J-INTECH (Journal of Information and Technology)
Terakreditasi Kemendikbud SK No. 204/E/KPT/2022
E-ISSN: 2580-720X || P-ISSN: 2303-1425

Penerapan Metode *Cosine Similarity* Pada Sistem Informasi *Retrival* Pencarian Terjemahan Ayat-Ayat Suci Al-Qur'An

Evy Poerbaningtyas [1], Rakhmad Maulidi [2*]

[1,2]*Sekolah Tinggi Informatika & Komputer Indonesia, Program Studi Informatika, Jalan Tidar 100, Kota Malang*

Informasi Artikel
Diterima: 24-11-2023
Direvisi: 09-12-2023
Diterbitkan: 24-12-2023

Kata Kunci
Informasi; Software; Aplikasi; Sistem

***Email Korespondensi:**
maulidi@stiki.ac.id

Abstrak
Ayat dalam Al-Qur'an adalah kalimat yang diatur secara ketat (tauqifi) berdasarkan petunjuk langsung dari Rasulullah saw., bukan melalui analogi (qiasi). Oleh karena itu, setiap gabungan huruf seperti (Alif Lam Mim) dan (Alif Lam Mim Shod Nun) dianggap sebagai satu ayat, sedangkan kombinasi seperti (Alif Lam Mim) dan (Kha Mim), (Alif Lam Mim Ra), serta (Ya Syin) dihitung sebagai satu ayat masing-masing. Namun, (Tho Syin Mim) tidak dihitung sebagai satu ayat. Ayat yang berisi ungkapan "Madahaa Mannaan" dianggap sebagai satu kata, dan secara umum disepakati bahwa shalat tidak sah jika hanya membaca separuh ayat. Dalam setiap ayat terdapat kata-kata yang membentuk kalimat, sehingga untuk menemukan kalimat yang diinginkan perlu dilakukan pencarian satu per satu pada ayat. Sebagai panduan hidup umat Islam, Al-Qur'an mengandung surat-surat dan ayat-ayat di dalamnya. Oleh karena itu, penelitian ini bertujuan untuk menciptakan mesin pencarian ayat yang menggunakan metode cosine similarity, sehingga dapat membantu mempercepat dan memudahkan proses pencarian bagi pembaca yang mencari ayat dalam kitab suci Al-Qur'an.

Abstract
Ayat is a sentence in the Qur'an, the verse is known as tauqifi (with the nash of the Messenger of Allah saw.), not qiasi (analogy). Since it counts (Alif Lam Mim) and (Alif Lam Mim Shod Nun) each as one verse but does not count (Alif Lam Mim) and (Kha Mim), (Alif Lam Mim Ra) and (Ya Syin) each count as one verse, but not so with (Tho Syin Mim), Ayat (Madahaa Mannaan) is one word, according to the agreement, invalid prayer by reading only half of the verse. In a verse some words make up a sentence, so to search for a desired sentence you have to look up 1 by 1 in the verse. Because the guidance of the life of Muslims is the holy book of the Qur'an, in which there are several letters and in the letter there are verses. From these conditions, in this study, a verse search engine was created to facilitate the search process using the cosine similarity method. The purpose of this study is to be able to help the search process quickly and easily, readers in the search for verses in the holy book of the Qur'an .

1. Pendahuluan

Generasi muda(Gen Z) mempunyai motivasi lebih tinggi untuk mempelajari bahasa asing, tidak hanya satu bahasa asing mereka cenderung ingin memiliki kemampuan beberapa beberapa bahasa(*polyglot*) (Calafato, 2020) . Bahasa arab di Indonesia banyak berkembang pesat di kalangan pondok pesantren atau perguruan tinggi agama islam (Sa'diyah & Abdurahman, 2021) . Sebagai negara dengan penduduk muslim terbanyak di dunia, muslim di Indonesia memiliki minat yang cukup tinggi untuk belajar Al-Quran tetapi banyak yang masih belum menguasai bahasa arab, sehingga mengandalkan terjemahan Al-Quran. Pada umumnya pencarian topik/ayat tertentu pada penggunaan terjemahan Al-Quran dilakukan dengan mencari satu persatu dari setiap surat. Padahal dalam satu surat di Al-Quran membahas beberapa topik tertentu, hal ini menyebabkan orang yang akan mempelajari Al-Quran mengalami kesulitan.

Sistem pencarian informasi konvensional hanya memanfaatkan isi dokumen untuk menghasilkan hasil dari pertanyaan. Information retrieval adalah suatu sistem yang memberikan akses cepat dan efisien sejumlah informasi besar yang tersimpan, sehingga meningkatkan produktivitas dan pengalaman pengguna sambil meminimalkan informasi yang tidak relevan. Pencarian informasi memiliki peran penting dalam mengubah data menjadi informasi dan pengetahuan yang berguna, ringkas, akurat, dan tepat waktu. Saat ini bidang ilmu komputer telah menjadikan information retrieval sebagai area penelitian yang penting (Anand, Sharma, & Kumar, 2020).

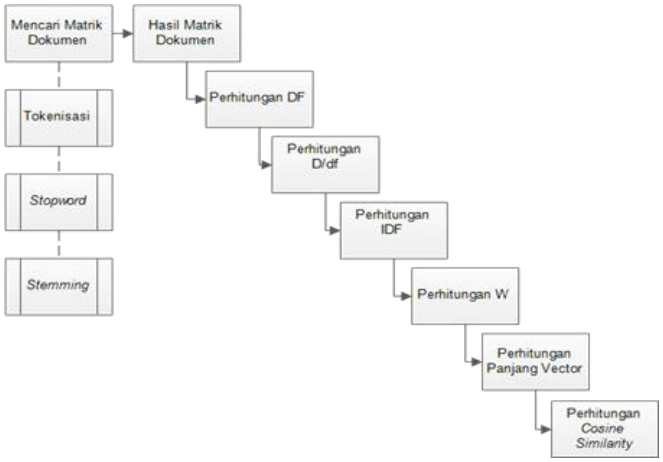
Dalam domain Information Retrieval, terdapat beberapa metode yang digunakan dalam proses pencarian, dan salah satu di antaranya adalah melalui penerapan Model Ruang Vektor ((Wahyudi, Akbar, & Suryamen, 2020) . Pendekatan ini didasarkan pada konsep bahwa esensi dari sebuah dokumen dapat ditentukan oleh kata-kata yang terkandung di dalamnya. Model ini mengevaluasi tingkat kemiripan (similarity) (Agustian & Sucipto, 2021) antara suatu dokumen dengan query dengan cara merepresentasikan keduanya dalam bentuk vektor. Setiap kata yang terdapat dalam dokumen dan query diberi bobot dan disimpan sebagai elemen vektor (Rasyid, Yudianto, Maimunah, & Purnomo, 2023)

Similaritas antar dokumen dijelaskan melalui representasi bag-of-words yang kemudian dikonversi ke dalam suatu model ruang vektor. Salah satu metode yang digunakan untuk mengukur tingkat kesamaan antara dua objek adalah Metode Cosine Similarity (Adiyanto & Handayani, 2022) . Metode ini menghitung similarity antara dua objek, misalnya D1 dan D2, yang direpresentasikan dalam bentuk vektor dengan menggunakan kata kunci dari dokumen sebagai parameter. Pengukuran ini memungkinkan penilaian peringkat dokumen berdasarkan tingkat kemiripannya (relevansi) terhadap suatu query. Setelah semua dokumen dinilai, sejumlah dokumen dengan peringkat tertinggi dikembalikan kepada pengguna. Metode cosine similarity mengukur nilai cosinus sudut antara dua vektor (Adiyanto & Handayani, 2022).

Penelitian terdahulu tentang penggunaaan Cosine Similarity tentang pencarian informasi berita pariwisata yang dilakukukan oleh (Suzanti , Husni, Awaldhia , & Syarief, 2023) , dalam penelitian tersebut dilakukan modifikasi pada model Cosine Similarity-nya dengan tahap pre-processing case folding, tokenizing, stop word, stemming feature selection yang digunakan adalah TF-IDF. Penelitian lain tentang pencarian informasi menggunakan Bahasa Gujari dengan menggunakan model Cosine similarity yang dilakukan oleh (Rakholia & Saini, 2017) , pada penelitian ini hanya menggunakan stop word removal pada tahap pre-processing, sedangkan untuk feature selection yang digunakan TF-IDF dan normalisasi. Perbedaan dalam penelitian ini adalah pada tahan pre-processing tidak dilakukan case folding dan pada feature selection tidak dilakukan normalisasi.

2. Metode Penelitian

Tahapan penelitian tentang sistem pencarian menggunakan algoritma cosine similarity sebagai berikut:



Gambar 1 Diagram Penerapan Algoritma Cosine Similarity

Sistem Indexing adalah bagian dari proses subsistem yang mengubah kumpulan dokumen menjadi suatu bentuk tertentu untuk mempermudah dan mempercepat pencarian serta penemuan kembali dokumen yang relevan (Davison, Avgil, Li, & Yang, 2022) . Tokenisasi adalah langkah dalam memisahkan serangkaian karakter berdasarkan karakter spasi, dan mungkin pada saat yang sama, melibatkan penghapusan karakter tertentu, seperti tanda baca (Anand, Sharma, & Kumar, 2020).

Penghapusan stopwords adalah tahap eliminasi term yang tidak memiliki makna atau relevansi yang signifikan, dan langkah ini dilakukan selama proses tokenisasi (Chanda & Pal, 2023). Proses penyaringan ini memanfaatkan daftar stopwords yang telah ditetapkan (Rinandyaswara, Sari, & Furqon, 2022), yang terdiri dari stopwords dalam bahasa Indonesia seperti: ada, yang, ke, kepada, misal, atau, dan sebagainya.

Stemming digunakan untuk mengubah term dengan menghilangkan awalan, sisipan, dan akhiran yang masih melekat pada term tersebut (Rinandyaswara, Sari, & Furqon, 2022). Proses stemming dilakukan dengan cara mengeliminasi semua jenis imbuhan, termasuk awalan (prefixes), sisipan (infixes), akhiran (suffixes), dan confixes (kombinasi dari awalan dan akhiran) pada kata turunan (Suzanti & Jauhari, 2021).

Setiap kata diberi nilai bobot sesuai dengan skema pembobotan yang dipilih, apakah itu pembobotan lokal, global, atau gabungan keduanya. Pembobotan kata memiliki dampak besar dalam menentukan sejauh mana kesamaan antara dokumen dan query. Faktor-faktor kunci dalam pembobotan kata melibatkan Term Frequency (TF), Inverse Document Frequency (IDF), dan TFIDF (Nugroho, Bachtiar, & Wihandika, 2022).

Metode cosine similarity mengukur tingkat kesamaan antara dua objek, seperti D1 dan D2, yang diwujudkan dalam dua vektor dengan menggunakan kata kunci dari sebuah dokumen sebagai indikator. Pengukuran ini memungkinkan pengurutan dokumen berdasarkan sejauh mana kesamaannya (relevansi) terhadap query. Setelah semua dokumen diberi peringkat, sejumlah tetap dokumen dengan peringkat tertinggi dikembalikan kepada pengguna (Yulianto, Budiharto, & Kartowisastro, 2017) . Penilaian dalam cosine similarity ini melibatkan perhitungan nilai cosinus sudut antara dua vektor. Jika terdapat dua vektor dokumen dj dan query q, serta term diekstrak dari lokasi dokumen, maka nilai cosinus antara dj dan q dapat diartikan sebagai berikut.

$$Similarity(D,Q) = \cos\theta \frac{\sum_{i=1}^{n}(WD_i \,.\, WQ_i)}{\sum_{i=1}^{n}(WD_i)^2 \,.\, \sum_{i=1}^{n}(WQ_i)^2 \,.}$$

3. Hasil dan Pembahasan

Penerapan algoritma cosine similarity untuk proses perhitungan bobot kata pada setiap kata (key) dalam ayat-ayat surat Al-Baqarah. Contoh studi kasus :

Q = kisah penyembelihan sapi

D1 = 'Sembelihlah seekor sapi dengan ciri-ciri kuning tua warnanya, lagi menyenangkan orang-orang yang memandangnya'.

D2 = 'Sesungguhnya sapi betina itu adalah sapi betina yang belum pernah dipakai membajak tanah dan tidak pula untuk mengairi tanaman, tidak bercacat, tidak ada belangnya.

D3 = 'Laksanakan perintah untuk menyembelih'.

Langkah-langkah :

a. Mencari matrik dari dokumen, dengan 3 tahap yaitu : Tokenisasi, *Stop Word Removal, Stemming*
b. Perhitungan df, D/df, dan Idf
c. Perhitungan W, dimana W = Tf*Idf
d. Perhitungan Panjang Vektor
e. Perhitungan Koefisien Cosinus

Tabel 1 Tekonisasi Dokumen

Sembelihlah	adalah
seekor	belum
sapi	pernah
dengan	dipakai
ciri	membajak
kuning	tanah
tua	dan
warnanya	tidak
lagi	pula
menyenangkan	mengairi
orang	tanaman
yang	bercacat
memandangnya	ada
sesungguhnya	belangnya
betina	laksanakan
itu	perintah

Proses tokenisasi dokumen adalah sebuah proses yang memecah dokumen, bisa berupa paragaraf, kalimat menjadi bagian-bagian tertentu. Pembagian dokumen berdasarkan separator/karakter tertentu, biasanya karakter spasi digunakan untuk memecah suatu kalimat menjadi kata-kata, karakter titik untuk memecah suatu paragraph menjadi kalimat dan karakter newline untuk memecah suatu dokumen menjadi paragraf. Pada table 1 merupakan hasil proses tokenisi suatu ayat terjemahan al-quran menjadi beberapa kata-kata dengan menggunakan karakter spasi sebagai pemisah. Setelah proses tokenisasi dilakukan, proses selanjutnya adalah memproses dokumen hasil tokenisasi ke dalam proses stopword removal yang merupakan rangkai proses pre-processing dalam penelitian ini.

Tabel 2. Stop word Removal

sembelihlah	warnanya	dipakai
seekor	menyenangkan	membajak
sapi	orang	tanah
ciri	memandangnya	mengairi
kuning	sesungguhnya	tanaman
tua	betina	bercacat

Proses *stop word removal* merupakan tahap pre-processing yang melakukan penghapusan kata-kata yang tidak terkait berdasarkan *whitelist* yang sudah ditentukan, misalnya kata dan, yang, adalah, dengan. Proses penghapusan ini tidak boleh berdampak perubahan makna dari rangkai kalimat. Tabel 2 merupakan dokumen hasil dari proses stopword removal yang sudah dilakukan. Terdapat 14 kata yang dihapus dari 32 kata dari hasil tokenisasi. Selain menghilangkan kata yang tidak terkait, hasil dari proses *stop word removal* juga berdampak pada jumlah kata yang akan diproses sehingga berdampak pada proses komputasi karena data yang diolah menjadi lebih sedikit. Proses selanjutnya setelah stop word removal adalah stemming.

Tabel 3 Stemming

sembelih	warna	pakai
ekor	senang	bajak
sapi	orang	tanah
ciri	pandang	air
kuning	sungguh	tanam
tua	betina	cacat

Stemming adalah sub proses dalam information retrieval dengan cara mengubah suatu kata menjadi kata dasar, perubahan ini menerapkan kaidah tata bahasa dari kata yang akan dibuah. Pada proses stemming dilakukan dengan cara menghilangkan awalan(*prefix*), akhiran(*suffix*), dan sisipan(*infix*). Contoh awalan dalam Bahasa Indonesia misalnya "ber, di,ke", contoh akhiran adalah "lah, an, nya". Tabel 3 merupakan hasil stemming, contohnya kata sembelihlah berubah menjadi sembelih dengan menghilangkan akhiran "lah", contoh lain kata membajak berubah menjadi bajak dengan menghilangkan awalan "mem".

Tabel 4 Perhitungan Document Frequency (DF)

Token	**tf**				**df**	**Token**	**tf**				**df**
	Q	***D1***	***D2***	***D3***			***Q***	***D1***	***D2***	***D3***	
sembelih	1	1	0	1	2	sungguh	0	0	1	0	1
ekor	0	1	0	0	1	betina	1	0	2	0	1
sapi	1	1	2	0	2	pakai	0	0	1	0	1
ciri	0	2	0	0	1	bajak	0	0	1	0	1
kuning	0	1	0	0	1	tanah	0	0	1	0	1
tua	0	1	0	0	1	air	0	0	1	0	1
warna	0	1	0	0	1	tanaman	0	0	1	0	1
senang	0	1	0	0	1	cacat	0	0	1	0	1
orang	0	2	0	0	1						
pandang	0	1	0	0	1						

Setelah proses pre-processing selesai dilakukan, proses selanjutnya adalah proses perhitungan tingkat kemunculan kata yang ingin dicari(Q) pada suatu kompulan dokumen(D_n). Tabel 4 merupakan hasil perhitungan Tingkat kemunculan kata pada dokumen, misalnya kata sembelih dan sapi muncul sekali pada dua dokumen. Sehingga nilai df-nya adalah dua, kata betina ditemukan sekali pada satu dokumen sehingga nilai df-nya adalah satu. Untuk masing-masing kata pada dokumen yang tidak ditemukan pada kata yang dicari nilai df-nya diberi nilai satu.

Tabel 5 Hasil Perhitungan Invers Document Frequency (Idf)

Token	tf				Df	D/df	Idf=log(D/df)
	Q	***D1***	***D2***	***D3***			
sembelih	1	1	0	1	2	1,5	0,176091259
ekor	0	1	0	0	1	3	0,477121255
sapi	1	1	2	0	2	1,5	0,176091259
ciri	0	2	0	0	1	3	0,477121255
kuning	0	1	0	0	1	3	0,477121255
tua	0	1	0	0	1	3	0,477121255
warna	0	1	0	0	1	3	0,477121255
senang	0	1	0	0	1	3	0,477121255
orang	0	2	0	0	1	3	0,477121255
pandang	0	1	0	0	1	3	0,477121255
sungguh	0	0	1	0	1	3	0,477121255
betina	1	0	2	0	1	3	0,477121255
pakai	0	0	1	0	1	3	0,477121255
bajak	0	0	1	0	1	3	0,477121255
tanah	0	0	1	0	1	3	0,477121255
air	0	0	1	0	1	3	0,477121255
tanaman	0	0	1	0	1	3	0,477121255
cacat	0	0	1	0	1	3	0,477121255

Proses perhitungan Inverse Document Frequency(Idf) adalah sebuah konsep yang digunakan dalam pengolahan bahasa alami dan sistem temu kembali informasi untuk mengevaluasi seberapa penting suatu kata kunci (term) dalam sebuah kumpulan dokumen. IDF digunakan untuk mengukur sejauh mana sebuah kata tertentu jarang muncul di seluruh dokumen dalam Kumpulan dokumen, dan digunakan untuk menilai tingkat keunikannya. Perhitungan ini dengan menggunakan persamaan logaritma berikut: $IDF(t) = \log\left(\frac{N}{df(t)}\right)$. Tabel 5 merupakan hasil perhitungan dari Idf.

Perhitungan bobot TF-Idf(w) bertujuan untuk mendapatkan nilai bobot keterkaitan/penting suatu kata terhadap suatu dokumen. Rumus perhitungannya sebagai berikut $w(t,d) = TF(t,d)x\ IDF(t)$. Jika nilai w dari suatu kata itu menunjukkan seberapa penting kata tersebut terhadap konteks pada Kumpulan dokumen. Tabel 6 merupakan hasil perhitungan bobot(w). Proses selanjutnya adalah proses perhitungan Panjang vector.

Tabel 6 Hasil Perhitungan Bobot TF-Idf(w)

Idf=log(D/df)	W			
	Q	***D1***	***D2***	***D3***
0,176091259	0,176091	0,1760913	0	0,176091
0,477121255	0	0,4771213	0	0
0,176091259	0,176091	0,1760913	0,3521825	0
0,477121255	0	0,9542425	0	0
0,477121255	0	0,4771213	0	0
0,477121255	0	0,4771213	0	0
0,477121255	0	0,4771213	0	0
0,477121255	0	0,4771213	0	0
0,477121255	0	0,9542425	0	0
0,477121255	0	0,4771213	0	0
0,477121255	0	0	0,4771213	0
0,477121255	0,477121	0	0,9542425	0
0,477121255	0	0	0,4771213	0
0,477121255	0	0	0,4771213	0
0,477121255	0	0	0,4771213	0
0,477121255	0	0	0,4771213	0
0,477121255	0	0	0,4771213	0
0,477121255	0	0	0,4771213	0
0,477121255	0	0	0,4771213	0
0,477121255	0	0	0	0,477121
0,477121255	0	0	0	0,477121

Panjang vector dalan information retrieval digunakan untuk mengukur relevansi suatu dokumen terhadap suatu query dalam bentuk vector. Metode yang umum digunakan untuk mengukur Panjang vector menggunakan Euclidean. Tabel 7 merupakan hasil perhitungan Panjang vector dari dokumen, panjang vector dihitung dengan menggunakan persamaan berikut. $\|V\| = \sqrt{v_1^2 + v_2^2 + \ldots + v_{n1}^2}$

Tabel 7 Hasil Panjang Vektor

W				Q^2	D1^2	D2^2	D3^2
Q	***D1***	***D2***	***D3***				
0,17	0,17	0	0,17	0,03	0,03	0	0,0
0	0,47	0	0	0	0,22	0	0
0,17	0,17	0,35	0	0,03	0,03	0,12	0
0	0,95	0	0	0	0,91	0	0
0	0,47	0	0	0	0,22	0	0
0	0,47	0	0	0	0,22	0	0
0	0,47	0	0	0	0,22	0	0
0	0,47	0	0	0	0,22	0	0
0	0,95	0	0	0	0,91	0	0

W				Q^2	D1^2	D2^2	D3^2
Q	*D1*	*D2*	*D3*				
0	0,47	0	0	0	0,22	0	0
0	0	0,47	0	0	0	0,22	0
0,47	0	0,95	0	0,22	0	0,91	0
0	0	0,47	0	0	0	0,22	0
0	0	0,47	0	0	0	0,22	0
0	0	0,47	0	0	0	0,22	0
0	0	0,47	0	0	0	0,22	0
0	0	0,47	0	0	0	0,22	0
0	0	0,47	0	0	0	0,22	0
0	0	0,47	0	0	0	0,22	0
0	0	0	0,47	0	0	0	0,22
0	0	0	0,477	0	0	0	0,22
			Jumlah	0,28		Sqrt (Di)	
			sqrt	0,53	1,80	1,68	0,69

Secara umum perhitungan koefisien cosinus dalam information retrieval terdiri dari tiga tahapan, tahap yang pertama disebut dengan representasi vector dalam ruang vektor dengan setiap dimensi mewakili kata atau konsep tertentu dan nilai vektor adalah nilai TF-IDF dari kata, tahap kedua perhitungan koefisien kosinus menggunakan rumus koefisien kosinus antara vektor kueri (QQ) dan vektor dokumen (DD) dan tahap yang terakhir perangkingan dokumen yang diurutkan berdasarkan nilai koefisien kosinus tersebut dalam urutan menurun. Semakin tinggi nilai koefisien kosinus, semakin relevan dokumen dengan kueri. Tabel 8 merupakan tahapan hasil perhitungan koefisien cosinus.

Tabel 8 Perhitungan Koefisien Cosinus

Q^2	D1^2	D2^2	D3^2	Q^2* D1^2	Q^2* D2^2	Q^2* D3^2
0,03	0,03	0	0,03	0,00	0	0,00
0	0,22	0	0	0	0	0
0,03	0,03	0,12	0	0,00	0,00	0
0	0,91	0	0	0	0	0
0	0,22	0	0	0	0	0
0	0,22	0	0	0	0	0
0	0,22	0	0	0	0	0
0	0,22	0	0	0	0	0
0	0,91	0	0	0	0	0
0	0,22	0	0	0	0	0
0	0	0,22	0	0	0	0
0,22	0	0,91	0	0	0,20	0
0	0	0,22	0	0	0	0
0	0	0,22	0	0	0	0
0	0	0,22	0	0	0	0
0	0	0,22	0	0	0	0
0	0	0,22	0	0	0	0

Q^2	D1^2	D2^2	D3^2	Q^2* D1^2	Q^2* D2^2	Q^2* D3^2
0	0	0,22	0	0	0	0
0	0	0,22	0	0	0	0
0	0	0	0,22	0	0	0
0	0	0	0,2	0	0	0
0,28		Sqrt (Di)		0,001	0,21	0,00
0,53	1,80	1,68	0,69			
	Rank Score			0,00	0,23	0,00

Pengujian akurasi dalam Information Retrieval (IR) dilakukan untuk mengevaluasi sejauh mana sistem temu kembali informasi berhasil dalam menemukan dan mengurutkan dokumen yang relevan untuk kueri pengguna. Pengujian dilakukan dengan menentukan topik pencarian sesuai kebutuhan, dalam penilitian ini dilakukan 15 pencarian topik dalam terjemaan al-quran yang bervariasi. Detail hasil pengujian terdapat pada table 9.

Tabel 9 Pengujian Akurasi

No	Topik	Hasil Cosine Similarity	TP	FN
1	Kisah Penyembelihan Sapi Betina	67,68,69,70,73	1	
2	Hukum Perceraian, Pernikahan, dan Menyusui	233, 229,231,235,102,232,237,236, 241,230,225	1	
3	Hukum Riba	233, 229, 231, 235, 102, 232, 237, 236, 241, 230, 225	1	
4	Menasakkan Susatu Ayat adalah Urusan Allah	187,61,221,242,219,252,266,129,151,41, 39,99,106,145		1
5	Anjuran Membelanjakan Harta	188,195,265,273,180,177,279,245,261,264, 215,254,262,155, 274	1	
6	Sekitar Pemindahan Kiblat	145,142,240,144,148		1
7	Makanan Halal dan Haram	194,144,187,35,229,169,173,230,198,191, 61,259,275,233,60,184,172,174	1	
8	Wasiat	180,132,181,182	1	
9	Hukum Perang	177,217,216,251,244	1	
10	Haji	197,128,158,189		1
11	Pokok-Pokok Kebajikan	177	1	
12	Kesaksian dalam Mu'amalah	283	1	
13	Kewajiban Berjihad	243,245,249		1
15	Pahala Orang Beriman	62	1	

Ringkasan hasil pengujian ditemukan jumlah True Positif(TP) adalah 11 dana jumlah False Negatif(FN) sejumlah 4. Pengujian akurasi dengan menghitung nilai presisinya dengan mengukur proporsi dokumen yang relevan yang ditemukan oleh sistem dari total dokumen yang diambil/ditemukan. Hasil perhitungan recall didapatkan nilai 73,33%, detail perhitungannya sebagai berikut. Nilai lebih rendah dibandingkan penelitian sebelumnya, ada kemungkinan tahapan case folding dan normalisasi yang tidak dilakukan pada proses pre-processing dan proses feature selection penelitian ini berpengaruh terhadap akurasi.

$$Recall = \frac{\text{TP}}{TP+FN}\ X\ 100 \frac{11}{11+4} X100 =\ 73{,}33\%$$

4. Kesimpulan

Setelah dilakukan analisa, perancangan, implementasi, pelatihan algoritma, dan pengujian pada aplikasi android pencarian ayat dengan menggunakan algoritma *Cosine Similarity* maka dapat disimpulkan bahwa penelitian ini dapat mempercepat proses pencarian ayat pada Al-Qur'an walaupun dari nilai akurasinya masih kurang, yakni dengan nilai recall 73.33%, nilai akurasinya masih dibawah penelitian yang dilakukan oleh (Suzanti , Husni, Awaldhia , & Syarief, 2023) yang menghasil nilai recall 84.83% atau yang dilakukan oleh (Rakholia & Saini, 2017) dengan hasil recall 86%, kedua penelitian tersebut melakukan sedikit modifikasi pada model Cosine Similarity sehingga menghasilkan nilai recall yang lebih baik jika dibandingkan penelitian ini. Berdasarkan penelitian yang telah dilakukan, sistem yang dibuat masih memiliki beberapa kekurangan. Saran yang dapat dijadikan acuan untuk penelitian lebih lanjut adalah melakukan modifikasi model *Cosine Similirity* adalah hanya mencari kata yang diinginkan (bukan arti atau pemahaan lainnya) atau menambahkan fiture selection. Saran lainnya bisa menggunakan alternatif metode yakni *Generalized Vector Space Model (GVSM).*

5. Referensi

Adiyanto, A. T., & Handayani, D. (2022). Information Retrieval Sistem Kearsipan Pencarian Dokumen Di Dinas Pemberdayaan Perempuan Dan Perlindungan Anak Kota Semarang Menggunakan Metode Vector Space Model. *Jurnal Mahajana Informasi, 7*(1).

Agustian, S., & Sucipto, A. (2021). Source Retrieval pada Deteksi Plagiarisme Berdasarkan Biword Fingerprint dengan Model Ruang Vektor. *Seminar Nasional Teknologi Informasi Komunikasi dan Industri(SNTIKI 2020).* Pekanbaru.

Anand, A., Sharma, U., & Kumar, D. (2020). Information Retrieval in Computing Model. *2019 International Conference on Intelligent Computing and Control Systems (ICCS).* Madurai. doi:10.1109/ICCS45141.2019.9065562

Calafato, R. (2020). Learning Arabic in Scandinavia: Motivation, metacognition, and autonomy. *Lingua, 246*.

Chanda, S., & Pal, S. (2023). The Effect of Stopword Removal on Information Retrievalfor Code- Mixed Data Obtained Via Social Media. *Springer Nature Computer Science, 4*. doi:https://doi.org/10.1007/s42979-023-01942-7

Davison, S., Avgil, D., Li, Y., & Yang, S. (2022). A Semi-automatic Indexing Pipeline for Medical Document Retrieval in Resource-constrained Settings. *Twenty-eighth Americas Conference on Information Systems.* Minneapolis.

Nugroho, S. A., Bachtiar, F. A., & Wihandika, R. C. (2022). ASPECT EXTRACTION IN E-COMMERCE USING LATENT DIRICHLET ALLOCATION (LDA) WITH TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF). *Jurnal Ilmiah Kursor, 11*(2).

Rakholia, R., & Saini, J. R. (2017). Information Retrieval for Gujarati Language Using Cosine Similarity Based Vector Space Model. *Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications.*

Rasyid, I., Yudianto, M. R., Maimunah, & Purnomo, T. A. (2023). Electronic Product Recommendation System Using the Cosine Similarity Algorithm and VGG-16. *Sinkron: Jurnal & Penelitian Teknik Informatika, 8*(4).

Rinandyaswara, R., Sari, Y. A., & Furqon, M. T. (2022). Pembentukan Daftar Stopword Menggunakan Term Based Random Sampling Pada Analisis Sentimen Dengan Metode Naïve Bayes (Studi Kasus: Kuliah Daring Di Masa Pandemi). *Jurnal Teknologi Informasi dan Ilmu Komputer, 9*(4).

Sa'diyah, H., & Abdurahman, M. (2021). Pembelajaran Bahasa Arab di Indonesia: PenelitianTerhadap Motivasi Belajar Bahasa Asing. *Lisanan Arabiya : Jurnal Pendidikan Bahasa Arab, 5*(1).

Suzanti , I. O., Husni, H., Awaldhia , B. S., & Syarief, M. (2023). Design and Implementation of Tourism News Information Retrieval System using Modified Cosine Similarity. *Technium: Romanian Journal of Applied Sciences and Technology, 16*(1).

Suzanti, I. O., & Jauhari, A. (2021). COMPARISON OF STEMMING AND SIMILARITY ALGORITHMS IN INDONESIAN TRANSLATED AL-QUR'AN TEXT SEARCH. *Jurnal Ilmiah Kursor, 11*(2).

Wahyudi, Akbar, F., & Suryamen, H. (2020). Perancangan Sistem Pemeriksaan Ujian Essay Menggunakan Sistem TemuKembali Informasi Model Ruang Vektor. *Jurnal Nasional Teknologi dan Sistem Informasi, 6*(3), 140.

Yulianto, B., Budiharto, W., & Kartowisastro, I. H. (2017). The Performance of Boolean Retrieval and Vector Space Model in Textual Information Retrieval. *COMMIT (COMMUNICATION AND INFORMATION TECHNOLOGY) JOURNAL, 11*(1).