KOMIK (Konferensi Nasional Teknologi Informasi dan Komputer) **ISSN 2597-4645 (media online)**
Volume 6, Nomor 1, November 2022 **ISSN 2597-4610 (media cetak)**
DOI: 10.30865/komik.v6i1.5738 **Page: 129-139**

Deteksi Berita Palsu Tentang Vaksinasi Covid-19 Dengan Menggunakan Text Mining Dan Algoritma Cosine Similarity

Diana Marta*, Guidio Leonarde Ginting, A.M Hatuaon Sihite

Fakultas Ilmu Komputer dan Teknologi Informasi, Program Studi Teknik Informatika, Universitas Budi Darma, Medan, Indonesia
Email: 1*dianamarta2408@gmail.com
Email Penulis Korespondensi: dianamarta2408@gmail.com

Abstrak-Indonesia sedang mengalami musibah wabah virus dunia yang disebut COVID- 19. Banyak berita palsu saat ini tersebar di media sosial adalah berita tentang vaksinasi. Penyebaran berita palsu hendaknya dapat dideteksi kebenarannya dengan menerapkan salah satu dari teknologi informasi yaitu *natural language processin*g (NLP). Algoritma yang Akan dipakai untuk mendukung deteksi berita palsu adalah algoritma text mining dan algoritma *cosine similarity*. Mencari referensi-referensi yang berkaitan dengan permasalahan, dimulai dari mencari buku-buku yang membahas tentang algoritma yang digunakan pada penelitian ini. Mencari data yang berhubungan dengan topik penelitian yaitu vaksinasi COVID-19. Pengolahan data dilakukan dengan menerapkan algoritma text mining sebagai *text processing*. Tahapan dari text mining yaitu *case folding, tokenizing, filtering, stemming*. Tahap analisis pengolahan dilakukan dengan menerapkan algoritma cosine similarity untuk analisis tingkat kesamaan antar data. Berdasarkan perhitungan dengan algoritma *cosine similarity* diatas, Q (berita yang di deteksi) memiliki nilai similaritas diatas 40% terhadap D31 dan D59, dimana pada label yang sudah dibuat pada data berita D31 dan D59 merupakan berita *hoax*. Probabilitas kemunculan berita dengan menggunakan rumus matematika untuk menghitung probabilitas, nilai probabilitas yang didapatkan Q (Berita yang di deteksi) hoax 100% dan fakta 0%. Berdasarkan hasil analisis terhadap data dengan menerapkan algoritma text mining dan *cosine similarity* diperoleh hasil yang menunjukkan berita yang diindetifikasi dinyatakan hoax sesuai dengan data yang dimiliki.

Kata Kunci: Berita; Vaksinasi; Covid-19**;** TF-IDF; Text Mining; *Cosine Similarity*

Abstract-Indonesia is currently experiencing an outbreak of a global virus called COVID-19. A lot of the fake news currently spreading on social media is news about vaccinations. The truth of spreading fake news should be detected by applying one of the information technologies, namely natural language processing (NLP). The algorithms that will be used to support fake news detection are text mining algorithms and cosine similarity algorithms. Looking for references related to the problem, starting from looking for books that discuss the algorithm used in this research. Looking for data related to the research topic, namely the COVID-19 vaccination. Data processing is done by applying a text mining algorithm as text processing. The stages of text mining are case folding, tokenizing, filtering, stemming. The processing analysis stage is carried out by applying the cosine similarity algorithm to analyze the level of similarity between data. Based on the calculation using the cosine similarity algorithm above, Q (detected news) has a similarity value of above 40% to D31 and D59, where the labels that have been made on D31 and D59 news data are hoax news. The probability of news appearing using a mathematical formula to calculate the probability, the probability value obtained by Q (news detected) is 100% hoax and 0% fact. Based on the results of the analysis of the data by applying the text mining and cosine similarity algorithms, the results show that the identified news is declared a hoax according to the data held.

Keywords: News; Vaccination; Covid-19; TF-IDF; Text Mining; Cosine Similarity

1. PENDAHULUAN

Perkembangan teknologi informasi bukan hal biasa lagi di era sekarang ini. Perkembangan teknologi informasi menjadikan manusia sebagai *user* menerima banyak keuntungan, beberapa diantaranya ialah dapat berhubungan jarak jauh dengan mudah, penyebaran dari informasi jauh lebih cepat, dunia pendidikan lebih maju dari yang dahulunya pembelajaran hanya dilakukan tatap muka sekarang pembelajaran dapat dilakukan secara online, dan masih banyak lagi keuntungan yang bisa peroleh dari perkembangan teknologi informasi.

Di sisi lain manusia sebagai *user* selain menerima banyak keuntungan dari perkembangan teknologi informasi juga menerima dampak negative yang ditimbulkan dari perkembangan teknologi informasi, beberapa diantaranya adalah semakin renggangnya kehidupan bersosial dalam dunia nyata, terjadinya kejahatan berbasis online seperti *hacking* dan *carding*, kemudian penyebaran ujaran kebencian terhadap kelompok melalui internet serta penyebaran berita *hoax* yang dapat merugikan banyak orang. Berita *hoax* merupakan salah satu dampak *negative* dari perkembangan teknologi informasi. Berita hoax itu sendiri dapat diartikan sebagai pesan atau suatu berita yang isinya memuat seakan-akan kebenarannya itu nyata, namun berita itu tidak dapat dibuktikan kebenarannya. Hingga saat ini berita *hoax* dengan cepat tersebar luas dikarenakan manusia sebagai *user* dari teknologi informasi dapat dengan mudah dan tanpa batas dalam mengakses informasi.

Adapun beberapa tujuan dari pembuatan dan penyebaran berita palsu (*hoax*) diantaranya ialah untuk keuntungan perorangan atau untuk kepentingan kelompok, dengan tujuan untuk memecah-mecahkan masyarakat, menakuti masyarakat sehingga membuat masyarakat merasa tidak aman di karenakan tergangu dengan adanya berita palsu (hoax) tersebut. Berita palsu (hoax) menjadi masalah yang serius di Indonesia, ada beberapa peraturan perundang – undangan yang mengatur mengenai berita palsu, salah satunya Undang- undang nomor 11 tahun 2008 tentang informasi dan transaksi elektronik (UU ITE). Dimana isi dari UU tersebut ialah Pelaku penyebar berita palsu akan dijerat hukuman 6 (enam) tahun penjara atau denda paling banyak Rp 1.000.000.000 (satu miliar).

Dimana pada saat ini, seperti yang di ketahui Indonesia sedang mengalami suatu musibah wabah virus dunia yang disebut COVID- 19. Banyak berita palsu yang saat ini tersebar di media sosial adalah berita tentang vaksinasi.

ISSN 2597-4645 (media online)
ISSN 2597-4610 (media cetak)
Page: 129-139

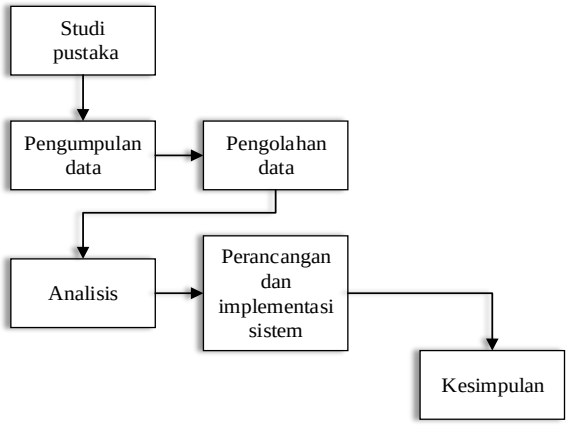
Dengan memanfaatkan perkembangan dari teknologi informasi, penyebaran berita palsu hendaknya dapat dideteksi kebenarannya dengan menerapkan salah satu dari teknologi informasi yaitu natural language processing (NLP). Salah satu algoritma yang Akan dipakai untuk mendukung mendeteksi berita palsu adalah algoritma text mining dan algoritma cosine similarity. Text mining adalah variasi dari data mining yang dapat mengestraksi informasi yang berguna dengan cara mengidentifikasi dan mengeksplorasi pola yang menarik dari sekumpulan sumber data tekstual yang tidak terstruktur. Ada beberapa tahapan pada text mining yaitu tahap tokenizing, filtering, stemming, tagging, analyzing. Algoritma *cosine similarity* merupakan metode untuk menghitung kesamaan antara dua buah objek yang dinyatakan dalam dua buah *vector* dengan menggunakan *keywords* (kata kunci) dari sebuah dokumen sebagai ukuran [1].

Dengan berdasarkan permasalahan yang disebutkan maka pada penelitian ini menggunakan algoritma text mining dan *cosine similarity* yang berguna untuk mengukur tingkat kesamaan berita palsu dengan berita benar atau fakta. Penelitian ini dilakukan untuk mendeteksi berita *hoax* tentang vaksinasi COVID-19 dengan membandingkan berita yang diindentifikasi dengan berbagai berita yang telah dikumpulkan, dengan hasil akhir yang diharapkan didapatkan sebuah hasil nilai similaritas dari berita dan *persentase hoax* dan fakta dari berita tersebut.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Tahapan penelitian meliputi studi pustaka, pengumpulan data, pengolahan data, analisis, perancangan dan implementasi sistem dan kesimpulan. Adapun tahapan penelitian ini dapat dilihat pada gambar berikut:



Gambar 1. Tahapan Penelitian

a. Studi Pustaka
Studi pustaka dalam penulisan penelitian ini adalah mencari referensi-referensi yang berkaitan dengan permasalahan, dimulai dari mencari buku-buku yang membahas tentang algoritma yang digunakan pada penelitian ini, mencari jurnal-jurnal dan artikel-artikel yang terkait yang terdapat di internet. Tujuan dari mencari referensi tersebut adalah untuk menambah wawasan peneliti dalam menyusun penelitian ini. Cara yang di terapkan peneliti dalam mencari referensi-referensi adalah dengan mengetikkan *keyword* topik yang diangkat di google scholar atau mencari buku – buku yang berisikan tentang algoritma text mining dan algoritma *cosine similarity*.

b. Pengumpulan Data
Sebelum melakukan pengolahan data pada penelitian ini, langkah pertama dilakukan pengumpulan data terlebih dahulu. Pengumpulan data dilakukan dengan mencari data yang berhubungan dengan topik penelitian yaitu vaksinasi COVID-19. Tahap pengumpulan data ini tidak boleh dilewatkan karena jika tidak ada data yang terkumpul maka tahapan pengolahan data dan analisis tidak dapat dilakukan. Proses pengumpulan data dilakukan dengan mencari berita tentang vaksinasi COVID-19 melalui mesin pencarian google dari website kominfo.go.id dan covid19.go.id, ini merupakan website milik Kementerian Komunikasi dan Informatika Republik Indonesia (KOMINFO) dan Satuan Tugas Penanganan COVID-19.

c. Pengolahan Data
Pengolahan data adalah merupakan tahap lanjutan dari pengumpulan data. Pengolahan data dilakukan dengan menerapkan algoritma text mining sebagai *text processing.* Tahapan dari text mining yaitu *case folding, tokenizing, filtering, stemming* dan *analyzing.* Proses text mining ini dibantu dengan menggunakan aplikasi orange.

ISSN 2597-4645 (media online)
ISSN 2597-4610 (media cetak)
Page: 129-139

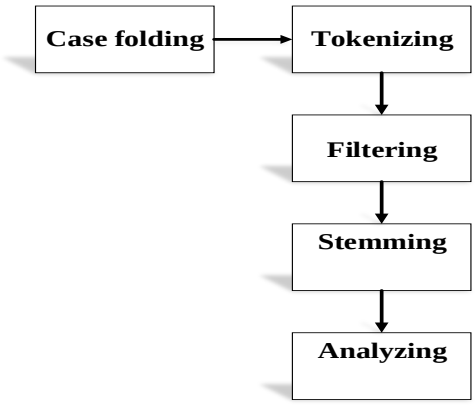
d. Analisis
Setelah tahap pengolahan data selesai di lakukan, maka data yang sudah diperoleh akan di lanjutkan ke tahap analisis. Pada tahap analisis pengolahan dilakukan dengan menerapkan algoritma *cosine similarity* untuk analisis tingkat kesamaan antar data. Setelah diperoleh hasil dari perhitungan dengan menggunakan algoritma *cosine similarity.* Selanjutnya dilakukan perhitungan probabilitas kemunculan berita dengan label hoax dan fakta. Dengan ketentuan jika nilai similaritas yang di dapat adalah >40%.

e. Perancangan Dan Implementasi Sistem
Pemodelan dan Implementasi sistem merupakan tahap dari proses merancang dan mengimplementasi dari program yang sudah dirancang.

f. Kesimpulan
Pada tahap ini peneliti menarik sebuah kesimpulan tentang penerapan text mining dan algoritma *cosine similarity* dalam deteksi berita palsu tentang vaksinasi COVID-19.

2.2 Text Mining

Natural language processing (NLP) adalah Salah satu cabang dari artificial intelegensi (kecerdasan buatan). Natural language (bahasa alami) dapat diartikan sebagai bahasa yang sering digunakan oleh manusia dalam berkomunikasi dan processing dapat diartikan sebagai bahasa pemrograman bentuk visual arts dan design yang digunakan dalam mengajar mengenai dasar pemrograman komputer kepada non-programmer. Maka dari itu natural language processing atau yang sering sebut dengan NLP biasanya digunakan dalam mendeskripsikan atau menggambarkan sebuah fungsi dari komponen software atau hardware pada sistem komputer yang dapat menganalisis atau menyintesis bahasa alami, baik secara lisan maupun tulisan[2].



Gambar 2. Proses Text Mining

Case folding adalah tahap mengubah semua huruf menjadi normal yaitu semua dokumen menjadi huruf kecil. Hanya huruf a sampai z yang di terima, karakter selain huruf akan di hilangkan dan dianggap *delimiter* (karakter yang digunakan untuk memisahkan text) [4][5][6].

Tokenizing adalah proses penguraian deskripsi yang semula berupa kalimat menjadi kata dengan cara dipotong berdasarkan kata yang menyusunnya. *Filtering* merupakan tahap lanjutan sesudah proses t*okenizing, filtering* dapat diartikan sebagai proses mengambil kata-kata penting dari proses token , dengan menggunakan algoritma stop list (membuat kata yang tidak penting) atau word list(menyimpan kata penting). Stopwords merupakan kosakata yang bukan merupakan ciri (kata unik) dari suatu dokumen. Pada proses filtering ini menggunakan stopwords yang sudah tersedia seperti stopwords by tala.

Stemming merupakan tahap lanjutan dari proses filtering, untuk mencari *root* kata dari hasil f*iltering*. *Stemming* adalah proses membentuk suatu kata menjadi kata dasar. *Tagging* dan *stemming* melakukan proses yang sama yaitu mencari *root* kata, namun disini tagging lebih berfokus mencari root kata dari tiap kata lampau. Contoh dari proses *tagging:* gone *menjadi* go, eaten *menjadi* eat. Proses mencari root kata disini akan menggunakan Kamus Besar Bahasa Indonesia. Tahap *analyzing* merupakan tahap penentuan seberapa jauh keterhubungan antar kata-kata terhadap suatu dokumen dengan menghitung nilai/bobot keterhubungan.

2.3 TF-IDF (*Term Freuency Inverse Document Freuency*)

Untuk dapat menghitung nilai similaritas menggunakan cosine similarity terlebih dahulu dilakukan proses pembobotan TF-IDF. Pembobotan TF-IDF adalah suatu hubungan kata (term) yang berada pada dokumen yang akan diberikan suatu nilai bobot (Robertson, 2005) [7]. Daftar term hasil dari proses stemming kemudian dilakukan perhitungan dengan menghitung jumlah term frequency data (TF) terlebih dahulu. Dan kemudian menghitung nilai jumlah data yang memiliki term (DF) dan selanjutnya menghitung nilai IDF dengan rumus log= (N/DF) +1, dimana N merupakan jumlah seluruh data yang ada. Setelah nilai dari TF dan IDF didapatkan kemudian tahap selanjutnya menentukan bobot kata dengan mengalikan TF dan IDF dengan menggunakan rumus Wdt= TFdt X IDFt [7].

ISSN 2597-4645 (media online)
ISSN 2597-4610 (media cetak)
Page: 129-139

2.4 *Cosine Similitary*

Cosine similarity merupakan metode yang dipakai untuk mengukur tingkat kemiripan atau kesamaan antar dua data. Tujuan dari metode ini adalah untuk membandingkan dua buah objek untuk di hitung tingkat kemiripannya sehingga dapat dilihat berapa tingkat kemiripannya[6]. Perhitungan dengan menggunakan *cosine similarity* didasarkan pada *vector space similarity measure*. Algoritma *cosine similarity* ini menghitung similaritas antara dua objek (misalkan D1 dan D2) dinyatakan menggunakan *keywords* (kata kunci) dari sebuah dokumen/data sebagai ukuran.
Rumus yang digunakan pada metode cosine similarity dapat dilihat sebagai berikut:

$$CosSim(d_j, q_k) = \frac{\sum_{i=1}^{n}(td_{ij} \times tq_{ik})}{\sqrt{\sum_{i=1}^{n} td_{ij}^2 \times \sum_{i=1}^{n} td_{ik}^2}} \quad (1)$$

Keterangan:
CosSim (dj, qk) : tingkat kesamaan data dengan query tertentu
Tdij : term ke-i dalam vektor untuk dokumen ke-j
Tqik : term ke-i dalam vektor untuk query ke-k
n : jumlah term yang unik dalam data set.[5]

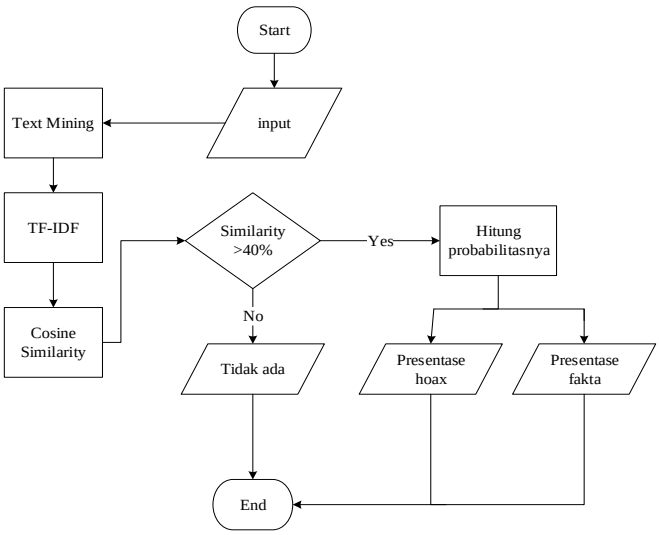
Jika nilai yang dihitung berasal dari metode *cosine similarity* lebih besar dan mendekati 1, dapat dikatakan bahwa dua vektor memiliki kemiripan yang tinggi, sebaliknya jika nilai yang dihitung lebih kecil dan mendekati 0 dapat dikatakan bahwa dua vektor memiliki kesamaan yang rendah. Nilai perhitungan *range* dimulai dari 0-1. Nilai 0 jika kedua vektor pada perhitungannya tidak sama, sedangkan nilai 1 jika kedua vektor adalah sama[5].

3. HASIL DAN PEMBAHASAN

3.1 Analisa

Analisa adalah sebuah proses dalam mendapatkan sebuah pemahaman, dengan cara mengindentifikasi dan menjabarkan permasalahan yang ada serta menentukan kebutuhan yang diperlukan. Analisa sistem perlu dilakukan agar pembuatan sebuah sistem dapat berjalan dengan baik dan lebih terstruktur sehingga dapat menghasilkan sebuah perangkat lunak yang dapat membantu dalam proses analisis. Sebagai langkah awal pada penelitian deteksi berita hoax dengan topik vaksinasi COVID-19 maka penulis telah mengumpulkan data melalui website kominfo.go.id dan covid-19.co.id website ini merupakan website milik Kementerian Komunikasi dan Informatika Republik Indonesia (KOMINFO) dan Satuan Tugas Penanganan COVID-19. Untuk menganalisa suatu berita hoax atau fakta, maka penulis memanfaatkan algoritma text mining dan algoritma Cosine Similarity dimana algoritma text mining berfungsi sebagai pengolahan text sedangkan algoritma cosine similarity berfungsi sebagai perbandingan similaritas antara berita yang akan diindentifikasi dengan data yang sebelumnya telah dikumpulkan.

Hasil akhir dari penelitian tersebut akan menunjukkan persentase kemiripan antara berita yang diidentifikasi dengan dataset berita yang telah dikumpulkan. Penulis mengasumsikan berita yang diindentifikasi jika memiliki similaritas diatas >40% terhadap data berita, akan dilanjutkan dihitung probabilitas similaritas berita yang diindentifikasi hoax atau fakta sehingga akan didapatkan hasil berupa presentase nilai hoax atau fakta dari berita yang di indentifikasi. Gambaran tahapan dari penerapan algoritma text mining dan algoritma cosine similarity dapat gambar berikut ini:



Gambar 3. Alur Penerapan Algoritma Text Mining Dan *Cosine Similitary*

ISSN 2597-4645 (media online)
ISSN 2597-4610 (media cetak)
Page: 129-139

3.2 Penerapan Algoritma Text Mining

Pada penelitian ini digunakan sampel data sebanyak 116 berita, contoh sampel dapat dilihat pada tabel berikut ini:

Tabel 1. Sampel Data

No.	Berita	
1	Vaksin COVID-19 Turunkan Fertilitas Pria.	Hoax
2	WHO ungkapkan jika vaksin tidak menimbulkan masalah kesuburan pada laki-laki maupun perempuan. Belum ada bukti nyata yang sebutkan vaksin COVID-19 timbulkan masalah pada organ reproduksi manusia.	Fakta
3	Vaksin COVID-19 Berisi Cairan Beracun Memberikan Banyak Penyakit.	Hoax
4	Hingga Senin, 11 Juli 2022, 66,68 persen atau 12,1 milliar populasi dunia telah menerima setidaknya satu dosis vaksin COVID-19. Sementara, di Indonesia sudah 61,6% yang sudah divaksinasi lengkap atau 169 juta orang. Vaksin COVID-19 telah terbukti mengurangi tingkat keparahan dan kematian dari orang yang terinfeksi COVID-19.	Fakta
5	Efektivitas Vaksin Pfizer Hanya Sebesar 12%.	Hoax
6	Laporan Badan Keamanan Kesehatan Inggris (UKHSA) pada 12 Mei 2022 tentang efektivitas vaksin dalam melawan penyakit simptomatik menyatakan: “Dengan 2 dosis, efektivitas Pfizer atau Moderna turun dari sekitar 65 sampai 70% menjadi sekitar 15% pada 25 minggi setelah dosis kedua”.	Fakta
7	Melansir Eastmojo.com, Vineeta Bal seorang ahli imunologi mengatakan bahwa vaksin COVID-19 sama sekali tidak meningkatkan kerentanan terhadap infeksi HIV.	Fakta
8	Beberapa vaksin COVID-19 dapat meningkatkan resiko HIV.	Hoax
9	Vaksin COVID-19 fungsinya untuk mengontrol manusia bukan untuk kesehatan.	Hoax
10	Vaksin COVID-19 Mengandung Microchip Magnetik.	Hoax
…	…	…
115	Pihak World Health Organization (WHO) dan European Medicines Agency (EMA) pada 18 Maret 2021 mengeluarkan pernyataan untuk merekomendasikan pemakaian vaksin AstraZeneca dilanjutkan.	Fakta
116	Lebih Mudah Terinfeksi COVID-19 Setelah Divaksin.	Hoax

Pada penelitian tersebut dilakukan pengolahan data dari setiap data berita dan berita yang akan diidentifikasi, pengolahan data dengan menggunakan algoritma text mining dibantu dengan menggunakan aplikasi orange dengan tujuan untuk mempermudah pengerjaan. Untuk menjelaskan detail alur proses text mining maka dilakukan pengolahan data secara manual terhadap 10 data berita dari 116 data berita yang tersedia yang dapat dilihat pada tahapan berikut:

3.2.1 *Case Folding*

Case folding merupakan tahap awal dalam proses text mining. Pada tahap *case folding* semua huruf dalam teks berita dirubah menjadi huruf kecil (*lowercase*) dan menghilangkan karakter selain huruf a-z serta angka. Proses *case folding* dapat dilihat pada tabel berikut ini:

Tabel 2. Tahap *Case Folding* (Sebelum Proses *Case Folding*)

No.	Berita	Label
1	Vaksin COVID-19 Turunkan Fertilitas Pria.	Hoax
2	WHO ungkapkan jika vaksin tidak menimbulkan masalah kesuburan pada laki-laki maupun perempuan. Belum ada bukti nyata yang sebutkan vaksin COVID-19 timbulkan masalah pada organ reproduksi manusia.	Fakta
3	Vaksin COVID-19 Berisi Cairan Beracun Memberikan Banyak Penyakit.	Hoax
4	Hingga Senin, 11 Juli 2022, 66,68 persen atau 12,1 milliar populasi dunia telah menerima setidaknya satu dosis vaksin COVID-19. Sementara, di Indonesia sudah 61,6% yang sudah divaksinasi lengkap atau 169 juta orang. Vaksin COVID-19 telah terbukti mengurangi tingkat keparahan dan kematian dari orang yang terinfeksi COVID-19.	Fakta
5	Efektivitas Vaksin Pfizer Hanya Sebesar 12%.	Hoax
6	Laporan Badan Keamanan Kesehatan Inggris (UKHSA) pada 12 Mei 2022 tentang efektivitas vaksin dalam melawan penyakit simptomatik menyatakan: “Dengan 2 dosis, efektivitas Pfizer atau Moderna turun dari sekitar 65 sampai 70% menjadi sekitar 15% pada 25 minggi setelah dosis kedua”.	Fakta
7	Melansir Eastmojo.com, Vineeta Bal seorang ahli imunologi mengatakan bahwa vaksin COVID-19 sama sekali tidak meningkatkan kerentanan terhadap infeksi HIV.	Fakta
8	Beberapa vaksin COVID-19 dapat meningkatkan resiko HIV.	Hoax
9	Vaksin COVID-19 fungsinya untuk mengontrol manusia bukan untuk kesehatan.	Hoax

ISSN 2597-4645 (media online)
ISSN 2597-4610 (media cetak)
Page: 129-139

No.	Berita	Label
10	Vaksin COVID-19 Mengandung Microchip Magnetik.	Hoax

Tabel 3. Tahap *Case Folding* (Sesudah Proses *Case Folding*)

No.	Berita	Label
1	Vaksin COVID-19 Turunkan Fertilitas Pria.	Hoax
2	WHO ungkapkan jika vaksin tidak menimbulkan masalah kesuburan pada laki-laki maupun perempuan. Belum ada bukti nyata yang sebutkan vaksin COVID-19 timbulkan masalah pada organ reproduksi manusia.	Fakta
3	Vaksin COVID-19 Berisi Cairan Beracun Memberikan Banyak Penyakit.	Hoax
4	Hingga Senin, 11 Juli 2022, 66,68 persen atau 12,1 milliar populasi dunia telah menerima setidaknya satu dosis vaksin COVID-19. Sementara, di Indonesia sudah 61,6% yang sudah divaksinasi lengkap atau 169 juta orang. Vaksin COVID-19 telah terbukti mengurangi tingkat keparahan dan kematian dari orang yang terinfeksi COVID-19.	Fakta
5	Efektivitas Vaksin Pfizer Hanya Sebesar 12%.	Hoax
6	Laporan Badan Keamanan Kesehatan Inggris (UKHSA) pada 12 Mei 2022 tentang efektivitas vaksin dalam melawan penyakit simptomatik menyatakan: “Dengan 2 dosis, efektivitas Pfizer atau Moderna turun dari sekitar 65 sampai 70% menjadi sekitar 15% pada 25 minggi setelah dosis kedua”.	Fakta
7	Melansir Eastmojo.com, Vineeta Bal seorang ahli imunologi mengatakan bahwa vaksin COVID-19 sama sekali tidak meningkatkan kerentanan terhadap infeksi HIV.	Fakta
8	Beberapa vaksin COVID-19 dapat meningkatkan resiko HIV.	Hoax
9	Vaksin COVID-19 fungsinya untuk mengontrol manusia bukan untuk kesehatan.	Hoax
10	Vaksin COVID-19 Mengandung Microchip Magnetik.	Hoax

3.2.2 *Tokenizing*

Tokenizing merupakan tahap lanjutan dari proses *case folding*, *tokenizing* merupakan pemotongan kalimat menjadi kata berdasarkan kata penyusunnya. Proses dari tahapan tokenizing dapat dilihat pada tabel 3.

Tabel 4. Tahap *Tokenizing*

No.	Narasi Berita Hasil *Tokenizing*	Label
1	[vaksin] [covid] [turunkan] [fertilitas] [pria]	hoax
2	[who] [ungkapkan] [jika] [vaksin] [tidak] [menimbulkan] [masalah] [kesuburan] [pada] [laki-laki] [maupun] [perempuan] [belum] [ada] [bukti] [nyata] [yang] [sebutkan] [vaksin] [covid] [timbulkan] [masalah] [pada] [organ] [reproduksi] [manusia]	fakta
3	[vaksin] [covid] [berisi] [cairan] [beracun] [memberikan] [banyak] [penyakit]	hoax
4	[hingga] [senin] [juli] [persen] [atau] [lengkap] [atau] [juta] [orang] [vaksin] [covid] [telah] [terbukti] [mengurangi] [tingkat] [keparahan] [dan] [kematian] [dari] [orang] [yang] [terinfeksi] [covid]	fakta
5	[efektivitas] [vaksin] [pfizer] [hanya] [sebesar]	hoax
6	[laporan] [badan] [keamanan] [kesehatan] [inggris] ([ukhsa]) [pada] [mei] [tentang] [efektivitas] [vaksin] [dalam] [melawan] [penyakit] [simptomatik] [menyatakan][dengan] [dosis] [efektivitas] [pfizer] [atau] [moderna] [turun] [dari] [sekitar] [sampai] [menjadi] [sekitar] [pada] [minggu] [setelah] [dosis] [kedua]	fakta
7	[melansir] [eastmojo] [com] [vineeta] [bal] [seorang] [ahli] [imunologi] [mengatakan] [bahwa] [vaksin] [covid] [sama] [sekali] [tidak] [meningkatkan] [kerentanan] [terhadap] [infeksi] [hiv]	fakta
8	[beberapa] [vaksin] [covid] [dapat] [meningkatkan] [resiko] [hiv]	hoax
9	[vaksin] [covid] [fungsinya] [untuk] [mengontrol] [manusia] [bukan] [untuk] [kesehatan]	hoax
10	[vaksin] [covid] [mengandung] [microchip] [magnetik]	hoax

3.2.3 *Filtering*

Setelah tahap tokenizing di selesai, di lanjutkan ke proses *filtering*, tahap *filtering* merupakan tahap mengambil kata-kata penting dari proses token , dengan menggunakan algoritma *stop list* (membuat kata yang tidak penting) atau (menyimpan kata penting). Contoh dari stopwords yaitu: “dan”, “di”,”yaitu”,”dari” dan seterusnya. Pada proses *filtering* dalam penelitian ini menggunakan data *stopwords* dari kamus tala.

ISSN 2597-4645 (media online)
ISSN 2597-4610 (media cetak)
Page: 129-139

Tabel 5. Tahap *Filtering*

No.	Narasi Berita Hasil *Filtering*	Label
1	[vaksin] [covid] [turun] [fertilitas] [pria]	hoax
2	[who] [ungkap] [vaksin] [timbul] [masalah] [subur] [laki-laki] [mau] [perempuan] [nyata] [sebut] [vaksin] [covid] [timbul] [masalah] [organ] [reproduksi] [manusia]	fakta
3	[vaksin] [covid] [isi] [cair] [racun] [beri] [penyakit]	hoax
4	[hingga] [persen] [lengkap] [juta] [vaksin] [covid] [bukti] [kurang] [tingkat] [parah] [mati] [infeksi] [covid]	fakta
5	[efektivitas] [vaksin] [pfizer] [besar]	hoax
6	[lapor] [badan] [aman] [sehat] [inggris] ([ukhsa]) [efektivitas] [vaksin] [lawan] [penyakit] [simptomatik] [nyata] [dosis] [efektivitas] [pfizer] [moderna] [turun] [dosis] [dua]	fakta
7	[melansir] [eastmojo] [com] [vineeta] [bal] [seorang] [ahli] [imunologi] [mengatakan] [bahwa] [vaksin] [covid] [sama] [sekali] [tidak] [meningkatkan] [kerentanan] [terhadap] [infeksi] [hiv]	fakta
8	[berapa] [vaksin] [covid] [tingkat] [resiko] [hiv]	hoax
9	[vaksin] [covid] [fungsi] [kontrol] [manusia] [sehat]	hoax
10	[vaksin] [covid] [kandung] [microchip] [magnetik]	hoax

3.2.4 *Stemming*

Setelah tahap *filtering* dilanjutkan ke proses *stemming*, stemming merupakan proses mencari *root* kata (kata dasar) dari tiap kata hasil dari proses *filtering*. Berikut hasil dari tahap *stemming* untuk sepuluh dataset berita dapat di lihat pada tabel berikut ini:

Tabel 6. Tahap *Stemming*

No.	Narasi Berita Hasil *Stemming*	Label
1	[vaksin] [covid] [turun] [fertilitas] [pria]	hoax
2	[who] [ungkap] [vaksin] [timbul] [masalah] [subur] [laki-laki] [mau] [perempuan] [nyata] [sebut] [vaksin] [covid] [timbul] [masalah] [organ] [reproduksi] [manusia]	fakta
3	[vaksin] [covid] [isi] [cair] [racun] [beri] [penyakit]	hoax
4	[hingga] [persen] [lengkap] [juta] [vaksin] [covid] [bukti] [kurang] [tingkat] [parah] [mati] [infeksi] [covid]	fakta
5	[efektivitas] [vaksin] [pfizer] [besar]	hoax
6	[lapor] [badan] [aman] [sehat] [inggris] ([ukhsa]) [efektivitas] [vaksin] [lawan] [penyakit] [simptomatik] [nyata] [dosis] [efektivitas] [pfizer] [moderna] [turun] [dosis] [dua]	fakta
7	[melansir] [eastmojo] [com] [vineeta] [bal] [seorang] [ahli] [imunologi] [mengatakan] [bahwa] [vaksin] [covid] [sama] [sekali] [tidak] [meningkatkan] [kerentanan] [terhadap] [infeksi] [hiv]	fakta
8	[berapa] [vaksin] [covid] [tingkat] [resiko] [hiv]	hoax
9	[vaksin] [covid] [fungsi] [kontrol] [manusia] [sehat]	hoax
10	[vaksin] [covid] [kandung] [microchip] [magnetik]	hoax

3.3 Penerapan TF-IDF

Setelah tahapan text mining selesai dilakukan, maka tahapan selanjutnya adalah melakukan penghitungan bobot, pada penelitian tersebut penulis memanfaatkan algoritma TF-IDF dimana *term* yang digunakan merupakan berita yang akan diidentifikasi sehinga ditemukan bobot dari setiap berita terpercaya dengan berita yang diidentifikasi. Namun sebelum menerapkan algoritma TF-IDF tersebut maka sebaiknya dilakukan tahapan text mining terhadap berita yang akan diidentifikasi untuk mengurangi terjadinya pembobotan yang tidak sesuai. Adapun hasil dari tahapan text mining terhadap data berita yang akan diidentifikasi dapat dilihat pada tabel 3.6.

Tabel 7. Data Uji

Data Uji	Setelah Proses Text Mining
vaksin booster covid-19 adalah cairan vaksin yang berbahaya dapat menyebabkan HIV dan AIDS	[vaksin] [booster] [covid] [cair]] [vaksin] [bahaya] [sebab] [hiv] [aids]

Pada proses pembobotan TF-IDF berita yang diindentifikasi diinisialkan sebagai berikut:

a. Term q (*query*) = Berita yang diindentifikasi
b. Term data (D1-D116) = Data berita yang telah dikumpulkan dan diolah dengan algoritma text mining.

Hasil dari pembobotan TF-IDF dapat dilihat pada tabel 8.

ISSN 2597-4645 (media online)
ISSN 2597-4610 (media cetak)
Page: 129-139

Tabel 8. Hasil TF-IDF

Term q	TF	DF	IDF	TF* IDF	Term Data	TF				DF	IDF	TF*IDF			
						D1	…	D115	D116			D1	…	D115	D116
Vaksin	2	1	1	2	bnar		…			1	3.0645	0	…	0	0
Cair	1	1	1	1	air		…			1	3.0645	0	…	0	0
Racun	1	1	1	1	negara		…			1	3.0645	0	…	0	0
Booster	1	1	1	1	agency		…	1		1	3.0645	0	…	3.064458	0
Penyakit	1	1	1	1	ahli		…			2	3.0645	0	…	0	0
Sebab	1	1	1	1	aman		…			1	3.0645	0	…	0	0
Hiv	1	1	1	1	anak		…			2	2.7634	0	…	0	0
Vaids	1	1	1	1	ardy anto		…			1	3.0645	0	…	0	0
Bahaya	1	1	1	1	astrezeneca		…			1	3.0645	0	…	0	0
Dapat	1	1	1	1	aids		…			3	2.5873	0	…	0	0
					…	…	…	…	…	…	…	…	…	…	…
					vagina		…			1	3.0645	0	…	0	0
					tahap		…			1	3.0645	0	…	0	0
					talabani		…			1	3.0645	0	…	0	0
					uji		…			2	2.7634	0	…	0	0
					racun		…			1	3.0645	0	…	0	0
					dapat					1	3.0645	0	…	0	0

3.4 Penerapan *Cosine Similarity*

Berikut hasil perkalian *query term* dengan *term* data adalah sebagai berikut:

Tabel 9. Hasil Hasil Perkalian *Query Term* Dengan *Term Data*

Term Dataset	Q * D																			
	D1	D2	D3	D4	D5	D29	…	D30	D31	D48	D49	D55	D59	D60	D77	D78	D79	D114	D115	D116
bnar	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0	0	0	0	0	0
air	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0	0	0	0	0	0
negara	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0	0	0	0	0	0
agency	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0	0	0	0	0	0
ahli	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0	0	0	0	0	0
aman	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0	0	0	0	0	0
anak	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0	0	0	0	0	0
ardyanto	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0	0	0	0	0	0
astrazeneca	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0	0	0	0	0	0
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
rupa	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0	0	0	0	0	0
sumber	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0	0	0	0	0	0
dapat	0	0	0	0	0	0	…	0	3.064458	0	0	0	0	0	0	0	0	0	0	0
Sum																				
	4.3965	3.7767	8.951	8.17	6.364	25.377	…	12.891	13.90423	14.49	10.76066	0	11.015	14.001	3.06	3.78	3.78	3.777	3.777	5.8869

Kemudian untuk menghitung nilai panjang setiap data berita termasuk *query* dengan mengkuadratkan hasil pembobotan TF-IDF untuk *query* dan hasil pembobotan TF-IDF untuk dataset, kemudian hasil dari nilai kuadrat dijumlah dan kemudian diakarkan. Hasil perhitungan pengkuadratan dari pembobotan TF-IDF *query* dan data berita dapat dilihat pada tabel 10.

Tabel 10. Hasil Perhitungan Pengkuadratan dari pembobotan TF-IDF *Query* dan data

term q	Kuadratka	Term	Kuadratkan bobot dari dataset TF-IDF																			
			D1	D	D	D	D	D	D	D	…	D	D	D	D	D	D	D	D	D	D1	D

ISSN 2597-4645 (media online)
ISSN 2597-4610 (media cetak)
Page: 129-139

	n bobot q dari TF-IDF			2	3	4	5	29	30	31		48	49	55	59	60	77	78	79	114	15	116
Vaksin	4	bnar	0	0	0	0	0	0	0	0	…	9.39	0	0	0	0	0	0	0	0	0	0
Covid	1	air	0	0	0	0	0	0	37.6	0	…	0	0	0	0	0	0	0	0	0	0	0
Racun	1	negara	0	0	0	0	0	9.39	0	0	…	0	0	0	0	0	0	0	0	0	0	0
Cair	1	agency	0	0	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0	0	9.390903	0
Booster	1	ahli	0	0	0	7.64	0	7.64	0	0	…	0	0	0	0	0	0	0	0	0	0	0
Penyakit	1	aman	0	0	0	0	9.39	0	0	0	…	0	0	0	0	0	0	0	0	0	0	0
Sebab	1	anak	0	0	0	0	0	0	0	0	…	0	0	0	0	0	30.5	30.5	0	0	0	0
Hiv	1	ardyanto	0	0	0	0	0	0	0	0	…	0	0	9.39	0	0	0	0	0	0	0	0
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
Hasil dari kuadrat diatas dijumlahkan		racun	0	0	9.39	0	0	0	0	0	…	0	0	0	0	0	0	0	0	0	0	0
14			Hasil dari kuadrat diatas dijumlahkan																			
Akarkan hasil dan penjumlahan diatas			34.058	27.3	125	82.7	139	246	143	36.4	…	120	53.6	101	30.6	267	63.7	76.6	334	132.3	12.95683	42.3
3.74165657387			Akarkan hasil dari penjumlahan diatas																			
			5.8359	5.22	11.2	9.09	11.8	15.7	11.9	6.03	…	10.9	7.32	10	5.54	16.4	7.98	8.75	18.3	11.5	3.59956	6.5

Tahap selanjutnya setelah mengetahui nilai dari masing- masing *query* dan data berita, dilanjutkan dengan menerapkan rumus *cosine similarity*. Dengan membagi nilai hasil dari penjumlahan dari q*D (D1 – D116) dibagi hasil akar q*D1 dan seterusnya.

Perhitungan dengan menggunakan rumus *cosine similarity* untuk menghitung nilai similaritas q dengan data (D1-D116) dapat dilihat di bawah ini:

a. $$CosSim(q, D1) = \frac{\sum_{i=1}^{n}(td_{ij} \times tq_{ik})}{\sqrt{\sum_{i=1}^{n} td_{ij}^2 \times \sum_{i=1}^{n} td_{ik}^2}} \quad (1)$$

$$= \frac{4.3965}{3.741 * 5.836} = 0.20$$

$$= 0.20 * 100\% = 20\%$$

Perhitungan similaritas antara q (berita yang dideteksi) dengan D1 didapatkan nilai similaritas sebesar 20%.

b. $$CosSim(q, D31) = \frac{\sum_{i=1}^{n}(td_{ij} \times tq_{ik})}{\sqrt{\sum_{i=1}^{n} td_{ij}^2 \times \sum_{i=1}^{n} td_{ik}^2}} \quad (2)$$

ISSN 2597-4645 (media online)
ISSN 2597-4610 (media cetak)
Page: 129-139

$$= \frac{13.9}{3.741 * 6.03} = 0.62$$
$$= 0.62 * 100\% = 62\%$$

Perhitungan similaritas antara q (berita yang dideteksi) dengan D31 didapatkan nilai similaritas sebesar 62%.

c. $$CosSim(q, D59) = \frac{\sum_{i=1}^{n}(td_{ij} \times tq_{ik})}{\sqrt{\sum_{i=1}^{n} td_{ij}^2 \times \sum_{i=1}^{n} td_{ik}^2}} \quad (3)$$
$$= \frac{11.0}{3.741 * 5.54} = 0.53$$
$$= 0.53 * 100\% = 53\%$$

Perhitungan similaritas antara q (berita yang dideteksi) dengan D59 didapatkan nilai similaritas sebesar 53%.

d. $$CosSim(q, D116) = \frac{\sum_{i=1}^{n}(td_{ij} \times tq_{ik})}{\sqrt{\sum_{i=1}^{n} td_{ij}^2 \times \sum_{i=1}^{n} td_{ik}^2}} \quad (4)$$
$$= \frac{5.88}{3.741 * 6.5} = 0.24$$
$$= 0.24 * 100\% = 24\%$$

Perhitungan similaritas antara q (berita yang dideteksi) dengan D124 didapatkan nilai similaritas sebesar 24%.

Berdasarkan perhitungan dengan algoritma *cosine similarity* diatas, Q (berita yang di deteksi) memiliki nilai similaritas diatas 40% terhadap D31 dan D59, dimana pada label yang sudah dibuat pada data berita D31 dan D59 merupakan berita **hoax**. Setelah hasil nilai *cosine similarity* didapatkan kemudian dilanjutkan untuk menghitung probabilitas kemunculan berita yang memiliki nilai similaritas diatas 40% dengan tujuan untuk mengetahui berapa persentase berita *hoax* dan fakta dari berita yang diindentifikasi yang didapatkan berdasarkan probabilitas kemunculan berita *hoax* dan fakta setelah proses analisis berita yang diindentifikasi dengan data yang telah dikumpulkan dengan menggunakan algoritma *cosine similarity.* Berikut perhitungan probabilitas kemunculan berita:
Dengan menggunakan rumus matematika untuk menghitung probabilitas,

P (A) = n (A) / n (S)

= Banyak data label hoax (2) / jumlah data yang memiliki nilai similaritas diatas 40% (2) = 1

1*100% = 100%

= Banyak data label fakta (0)/jumlah data yang memiliki nilai similaritas diatas 40% (2) = 0

0 *100% = 0%

Berdasarkan nilai probabilitas yang didapatkan Q (Berita yang di deteksi) **hoax 100%** dan **fakta 0%.**

4. KESIMPULAN

Berdasarkan hasil analisis terhadap data dengan menerapkan algoritma text mining dan *cosine similarity* diperoleh hasil yang menunjukkan berita yang diindetifikasi dinyatakan *hoax* sesuai dengan data yang dimiliki dengan presentasi berita *hoax* 100% berdasarkan perhitungan nilai probabilitas kemunculan berita. Berdasarkan hasil penelitian langkah cara yang diterapkan dalam mendeteksi berita *hoax* dengan menerapkan algoritma text mining dan *cosine similarity* ini merupakan urutan langkah yang sesuai karena hasil output sesuai dengan data yang dikumpulkan. Aplikasi deteksi berita *hoax* yang dirancang dan dibangun dapat membantu pihak yang membutuhkan untuk mendeteksi berita hoax vaksinasi COVID-19.

REFERENCES

[1] R. T. Wahyuni, D. Prastiyanto, and E. Supraptono, "Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi," vol. 9, no. 1, 2017.
[2] muhammad naufal Alfareza, "pembangunan chatboot menggunakan natural language processing di jurusan teknik industri." pp. 15–19, 2020.
[3] . M. Yulian Findawati, S.T. and M. K. Muhammad Alfan Rosid, S.Kom., *BUKU AJAR TEXT MINING*, Cetakan pe. UMSIDA Press Anggota IKAPI No. 218/Anggota Luar Biasa/JTI/2019 Anggota APPTI No. 002 018 1 09 2017, 2020.
[4] N. Silalahi and G. L. Ginting, "Analisa Sentimen Masyarakat Dalam Penggunaan Vaksin Sinovac Dengan Menerapkan Algoritma Term Frequence – Inverse Document Frequence (TF-IDF) dan Metode Deskripsi," vol. 3, no. 3, pp. 206–217, 2022, doi: 10.47065/josh.v3i3.1441.
[5] R. Saptono, H. Prasetyo, and A. Irawan, "Combination of Cosine Similarity Method and Conditional Probability for Plagiarism Detection in the Thesis Documents Vector Space Model," no. July, 2018.

ISSN 2597-4645 (media online)
ISSN 2597-4610 (media cetak)
Page: 129-139

[6] A. SALEH, "ANALISIS ACCURATE LEARNING RADIAL BASIS FUNCTION NEURAL NETWORK MENGGUNAKAN COSINE SIMILARITY PADA PENGENALAN DAUN," UNIVERSITAS SUMATERA UTARA, 2018.
[7] "PENERAPAN METODE TERM FREQUENCY INVERSE DOCUMENT FREQUENCY (TF-IDF) DAN COSINE SIMILARITY PADA SISTEM TEMU KEMBALI INFORMASI U - 8623-27168-2-PB.pdf." .
[8] A. E. M Syukron Nawawi, Falentino Sembiring, "Implementasi Algoritma K-Means Clusterinf Menggunaka Orange Untuk Penentuan Produk Busana Muslim Terlaris," pp. 1–9, 2021.
[9] V. Michael Sitorus, "Penerapan Algoritma C4.5 terhadap pengaruh VSTOCK.ID pada masyarakat di masa pandemi covid-19," *J. Inf. Adv. Comput.*, vol. .2, no.2, pp. 1–5, 2021.
[10] U. S. Rut Samuel, Ripa Natan, Fitria, "Penerapan Cosine Similarity dan K-Nearest Neighbor (K-NN) pada Klasifikasi dan Pencarian Buku," *J. big data Anal. artifical Intell.*, vol. . 1, No. 1, pp. 2–6, 2018.
[11] M. D. R. Wahyudi, "Penerapan Algoritma Cosine Similarity pada Text Mining Terjemah Al-Qur'an berdasarkan keterkaitan topik," *semesta Tek.*, vol. 1,41-50, pp. 1–10, 2019, doi: 10.18196/st.221235.
[12] D. Kurniadi, S. Farisa, C. Haviana, and A. Novianto, "Implementasi Algoritma Cosine Similarity pada sistem arsip dokumen di Universitas Islam Sultan Agung," vol. 17, no. 2, pp. 124–132, 2020.
[13] J. Jtik, J. Teknologi, Y. Lasena, and M. Hasan, "Text Mining Analysis untuk Identifikasi Artikel Hoax Menggunakan Algoritma Cosine Similarity," vol. 4, no. 2, pp. 0–5, 2020.