier antar variabel. Dimana nilai r hubungan korelasi William Shakespeare tempest dan julius caesar sebesar = 0,8775 maka dari itu penulis membuat interval kategorisasi kekuatan hubungan korelasi sebagai berikut :

0	:	Tidak ada korelasi
0,00 – 0,25	:	Korelasi sangat lemah
0,25 – 0,50	:	Korelasi cukup
0,50 – 0,75	:	Korelasi kuat
0,75 – 0,99	:	Korelasi sangat kuat
1	:	Korelasi sempurna

Berdasarkan nilai korelasi antara dua sampel sebesar 0,8775 setelah dicocokkan dengan interval kategorisasi kekuatan hubungan korelasi menyatakan bahwa hubungan korelasi William Shakespeare tempest dan julius caesar termasuk kategori Korelasi Sangat Kuat.

- **Blog Raditya Dika**
  - Artikel 3 dan Artikel 2

	artikel 3	*artikel 2*
Mean	0.454545	0.4545455
Variance	0.309219	0.3883494
Observations	220	220
Pearson Correlation	0.657618	

Korelasi antara dua sampel (Pearson Correlation) adalah suatu ukuran hubungan linier antar variabel. Dimana nilai r hubungan korelasi Blog Raditya Dika artikel 3 dan artikel 2 sebesar = 0,657 maka dari itu penulis membuat interval kategorisasi kekuatan hubungan korelasi sebagai berikut :

DOI: https://doi.org/10.31004/riggs.v4i3.2532

0	:	Tidak ada korelasi
0,00 – 0,25	:	Korelasi sangat lemah
0,25 – 0,50	:	Korelasi cukup
0,50 – 0,75	:	Korelasi kuat
0,75 – 0,99	:	Korelasi sangat kuat
1	:	Korelasi sempurna

Berdasarkan nilai korelasi antara dua sampel sebesar 0,657 setelah dicocokkan dengan interval kategorisasi kekuatan hubungan korelasi menyatakan bahwa hubungan korelasi Blog Raditya Dika artikel 3 dan artikel 2 termasuk kategori Korelasi Kuat.

- Artikel 3 dan Artikel 1

	artikel 3	*artikel 1*
Mean	0.454545	0.4545455
Variance	0.309219	0.4256462
Observations	220	220
Pearson Correlation	0.692346	

Korelasi antara dua sampel (Pearson Correlation) adalah suatu ukuran hubungan linier antar variabel. Dimana nilai r hubungan korelasi Blog Raditya Dika artikel 3 dan artikel 1 sebesar = 0,692 maka dari itu penulis membuat interval kategorisasi kekuatan hubungan korelasi sebagai berikut :

0	:	Tidak ada korelasi
0,00 – 0,25	:	Korelasi sangat lemah
0,25 – 0,50	:	Korelasi cukup
0,50 – 0,75	:	Korelasi kuat
0,75 – 0,99	:	Korelasi sangat kuat
1	:	Korelasi sempurna

Berdasarkan nilai korelasi antara dua sampel sebesar 0,692 setelah dicocokkan dengan interval kategorisasi kekuatan hubungan korelasi menyatakan bahwa hubungan korelasi Blog Raditya Dika artikel 3 dan artikel 1 termasuk kategori Korelasi Kuat.

- Artikel 2 dan Artikel 1

	artikel 2	*artikel 1*
Mean	0.454545	0.454545
Variance	0.388349	0.425646
Observations	220	220
Pearson Correlation	0.593203	

Korelasi antara dua sampel (Pearson Correlation) adalah suatu ukuran hubungan linier antar variabel. Dimana nilai r hubungan korelasi Blog Raditya Dika artikel 2 dan artikel 1 sebesar = 0,5932 maka dari itu penulis membuat interval kategorisasi kekuatan hubungan korelasi sebagai berikut :

DOI: https://doi.org/10.31004/riggs.v4i3.2532

0	:	Tidak ada korelasi
0,00 – 0,25	:	Korelasi sangat lemah
0,25 – 0,50	:	Korelasi cukup
0,50 – 0,75	:	Korelasi kuat
0,75 – 0,99	:	Korelasi sangat kuat
1	:	Korelasi sempurna

Berdasarkan nilai korelasi antara dua sampel sebesar 0,5932 setelah dicocokkan dengan interval kategorisasi kekuatan hubungan korelasi menyatakan bahwa hubungan korelasi Blog Raditya Dika artikel 2 dan artikel 1 termasuk kategori Korelasi Kuat.

4. Kesimpulan

N-Gram	Analisis Data	Normalisasi Percentage	Hasil
2 N-gram	Oprah Winfrey	Facebook vs twitter	Sama
	William Shakespeare	Othello vs tempest	Sama
	Oprah Winfrey	Facebook vs blog	Sama
	William Shakespeare	Othello vs Julius Caesar	Sama
	Oprah Winfrey	Twitter vs blog	Sama
	William Shakespeare	Tempest vs Julius Caesar	Sama
	Blog Raditya Dika	Artikel 3 dan Artikel 2	Sama
	Blog Raditya Dika	Artikel 3 dan Artikel 1	Sama
	Blog Raditya Dika	Artikel 2 dan Artikel 1	Sama

Dari tabel kesimpulan di atas dapat di disimpulkan persentasi keberhasilan dalam mengidentifikasi user analisis data Oprah Winfrey, William Shakespeare dan Blog Raditya Dika pada 2 ngram dapat disimpulkan sama atau mirip baik kebenarannya percentage normalisasi. Penelitian ini menggunakan N-gram analisis yang mengambil sampel pada novel William Shakespeare, Blog Oprah Winfrey dan Blog Raditya Dika, penggunaan analisis N-gram tidak terlalu sensitif terhadap kesalahan penulisan yang terdapat pada suatu dokumen. Penelitian ini untuk menganalisa gaya penulisan pada *Natural language* dengan tujuan untuk dapat memungkinkan mendeteksi penyusup yang menyamar sebagai pengguna yang sebenarnya.

Referensi

1. Abou-Assaleh, T., Cercone, N., Kešelj, V., & Sweidan, R. (2004). N-gram-based detection of new malicious code. *Proceedings - International Computer Software and Applications Conference*, *2*(1), 41–42. https://doi.org/10.1109/cmpsac.2004.1342667
2. F Luo, Q OU, G. W. (2010). *Research on n-gram-based malicious code feature extraction algorithm. Computer Application and System Modeling (ICCASM)*. V6-89-V6-92.
3. G Pannell, H. A. (2010). *User modelling for exclusion and anomaly detection: a behavioural intrusion detection system*. 207–218. https://link.springer.com/chapter/10.1007/978-3-642-13470-8_20
4. Hubballi, N., Biswas, S., & Nandi, S. (2011). Sequencegram: n-gram modeling of system calls for program based anomaly detection. *2011 Third International Conference on Communication Systems and Networks (COMSNETS 2011)*, 1–10. https://doi.org/10.1109/COMSNETS.2011.5716416
5. Mustafa, S. H. (2005). Character contiguity in N-gram-based word matching: the case for Arabic text searching. *Information Processing & Management*, *41*(4), 819–827. https://doi.org/10.1016/j.ipm.2004.02.003
6. Raad, E., Chbeir, R., Dipanda, A., & Raad, E. (2014). *User profile matching in social networks To cite this version : User Profile Matching in Social Networks*. 297–304.

DOI: https://doi.org/10.31004/riggs.v4i3.2532

7. SA Sugianto, L Liliana, S. R. (2013). *Pembuatan Aplikasi Predictive Text Menggunakan Metode N-gram-based*. https://www.neliti.com/publications/105718/pembuatan-aplikasi-predictive-text-menggunakan-metode-n-gram-based
8. VNP Dao, R Vemuri, S. T. (2000). *Profiling users in the UNIX OS environment*. http: / /www.doc.pov/bridge
9. Vosecky, J., Hong, D., & Shen, V. Y. (2009). User identification across multiple social networks. *2009 First International Conference on Networked Digital Technologies*, 360–365. https://doi.org/10.1109/NDT.2009.5272173
10. Zhang, B., Yin, J., Hao, J., Wang, S., Zhang, D., & Tang, W. (2006). New malicious code detection based on N-gram analysis and rough set theory. *2006 International Conference on Computational Intelligence and Security, ICCIAS 2006*, *2*, 1229–1232. https://doi.org/10.1109/ICCIAS.2006.295252

DOI: https://doi.org/10.31004/riggs.v4i3.2532