KOMPUTA : Jurnal Ilmiah Komputer dan Informatika
Vol. xx, No. x, mmmm yyyy, ISSN : 2089-9033, e- ISSN : 2715-7849
DOI 10.34010/komputa.xxxxxx

Implementasi BERT dan SNA dalam Sistem Tanya Jawab untuk Survei *Demam Berdarah Dengue*

Rofiatun Nadifah[1], Rindra Syaifullah[2], Nasywaa Almaasah Zatri[3] , Regita Putri Permata[4]
[1234]Program Studi Sains Data, Telkom University Surabaya
E-mail : regitapermata@telkomuniversity.ac.id

Abstrak

Sebagai tantangan kesehatan masyarakat global, Demam Berdarah Dengue (DBD) menuntut adanya analisis mendalam mengenai pemahaman dan persepsi di tingkat publik. Dalam upaya mengendalikan penyebaran DBD, data survei menjadi instrumen kunci untuk mengungkap persepsi ini. Penelitian ini dilakukan untuk menganalisis secara mendalam pemahaman masyarakat tentang DBD dengan mengkaji respons tekstual dari survei. Data yang dianalisis merupakan jawaban dari 33 responden terhadap lima pertanyaan kunci mengenai definisi, gejala, penyebab, penularan, dan pencegahan DBD. Metode yang digunakan adalah analisis kemiripan makna menggunakan model *pre-trained* BERT, perhitungan *cosine similarity*, dan SNA digunakan untuk mengidentifikasi responden kunci sebagai pusat opini dan memetakan pola penyebaran pemahaman. Pada model ini, diperoleh tingkat pemahaman yang sangat konsisten, dengan *F1-Score* untuk kategori "Sangat Mirip" mencapai 0.94 dan nilai *cosine similarity* tertinggi 0.995.

Kata kunci : BERT*, Cosine Similarity, Dengue Demam Berdarah,* SNA.

Implementation of BERT and SNA in Question and Answer System for Dengue Hemorrhagic Fever (DHF)

Abstract

As a global public health challenge, Dengue Hemorrhagic Fever (DHF) necessitates an analysis of public understanding and perception. For controlling the spread of DHF, survey data is instrumental in uncovering these perceptions. This study was conducted to perform an in-depth analysis of the public's understanding of DHF by analyzing textual responses from surveys. The analyzed data consists of answers from 33 respondents to five key questions concerning the definition, symptoms, causes, transmission, and prevention of DHF. The methods used were semantic similarity analysis with a pre-trained BERT model, cosine similarity calculation, and Social Network Analysis (SNA) to identify key respondents as opinion leaders and to map the patterns of understanding dissemination. In this model, a highly consistent level of understanding was obtained, with an F1-Score for the "Very Similar" category reaching 0.94 and a highest cosine similarity value of 0.995.

Keywords : BERT*, Cosine Similarity, Dengue Hemorrhagic Fever,* SNA.

1. Pendahuluan

Demam Berdarah Dengue (DBD) adalah penyakit infeksi menular yang menjadi isu kesehatan masyarakat secara global, khususnya di daerah tropis dan subtropis [1]. Penyakit akibat virus *dengue* ini telah menginfeksi jutaan orang di berbagai belahan dunia dan menyebabkan angka kesakitan serta kematian yang cukup tinggi [2]. Indonesia, sebagai negara dengan iklim tropis, menghadapi tantangan besar dalam pengendalian dan pencegahan DBD dengan kasus yang terus meningkat setiap tahunnya. Dalam pengendalikan penyebaran DBD, data survei memainkan peran penting dalam mengungkap pola persebaran penyakit, faktor-faktor risiko, serta persepsi masyarakat terhadap DBD. Survei yang melibatkan masyarakat umum, mahasiswa, dan pemangku kepentingan terkait menghasilkan kumpulan data yang besar dan kompleks. Namun demikian, pendekatan analisis survei secara konvensional sering kali mengalami kendala dalam menggali makna yang lebih dalam dari respons teks yang bersifat kualitatif dan semi-terstruktur.

Kemajuan teknologi *Natural Language Processing* (NLP) telah menghadirkan berbagai peluang dalam pengolahan dan analisis data berbasis teks. Salah satu model bahasa terdepan, yaitu *Bidirectional Encoder*

Naskah dikirim: dd-mm-yyyy; direvisi dd-mm-yyyy; diterima: dd-mm-yyyy

Representations from Transformers (BERT), dikenal memiliki kemampuan unggul dalam memahami konteks dan makna semantik suatu teks secara akurat. Dalam implementasinya, berbagai algoritma BERT menjadi salah satu pilihan utama karena memiliki tingkat akurasi yang tinggi, terutama dalam pengembangan sistem tanya jawab otomatis [3]. Fungsi utama BERT terletak pada popularitasnya dan efektivitasnya dalam berbagai tugas NLP, terutama karena kemampuannya menangkap makna kata secara mendalam melalui arsitektur *bidirectional* yang memahami konteks secara menyeluruh [3]. Model BERT menunjukkan efektivitas yang luar biasa dalam menganalisis teks kompleks, termasuk judul penelitian dan abstrak akademik, karena kapasitasnya dalam memahami konteks secara mendalam [4].

Cosine similarity merupalan suatu metode yang umum digunakan dalam analisis teks untuk menghitung seberapa mirip dua representasi dalam bentuk vektor. Dalam penerapannya pada data survei, metode ini dapat dimanfaatkan untuk mengenali kesamaan antar jawaban responden, mengelompokkan respons yang memiliki makna serupa, serta mengeksplorasi pola tersembunyi dalam kumpulan data survei terkait DBD. Perhitungan *cosine similarity* didasarkan pada prinsip pengukuran kemiripan ruang vektor, di mana kata kunci dari suatu dokumen digunakan sebagai dasar pembentukan vektor [5]. Kelebihan dari metode ini terletak pada kemudahan implementasinya, efisiensi dalam komputasi, serta konsep dasarnya yang mudah dipahami [6]. *Cosine similarity* berfungsi untuk menghitung tingkat kesamaan antara dua representasi vektor berdasarkan kata kunci yang relevan dari masing-masing dokumen.

Analisis Jaringan Sosial *(Social Network Analysis atau SNA)* merupakan metode yang digunakan untuk menganalisis pola interaksi dan keterkaitan antar individu dalam suatu struktur jaringan sosial. SNA mengkaji struktur interaksi sosial dengan memanfaatkan prinsip-prinsip *network science*, yakni cabang ilmu yang berfokus pada studi jaringan kompleks seperti jaringan sosial, komputer, biologis, dan komunikasi, yang secara konseptual bertumpu pada teori graf [7]. Dalam penerapannya, SNA biasanya menggunakan pemodelan graf tak berarah untuk merepresentasikan hubungan antar aktor atau simpul dalam jaringan. Dalam konteks pengendalian penyakit seperti DBD, SNA memiliki potensi besar dalam membantu mengidentifikasi pola penyebaran informasi kesehatan, pemetaan aktor yang berperan penting dalam proses komunikasi, serta memahami dinamika interaksi dalam komunitas. Melalui pemahaman tersebut, SNA dapat dimanfaatkan untuk merancang strategi penyebaran informasi yang lebih efektif di berbagai wilayah [8].

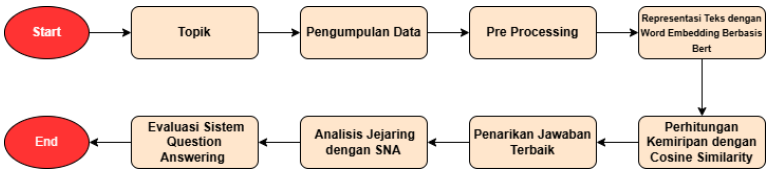
Beberapa penelitian sebelumnya telah menunjukkan efektivitas penerapan teknologi NLP dan analisis jaringan sosial dalam berbagai konteks. Penelitian mengenai *"Pengukuran Kemiripan Makna Menggunakan Cosine Similarity dan Basis Data Sinonim Kata"* [9]. Penelitian tersebut menghasilkan persentase kemiripan, mencapai rata-rata nilai kemiripan 94.48%. Penelitian selanjutnya berjudul *"Sistem Penilaian Jawaban Singkat Otomatis pada Ujian Online Berbasis Komputer Menggunakan Algoritma Cosine Similarity"* [10]. Menggunakan metode *cosine similarity* dengan menghasilkan *similarity* rata-rata sebesar 85.4% antara jawaban peserta didik dengan kunci jawaban serta nilai korelasi tinggi w=0.885 dan w=0.883 dengan instruktur pembanding.

Implementasi BERT dalam sistem rekomendasi dalam penelitian *"Implementasi BERT dan Cosine Similarity untuk Rekomendasi Dosen Pembimbing berdasarkan Judul Tugas Akhir"* [4]. Penelitian ini mengimplementasikan model BERT dan *cosine similarity,* yang menghasilkan tingkat akurasi sebesar 90%. Pengembangan sistem question answering menggunakan BERT dalam penelitian *"Model Penjawab Pertanyaan Otomatis Berdasarkan Peringkat Relevansi Kalimat Menggunakan Model BERT"* [11]. Sistem yang dikembangkan menggunakan pemeringkatan relevansi kalimat berbasis model BERT dan dataset SQuAD menunjukkan hasil evaluasi kinerja dengan *F1 Score* sebesar 0.6 dan *Exact Match* sebesar 0.5.

Penelitian ini menggunakan BERT sebagai metode embedding, cosine similarity untuk mengukur kemiripan antar respons, serta pendekatan Social Network Analysis (SNA). Diharapkan, pendekatan ini dapat memberikan kontribusi dalam pengembangan metodologi analisis survei berbasis teks yang lebih akurat dan menyeluruh. Selain itu, integrasi sistem question answering juga diharapkan dapat mempermudah pemahaman dan pengambilan keputusan terkait respons masyarakat terhadap isu Demam Berdarah Dengue (DBD) secara lebih efektif.

2. Metodologi

Tahapan ini bertujuan untuk menyusun desain sistem dan alur pengembangan program yang akan menjadi acuan atau pedoman selama proses penelitian berlangsung. Perancangan tersebut dapat berbentuk diagram alur program, prosedur tahapan kerja, atau diagram alir (*flowchart*) [12].



Gambar 1. *Flowchart* Penelitian

2.1 Pengumpulan Data

Dalam penelitian ini, data primer dikumpulkan melalui metode survei *online* yang dirancang pada platform *Google Forms*. Tautan *(link)* survei kemudian disebarkan secara digital kepada responden untuk diisi. Semua jawaban dari responden otomatis terekam dan terorganisir di dalam platform tersebut untuk dianalisis lebih lanjut. Survei ini terdiri dari lima pertanyaan utama yang dirancang untuk menggali pemahaman dan persepsi responden mengenai penyakit DBD. Jumlah total responden yang berhasil dikumpulkan sebanyak 33 orang. Pertanyaan dan Kunci Jawaban tertera seperti ditabel di bawah ini:

Tabel 1. Daftar Pertanyaan dan Jawaban

No	Pertanyaan	Kunci Jawaban
1	Apa yang dimaksud dengan DBD?	DBD adalah penyakit infeksi akut yang disebabkan oleh virus dengue dan ditularkan melalui gigitan nyamuk Aedes aegypti. Penyakit ini ditandai dengan demam mendadak, nyeri otot, sakit kepala, dan dapat menyebabkan perdarahan serta syok yang berpotensi fatal jika tidak ditangani dengan tepat
2	Apa gejala utama DBD!	Gejala utama DBD meliputi demam tinggi mendadak (38-40°C), sakit kepala hebat, nyeri otot dan sendi, mual muntah, ruam kemerahan pada kulit, nyeri perut, dan tanda perdarahan seperti mimisan, gusi berdarah, atau bintik merah di kulit.
3	Apa penyebab utama penyakit DBD?	DBD disebabkan oleh virus dengue yang memiliki empat serotipe yaitu DEN-1, DEN-2, DEN-3, dan DEN-4. Virus ini ditularkan melalui gigitan nyamuk Aedes aegypti yang terinfeksi.
4	Bagaimana cara penularan penyakit DBD terjadi?	DBD ditularkan melalui gigitan nyamuk Aedes aegypti yang aktif pada siang dan sore hari. Nyamuk ini berkembang biak di air jernih tergenang seperti bak mandi, kaleng bekas, dan ban mobil. DBD tidak menular langsung antarmanusia
5	Apa solusi untuk meningkatkan kepedulian masyarakat terhadap pencegahan DBD?	Solusi meningkatkan kepedulian meliputi sosialisasi intensif tentang dampak fatal DBD, melibatkan tokoh masyarakat dalam kampanye 3M Plus, edukasi melalui media massa, dan demonstrasi langsung cara pemberantasan jentik nyamuk

2.2 Preprocessing

Setelah data primer hasil survei tentang DBD berhasil dikumpulkan, data tersebut masih berupa data mentah (*raw data*) yang perlu melalui serangkaian tahap *preprocessing*. Tujuan dari tahap ini adalah untuk membersihkan, menstandarisasi, dan mentransformasi data teks agar siap dan optimal untuk diolah oleh model BERT serta untuk diekstraksi fiturnya bagi kebutuhan SNA.

a. *Case Folding:* Proses ini bertujuan untuk menyeragamkan seluruh data teks dengan mentransformasikannya ke dalam format huruf kecil (*lowercase*). Langkah ini esensial untuk memastikan konsistensi data dan mencegah model menginterpretasikan kata yang sama dengan kapitalisasi berbeda (misalnya, "Nyamuk" dan "nyamuk") sebagai dua entitas yang berlainan.
b. Normalisasi Kata: Tahap normalisasi diaplikasikan untuk mengonversi kata-kata tidak baku, singkatan, atau istilah informal ke dalam bentuk standarnya sesuai kaidah bahasa Indonesia. Implementasinya memanfaatkan sebuah kamus normalisasi yang telah didefinisikan untuk melakukan substitusi pada setiap kata yang tidak sesuai standar.

c. *Stopword Removal:* Tahap ini melibatkan eliminasi *stopwords* atau kata-kata umum yang memiliki frekuensi kemunculan tinggi namun kontribusi makna yang rendah dalam konteks analisis (contoh: di, dan, dari, yang). Proses ini menggunakan daftar *stopword* untuk Bahasa Indonesia dari pustaka *Sastrawi* untuk menyaring teks, sehingga model dapat lebih fokus pada *term-term* yang signifikan.
d. *Tokenize dan Stemming:* Tahap ini bertujuan untuk memecah teks menjadi kata-kata (tokenisasi) dan mengubah setiap kata ke bentuk dasarnya (stemming), misalnya "berlari" menjadi "lari". Proses ini penting untuk mereduksi keragaman bentuk kata yang memiliki makna serupa, sehingga memudahkan analisis lebih lanjut, terutama dalam proses pencocokan atau pembobotan kata.

2.3 Word Embedding BERT

Setelah melewati tahap *preprocessing*, jawaban responden dalam bentuk teks diubah menjadi representasi vektor numerik menggunakan model BERT. BERT dirancang untuk menghasilkan vektor yang mempertimbangkan konteks kata secara dua arah (*bidirectional*), sehingga mampu merepresentasikan makna semantik dari sebuah respons dengan lebih mendalam. Vektor-vektor ini kemudian dimanfaatkan dalam proses analisis lanjutan. BERT mengandalkan arsitektur *transformer*, yang dikenal karena kemampuannya dalam memahami konteks antar kata dalam sebuah kalimat. Dengan memanfaatkan sejumlah *layer* dalam arsitektur tersebut, BERT dapat menangkap hubungan konteks secara menyeluruh, menghasilkan representasi kata yang lebih presisi dan sesuai dengan makna kalimat secara keseluruhan..

2.4 Cosine Similarity

Cosine similarity merupakan teknik yang sering dimanfaatkan secara luas dalam analisis teks untuk menghitung seberapa mirip dua representasi dalam bentuk vektor. Perhitungan ini didasarkan pada prinsip pengukuran kemiripan ruang vektor untuk menentukan derajat kemiripan antara dua objek berbasis vektor.

$$Cosine\ Similarity(A,B) = \frac{A.B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

Dalam evaluasi otomatis terhadap jawaban esai, metode *cosine similarity* kerap digunakan untuk mengukur tingkat kemiripan antar jawaban. Berdasarkan temuan penelitian [13] Sistem penilaian esai otomatis yang menggunakan pendekatan *cosine similarity* terbukti efektif dalam mengevaluasi esai berbahasa Inggris. Hasil pengujian menunjukkan bahwa skor yang dihasilkan oleh sistem memiliki tingkat kecocokan rata-rata sebesar 89.48% dibandingkan dengan penilaian yang diberikan oleh pengajar. Hasil perhitungan kemudian digunakan untuk mengeksplorasi pola tersembunyi dalam kumpulan data survei terkait DBD , dengan mengategorikan jawaban berdasarkan tingkat kemiripannya seperti yang ditunjukkan pada hasil penelitian. Berikut adalah tabel yang mendefinisikan kategori kemiripan berdasarkan nilai *cosine similarity* yang digunakan dalam penelitian ini:

Tabel 2. Justifikasi Kategori *Cosine Similarity*

Kategori	**Rentang Nilai**	**Justifikasi Singkat**
Sangat Mirip	> 0.85	Menunjukkan kesamaan makna yang sangat tinggi, hampir identik dalam konten atau gagasan utama.
Cukup Mirip	> 0.70	Menunjukkan adanya kesamaan topik atau gagasan inti, meskipun terdapat variasi dalam detail atau cara penyampaian.
Kurang Mirip	≤ 0.70	Menunjukkan adanya beberapa kata kunci atau konsep yang tumpang tindih, namun secara keseluruhan membahas aspek berbeda.

2.5 Social Network Analysis (SNA)

Social Network Analysis adalah bidang ilmu yang mempelajari hubungan antar individu dengan memanfaatkan teori graf sebagai dasar analisis [14]. Dalam konteks pengendalian DBD, SNA memiliki potensi besar untuk membantu mengidentifikasi pola penyebaran informasi kesehatan, memetakan aktor yang berperan penting dalam proses komunikasi, serta memahami dinamika interaksi dalam komunitas. Pendekatan ini juga dapat digunakan untuk melacak arah dan sumber narasi dalam jaringan serta mengetahui posisi strategis aktor dalam menyebarkan informasi.

2.6 Evaluasi

Tahap evaluasi dilakukan untuk mengukur kinerja dari sistem yang dikembangkan. Merujuk pada penelitian sebelumnya, evaluasi kinerja dapat dilakukan dengan menggunakan metrik seperti *F1 Score* dan *Exact Match* untuk sistem Tanya Jawab atau mengukur tingkat akurasi untuk sistem rekomendasi. Berikut adalah tabel yang mendefinisikan metrik evaluasi yang digunakan untuk sistem QA:

Tabel 3. Definisi Metrik Evaluasi

Nama Metrik	Deskripsi Singkat	Formula
Precision	Menilai proporsi prediksi positif yang memang sesuai dengan kondisi sebenarnya. Ini menunjukkan tingkat keakuratan dari klaim positif yang dibuat.	$\frac{TP}{TP + FP}$
Recall	Mengukur seberapa besar jumlah sampel positif yang sebenarnya berhasil dikenali secara tepat oleh model.	$\frac{TP}{TP + FN}$
F1-Score	Merupakan rata-rata harmonis dari *Precision* dan *Recall* yang digunakan untuk menilai keseimbangan keduanya, dengan mempertimbangkan tingkat kesamaan token yang tumpang tindih.	$2\ x\ \frac{presisi\ x\ recall}{presisi\ +\ recall}\ x\ 100\%$
Akurasi	Menilai persentase keseluruhan prediksi yang tepat, mencakup prediksi positif dan negatif, terhadap seluruh data yang dianalisis.	$\frac{TP\ +\ TN}{TP\ +\ TN\ +\ FP\ +\ FN}$
Exact Match (EM)	Persentase jawaban prediksi yang sama persis (setelah normalisasi) dengan jawaban referensi.	EM = 1 jika cocok persis; 0 jika tidak. (Dihitung rata-rata untuk seluruh set evaluasi)

Tahap evaluasi dilakukan untuk mengukur kinerja model setelah proses pelatihan usai. Instrumen fundamental dalam tahap ini adalah *confusion matrix*, yang berfungsi untuk menganalisis dan menilai keluaran prediksi dari model. Berdasarkan *confusion matrix*, dapat dikalkulasi berbagai metrik krusial seperti *precision*, *accuracy*, *f1-score*, dan *recall*. Kumpulan metrik ini menyajikan informasi mendetail mengenai tingkat keandalan sistem klasifikasi dalam mengidentifikasi setiap kelas secara tepat.

3. Hasil Dan Pembahasan

3.1 Hasil Preprocessing

Tahapan *preprocessing* berhasil menyaring dan menstandarkan data teks dari jawaban responden terkait survei DBD. Proses *preprocessing* mencakup *case folding,* normalisasi, *stopword, tokenize dan stemming*.

Tabel 4. Sampel Hasil *Tokenize* dan *Stemming*

Pertanyaan	R1	R1_token	R1_stem
1	demam berdarah dengue penyakit infek...	[demam, berdarah, dengue, penyakit, infeksi, v...	[demam, darah, dengue, sakit, infeksi, virus, ...
2	gejala utama demam berdarah deng...	[gejala, utama, demam, berdarah, dengue, melip.	[gejala, utama, demam, darah, dengue, liput, d..
3	penyebab utama penyakit dem...	[penyebab, utama, penyakit, demam....	[sebab, utama, sakit, demam, darah, dengue, in..
4	penularan demam berdarah dengue terjad...	[penularan, demam, berdarah, deng ...	[tular, demam, darah, dengue, jadi, nyamuk, ae...
5	solusi meningkatkan kepedu....	[solusi, meningkatkan, kepedulian....	[solusi, tingkat, peduli, masyarakat, cegah, d...

Salah satu hasil visualisasi yang digunakan adalah *Wordcloud* merupakan salah satu bentuk visualisasi yang digunakan untuk memahami sebaran kata-kata yang paling sering muncul dalam sebuah korpus teks. Visualisasi ini menampilkan kata-kata dengan ukuran huruf yang bervariasi, di mana semakin besar ukuran font suatu kata, semakin sering kata tersebut muncul dalam data teks yang telah diproses.

Gambar 2. *Wordcloud* Hasil *Preprocessing*

Berdasarkan hasil *Wordcloud* yang dihasilkan dari teks jawaban responden setelah dilakukan proses tokenisasi dan stemming, terdapat beberapa kata dominan yang muncul dengan frekuensi tinggi, seperti:

- (“demam”, “darah”, “dengue”) Tiga kata ini secara konsisten mendominasi karena merupakan topik utama dari survei, menunjukkan fokus responden terhadap penyakit DBD secara umum.
- (“nyamuk”) Kata ini menonjol karena berkaitan langsung dengan vektor penular DBD, yaitu nyamuk *Aedes aegypti.* Frekuensi tinggi kata ini menunjukkan kesadaran responden terhadap faktor penyebab DBD.
- (“virus”, “infeksi”, “penyakit”) Menandakan adanya pemahaman responden bahwa DBD merupakan penyakit yang timbul akibat infeksi virus.
- (“gejala”, “tinggi”, “panas”, “kepala”) Merupakan kata-kata yang umum ditemukan pada respons mengenai gejala DBD, mencerminkan persepsi umum masyarakat terhadap tanda-tanda penyakit.
- (“bersih”, “fogging”, “cegah”, “tempat”, “air”) Kata-kata ini umumnya muncul dalam konteks pencegahan dan solusi. Hal ini mengindikasikan bahwa mayoritas responden memahami pentingnya menjaga kebersihan lingkungan serta mengambil langkah-langkah pencegahan seperti *fogging* dan pengelolaan tempat penampungan air.

Kehadiran kata-kata tersebut menunjukkan bahwa mayoritas responden memiliki pemahaman dasar yang baik terkait DBD, termasuk penyebab, gejala, dan cara pencegahannya. *Wordcloud* ini menjadi alat awal yang efektif untuk melihat topik atau konsep utama yang banyak dibahas responden sebelum dilakukan analisis lebih mendalam menggunakan BERT dan *cosine similarity*.

3.2 Hasil Implementasi Model

Setelah data teks hasil survei di*preprocessing*, representasi kata-kata yang telah di*stemming* selanjutnya dikonversi ke dalam bentuk vektor numerik dengan memanfaatkan model BERT. Model ini memungkinkan representasi semantik yang kontekstual dan kaya makna, sehingga respons yang memiliki makna serupa dapat dikenali walaupun ditulis dengan gaya bahasa yang berbeda [15].

Tabel 5. Sampel Hasil BERT *Embedding*

Responden	Pertanyaan 1	Pertanyaan 2	Pertanyaan 3	Pertanyaan 4	Pertanyaan 5
r1	[0.15549910068511963, 0.66...	[0.02108485251665 1154, 0.9...	[0.039561085402 965546, 0.6...	[0.12920063734054 565, 0.57...	[-0.77701807022094 73, 1.121...
....	....	....	.....	....	....
r33	[-0.04375592246651 6495, 0..	[0.22548431158065 796, 0.51...	[-0.1775466054677 9633, 0....[-0.1048910543322 5632,	[0.13485628366470 337, 1.12... [-0.15716218948364 258,	[-0.87503194808959 96, 0.926...

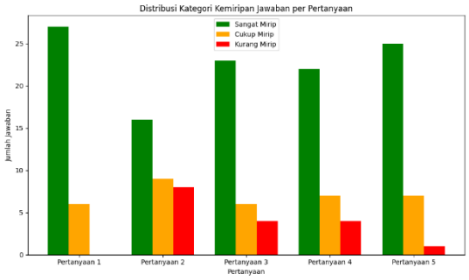
Hasil vektorisasi dari BERT menghasilkan nilai-nilai *embedding* untuk setiap respons yang diberikan oleh responden terhadap lima pertanyaan survei. Masing-masing jawaban direpresentasikan

sebagai vektor berdimensi tinggi, dan kesamaan antar jawaban diukur menggunakan metode *cosine similarity*. *Cosine similarity* ini berfungsi untuk menilai seberapa sejajar arah dua vektor *embedding,* yang mencerminkan kesamaan makna antar jawaban, tanpa mempertimbangkan panjang atau magnitudo dari vektor tersebut.

Tabel 6. Sampel Hasil *Cosine Similarity*

Pertanyaan	Responden	Respon Jawaban	*Cosine Similarity*	Hasil Kategori
1	r2	[demam, darah, deng..	0.993777	Sangat Mirip
2	r1	[gejala, utama, demam, dar....	0.989861	Sangat Mirip
3	r1	[sebab, utama, sakit, demam, dar...	0.982378	Sangat Mirip
4	r32	[gigit, nyamuk, aed...]	0.828755	Cukup Mirip
5	r15	[terap, hidup, bersih]	0.715730	Cukup Mirip

Berdasarkan hasil perhitungan *cosine similarity*, setiap jawaban responden kemudian dikategorikan ke dalam tiga kelas tingkat kemiripan, sebagaimana ditunjukkan pada Tabel 2. Sebagian besar jawaban responden terhadap pertanyaan-pertanyaan fundamental seperti penyebab, gejala, dan penularan DBD termasuk dalam kategori "Sangat Mirip", menunjukkan tingkat pemahaman yang seragam terhadap informasi dasar mengenai DBD. Sebagai contoh, respons terhadap pertanyaan tentang penyebab DBD menghasilkan nilai *cosine similarity* antara 0.97 hingga 0.99, yang mencerminkan bahwa responden memiliki pandangan yang hampir identik mengenai virus dan peran nyamuk dalam penyebaran penyakit.



Gambar 3. Distribusi Kategori *Cosine Similarity*

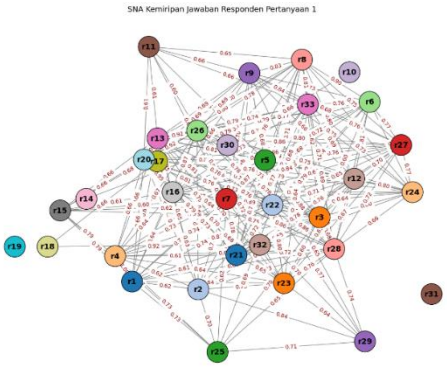
Nilai *cosine similarity* yang tinggi merepresentasikan bahwa mayoritas responden memiliki pemahaman yang seragam dan konsisten, terutama pada pertanyaan-pertanyaan faktual seperti *"Apa itu DBD?"*, *"Apa penyebab utama DBD?"*, dan *"Bagaimana penularan DBD terjadi?"*. Namun, terjadi sedikit penurunan tingkat similarity pada pertanyaan yang bersifat opini atau solutif, seperti *"Bagaimana cara meningkatkan kepedulian masyarakat terhadap pencegahan DBD?"*. Misalnya, jawaban seperti "terap hidup bersih" (similarity: 0.715) dan "hindari gigitan nyamuk, gunakan fogging" (similarity: 0.905) diklasifikasikan sebagai *Cukup Mirip*. Hal ini menunjukkan adanya variasi pendekatan yang diusulkan oleh responden dalam upaya pencegahan DBD, yang mencerminkan keberagaman pemikiran serta pengalaman pribadi masing-masing individu.

Tabel 7. Most Jawaban

Most Jawaban	
Responden	**Cosine similarity**
r2	0.993777
r3	0.995322
r28	0.985774
r2	0.981760

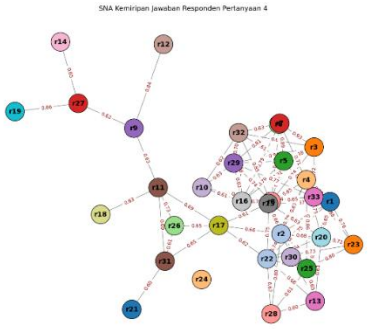
r5	0.978627

Tabel 7 menunjukkan lima responden dengan nilai *cosine similarity* tertinggi, yang berarti jawaban mereka memiliki tingkat kemiripan makna paling dekat terhadap responden acuan atau pusat analisis. Responden r3 menempati posisi teratas dengan nilai 0.995, diikuti oleh r2 dengan dua entri (nilai 0.993 dan 0.981), kemudian r28 (0.985) dan r5 (0.978). Nilai-nilai yang mendekati angka 1 ini menunjukkan bahwa responden-responden tersebut memberikan jawaban yang sangat serupa, baik dalam struktur maupun kandungan makna, sehingga dapat diasumsikan memiliki pemahaman atau persepsi yang sejalan terhadap topik yang ditanyakan.



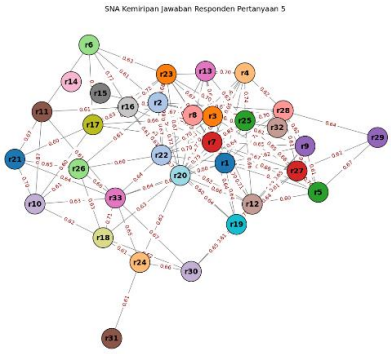
Gambar 4. Graph SNA Pertanyaan 1

Gambar 4 merupakan visualisasi SNA yang menunjukkan kemiripan jawaban responden terhadap Pertanyaan 1. Node dengan ukuran yang lebih besar (seperti r7) menunjukkan tingkat keterhubungan yang lebih tinggi, menandakan bahwa responden tersebut memiliki kemiripan jawaban dengan banyak responden lainnya. Hal ini menunjukkan bahwa r7 berpotensi merepresentasikan pendapat yang paling umum atau paling banyak disetujui di antara responden lain. Dan Warna yang berbeda digunakan untuk membedakan kelompok dalam jaringan, yang mengindikasikan adanya klaster atau pengelompokan responden dengan jawaban yang serupa.



Gambar 5. Graph SNA Pertanyaan 4

Gambar 5 menunjukkan untuk Pertanyaan 4. Responden seperti r1, r16, dan r30 menjadi pusat dalam kelompok ini dan merepresentasikan pandangan mayoritas. Di sisi lain, terdapat cabang minoritas yang terhubung secara bertahap melalui responden seperti r11 dan r17, yang berperan sebagai penghubung antara kelompok utama dan responden dengan jawaban yang lebih unik, seperti r12 dan r14. Hal ini mengindikasikan tingkat kesepakatan yang tinggi terhadap Pertanyaan 4, meskipun masih ada variasi pandangan yang terstruktur.



Gambar 6. Graph SNA Pertanyaan 5

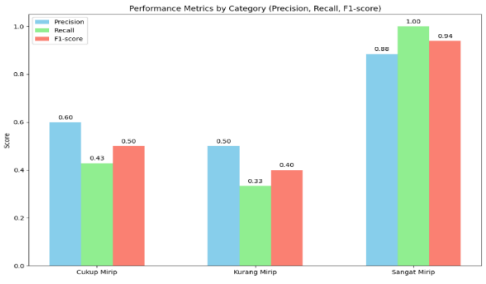
Grafik diatas untuk Pertanyaan 5 menunjukkan struktur jaringan yang sangat terhubung dan menyebar luas. Hampir semua responden saling terhubung, membentuk pola seperti jaring laba-laba, yang mencerminkan bahwa jawaban mereka saling berbagi elemen meskipun tidak identik. Responden seperti r1, r7, dan r22 berada di pusat jaringan karena jawaban mereka dianggap paling umum atau representatif, sementara responden seperti r31 dan r29 berada di tepi jaringan dengan jawaban yang lebih unik. Hal ini menandakan bahwa Pertanyaan 5 memunculkan jawaban yang kompleks, beragam, dan bernuansa.

Secara keseluruhan, hasil implementasi model menunjukkan bahwa sistem berbasis BERT dan *cosine similarity* mampu mengelompokkan jawaban secara efektif berdasarkan kesamaan makna. Hal ini mendukung tujuan utama penelitian, yaitu menyediakan dasar sistem tanya jawab dan analisis survei yang lebih cerdas, efisien, dan berbasis makna semantik [16]. Selain itu, penerapan SNA turut memperkuat sistem dengan mengungkap pola hubungan antar entitas dalam data, memungkinkan identifikasi topik sentral dan distribusi opini.

3.3 Hasil Evaluasi

Tabel 8. Hasil *Performance Matrix*

Parameter	Kategori		
	Cukup Mirip	**Kurang Mirip**	**Sangat Mirip**
Precision	0.60	0.50	0.88
Recall	0.43	0.33	1.00
F1-Score	0.50	0.40	0.94
Akurasi	0.76		



Gambar 7. Barchart Hasil *Performance Matrix*

Berikut hasil dari *Performance Metrix*:

a. Kategori "Sangat Mirip" menunjukkan performa model yang paling unggul. Dengan presisi sebesar 0.88, model mampu mengklasifikasikan 88% item dengan benar sebagai "Sangat Mirip", menunjukkan tingkat kesalahan yang rendah. *Recall* mencapai skor sempurna 1.00, menandakan bahwa seluruh data yang benar-benar termasuk dalam kategori ini berhasil teridentifikasi tanpa ada yang terlewat. *F1-score* yang tinggi sebesar 0.94 mencerminkan keseimbangan optimal antara *precision* dan *recall*, menandakan kinerja model yang sangat andal dan efektif dalam mengenali kategori ini.

b. Performa model pada kategori "Cukup Mirip" berada pada tingkat sedang. Nilai *precision* sebesar 0.60 mengindikasikan bahwa hanya 60% prediksi pada kategori ini benar, prediksi kategori ini yang benar, sementara sisanya merupakan kesalahan klasifikasi. *Recall* yang lebih rendah, yakni 0.43, menunjukkan bahwa model tidak berhasil mengenali sebagian besar item yang seharusnya terklasifikasi. Hal ini tercermin dalam *F1-score* sebesar 0.50 yang menunjukkan keseimbangan antara *precision* dan *recall* masih perlu ditingkatkan agar model dapat mengenali kategori ini dengan lebih akurat.

c. Kategori "Kurang Mirip" merupakan kategori dengan performa model yang paling lemah. *Precision* sebesar 0.50 menunjukkan bahwa hanya separuh dari prediksi yang dibuat untuk kategori ini benar, sementara sisanya merupakan kesalahan klasifikasi. Dengan *recall* yang rendah yaitu 0.33, model hanya mampu mengidentifikasi sepertiga dari seluruh item yang sebenarnya termasuk dalam kategori ini. Hal ini mencerminkan kesulitan model dalam mengenali pola-pola yang sesuai. *F1-score* sebesar 0.40 semakin menegaskan bahwa kemampuan model dalam mengklasifikasikan item ke dalam kategori "Kurang Mirip" masih jauh dari optimal, baik dari segi ketepatan maupun kelengkapan. Secara keseluruhan, model ini menunjukkan kinerja yang sangat baik untuk kategori "Sangat Mirip", namun performanya menurun secara signifikan untuk kategori "Cukup Mirip" dan paling rendah pada kategori "Kurang Mirip". Meskipun demikian, model secara umum mencapai akurasi sebesar 0.76, yang menunjukkan performa klasifikasi yang cukup baik dalam konteks analisis survei berbasis teks.

Tabel 9. Hasil *Exact Match*

Exact Match	
Total *Comparations*	165
Total *Exact Matches*	0
Proporsi *Exact Match (Pseudo-Accuracy)*	0%

Interpretasi dari hasil perbandingan untuk menemukan kecocokan persis (*Exact Match*):

- Total *Comparisons*: 165 Ini menunjukkan bahwa proses evaluasi telah melakukan sebanyak 165 kali perbandingan antara pasangan item data.
- Total *Exact Matches*: 0 Dari 165 perbandingan tersebut, tidak ditemukan satupun (0) pasangan yang cocok secara persis. Ini berarti tidak ada dua item yang dibandingkan memiliki isi atau nilai yang 100% identik.
- Proporsi *Exact Match (Pseudo-Accuracy):* 0.0000 Angka ini adalah rasio dari jumlah kecocokan persis dibagi dengan total perbandingan (0 / 165). Nilai 0.0000 mengonfirmasi bahwa tingkat akurasi untuk menemukan kecocokan yang identik dalam set data ini adalah nol.

Analisis *Exact Match* menunjukkan bahwa di antara set data yang diuji, tidak ada entri yang sama persis atau duplikat secara mutlak. Semua 165 item yang diperbandingkan memiliki setidaknya sedikit perbedaan satu sama lain.

4. Kesimpulan

Penelitian ini berhasil mengimplementasikan kombinasi model BERT dan SNA dalam sistem Tanya Jawab untuk menganalisis data survei DBD dengan 33 responden. Hasil *preprocessing* data menunjukkan bahwa kata-kata dominan seperti "demam", "darah", "dengue", "nyamuk", dan "virus" mencerminkan pemahaman dasar responden terhadap penyakit DBD. Berdasarkan hasil perhitungan *cosine similarity*, mayoritas jawaban responden terhadap pertanyaan fundamental mengenai DBD (definisi,

penyebab, gejala, dan penularan) menunjukkan nilai kemiripan yang sangat tinggi (>0.85), dengan responden r3 memperoleh nilai tertinggi sebesar 0.995. Ini menunjukkan bahwa mayoritas responden memiliki persepsi yang sejalan dan pemahaman yang konsisten terhadap informasi fundamental mengenai DBD. Namun, pada pertanyaan yang bersifat opini seperti solusi pencegahan, terdapat variasi yang lebih besar dalam jawaban responden, menunjukkan keberagaman pendekatan dan pengalaman pribadi mereka. Analisis SNA mengungkap pola hubungan yang menarik antar responden, dimana beberapa responden seperti r7, r27, dan r28 berperan sebagai pusat opini karena jawaban mereka memiliki kemiripan tinggi dengan banyak responden lainnya. Visualisasi grafik SNA menunjukkan bahwa pertanyaan faktual menghasilkan struktur jaringan yang lebih terpusat, sementara pertanyaan opini menghasilkan jaringan yang lebih tersebar dan kompleks.

Secara keseluruhan, Penelitian ini membuktikan bahwa integrasi BERT dan SNA dapat menjadi pendekatan yang efektif untuk menganalisis data survei berbasis teks. Model menunjukkan kinerja yang sangat baik dalam mengidentifikasi jawaban dengan kemiripan tinggi (kategori "Sangat Mirip"), namun mengalami penurunan performa pada kategori kemiripan yang lebih rendah. Meskipun demikian, sistem ini berhasil memberikan wawasan mendalam tentang distribusi pemahaman masyarakat terhadap DBD dan menyediakan dasar yang kuat untuk pengembangan sistem tanya jawab yang lebih cerdas dalam konteks ekspresi jawaban responden, yang tercermin dari hasil *Exact Match* 0%, justru menunjukkan kekayaan perspektif masyarakat dalam memahami isu kesehatan, sambil tetap mempertahankan konsistensi dalam substansi pemahaman dasar tentang DBD.

Daftar Pustaka

[1] H. Gholami, M. Gachpazan, and M. Erfanian, “Mathematical modeling and dynamic analysis of dengue fever: examining economic and psychological impacts and forecasting disease trends through 2030—a case study of Nepal,” *Sci Rep*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-94527-8.

[2] M. Ompusunggu *et al.*, “IMPLEMENTASI SENTENCE TRANSFORMER BERBASIS BERT UNTUK EDUKASI DAN INFORMASI TENTANG HIV/AIDS,” *Prosiding Seminar Nasional Sains dan Teknologi Seri III Fakultas Sains dan Teknologi*, vol. 2, no. 1, 2025.

[3] F. R. Fatonah, D. S. Maylawati, and E. Nurlatifah, “Chatbot Edukasi Pra-Nikah berbasis Telegram Menggunakan Bidirectional Encoder Representations From Transformers (BERT),” *Jurnal Algoritma*, vol. 21, no. 2, pp. 29–40, Nov. 2024, doi: 10.33364/algoritma/v.21-2.1657.

[4] F. T. Sabilillah, S. Winarno, and R. B. Abiyyi, “Implementasi BERT dan Cosine Similarity untuk Rekomendasi Dosen Pembimbing berdasarkan Judul Tugas Akhir,” *Edumatic: Jurnal Pendidikan Informatika*, vol. 8, no. 2, pp. 585–594, Dec. 2024, doi: 10.29408/edumatic.v8i2.27791.

[5] Rio Feriangga Kurniawan, “IMPLEMENTASI TEXT MINING MENGGUNAKAN METODE COSINE SIMILARITY UNTUK KLASIFIKASI KONTEN BERITA DI POSTINGAN GRUP FACEBOOK INFO LANTAS DAN KRIMINAL PASURUAN,” *JAMI: Jurnal Ahli Muda Indonesia*, vol. 3, no. 1, pp. 9–17, Jun. 2022, doi: 10.46510/jami.v3i1.41.

[6] D. Marta, G. Leonarde Ginting, and A. Hatuaon Sihite, “Deteksi Berita Palsu Tentang Vaksinasi Covid-19 Dengan Menggunakan Text Mining Dan Algoritma Cosine Similarity,” *Nasional Teknologi Informasi dan Komputer)*, vol. 6, no. 1, 2022, doi: 10.30865/komik.v6i1.5738.

[7] T. P. Lestari, “Analisis Text Mining pada Sosial Media Twitter Menggunakan Metode Support Vector Machine (SVM) dan Social Network Analysis (SNA),” *Jurnal Informatika Ekonomi Bisnis*, pp. 65–71, Aug. 2022, doi: 10.37034/infeb.v4i3.146.

[8] A. Syahid and A. Nurbahri Akhmad, “ANALISIS JARINGAN SOSIAL DALAM PENYEBARAN INFORMASI TERKAIT MENTAL HEALTH DI MEDIA SOSIAL: STUDI KASUS X,” Nov. 2024, [Online]. Available: http://jurnal.uts.ac.id/index.php/KAGANGA

[9] A. Sanjaya, A. Bagus Setiawan, U. Mahdiyah, I. Nur Farida, A. Risky Prasetyo, and U. Nusantara PGRI Kediri, “PENGUKURAN KEMIRIPAN MAKNA MENGGUNAKAN COSINE SIMILARITY DAN BASIS DATA SINONIM KATA MEASUREMENT OF MEANING SIMILARITY USING COSINE SIMILARITY AND WORD SYNONYMS DATABASE,” vol. 10, no. 4, 2023, doi: 10.25126/jtiik.2023106864.

[10] D. Darwis, E. Shintya Pratiwi, A. Ferico, and O. Pasaribu, "PENERAPAN ALGORITMA SVM UNTUK ANALISIS SENTIMEN PADA DATA TWITTER KOMISI PEMBERANTASAN KORUPSI REPUBLIK INDONESIA," 2020.
[11] J. Sasongko Wibowo, H. Februariyanti, and H. Listiyono, "Model Penjawab Pertanyaan Otomatis Berdasarkan Peringkat Relevansi Kalimat Menggunakan Model BERT," 2024.
[12] Z. Alhaq, A. Mustopa, and J. D. Santoso, "PENERAPAN METODE SUPPORT VECTOR MACHINE UNTUK ANALISIS SENTIMEN PENGGUNA TWITTER."
[13] R. Fitri, A. Noor Asyikin, and S. Pengajar Jurusan Teknik Elektro Politeknik Negeri Banjarmasin Ringkasan, "APLIKASI PENILAIAN UJIAN ESSAY OTOMATIS MENGGUNAKAN METODE COSINE SIMILARITY," vol. 7, no. 2, pp. 54–105, 2015.
[14] M. A. Akbar, Masniarara Aziza Balfas Amril, Raiza Syahira, Fahrein Rachel Latisha, and Noor Jihan, "ANALISIS STRUKTUR JARINGAN KOMUNIKASI #SEAGAMES2022 DI TWITTER MENGGUNAKAN PENDEKATAN SOCIAL NETWORK ANALYSIS (SNA)," *Jurnal Studi Komunikasi dan Media*, vol. 26, no. 1, pp. 1–16, Dec. 2022, doi: 10.17933/jskm.2022.4780.
[15] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," Aug. 2019, [Online]. Available: http://arxiv.org/abs/1908.10084
[16] J. Wang and Y. Dong, "Measurement of text similarity: A survey," Sep. 01, 2020, *MDPI AG*. doi: 10.3390/info11090421.