IF185401 TESIS – SIDANG AKHIR

PENGUKURAN KEMIRIPAN BERBASIS LEKSIKAL DAN SEMANTIK UNTUK PERANGKINGAN DOKUMEN BERBAHASA ARAB

SYADZA ANGGRAINI
NRP. 05111850010031

Dosen Pembimbing
Dr. Diana Purwitasari, S.Kom., M.Sc.

Departemen Teknik Informatika
Fakultas Teknologi Elektro dan Informatika Cerdas
Institut Teknologi Sepuluh Nopember
2022

[Halaman ini sengaja dikosongkan]

LEMBAR PENGESAHAN TESIS

Tesis disusun untuk memenuhi salah satu syarat memperoleh gelar
Magister Komputer (M.Kom.)
di
Institut Teknologi Sepuluh Nopember

Oleh:
Syadza Anggraini
05111850010031

Tanggal Ujian: 28 Juni 2022
Periode Wisuda: September 2022

Disetujui oleh:

Pembimbing

1. Dr. Diana Purwitasari, S.Kom., M.Sc.
   NIP. 19780410 200312 2 001 .....

Penguji

1. Dr. Eng. Chastine Fatichah, S.Kom., M.Kom.
   NIP. 19751220 200112 2 002 .....
2. Dr. Ahmad Saikhu, S.Si., M.T.
   NIP. 19710718 200604 1 001 .....
3. Ary Mazharuddin Shiddiqi, S.Kom., M.Comp.Sc., Ph.D.
   NIP. 19810620 200501 1 003 .....

Kepala Departemen Teknik Informatika
Fakultas Teknologi Elektro dan Informatika Cerdas

Dr. Eng. Chastine Fatichah, S.Kom., M.Kom.
NIP. 19751220 200112 2 002

[Halaman ini sengaja dikosongkan]

PERNYATAAN ORISINALITAS

Yang bertanda tangan di bawah ini:

Nama mahasiswa/NRP	: Syadza Anggraini/05111850010031
Departemen	: Teknik Informatika
Dosen Pembimbing/NIP	: Dr. Diana Purwitasari, S.Kom., M.Sc. / 19780410 200312 2 001

dengan ini menyatakan bahwa Tugas Akhir dengan judul "Pengukuran Kemiripan Berbasis Leksikal dan Semantik untuk Perangkingan Dokumen Berbahasa Arab" adalah hasil karya sendiri, bersifat orisinal, dan ditulis dengan mengikuti kaidah penulisan ilmiah.

Bilamana di kemudian hari ditemukan ketidaksesuaian dengan pernyataan ini, maka saya bersedia menerima sanksi sesuai dengan ketentuan yang berlaku di Institut Teknologi Sepuluh Nopember.

Surabaya, 27 Juli 2022

Mengetahui

Mahasiswa,

Syadza Anggraini

NRP. 05111850010031

[Halaman ini sengaja dikosongkan]

PENGUKURAN KEMIRIPAN BERBASIS LEKSIKAL DAN SEMANTIK UNTUK PERANGKINGAN DOKUMEN BERBAHASA ARAB

Nama	: Syadza Anggraini
NRP	: 05111850010031
Pembimbing	: Dr. Diana Purwitasari, S.Kom., M.Sc.

ABSTRAK

Perangkingan dokumen merupakan salah satu topik dalam sistem temu kembali informasi. Dalam menghasilkan dokumen yang relevan, pengukuran kemiripan antara *query* dan dokumen menjadi faktor penting terhadap dokumen yang dirangking. Pengukuran kemiripan dapat dihitung berdasarkan bobot kata antara *query* dan dokumen.

Namun, pengukuran kemiripan menggunakan bobot kata dimungkinkan adanya lafal kata yang berbeda tetapi memiliki makna kata yang sama. Selain itu, hasil dokumen pencarian suatu teks berbahasa Arab dipengaruhi oleh beragamnya penguasaan pengguna dalam memahami bahasa Arab.

Oleh sebab itu, penelitian ini mengembangkan pengukuran kemiripan secara leksikal untuk mengatasi lafal kata dan pengukuran kemiripan secara semantik untuk mengatasi makna kata. Penggabungan perhitungan kemiripan leksikal dan semantik dihitung berdasarkan bobot kata (leksikal) dan digabungkan dengan *word embedding* (semantik).

Berdasarkan hasil uji coba pada 2900 kitab berbahasa Arab, metode usulan memiliki rata-rata *recall*, *precision*, dan *f-measure* tertinggi daripada metode lainnya sebesar 72.42%, 65.83%, 64.2% pada *all query*, kemudian 73.2%, 63.15%, 63.1% pada *short query*, serta 71.31%, 69.86%, 65.7% pada *long query*. *Short query* adalah *query* dengan frekuensi sebanyak 1-2 kata sedangkan *long query* adalah *query* dengan frekuensi lebih dari 2 kata.

Kata kunci: kemiripan leksikal, kemiripan semantik, perangkingan dokumen, pengukuran kemiripan

[Halaman ini sengaja dikosongkan]

SIMILARITY MEASUREMENT BASED ON LEXICAL AND SEMANTIC FOR ARABIC DOCUMENT RANKING

Student Name	: Syadza Anggraini
NRP	: 05111850010031
Supervisor	: Dr. Diana Purwitasari, S.Kom., M.Sc.

ABSTRACT

Document ranking is one of the topics in the information retrieval. In producing relevant documents, measuring the similarity between the query and the document becomes an important factor for the ranked documents. The similarity measurement can be calculated based on the term weight between the query and the document.

However, the calculation of similarity based on the term weights has the possibility of differences in the calculation of weights on terms that are written differently but have the same word meaning. In addition, the results of searching documents for an Arabic text are influenced by the variety of user mastery in understanding Arabic.

Therefore, a lexical similarity measurement is developed to overcome word pronunciation and a semantic similarity measurement is developed to overcome word meaning. The combination of lexical and semantic similarity calculations is calculated based on term weights (lexical) and combined with word embedding (semantic).

Based on the results of evaluation on 2900 Arabic books, this research method has the highest average recall, precision, and f-measure compared to other methods of 72.42%, 65.83%, 64.2% for all queries, then 73.2%, 63.15%, 63.1% for short queries, and also 71.31%, 69.86%, 65.7% on long queries. A short query is a query with a word frequency of 1-2 words, while a long query is a query with frequency of more than 2 words.

Keywords: lexical similarity, semantic similarity, document ranking, similarity measurement

[Halaman ini sengaja dikosongkan]

KATA PENGANTAR

Assalamu'alaikum Wr. Wb. Puji syukur ke hadirat Allah SWT atas limpahan rahmat dan karunia-Nya lah penulis dapat menyelesaikan tesis yang berjudul "Pengukuran Kemiripan Berbasis Leksikal dan Semantik untuk Perangkingan Dokumen Berbahasa Arab". Penulis menyadari bahwa dapat terselesainya tesis ini dengan baik tidak lepas dari dukungan moral maupun material dari berbagai pihak. Oleh karena itu, penulis ingin menyampaikan ucapan terima kasih yang mendalam kepada berbagai pihak atas dukungan yang diberikan di antaranya:

1. Kedua orang tua, Umi dan Abah yang senantiasa memberikan dukungan baik berupa do'a, semangat, motivasi, dan nasehat untuk kesuksesan penulis.
2. Ibu Dr. Diana Purwitasari, S.Kom., M.Sc. selaku dosen pembimbing yang telah membimbing dan memotivasi penulis dalam menyelesaikan penulisan tesis.
3. Bapak Prof. Dr. Agus Zainal Arifin, S.Kom., M.Kom. yang telah memberikan bimbingan kepada penulis pada masa perkuliahan dan dalam menyelesaikan penulisan tesis.
4. Ibu Dr. Eng. Chastine Fatichah, S.Kom., M.Kom., Bapak Dr. Ahmad Saikhu, S.Si., M.T., dan Bapak Ary Mazharuddin Shiddiqi, S.Kom., M.Comp.Sc., Ph.D. selaku dosen penguji yang telah memberikan kritik dan masukan atas penulisan tesis.
5. Teman-teman S2 Teknik Informatika Angkatan 2018 yang telah memberikan bantuan selama perkuliahan.
6. Pihak-pihak lain yang tidak dapat disebutkan satu persatu, terima kasih atas segala bantuan yang diberikan sehingga penulis dapat menyelesaikan pendidikan magister.

Penulis menyadari bahwa masih terdapat banyak kekurangan dalam penulisan tesis ini. Oleh sebab itu, penulis membutuhkan kritik dan saran yang membangun dari semua pihak agar dapat berkembang lebih baik lagi di masa

depan. Semoga dengan adanya penelitian tesis ini dapat memberikan ilmu yang bermanfaat bagi perkembangan ilmu pengetahuan khususnya di bidang sistem temu kembali informasi. Akhir kata penulis ingin mengucapkan terima kasih atas segala dukungan yang diberikan dan mohon maaf yang sebesar-besarnya atas segala kekurangan. *Wassalamu'alaikum Wr. Wb.*

Surabaya, Juli 2022

Syadza Anggraini

DAFTAR ISI

DAFTAR GAMBAR

[Halaman ini sengaja dikosongkan]

DAFTAR TABEL

BAB 1
PENDAHULUAN

Bab ini berisi penjelasan mengenai latar belakang, perumusan masalah, tujuan penelitian, manfaat penelitian, kontribusi penelitian, dan batasan masalah.

1.1 Latar Belakang

Informasi teks berbahasa Arab semakin diminati oleh banyak pengguna karena telah tersedia dalam bentuk *digital*, karena memudahkan penggunanya dalam melakukan pencarian informasi secara cepat (Bounhas, 2019). Sistem temu kembali informasi menjadi sarana bagi pengguna dalam menemukan informasi teks berbahasa Arab tersebut (Al-barhamtoshy and Jambi, 2021). Pencarian informasi dilakukan dengan memasukkan kata kunci (*query*) ke dalam suatu sistem agar menampilkan informasi (dokumen) yang sesuai. Kesesuaian informasi terhadap *query* yang dimasukkan oleh pengguna bergantung pada pengukuran kemiripan yang digunakan oleh suatu sistem. Hasil dokumen yang tidak sesuai atau relevan dapat disebabkan oleh pengukuran kemiripan yang kurang akurat.

Salah satu metode pengukuran kemiripan antara *query* dan dokumen dalam perangkingan dokumen adalah kemiripan leksikal. Penelitian yang dilakukan oleh (Hajeer *et al.*, 2017) menggunakan dokumen berupa halaman *website* sebagai dataset perangkingan pada mesin pencarian bahasa Arab. Perangkingan tersebut menggunakan algoritma yang disebut EHURA (*new Hybrid Usage-Based Ranking*). Salah satu task EHURA adalah *Click-through* yang merupakan *click history* yang dijadikan sebagai bobot kuantitatif tiap halaman yang di klik oleh *user*. Dokumen dan *query* dilakukan proses tokenisasi, *data cleaning*, *stemming*, *indexing* dan *ranking*. Pada proses *indexing* dan *ranking* peneliti mencocokkan indeks *term* yang ada pada *query* dan dokumen dengan menggunakan perhitungan *Cosine Similarity*. Perankingan dokumen oleh Khadijah, dkk (Holle, Arifin and Purwitasari, 2015) mengusulkan metode pembobotan (*term weighting*) *Invers Preference Frequency with Value* (IPFα)

untuk melakukan perankingan dokumen fiqih berbahasa Arab dari segi aspek preferensi (keutamaan). Perangkingan dokumen dimulai dari tahap *preprocessing*, pembobotan *Term Frequency* (TF), *Invers Document Frequency* (IDF), dan *Invers Book Frequency* (IBF) serta *Invers Preference Frequency* (IPF) yang menjadi metode usulan. Hasil dari tiap pembobotan (TF, IDF, IBF dan IPF) dikombinasikan dengan mengalikan semua hasilnya. Kemudian proses selanjutnya dilakukan perhitungan similaritas antara *query* dan dokumen menggunakan *Cosine Similarity*. Selain itu, perangkingan dokumen yang dilakukan oleh (Fauzi, Arifin and Yuniarti, 2017) menggunakan *Term Frequency Inverse Document Frequency* (TF.IDF) yang digabung dengan *Inverse Book Frequency* (IBF). IBF merupakan pembobotan yang dihitung berdasarkan *book* atau buku pada dokumen berbahasa Arab yang digunakan. Sehingga untuk mendapatkan dokumen yang relevan, dokumen yang telah dilakukan pembobotan akan diukur kemiripannya menggunakan cosine *similarity* terhadap *query*. Serta, penelitian lain yang membahas perangkingan dokumen fiqih berbahasa Arab (Sholikah *et al.*, 2017) menggunakan *Positif Impact Factor Query* (PIFQ) sebagai metode usulan. Metode yang diusulkan merupakan gabungan metode antara *Term Frequency Inverse Document Frequency* (TF.IDF) dan *Positif Impact Factor Query* (PIFQ). PIFQ merupakan kemunculan *key term* pada setiap kategori (mazhab) untuk meningkatkan relevansi antara *query* dan dokumen hasil pencarian. Selain itu, PIFQ berguna untuk meningkatkan similaritas antara vektor *query* dan vektor dokumen dengan menaikkan bobot dokumen menggunakan *key term*. Untuk mendapatkan nilai akhir dari tiap dokumen, masing-masing hasil pembobotan diintegrasikan dengan menggunakan suatu rumus, dimana rumus tersebut mengalikan hasil dari TF.IDF dan hasil dari PIFQ.

Selain kemiripan leksikal, penelitian-penelitian berikut menggunakan kemiripan semantik dalam pengukuran kemiripan antar dua objek. Penelitian yang dilakukan oleh (Almarwani and Diab, 2017) menggunakan *Weighted Textual Matrix Factorization* (WTMF) sebagai kemiripan semantik pada sistem *Question Answering*. WTMF difokuskan pada *missing words* (kata-kata hilang) serta *unlabeled data* (data tidak berlabel) pada dokumen teks-teks pendek yang digunakan dan tidak difokuskan pada pemrosesan makna kata dari suatu kata.

Penelitian oleh (Abdi, Mariyam and Aliguliyev, 2018) melakukan perangkingan terhadap kalimat yang akan dijadikan sebagai bahan peringkasan dokumen *review*. Perangkingan diproses menggunakan kemiripan semantik yang disebut *semantic similarity measurement* (SSM). SSM menggunakan *WordNet* guna mendapatkan kumpulan *synset* (kumpulan sinonim kata) dari suatu *term*. Pengukuran kemiripan semantik antar term menggunakan *synset- synset WordNet,* bahwa yang memiliki kemiripan adalah antar *synset*-nya bukan untuk mengukur antar *term*-nya. Jika terdapat *synset-synset* antar *term* yang saling beririsan (*overlap*) maka antar *term* tersebut memiliki kemiripan. Namun jika tidak terdapat *synset* yang saling *overlap* maka tidak ada kemiripan antar term tersebut. Selain penelitian-penelitian yang telah dijelaskan sebelumnya di mana menggunakan kemiripan semantik sebagai pengukuran, adapun cara lain yaitu menggunakan *word embedding* (Mahdaouy *et al.*, 2018). *Word embedding* atau *word vector representation* merupakan pengukuran kemiripan semantik yang mengukur antara dua objek kata dengan memperhatikan makna kata (Moatez *et al.*, 2017). Makna kata didapatkan melalui kata-kata terdekatnya (Suleiman and Awajan, 2018; Othman, Faiz and Smaili, 2019).

Pembobotan kata yang digunakan pada perhitungan kemiripan secara leksikal memiliki kemungkinan adanya perbedaan perhitungan bobot pada kata yang secara lafal berbeda namun memiliki makna kata yang sama. Selain itu, penguasaan pengguna terhadap bahasa Arab yang beragam dapat mempengaruhi hasil pencarian dokumen teks berbahasa Arab. Oleh karena itu dibutuhkan pertimbangan selain hanya menggunakan pembobotan (kemiripan leksikal) tetapi juga menggunakan kemiripan semantik yang diproses berdasarkan kemiripan antar kata menggunakan *word embedding*.

Penelitian ini mengembangkan metode penggabungan kemiripan leksikal dan kemiripan semantik guna merangking dokumen berbahasa Arab. Kemiripan semantik yang digunakan dalam penelitian adalah *word embedding*, di mana *word embedding* mampu menangkap makna kata yang tidak dapat dilakukan oleh kemiripan leksikal. Sehingga, metode ini mampu menampilkan hasil pencarian dokumen yang relevan terhadap *query* yang dimasukkan oleh pengguna daripada yang hanya menggunakan kemiripan leksikal.

1.2 Rumusan Masalah

Rumusan masalah pada penelitian ini adalah sebagai berikut:

1. Bagaimana cara melakukan perhitungan kemiripan dokumen berbahasa Arab secara leksikal?
2. Bagaimana cara melakukan perhitungan kemiripan dokumen berbahasa Arab secara semantik?
3. Bagaimana cara melakukan perhitungan kemiripan gabungan antara leksikal dan semantik pada dokumen berbahasa Arab?

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah mengembangkan pengukuran kemiripan antara *query* dan dokumen berbasis leksikal dan semantik untuk perangkingan dokumen berbahasa Arab.

1.4 Manfaat Penelitian

Penelitian ini diharapkan mampu membantu pengguna sebagai perangkat dalam mencari dan menampilkan informasi relevan terkait kajian Islam sesuai dengan *query* yang dimasukkan oleh pengguna.

1.5 Kontribusi Penelitian

Kontribusi penelitian ini adalah melakukan perhitungan kemiripan dengan menggabungkan kemiripan leksikal dan semantik menggunakan *word embedding* pada perangkingan dokumen berbahasa Arab.

1.6 Batasan Masalah

Penelitian ini memiliki beberapa batasan antara lain:

1. Penelitian ini menggunakan dataset berbahasa Arab.
2. Dataset Maktabah Syamilah sebanyak 2900 kitab.
3. Topik-topik *dataset* beragam dan tidak digolongkan ke dalam topik tertentu.
4. Topik pada *query* dibuat secara acak.

BAB 2
KAJIAN PUSTAKA DAN DASAR TEORI

Bab ini akan menjelaskan tentang beberapa referensi yang berkaitan dengan penelitian. Referensi tersebut meliputi kajian pustaka dan dasar teori mengenai kemiripan leksikal, kemiripan semantik hingga *word embedding*.

2.1 Kajian Pustaka

2.1.1 Kemiripan Leksikal

Kemiripan secara leksikal diukur menggunakan konsep *bag-of-words* atau kumpulan kata. Pengukuran leksikal dianalisa dengan memecah kalimat menjadi sekumpulan kata, yang selanjutnya menghapus kata-kata tidak penting (*stopwords removal*). *Output* yang dihasilkan dari analisa leksikal adalah sekumpulan kata yang telah dilakukan *stemming*, dimana *stemming* merupakan perubahan suatu kata menjadi kata dasarnya.

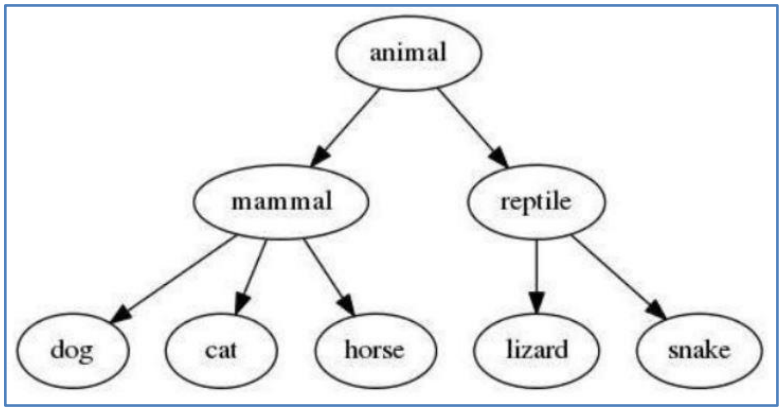
Kemiripan antar dua kata dapat dikatakan mirip secara leksikal apabila masing-masing dari kata tersebut memiliki penyusun karakter dengan urutan yang sama (Kaur and Maini, 2018). Kemiripan leksikal atau *string-based similarity*, terbagi ke dalam dua kategori yaitu pengukuran berdasarkan karakter (*character-based*) dan pengukuran berdasarkan *term* (*term-based*). Pengukuran berdasarkan karakter terdiri dari *Longest Common SubString* (LCS), *Damerau-Levenshtein*, *Jaro*, *Jaro-Winkler*, *Needleman-Wunsch*, *Smith-Waterman*, dan *N-gram*. Sedangkan pengukuran berdasarkan *term* terdiri dari *Block Distance*, *Cosine Similarity*, *Dice's Coefficient*, *Euclidean Distance*, *Jaccard Similarity*, *Matching Coefficient*, dan *Overlap Coefficient*. LCS merupakan salah satu contoh dari character based similarity. Algoritma *Longest Common SubString* (LCS) mempertimbangkan panjang maksimum rantai karakter yang muncul di antara kedua string. Contoh lain dari *character based similarity* adalah *Levenshtein edit distance*. *Levenshtein edit distance* menentukan jarak antara kedua *string* dengan menghitung jumlah operasi minimum seperti: *insertion*, *deletion* atau *substitution* dari sebuah karakter tunggal atau perpindahan antara dua karakter. Perhitungan

jumlah operasi minimum tersebut diperlukan untuk mengubah satu *string* menjadi *string* lainnya. Pengukuran *term based similarity* memiliki sejumlah pengukuran kemiripan di antaranya *Cosine Similarity*. *Cosine Similarity* mengukur kemiripan antara dua vektor yang mengukur sudut *cosine* di antara keduanya. Selain itu, *Euclidean Distance* atau L2 *distance* yang merupakan akar kuadrat dari jumlah selisih kuadrat antara elemen-elemen yang bersesuaian di antara dua vektor.

2.1.2 Kemiripan Semantik

Kemiripan secara semantik digunakan untuk mengukur makna suatu kalimat. Suatu kata yang secara penyusun karakternya berbeda dari kata lainnya bisa jadi memiliki makna kata yang sama (Abriania and Yaqin, 2019). Sebagai contoh, kata “pergi” dan “berangkat” masing-masing memiliki penyusun karakter yang berbeda, namun kedua kata tersebut memiliki makna yang sama.

Kemiripan secara semantik dapat disebut sebagai *corpus-based similarity* karena kemiripan antar kata-nya diukur berdasarkan informasi yang didapatkan dari suatu korpus besar. Selain *corpus-based similarity*, kemiripan semantik juga dapat disebut sebagai *knowledge-based similarity*. *Knowledge-based similarity* adalah kemiripan semantik yang mengukur tingkat kemiripan antar kata menggunakan informasi jaringan semantik (Sunilkumar and Shaji, 2019). *Knowledge based similarity* mengukur tingkat kesamaan antara kata-kata dengan menggunakan informasi yang berasal dari jaringan semantik dalam satu korpus. Salah satu contoh jaringan semantik ialah *WordNet*. *WordNet* merupakan suatu database yang besar di mana memuat kata-kata bahasa Inggris seprti kata benda, kata kerja, kata sifat, dan kata keterangan. Kata-kata tersebut digolongkan menjadi sekumpulan sinonim atau yang disebut dengan *synsets*, di mana setiap *synset* berbeda satu sama lain. Sedangkan *corpus based similarity* menentukan kemiripan semantik antar kata berdasarkan informasi yang dikumpulkan dari korpus yang sangat besar. Salah satu contohnya adalah *Pointwise Mutual Information - Information Retrieval* (PMI-IR). PMI-IR adalah sebuah metode menghitung kemiripan antar pasang kata, di mana semakin sering dua pasang kata muncul secara bersamaan dalam sebuat *website*, maka semakin tinggi pula skor PMI-IR-nya.



Gambar 2.1 Representasi *Knowledge Based* (Sunilkumar and Shaji, 2019)

Gambar 2.1 menunjukkan representasi dari *knowledge based similarity*. Selain kemiripan semantik yang diukur dari tingkat kemiripan antar kata menggunakan informasi jaringan semantik, *knowledge base similarity* juga mengukur tingkat kemiripan semantik menggunakan *knowledge graphs*. *Knowledge graphs* adalah graf berarah, G = (V, E, τ), di mana V merupakan sekumpulan node, E adalah sekumpulan *edge* yang saling berhubungan dan τ adalah fungsi dari V × V → E yang menghubungkan antar node, di mana hubungan node-node tersebut menjadi tiga rangkap seperti pada Gambar 2.1.

2.2 Dasar Teori

2.2.1 Bahasa Arab

Bahasa Arab merupakan bahasa yang termasuk kedalam rumpun bahasa Semitik yang dituturkan oleh lebih dari 300 juta orang dan menjadikannya sebagai bahasa utama oleh masyarakat yang tinggal di wilayah Timur Tengah dan Afrika. Terdapat dua tipe bahasa Arab yaitu: bahasa Arab fasih dan Bahasa Arab yang dipakai sehari-hari. Bahasa Arab fasih terdiri dari bahasa Arab klasik dan bahasa Arab modern, dimana bahasa Arab klasik biasa ditemui pada syair-syair Arab serta Al-Qur'an diturunkan dari bahasa Arab klasik tersebut.

Penulisan bahasa Arab berbeda dari abjad pada umumnya yang biasa ditulis dari kiri ke kanan. Sedangkan bahasa Arab ditulis dari kanan ke kiri. Bahasa Arab itu sendiri memiliki dua jenis kata yang dibedakan menjadi *mudzakkar* (laki-laki) dan *muannats* (perempuan). Kemudian, berdasarkan jumlahnya bahasa Arab dibagi kedalam tiga jenis yaitu: *mufrad* (tunggal),

mutsanna (dua) dan *jama'* (lebih dari dua). Berdasarkan unsur pembentuknya, dibedakan menjadi kata benda (*isim*), kata kerja (*fi'il*), dan huruf (Hamsiati, 2019).

2.2.2 *Preprocessing*

Preprocessing atau praproses merupakan salah satu tahapan penting sebelum melakukan pengolahan pada dokumen teks. Tahapan preprocessing diantaranya meliputi tokenisasi terhadap string serta normalisasi, menghapus kata-kata yang tidak penting (*stopwords removal*), dan *stemming* (Alhanjouri, 2017). Tokenisasi merupakan proses memecah suatu dokumen baik dalam bentuk paragraf atau kalimat menjadi sekumpulan kata-kata yang terpisah (*token*). Selain itu, proses normalisasi melakukan penghapusan terhadap karakter-karakter yang tidak diperlukan seperti tanda baca dan angka. Tanda baca tersebut seperti: tanda baca titik (.), tanda baca koma (,), dan tanda baca lainnya. Setelah dilakukan tokenisasi, selanjutnya dokumen dilakukan proses *stopwords removal*. *Stopwords removal* adalah sekumpulan kata (*bag of words*) yang dimana kata-kata tersebut merupakan kata-kata yang tidak penting. Kata-kata tidak penting tersebut dapat mempengaruhi proses pembobotan pengolahan teks dikarenakan frekuensi kemunculannya pada sekumpulan dokumen. Semakin sering suatu kata tersebut muncul, maka kata tersebut dianggap tidak penting dan memiliki bobot yang rendah.

Dokumen-dokumen yang telah melalui proses *stopwords removal* merupakan dokumen yang hanya berisi kumpulan kata-kata penting yang selanjutnya dilakukan proses *stemming*. *Stemming* merupakan proses mengubah suatu kata menjadi kata dasar (*root*) dari kata tersebut. Berikut dibawah ini merupakan contoh dari serangkaian tahapan *preprocessing*:

1. Tokenisasi

   Dokumen asli:

   يجوز تقديم زكاة الفطر قبل العيد بيوم أو يومين

   Kalimat tersebut di atas berarti "*Dibolehkan mengeluarkan zakat fitrah satu atau dua hari sebelum Idul Fitri*". Setelah dilakukan tokenisasi maka akan menjadi sekumpulan kata-kata yang dapat dilihat pada Tabel 2.1.

Tabel 2.1 Kata-Kata Hasil Tokenisasi

Arab	Latin	Terjemahan
يجوز	yajuz	dibolehkan
تقديم	taqdim	membayar
زكاة	zaka	zakat
الفطر	alfitr	fitrah
قبل	qabl	sebelum
العيد	aleid	idul fitri
بيوم	byum	satu
أو	'aw	atau
يومين	yawmayn	dua hari

Berdasarkan hasil dari contoh diatas, proses normalisasi juga dapat diartikan sebagai penghapusan *diacritic* atau harokat dan menjadikan kata-kata tersebut kedalam bentuk dasar huruf hijaiyah.

2. *Stopwords Removal*

Hasil dokumen yang akan dilakukan *stopwords removal* dapat dilihat pada Tabel 2.3:

Tabel 2.2 Kata-Kata yang akan Dilakukan *Stopwords Removal*

Arab	Latin	Terjemahan
يجوز	yajuz	dibolehkan
تقديم	taqdim	membayar
زكاة	zaka	zakat
الفطر	alfitr	fitrah
قبل	**qabl**	**sebelum**
العيد	aleid	idul fitri
بيوم	byum	satu
أو	**'aw**	**atau**
يومين	yawmayn	dua hari

Berikut Tabel 2.3 merupakan hasil setelah dilakukan *stopwords removal*. Kata-kata tersebut setelah dilakukan penghapusan kata tidak penting akan

berkurang jumlahnya. Kata-kata yang dihapus tersebut yang bercetak tebal pada tabel.

Tabel 2.3 Kata-Kata Hasil *Stopwords Removal*

Arab	Latin	Terjemahan
يجوز	yajuz	dibolehkan
تقديم	taqdim	membayar
زكاة	zaka	zakat
الفطر	alfitr	fitrah
العيد	aleid	idul fitri
بيوم	byum	satu
يومين	yawmayn	dua hari

Berdasarkan hasil *stopwords removal* (penghapusan kata-kata tidak penting) diatas, terdapat kata "قبل" yang berarti "sebelum" dan juga kata "أو" yang berarti "atau". Kata-kata tersebut merupakan kata-kata kurang penting sehingga dilakukan penghapusan.

3. Stemming

Proses *stemming* dilakukan setelah mendapatkan hasil pengolahan dari *stopwords removal*. *Stemming* merupakan proses mengubah suatu kata menjadi kata dasarnya. Berikut Tabel 2.4 adalah hasil dari *stopwords removal*:

Tabel 2.4 Kata-Kata yang akan Dilakukan *Stemming*

Arab	Latin	Terjemahan
يجوز	yajuz	dibolehkan
تقديم	taqdim	membayar
زكاة	zaka	zakat
الفطر	**alfitr**	**fitrah**
العيد	**aleid**	**idul fitri**
بيوم	byum	satu
يومين	yawmayn	dua hari

Setelah dilakukan *stemming* (pengubahan suatu kata kedalam kata dasarnya) terhadap sekumpulan kata diatas maka kata-kata tersebut berubah menjadi kata-kata dasarnya. Berikut Tabel 2.5 adalah daftar kata setelah dilakukan *stemming*. Kata-kata yang telah dilakukan *stemming* bercetak tebal pada tabel.

Tabel 2.5 Kata-Kata Hasil *Stemming*

Arab	Latin	Terjemahan
يجوز	yajuz	dibolehkan
تقديم	taqdim	membayar
زكاة	zaka	zakat
فطر	**fitr**	**fitrah**
عيد	**eid**	**idul fitri**
بيوم	byum	satu
يومين	yawmayn	dua hari

2.2.3 *Term Frequency Inverse Document Frequency (TF.IDF)*

Term Frequency Inverse Document Frequency (TF.IDF) adalah pembobotan yang terdiri dari *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF). *Term Frequency* (TF) adalah frekuensi banyaknya kata yang muncul dalam suatu dokumen. Sedangkan *Inverse Document Frequency* (IDF) adalah untuk mengukur seberapa penting sebuah kata dalam suatu dokumen yang dilihat dari kumpulan dokumen secara keseluruhan: Berikut persamaan dari TF.IDF:

$$IDF(t) = \log \frac{N}{df(t)} \quad (2.1)$$

N = jumlah dokumen yang ada didalam koleksi

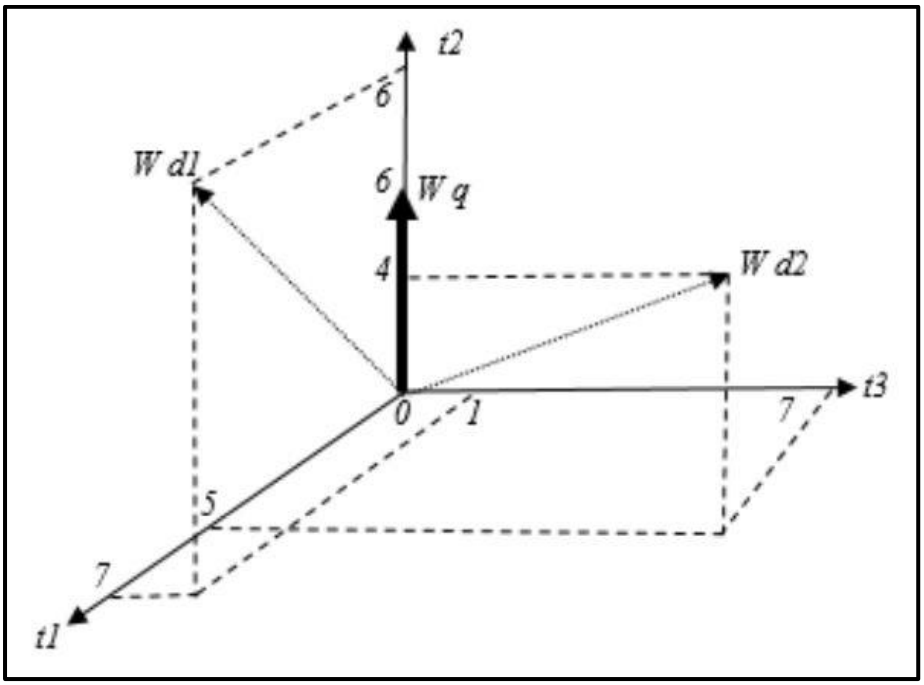
df(t) = banyaknya dokumen yang berisi term t

$$TF.IDF = TF(dt) * IDF(t) \quad (2.2)$$

Jadi TF.IDF merupakan pembobotan suatu dokumen dengan mengalikan antara *term frequency* (TF) dan *inverse document frequency* (IDF).

2.2.4 *Vector Space Model*

Vector Space Model (VSM) atau yang biasa disebut sebagai model ruang vektor merupakan representasi sejumlah dokumen dalam ruang vektor yang biasanya digunakan dalam *information retrieval* untuk memberi skor pada dokumen, klasifikasi dokumen hingga klaster dokumen. Dokumen direpresentasikan ke dalam ruang vektor multidimensi yang ditentukan dengan membandingkan jarak antara vektor-vektor. Tahapan model ruang vektor terbagi menjadi tiga tahap. Tahap pertama adalah *document indexing* (pengindeksan dokumen) dengan mengekstrak *term* paling relevan. Tahap kedua adalah didasarkan pada bobot yang terkait dengan indeks *term-term* untuk meningkatkan hasil *retrieve* suatu informasi. Tahap ketiga mengklasifikasikan dokumen dengan pengukuran kemiripan tertentu (Alhanjouri, 2017). Gambar 2.2 menunjukkan ilustrasi dari *Vector Space Model* (VSM).



Gambar 2.2 Ilustrasi *Vector Space Model* (Sholikah *et al.*, 2017)

Berdasarkan Gambar 2.2, Sebagai contoh, *query* dan dokumen adalah vektor-vektor dalam suatu ruang n-dimensi, di mana *k* adalah jumlah kata yang ada di dalam leksikon. Leksikon merupakan kumpulan kata pada suatu indeks.

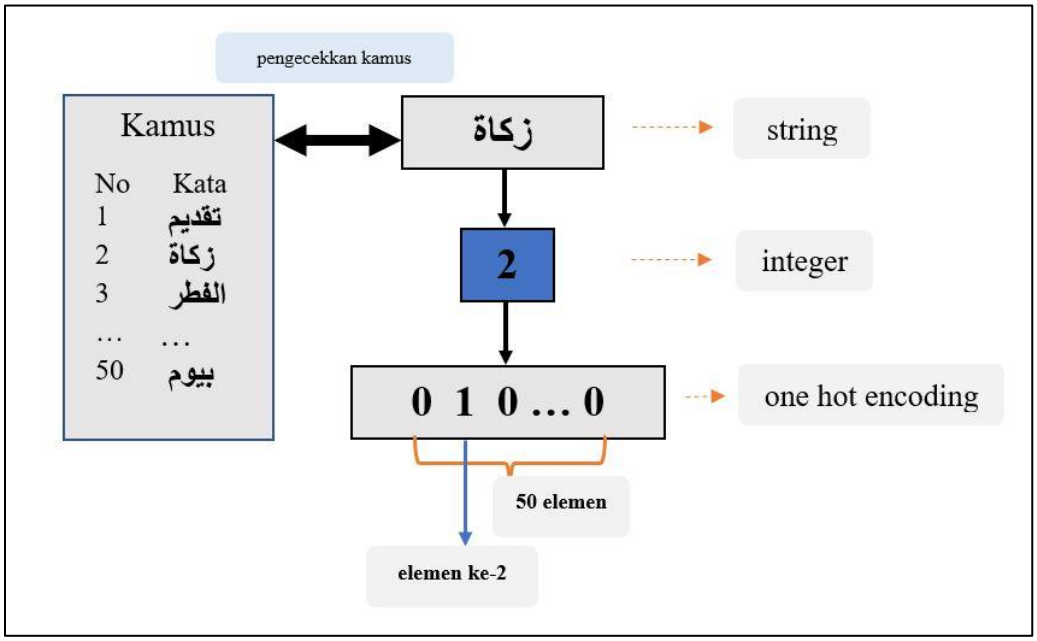
Sudut *cosine* antara dua vektor, di mana *Wq* sebagai *query* dan *Wd* sebagai tiap dokumen akan dihitung. Setiap kata *t* di dalam sebuah dokumen dianggap sebagai 1 dimensi, sehingga jika terdapat tiga kata maka akan membentuk 3 dimensi (Sholikah *et al.*, 2017).

2.2.5 *Word Embedding*

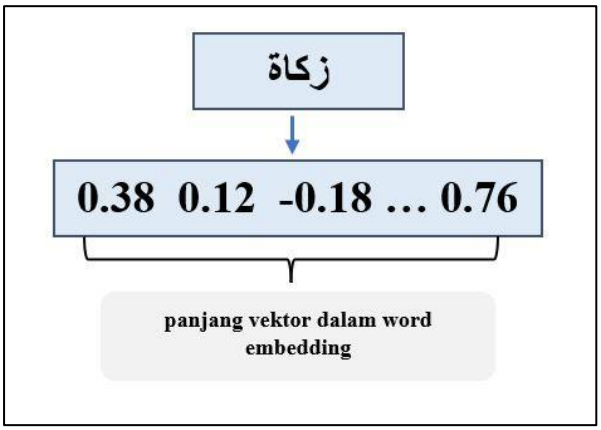
Word embedding (*word vector representation*) merupakan teknik atau proses yang mengubah teks menjadi angka atau dengan kata lain merepresentasikan kata ke dalam suatu vektor numerik (Othman, Faiz and Smaili, 2019). Selain itu, *word embedding* merupakan representasi suatu teks di mana kumpulan kata di dalamnya yang memiliki makna serupa akan direpresentasikan memiliki kemiripan yang sama. Oleh karena itu *word embedding* mampu menangkap konteks suatu dokumen di mana konteks tersebut berkaitan dengan kemiripan semantik, hubungan antar kata dan lain sebagainya. Teknik yang digunakan dalam *word embedding* di mana per individu kata akan direpresentasikan ke dalam nilai vektor dalam suatu ruang vektor (*vector space*). Setiap kata dipetakan menjadi satu vektor dan nilai dari vektor tersebut dipelajari dengan cara menyerupai jaringan saraf (*neural network*). Sehingga, *word embedding* sering kali dimasukkan ke dalam kategori *deep learning*.

Secara umum, ketika suatu model *machine learning* akan mengolah suatu input berupa teks, maka teks tersebut tidak dapat langsung diolah begitu saja. Dalam mengolah teks, secara tradisional teks tersebut akan dipecah menjadi sekumpulan kata, di mana kumpulan kata tersebut membentuk suatu kamus (*dictionary*), dengan kata lain mengolah daftar kata selayaknya daftar kata di dalam sebuah kamus. Sebagai contoh, setiap kata (*string*) akan diubah menjadi angka *integer* dengan memberikan nomor pada *string-string* tersebut sesuai dengan urutannya di dalam kamus. Kemudian, angka-angka *integer* tersebut kembali diubah menjadi sebuah vektor dalam ini *array* satu dimensi, di mana hanya ada angka 1 atau 0 di dalamnya. Nilai 1 diletakkan pada *string* yang terindeks di dalam daftar kamus, dan *string-string* lainnya mendapat nilai 0. Teknik tersebut juga dikenal dengan *one hot encoding*. Teknik tersebut menjadi kurang efisien karena hanya akan memuat banyak nilai 0 dan dari segi memori

tidak akan memuat banyak informasi. Namun, *word embedding* dapat merepresentasikan sebuah kata menjadi vektor yang berisi angka-angka cukup kecil, sehingga vektor tersebut mengandung banyak informasi sampai mampu memuat makna suatu kata. Gambar 2.3 menunjukkan ilustrasi *one hot encoding* dan Gambar 2.4 menunjukkan panjang dimensi vektor dalam *word embedding*.



Gambar 2.3 Ilustrasi *One Hot Encoding*

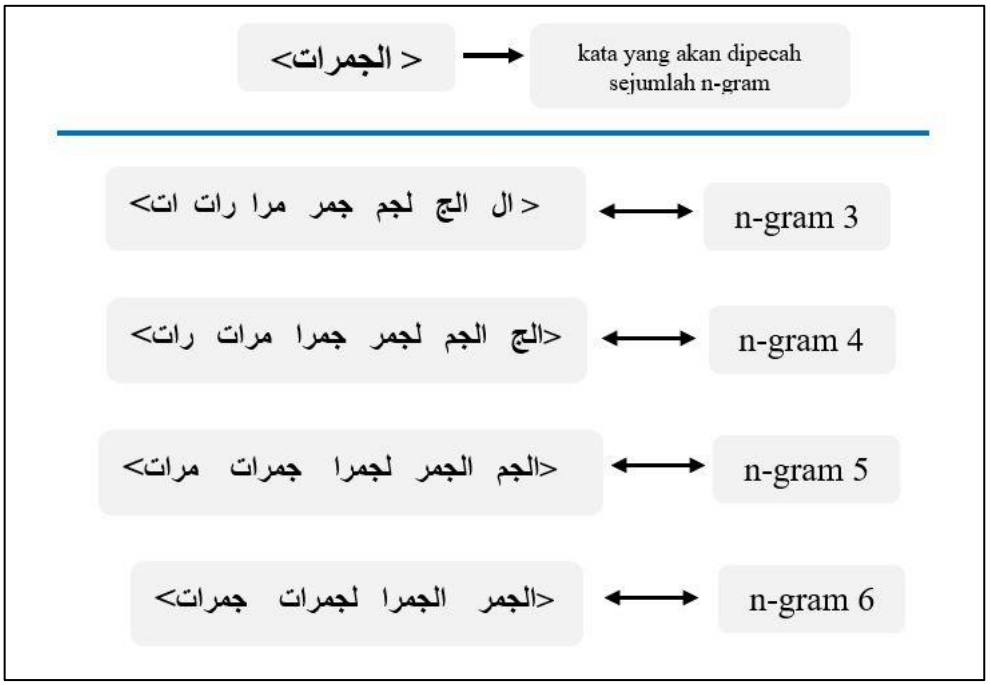


Gambar 2.4 Ilustrasi Panjang Dimensi Vektor *Word Embedding*

Fasttext merupakan salah satu teknik *word embedding* yang didasari pada prediksi (*prediction-based method*). *Fasttext* merupakan pengembangan dari model *word embedding* sebelumnya yaitu *word2vec* (Bojanowski *et al.*, 2017). Terdapat dua cara dalam mengimplementasikan *fasttext* yaitu: *continuous bag-of-*

words (CBOW) dan *skip-gram*(SG). Model *fasttext* dibangun untuk dapat mengatasi kata-kata di luar *vocabulary* yang ada (*out-of-vocabulary* (OOV)) dengan memperluas sub-kata informasi *word2vec skip-gram* (SG) (Alghamdi and Assiri, 2020). *Fasttext* menyediakan *pre-trained* model yang diolah ke dalam 157 bahasa (Grave *et al.*, 2018). Salah satu di antara nya adalah bahasa Arab. *Pre-trained* model tersebut diolah berdasarkan korpus *common crawl* dan *wikipedia*. Parameter pengolahan model-model tersebut menggunakan dimensi 300, dengan karakter *n-gram* sebanyak 5, serta *window size* sebesar 5. Namun, selain *pre-trained* model yang telah disediakan, model *fasttext* dapat dilatih sendiri menggunakan korpus yang disesuaikan dengan kebutuhan.

Penggunaan *fasttext* didasari pada *n-gram* karakter. *N-gram* karakter yang digunakan biasanya dimulai dari yang berjumlah 3 sampai 6. Gambar 2.5 menunjukkan contoh penggunaan tiap *n-gram* karakter pada sebuah kata.



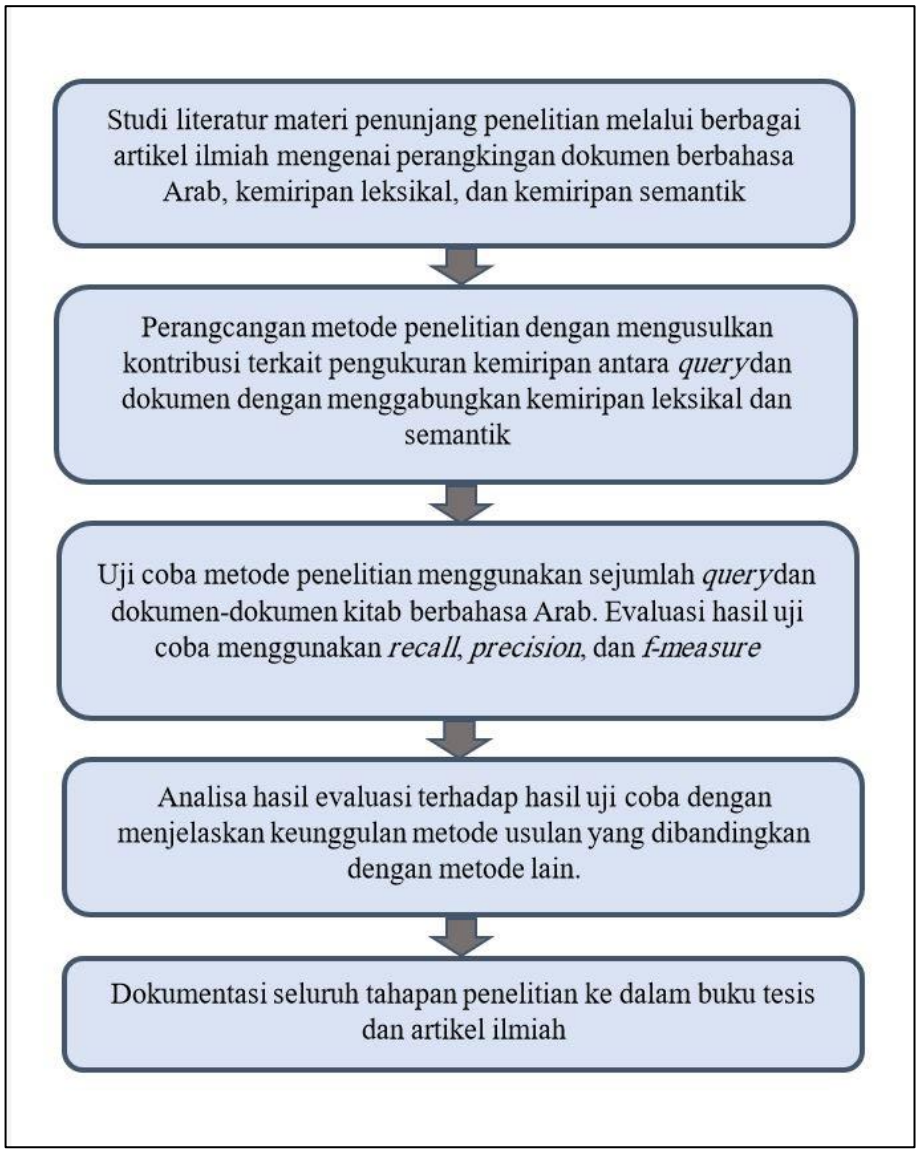
Gambar 2.5 Contoh Penggunaan *N-Gram* Pada Sebuah Kata

Berdasarkan Gambar 2.5, pada suatu kata untuk mendapatkan karakter sejumlah n-gram-nya maka kata tersebut dipecah-pecah. Pemecahan kata menjadi sejumlah karakter n-gram dilakukan agar kata tersebut akan semakin mudah untuk diukur kemiripannya dengan kata lainnya. Selain itu terdapat tanda kurung siku

pada awal dan akhir kata, di mana tanda kurung siku menunjukkan karakter awal dan akhir dari kata tersebut. Hal ini membantu dalam menangkap kata-kata yang lebih pendek serta membantu dalam memahami akhiran (*suffix*) dan awalan (*prefix*) pada suatu kata.

BAB 3
METODOLOGI PENELITIAN

Bab ini menjelaskan tahapan-tahapan penelitian yang terdiri dari studi literatur, perancangan algoritma yang digunakan, pengujian dan evaluasi, analisa hasil penelitian, serta dokumentasi keseluruhan penelitian. Gambar 3.1 merupakan alur metodologi penelitian secara keseluruhan.

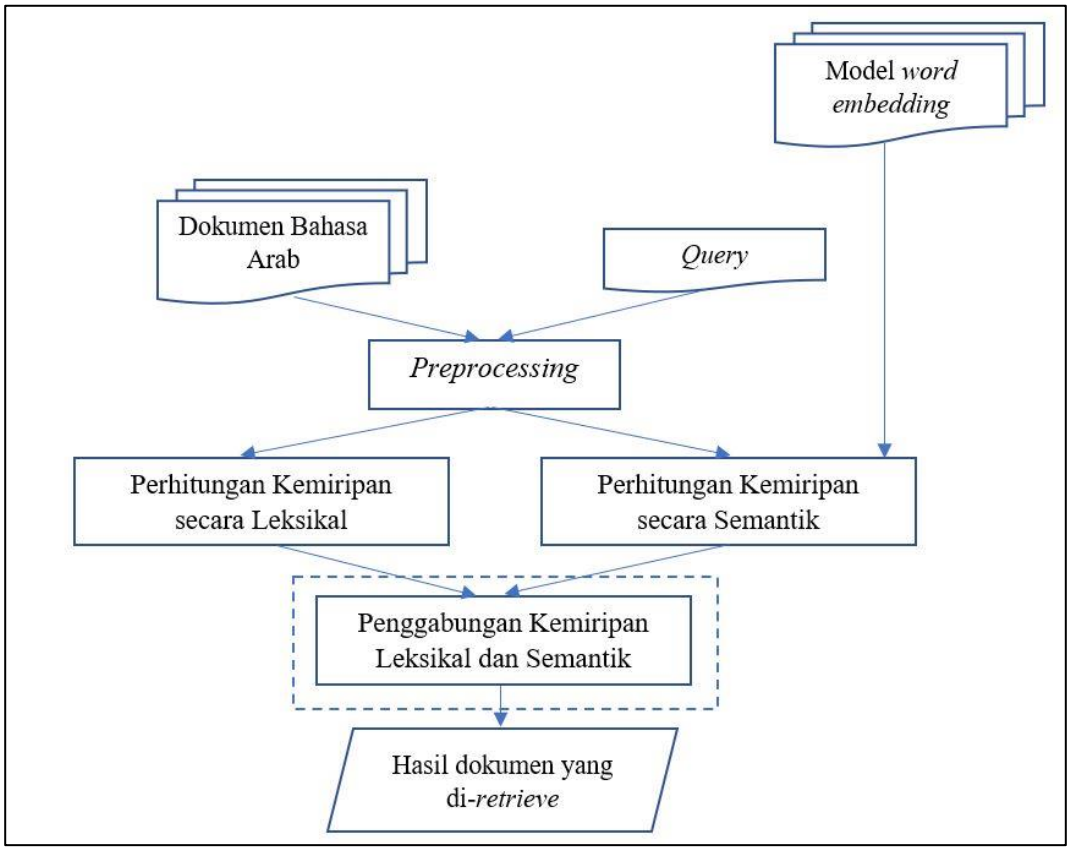


Gambar 3.1 Alur Metodologi Penelitian

Gambar 3.1 adalah alur metodologi penelitian yang dimulai dari tahap studi literatur hingga tahap dokumentasi. Tahap pertama adalah studi literatur dengan mencari dan mempelajari literatur yang berkaitan dengan penelitian dari berbagai sumber yaitu: buku, jurnal ilmiah, artikel ilmiah, dan prosiding. Materi-metari yang berkaitan dengan penelitian yaitu tentang perangkingan dokumen berbahasa Arab, pengukuran kemiripan leksikal hinggan pengukuran kemiripan semantik. Tahap kedua adalah perancangan metode penelitian yang merupakan bagian inti di mana memuat pemecahan masalah dari penelitian sebelumnya hingga kontribusi penelitian dalam memecahkan permasalahan tersebut. Kontribusi penelitian terkait dengan pengukuran kemiripan antara *query* dan dokumen dengan menggabungkan kemiripan leksikal dan kemiripan semantik. Tahap ketiga berisi uji coba dan evaluasi memuat hasil uji coba dan evaluasi terhadap hasil uji coba sehingga untuk mengetahui performa dari kontribusi penelitian. Uji coba dilakukan dengan menggunakan sejumlah *query* dan dokumen-dokumen kitab berbahasa Arab. Sedangkan evaluasi memuat penjelasan mengenai hasil uji coba yang dievaluasi menggunakan *recall*, *precision*, dan *f-measure*. Setelah uji coba dan evaluasi, selanjutnya adalah analisa hasil yang memuat penjelasan mengenai kelebihan dan kekurangan berdasarkan hasil uji coba yang telah dilakukan. Selain itu, analisa hasil dilakukan terhadap hasil uji coba perbandingan antara metode usulan dan metode-metode lainnya. Tahap kelima adalah dokumentasi berupa buku tesis dan artikel ilmiah yang merangkum studi literatur, perancangan metode penelitian, uji coba dan evaluasi serta analisa hasil penelitian.

3.1 Tahap Perangkingan Dokumen

Tahap kedua setelah dilakukan studi literatur adalah merancang algoritma yang digunakan. Gambar 3.2 merupakan proses perangkingan dokumen secara keseluruhan dimulai dari *preprocessing* dokumen dan *query*, hasil dari *preprocessing* dimasukan ke dalam perhitungan leksikal dan semantik, hingga menghasilkan dokumen yang telah dirangking. Gambar 3.2 akan dijelaskan lebih lanjut pada sub bab-sub bab selanjutnya.

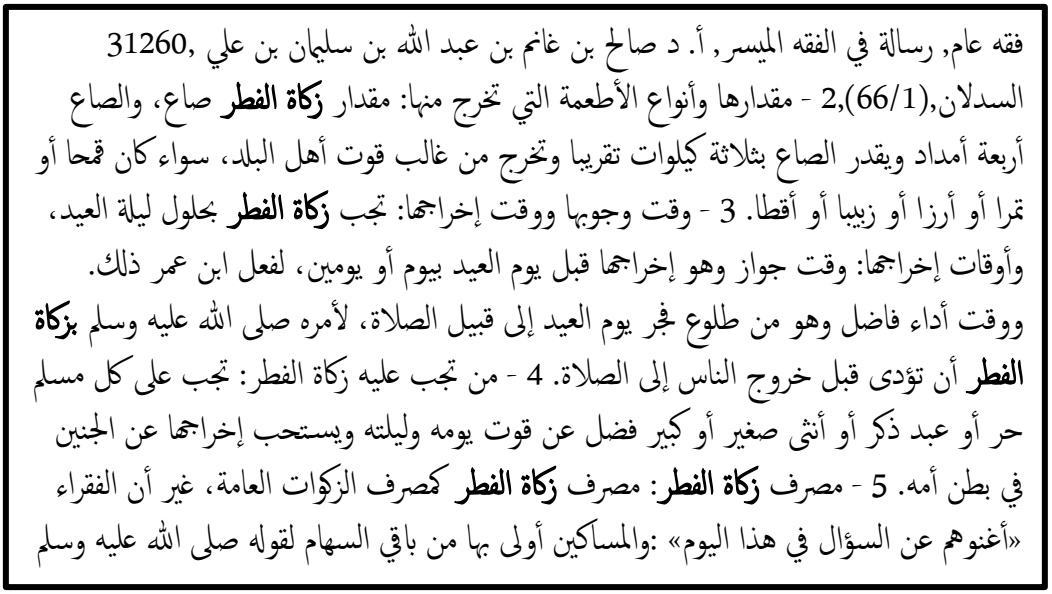


Gambar 3.2 Perangkingan Dokumen Keseluruhan

3.1.1 Data

Data dokumen yang digunakan pada penelitian ini merupakan kitab-kitab berbahasa Arab. Istilah dokumen dalam penelitian ini direpresentasikan oleh per halaman kitab (buku). Dataset diambil melalui kumpulan kitab Maktabah Syamilah (Aris, 2015). Maktabah Syamilah berisi ribuan kitab berbahasa Arab untuk mempermudah penggunanya dalam mencari informasi yang berkaitan dengan kajian agama Islam. Maktabah syamilah memiliki sejumlah bidang-bidang rumpun keilmuan (kategori) serta masing-masing jumlah kitab yang dimiliki dengan jumlah keseluruhan kitab adalah 2900 kitab. Beberapa contoh kategori di antara-nya yaitu "الأجزاء الحديثية" yang berarti "bagian-bagian hadis", di mana kategori ini memiliki jumlah kitab sebanyak 542 kitab (buku). Kemudian, kategori "العقيدة" yang berarti "akidah", di mana kategori ini memiliki jumlah kitab sebanyak 505 kitab (buku), kategori "فقه عام" yang berarti "fikih secara umum" memiliki jumlah kitab sebanyak 45 kitab (buku). Keseluruhan kategori beserta artinya dapat

dilihat pada lampiran 1. Berikut salah satu contoh isi dokumen Maktabah Syamilah dari kitab berjudul *risalah fi al-fiqh al-muyassar*.


فقه عام, رسالة في الفقه الميسر, أ. د صالح بن غانم بن عبد الله بن سليمان بن علي ,31260
السدلان,(66/1),2 - مقدارها وأنواع الأطعمة التي تخرج منها: مقدار **زكاة الفطر** صاع، والصاع
أربعة أمداد ويقدر الصاع بثلاثة كيلوات تقريبا وتخرج من غالب قوت أهل البلد، سواء كان قمحا أو
تمرا أو أرزا أو زبيبا أو أقطا. 3 - وقت وجوبها ووقت إخراجها: تجب **زكاة الفطر** بحلول ليلة العيد،
وأوقات إخراجها: وقت جواز وهو إخراجها قبل يوم العيد بيوم أو يومين، لفعل ابن عمر ذلك.
ووقت أداء فاضل وهو من طلوع فجر يوم العيد إلى قبيل الصلاة، لأمره صلى الله عليه وسلم **بزكاة**
الفطر أن تؤدى قبل خروج الناس إلى الصلاة. 4 - من تجب عليه زكاة الفطر: تجب على كل مسلم
حر أو عبد ذكر أو أنثى صغير أو كبير فضل عن قوت يومه وليلته ويستحب إخراجها عن الجنين
في بطن أمه. 5 - مصرف **زكاة الفطر**: مصرف **زكاة الفطر** كمصرف الزكوات العامة، غير أن الفقراء
«أغنوهم عن السؤال في هذا اليوم» :والمساكين أولى بها من باقي السهام لقوله صلى الله عليه وسلم

Gambar 3.3 Contoh Isi Kitab Maktabah Syamilah

Gambar 3.3 merupakan salah satu contoh isi sebuah halaman dari salah satu kitab pada al-majmu'ah (bidang keilmuan) atau kategori *فقه عام* (fikih secara umum) Maktabah Syamilah. Contoh isi kitab tersebut mengenai zakat fitrah. Isi dokumen kitab tersebut memiliki arti sebagai berikut: *32160 fikih, "risalah fi al-fiqh al-muyassar", Prof. DR. Saleh bin ghanim bin abdullah bin sulaiman bin ali bin as-sadlan, (66/1), 2- takaran dan jenis makanan yang dikeluarkan untuk zakat fitrah: takaran zakat fitrah adalah satu sha', nilai satu sha' sebesar empat mud, satu sha' diperkirakan senilai dengan berat tiga kilo kurang lebih, makanan yang dikeluarkan untuk zakat fitrah berasal dari makanan pokok suatu negeri yang dominan baik itu berupa gandum, kurma, beras, kismis, atau jameed. 3- waktu wajib dan waktu mengeluarkan zakat fitrah: zakat fitrah menjadi wajib saat munculnya malam eid, waktu-waktu diperbolehkan mengeluarkan zakat fitrah: waktu boleh mengeluarkannya adalah sehari ataupun dua hari sebelum hari eid, hal ini disebabkan karena adanya perbuatan ibnu umar akan hal tersebut, dan waktu utama untuk menunaikan zakat fitrah adalah saat fajar hari eid mulai menyingsing hingga tepat sebelum shalat eid dilaksanakan, hal ini disebabkan karena adanya perintah nabi saw untuk menunaikan zakat fitrah sebelum orang-*

orang keluar untuk melaksanakan shalat eid. 4- siapa yang diwajibkan mengeluarkan zakat fitrah: zakat fitrah diwajibkan bagi setiap muslim merdeka atau hamba baik itu pria, wanita, anak kecil, atau orang tua sebagai karunia atas makanan siang dan malamnya, dan disunnahkan untuk mengeluarkan zakat fitrah untuk janin yang ada di perut ibunya, 5- orang-orang yang diberikan zakat fitrah: orang-orang yang diberikan zakat fitrah sama seperti orang-orang yang diberikan zakat pada umumnya, namun orang fakir dan orang miskin lebih berhak mendapatkannya dibandingkan mustahiq zakat yang lain sebagaimana sabda rasulullah saw: "cukupilah (kebutuhan) mereka, agar mereka tidak meminta-minta pada hari seperti ini".

3.1.2 *Preprocessing* Dokumen

Dokumen-dokumen bahasa Arab sebelum dilakukan pemrosesan lebih lanjut, terlebih dahulu dilakukan *preprocessing*. *Preprocessing* ini dilakukan agar dokumen siap diproses. Oleh karena itu, dokumen dilakukan tiga tahap praproses yaitu: tokenisasi, *stopwords removal*, dan *stemming*. Tokenisasi merupakan proses pemecahan dokumen, dimana jika dokumen direpresentasikan sebagai satu halaman buku maka dokumen yang isinya berupa paragraf-paragraf dipecah menjadi kalimat, kemudian kalimat dipecah menjadi kumpulan kata. Selain itu, dokumen yang telah dipecah tadi, juga dilakukan penghapusan terhadap *diacritic* atau harokat yang tidak diperlukan pada bahasa Arab. Tahap kedua adalah *stopwords removal*. *Stopwords removal* merupakan proses penghapusan kata-kata tidak penting yang dapat mempengaruhi nilai bobot pada saat pembobotan. Kata-kata tidak penting tersebut akan muncul dengan frekuensi paling banyak daripada kata-kata penting lainnya. Tahap ketiga adalah stemming. *Stemming* merupakan proses pengubahan suatu kata menjadi kata dasarnya. *Stemming* yang digunakan adalah *Arabic Light Stemmer*.

3.1.3 Perhitungan Kemiripan Secara Leksikal

Perhitungan kemiripan leksikal merupakan perhitungan kemiripan yang dimulai dari pembobotan *term* menggunakan *Term Frequency Inverse Document Frequency* (TF.IDF) kemudian dilanjutkan dengan perkalian vektor antara *query* dan dokumen. Setelah TF.IDF, tahap selanjutnya menggunakan *Cosine Similarity*

untuk mendapatkan nilai kemiripan leksikal tiap dokumen. Gambar 3.4 menunjukkan alur dari proses perhitungan kemiripan secara leksikal.

a. Pembobotan *Term Frequency Inverse Document Frequency* (TF.IDF)

   Pembobotan menggunakan *Term Frequency Inverse Document Frequency* (TF.IDF) pada Persamaan 3.1 dihitung berdasarkan dua bagian. Bagian pertama menghitung frekuensi kemunculan kata menggunakan $W_{tf}(w_i, d_j)$. $W_{tf}(w_i, d_j)$ di mana w_i adalah kata ke-i dan d_j adalah dokumen ke-j. Kata-kata tersebut diberi bobot sesuai jumlah masing-masing kata.

   $$TF.IDF(w_i, d_j) = W_{tf}(w_i, d_j) * W_{idf}(w_i) \ , \quad (3.1)$$

   Bagian kedua menghitung *inverse document frequency* berdasarkan Persamaan 3.2.

   $$W_{idf}(w_i) = \log \frac{N}{df(w_i)}, \quad (3.2)$$

   dimana N merupakan jumlah dokumen yang ada didalam koleksi, sedangkan $df(w_i)$ merupakan banyaknya dokumen yang berisi kata ke-i.

b. Perkalian Vektor antara *Query* dan Dokumen

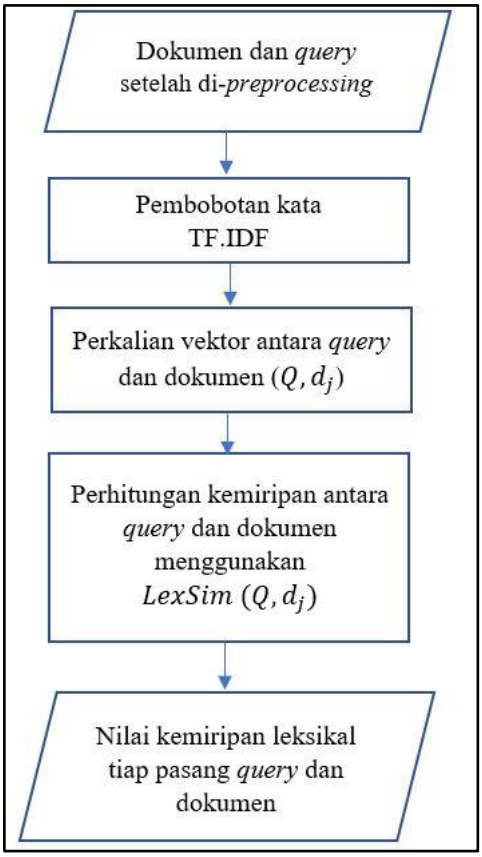
   Pada proses ini dilakukan perkalian vektor antara *query* dan dokumen untuk menghasilkan nilai bobot dari tiap kata dari suatu dokumen. Nilai bobot dari tiap kata pada tiap dokumen dijumlahkan, di mana nilai ini nantinya akan dimasukkan ke dalam perhitungan kemiripan leksikal.

c. Perhitungan Kemiripan Leksikal

   Perhitungan kemiripan leksikal didasarkan pada penggunaan *Cosine Similarity*, di mana perhitungan kemiripan memiliki nilai minimum yaitu 0 dan nilai maksimum yaitu 1. Pengukuran kemiripan antara dua objek dikatakan mirip jika memiliki nilai mendekati 1. Sebaliknya jika kedua objek tersebut semakin menjauhi 1 maka dikatakan tidak mirip. Berikut Persamaan 3.3 menunjukkan perhitungan kemiripan leksikal:

$$LexSim\ (Q, d_j) = \frac{\sum(TF.IDF(w_i,Q)).(TF.IDF(w_i,d_j))}{\sqrt{\sum|TF.IDF_Q|^2}.\sqrt{\sum|TF.IDF_{d_j}|^2}}\ , \quad (3.3)$$

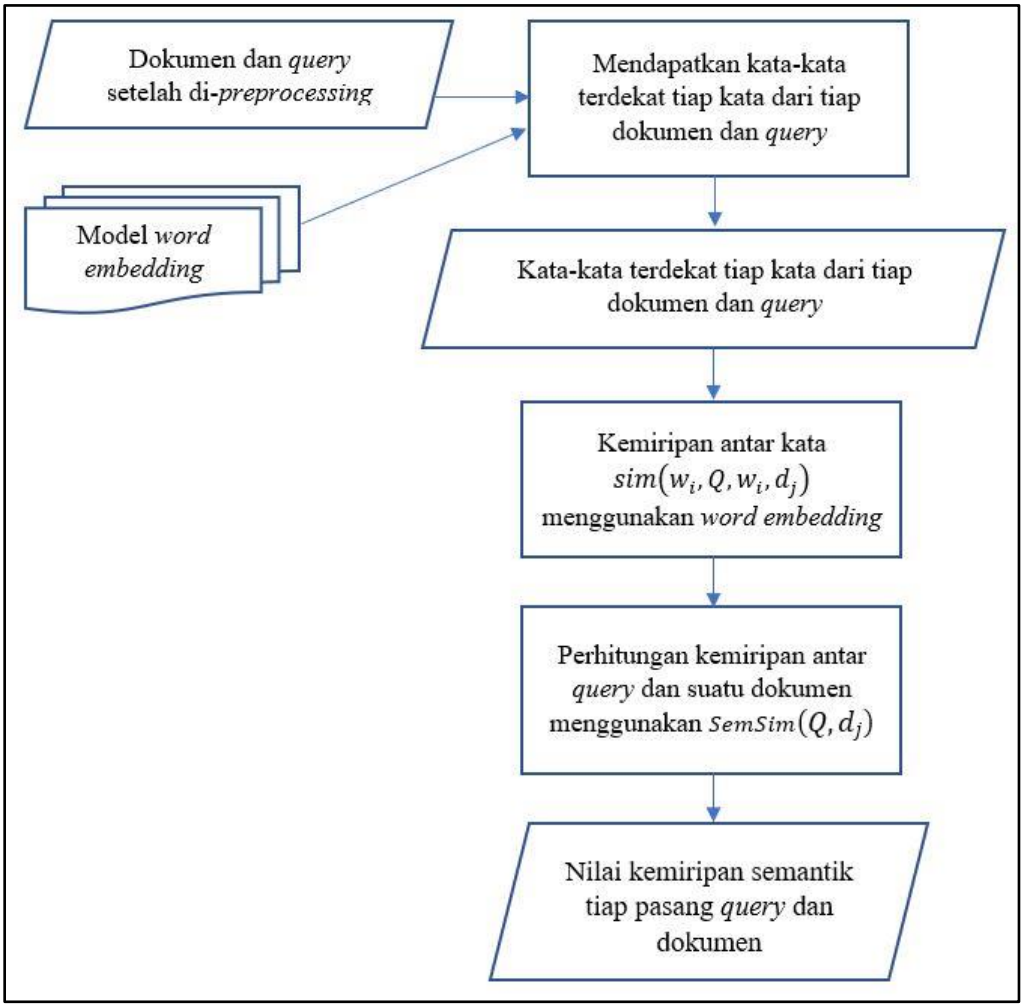
dimana Q adalah *query* sedangkan d_j adalah suatu dokumen ke-j. $TF.IDF(w_i, Q)$ dan $TF.IDF(w_i, d_j)$ adalah perhitungan TF.IDF antara *query* dan dokumen. $TFIDF_Q$ adalah jumlah kata pada *query*. Sedangkan $TFIDF_{d_j}$ adalah jumlah kata pada dokumen ke-j.



Gambar 3.4 Perhitungan Kemiripan Leksikal

3.1.4 Perhitungan Kemiripan Secara Semantik

Kemiripan semantik diproses berdasarkan kata-kata terdekat yang dimiliki tiap kata. Kata-kata terdekat tersebut dihasilkan melalui proses *word embedding*. Gambar 3.5 menunjukkan alur dari proses perhitungan kemiripan secara semantik.



Gambar 3.5 Perhitungan Kemiripan Semantik

Berdasarkan Gambar 3.5, perhitungan kemiripan secara semantik dimulai dari mendapatkan kata-kata terdekat tiap kata dari tiap dokumen dan *query*. Selanjutnya kata-kata terdekat tiap kata yang telah didapatkan, digunakan untuk menghitung kemiripan antar kata. Setelah dilakukan perhitungan kemiripan antar kata, kemudian kemiripan antar *query* dan dokumen juga dihitung. Terakhir, nilai kemiripan semantik dari tiap dokumen akan didapatkan.

a. Kata-kata terdekat tiap kata

   Kata-kata terdekat dari suatu kata didapatkan secara otomatis dari hasil *word embedding* berdasarkan pustaka *digital Arabic*. Tiap kata yang dihasilkan dari proses *word embedding* dengan *fasttext*, masing-masing akan memiliki bobot mulai dari bobot terbesar (memiliki kemiripan paling dekat) hingga bobot terkecil.

b. Kemiripan Antar Kata

Kata-kata yang terdapat pada *query* dan dokumen diukur kemiripannya menggunakan $sim\left(w_{iQ}, w_{id_j}\right)$, dimana w_{iQ} adalah kata ke-i dari *query*, w_{id_j} adalah kata-i dari suatu dokumen ke-j. $sim\left(w_{iQ}, w_{id_j}\right)$ merupakan representasi hasil kemiripan antara dua buah kata. Kemiripan antar kata tersebut berdasarkan proses *word embedding*, di mana pada *word embedding* tersebut menggunakan *Cosine Similarity* di dalam prosesnya. Sehingga pada akhirnya kemiripan antar dua buah kata memiliki bobot yang akan dipakai pada perhitungan selanjutnya. Sebagai contoh kata "**زكاة**" dan kata "**الفطر**". Kedua kata tersebut memiliki kemiripan sebesar 0.335. Penggunaan *fasttext* mewakili tiap kata sebagai *n-gram* karakter. Misalnya kata "**الفطر** " dan n=2 maka representasi *fasttext* terhadap kata "**الفطر** " adalah <**ال, ال, لف, فط, طر, طر**>, di mana tanda kurung siku menunjukkan karakter awal dan akhir dari kata tersebut. Hal ini membantu dalam menangkap kata-kata yang lebih pendek serta membantu dalam memahami akhiran (*suffix*) dan awalan (*prefix*) pada suatu kata.

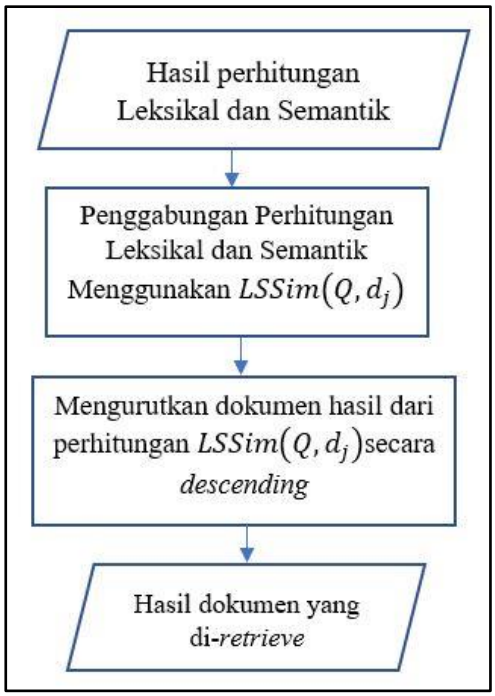
c. Perhitungan Kemiripan Semantik

Kemiripan antar *query* dan dokumen dihitung untuk mendapatkan nilai semantik dokumen. Perhitungan kemiripan dihitung berdasarkan Persamaan 3.4.

$$SemSim(Q, d_j) = \frac{\Sigma\left(sim\left(w_{iQ}, w_{id_j}\right)\right)}{\Sigma {Q_{term}}^2 + \Sigma {d_j}_{term}{}^2 - \Sigma\left(sim\left(w_{iQ}, w_{id_j}\right)\right)}, \quad (3.4)$$

$SemSim(Q, d_j)$ adalah perhitungan kemiripan antar *query* dan dokumen, yang dimana Q adalah *query* dan d_j adalah suatu dokumen ke-j. Kemudian, Q_{term} adalah jumlah kata pada *query*, sedangkan ${d_j}_{term}$ adalah jumlah kata suatu dokumen ke-j. $SemSim(Q, d_j)$didasarkan pada *Jaccard Measure* (Abdi, Mariyam and Aliguliyev, 2018) dengan nilai minimumnya adalah 0 dan nilai maksimumnya adalah 1.

3.1.5 Penggabungan Perhitungan Kemiripan Leksikal dan Semantik



Gambar 3.6 Penggabungan Perhitungan Kemiripan Leksikal dan Semantik

Nilai akhir didapatkan dengan menggabungkan nilai yang didapatkan dari perhitungan kemiripan leksikal dan perhitungan kemiripan semantik seperti yang tampak pada Gambar 3.6. Nilai akhir ini bertujuan untuk mendapatkan nilai kemiripan antara *query* dan dokumen, di mana nilai tersebut yang akan dirangking secara *descending* (dari yang terbesar hingga terkecil) untuk mendapatkan satu dokumen dengan skor tertinggi sebagai hasil perangkingan. Sebelum mendapatkan satu dokumen tertinggi sebagai hasil perangkingan, Persamaan 3.5 menunjukkan persamaan dalam mencari nilai gabungan hasil dari perhitungan leksikal dan semantik dari suatu *query* (Q) dan dokumen ke-j (d_j) .

$$LSSim(Q, d_j) = \alpha . LexSim(Q, d_j) + (1 - \alpha) . SemSim(Q, d_j) \quad , \qquad (3.5)$$

dimana $\alpha \in [0, 1]$ merupakan parameter untuk penggunaan efektif dari $LexSim(Q, d_j)$ dan $SemSim(Q, d_j)$.

$LexSim(Q, d_j)$ adalah perhitungan kemiripan leksikal. Sedangkan $SemSim(Q, d_j)$ adalah nilai kemiripan antara *query* dan suatu dokumen yang

didapatkan dari proses perhitungan kemiripan semantik. Nilai minimum dari perhitungan gabungan leksikal dan semantik adalah 0 dan nilai maksimumnya adalah 1. Sehingga semakin mendekati nilai 1 maka antara *query* dan dokumen dapat dikatakan semakin mirip. Setelah mendapatkan nilai $LSSim(Q, d_j)$, selanjutnya adalah mendapatkan sejumlah dokumen tertinggi sebanyak 5 sampai 15 dokumen, yang sebelumnya diurutkan terlebih dahulu dari yang tertinggi hingga terendah (*descending*). 5 sampai 15 dokumen tertinggi tersebut sebagai hasil perangkingan yang akan dijadikan sebagai hasil *retrieve* suatu *query*.

3.1.6 Contoh Ilustrasi Perhitungan Metode

Subbab berikut akan menjelaskan tahapan-tahapan contoh perhitungan metode secara keseluruhan. Tahapan perhitungan dimulai dari proses *preprocessing* sampai tahap penggabungan perhitungan kemiripan leksikal dan semantik.

a. *Preprocessing query* dan dokumen

Tahap preprocessing ini *query* dan sejumlah dokumen akan dilakukan sejumlah praproses agar data siap untuk dibawa ke proses selanjutnya. Tabel 3.1 menampilkan contoh sebuah *query* dan empat buah dokumen berbahasa Arab.

Tabel 3.1 Contoh *Query* dan Dokumen Ilustrasi Perhitungan Metode

Query	زكاة الفطر
Dokumen 1	مسح الرأس ركن من أركان الوضوء
Dokumen 2	حكم إخراج زكاة الفطر بعد خروج وقتها
Dokumen 3	كيفية صلاة التهجد و دفع زكاة
Dokumen 4	يجوز تقديم زكاة الفطر قبل العيد بيوم أو يومين

Berdasarkan Tabel 3.1, *query* "زكاة الفطر" berarti "zakat fitrah". Dokumen 1 mempunyai arti "mengusap kepala adalah salah satu rukun wudhu", Selanjutnya dokumen 2 memiliki terjemahan "hukum mengeluarkan zakat fitrah setelah lewat waktunya". Dokumen 3 berarti "tata cara shalat tahajud dan membayar zakat", serta dokumen 4 mempunyai terjemahan "dibolehkan mengeluarkan zakat fitrah satu atau dua hari sebelum Idul Fitri". *Query* dan sejumlah dokumen tersebut akan

dilakukan *preprocessing*, di mana hasil *preprocessing*-nya dapat dilihat pada Tabel 3.2.

Tabel 3.2 Contoh *Preprocessing Query* dan Dokumen Ilustrasi Perhitungan Metode

Praproses				
Query	Dokumen 1	Dokumen 2	Dokumen 3	Dokumen 4
زكاة	مسح	حكم	كيفية	يجوز
الفطر	الرأس	إخراج	صلاة	تقديم
	ركن	زكاة	التهجد	زكاة
		الفطر		الفطر
	ركان	بعد	دفع	
	الوضوء	خروج	زكاة	العيد
		وقتها		بيوم
				يومين

Setelah dilakukan *preprocessing*, berdasarkan Tabel 3.2 terlihat bahwa baik *query* atau dokumen-dokumen dipecah menjadi kumpulan kata. Selain itu, terdapat beberapa kata yang dihapus sebab kata-kata tersebut merupakan kata-kata tidak penting. Kata-kata tidak penting yang dihapus tersebut adalah "من" pada dokumen 1 yang berarti "dari". Kemudian kata "و" yang berarti "dan" pada dokumen 3. Selanjutnya, kata "قبل" yang berarti "sebelum" dan kata "أو" yang berarti "atau" pada dokumen 4.

b. Perhitungan Kemiripan Secara Leksikal

- Pembobotan Kata TF.IDF

Query dan dokumen-dokumen yang telah dilakukan *preprocessing* selanjutnya dibawa ke tahap perhitungan kemiripan leksikal. Pada perhitungan kemiripan leksikal, proses pertama ialah pembobotan kata menggunakan TF.IDF. Hal pertama yang dilakukan dalam pembobotan kata TF.IDF adalah menghitung *Term Frequency* (TF) yaitu menghitung frekuensi kemunculan kata pada setiap dokumen. Tabel 3.3 menampilkan contoh perhitungan TF. Pada gambar tersebut terlihat bahwa setiap kata dihitung kemunculan pada setiap dokumen. Sebagai contoh, kata "زكاة", kata tersebut muncul di beberapa dokumen yaitu dokumen 2, dokumen 3, dan dokumen 4. Adapun kata "الفطر" muncul di dua dokumen yaitu

dokumen 2 dan dokumen 4. Selain itu, kata "ركن" muncul dua kali pada dokumen 1.

Tabel 3.3 Contoh *Term Frequency* Ilustrasi Perhitungan Metode

Term	Dokumen 1	Dokumen 2	Dokumen 3	Dokumen 4
زكاة	0	1	1	1
الفطر	0	1	0	1
مسح	1	0	0	0
الرأس	1	0	0	0
ركن	2	0	0	0
الوضوء	1	0	0	0
حكم	0	1	0	0
إخراج	0	1	0	0
بعد	0	1	0	0
خروج	0	1	0	0
وقتها	0	1	0	0
كيفية	0	0	1	0
صلاة	0	0	1	0
التهجد	0	0	1	0
يجوز	0	0	0	1
تقديم	0	0	0	1
العيد	0	0	0	1
بيوم	0	0	0	1
يومين	0	0	0	1
دفع	0	0	1	0

Tabel 3.4 Contoh *Invers Document Frequency* Ilustrasi Perhitungan Metode

Term	Dokumen 1	Dokumen 2	Dokumen 3	Dokumen 4	Document Frequency(df)	idf log(n/df)
زكاة	0	1	1	1	3	0.125
الفطر	0	1	0	1	2	0.301
مسح	1	0	0	0	1	0.602
الرأس	1	0	0	0	1	0.602
ركن	2	0	0	0	1	0.602
الوضوء	1	0	0	0	1	0.602
حكم	0	1	0	0	1	0.602
إخراج	0	1	0	0	1	0.602
بعد	0	1	0	0	1	0.602
خروج	0	1	0	0	1	0.602
وقتها	0	1	0	0	1	0.602
كيفية	0	0	1	0	1	0.602
صلاة	0	0	1	0	1	0.602
التهجد	0	0	1	0	1	0.602
يجوز	0	0	0	1	1	0.602
تقديم	0	0	0	1	1	0.602
العيد	0	0	0	1	1	0.602
بيوم	0	0	0	1	1	0.602
يومين	0	0	0	1	1	0.602
دفع	0	0	1	0	1	0.602

Setelah menghitung frekuensi kemunculan kata pada setiap dokumen selanjutnya menghitung *Invers Document Frequency* (IDF). IDF menghitung jumlah dokumen di mana suatu kata yang sama terkandung di dalamnya. Tabel 3.4 menunjukkan contoh perhitungan IDF. Jika perhitungan TF dan IDF telah selesai dilakukan maka proses selanjutnya adalah menghitung TF.IDF. Tabel 3.5 menunjukkan hasil perhitungan antara TF dan IDF. Sehingga, tahap pembobotan kata menggunakan TF.IDF selesai.

Tabel 3.5 Contoh Perkalian *Term Frequency* dan *Invers Document Frequency* Ilustrasi Perhitungan Metode

tf*idf			
Dokumen 1	Dokumen 2	Dokumen 3	Dokumen 4
0	0.125	0.125	0.125
0	0.301	0	0.301
0.602	0	0	0
0.602	0	0	0
1.204	0	0	0
0.602	0	0	0
0	0.602	0	0
0	0.602	0	0
0	0.602	0	0
0	0.602	0	0
0	0.602	0	0
0	0	0.602	0
0	0	0.602	0
0	0	0.602	0
0	0	0	0.602
0	0	0	0.602
0	0	0	0.602
0	0	0	0.602
0	0	0	0.602
0	0	0.602	0

- Perkalian Vektor antara *Query* dan Dokumen

Perkalian vektor antara *query* dan dokumen dilakukan untuk mengetahui bobot dari tiap kata pada tiap dokumen. Tabel 3.6 menunjukkan hasil perkalian vektor antara *query* "زكاة الفطر" dan tiap dokumen. Berdasarkan hasil tersebut terlihat bahwa tiap kata yang ada pada *query* memiliki bobot yang berbeda di tiap dokumennya. Sebagai contoh kata "زكاة" memiliki bobot sama besar pada

dokumen 2, dokumen 3 dan dokumen 4. Sedangkan kata “زكاة” memiliki bobot yang berbeda dengan kata “الفطر” dan hanya dokumen 2 dan dokumen 4 yang memiliki nilai bobot dari kata tersebut. Hasil jumlah dari masing-masing dokumen pada perkalian vektor ditandai dengan yang berwarna kuning seperti yang terlihat pada tabel.

Tabel 3.6 Contoh Perkalian Vektor *Query* dan Dokumen Ilustrasi Perhitungan Metode

Perkalian vektor antara query dan dokumen				
Query	Dokumen 1	Dokumen 2	Dokumen 3	Dokumen 4
1	0	0.125	0.125	0.125
1	0	0.301	0	0.301
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
Jumlah	0	0.426	0.125	0.426

- Perhitungan Kemiripan Leksikal

Perhitungan kemiripan leksikal ini menggunakan rumus $LexSim(Q, d_j)$. Setelah perkalian vektor antara *query* dan dokumen selesai dilakukan, selanjutnya hasil dari perkalian vektor tersebut dimasukkan ke dalam perhitungan rumus untuk menghasilkan nilai kemiripan antara *query* dan tiap dokumen. Tabel 3.7 menampilkan nilai dari masing-masing kemiripan leksikal antara *query* dan tiap dokumen.

Tabel 3.7 Contoh Hasil Perhitungan Kemiripan Secara Leksikal Ilustrasi Perhitungan Metode

Leksikal	
LexSim(Q,D1)	0.000
LexSim(Q,D2)	0.154
LexSim(Q,D3)	0.052
LexSim(Q,D4)	0.154

c. Perhitungan Kemiripan Secara Semantik

- Kata-Kata Terdekat *Query* dan Dokumen

Tabel 3.8 Contoh Kata-Kata Terdekat *Query* Ilustrasi Perhitungan Metode

Query	
الفطر	زكاة
'الأضحى' '0.7083'	'وزكاة', '0.7777'
'فطر' '0.5927'	'الزكاة', '0.7558'
'والأضحى' '0.5892'	'الزروع', '0.7126'
'الفصح' '0.5802'	'وتجب', '0.6974'
'النوروز' '0.5791'	'الركاز', '0.6743'
'النيروز' '0.5629'	'الحلال', '0.6636'
'الطلع' '0.5599'	'تجب', '0.6568'
'الفطور' '0.5568'	'والزكاة', '0.6562'
'الهالووين' '0.5536'	'الربا', '0.6548'
'الاضحى' '0.5517'	'شرعا', '0.6363'

Tabel 3.9 Contoh Kata-Kata Terdekat Dokumen 1 Ilustrasi Perhitungan Metode

Dokumen 1				
الوضوء	ركان	ركن	الرأس	مسح
'الطهارة' '0.6942'	'وركن' '0.6499'	'وركن' '0.6499'	'العنق' '0.7337'	'فحص' '0.6338'
'السجود' '0.6862'	'الركن' '0.4973'	'الركن' '0.4973'	'الرقبة' '0.7196'	'المسح' '0.6322'
'الصلاة' '0.6813'	'أركان' '0.4835'	'أركان' '0.4835'	'والرأس' '0.6881'	'استقصاء' '0.5906'
'الغسل' '0.6794'	'لجلال' '0.4792'	'لجلال' '0.4792'	'للرأس' '0.6702'	'استطلاع' '0.5644'
'الإحرام' '0.6747'	'لعلاء' '0.4678'	'لعلاء' '0.4678'	'الذقن' '0.6601'	'استكشاف' '0.5593'
'الاغتسال' '0.6702'	'النور' '0.4654'	'النور' '0.4654'	'الأذنين' '0.6548'	'مسوحات' '0.5584'
'الأذان' '0.6547'	'محي' '0.4628'	'محي' '0.4628'	'الساقين' '0.6535'	'رصد' '0.5579'
'للصلاة' '0.6382'	'محيي' '0.4545'	'محيي' '0.4545'	'الذراعين' '0.6481'	'تنقيب' '0.5347'
'والصيام' '0.6302'	'محيي' '0.4533'	'محيي' '0.4533'	'والساقين' '0.6449'	'تحليل' '0.5281'
'والوضوء' '0.6279'	'وفخر' '0.4524'	'وفخر' '0.4524'	'الكتفين' '0.6443'	'استعراض' '0.5227'

Berdasarkan Tabel 3.8, gambar tersebut menunjukkan tiap kata pada *query* beserta kata-kata terdekatnya masing-masing. Selain pada *query* kata-kata yang terdapat pada tiap dokumen juga memiliki masing-masing kata-kata terdekatnya. Sebagai contoh dokumen 1 memiliki sejumlah kata-kata terdekat yang ditunjukkan pada Tabel 3.9. Kata-kata terdekat *query* dan tiap dokumen

digunakan untuk menghitung kemiripan antar kata-kata pada *query* dan kata-kata tiap dokumen. Kemiripan antar kata *query* dan dokumen dapat dilihat pada tahap proses selanjutnya.

- Kemiripan Antar Kata *Query* dan Dokumen

Tabel 3.10 Contoh Kemiripan Antar Kata *Query* dan Dokumen 1 Ilustrasi Perhitungan Metode

		Dokumen 1				
		مسح	الرأس	ركن	ركان	الوضوء
Query	زكاة	-0.06951	0.062209	0.179392	0.138652	0.491326
	الفطر	-0.0488	0.233473	0.025119	0.053836	0.345948
					Σ =	1.411649

Tabel 3.11 Contoh Kemiripan Antar Kata *Query* dan Dokumen 2 Ilustrasi Perhitungan Metode

		Dokumen 2						
		حكم	إخراج	زكاة	الفطر	بعد	خروج	وقتها
Query	زكاة	-0.04872	0.032509	1	0.335342	-0.07874	0.077363	-0.13165
	الفطر	-0.01559	-0.01689	0.335342	1	-0.0202	0.099756	-0.13165
							Σ =	2.436867

Tabel 3.12 Contoh Kemiripan Antar Kata *Query* dan Dokumen 3 Ilustrasi Perhitungan Metode

		Dokumen 3				
		كيفية	صلاة	التهجد	دفع	زكاة
Query	زكاة	0.029067	0.283702	-0.12569	0.128907	1
	الفطر	-0.00368	0.128907	-0.00368	0.283702	0.335342
					Σ =	2.056572

Tabel 3.13 Contoh Kemiripan Antar Kata *Query* dan Dokumen 4 Ilustrasi Perhitungan Metode

		Dokumen 4						
		يجوز	تقديم	زكاة	الفطر	العيد	بيوم	يومين
Query	زكاة	0.377608	0.128048	1	0.335342	0.252384	0.116681	0.039372
	الفطر	0.118809	0.016274	0.335342	1	0.548234	0.286638	0.180132
							Σ =	4.734864

Tabel 3.10, 3.11, 3.12 dan 3.13 menunjukkan kemiripan antar kata *query* dan tiap dokumen. Berdasarkan hasil kemiripan antar kata tersebut nilai sigma (yang ditunjukkan dengan blok berwarna kuning) atau jumlah nilai dari semua nilai kemiripan antar kata tiap dokumen akan dimasukkan ke dalam perhitungan kemiripan semantik.

- Perhitungan Kemiripan Semantik

Perhitungan kemiripan semantik menggunakan rumus $SemSim(Q, d_j)$ untuk mendapatkan nilai kemiripan semantik antara *query* dan dokumen. Tabel 3.14 menampilkan nilai dari masing-masing kemiripan semantik antara *query* dan tiap dokumen.

Tabel 3.14 Contoh Hasil Perhitungan Kemiripan Secara Semantik Illustrasi Perhitungan Metode

Semantik	
SemSim(Q,D1)	0.282
SemSim(Q,D2)	0.348
SemSim(Q,D3)	0.411
SemSim(Q,D4)	0.676

d. Penggabungan Perhitungan Kemiripan Leksikal dan Semantik

Setelah perhitungan kemiripan secara leksikal dan semantik selesai dilakukan, selanjutnya adalah menggabungkan hasil perhitungan kemiripan leksikal dan semantik ke dalam rumus $LSSim(Q, d_j)$. Tabel 3.15 menunjukkan nilai akhir dari kemiripan antara *query* dan dokumen. Selain itu, setelah mendapatkan nilai akhir, hasil kemiripan tiap dokumen diurutkan sesuai dengan nilainya dimulai dari yang tertinggi hingga terendah. Sehingga, tahap akhir dari metode telah selesai dilakukan.

Tabel 3.15 Contoh Hasil Penggabungan Perhitungan Kemiripan Leksikal dan Semantik Ilustrasi Perhitungan Metode

Gabungan Leksikal dan Semantik	
LSSim(Q,D1)	0.141
LSSim(Q,D2)	0.251
LSSim(Q,D3)	0.231
LSSim(Q,D4)	0.415
Setelah diurutkan	
Dokumen 4	0.415
Dokumen 2	0.251
Dokumen 3	0.231
Dokumen 1	0.141

Berdasarkan Tabel 3.15, nilai kemiripan antara *query* dan empat buah dokumen didapatkan dengan besaran nilai yang berbeda tiap pasang *query* dan dokumennya. Setelah diurutkan, dokumen 4 memiliki nilai tertinggi diikuti oleh dokumen 2, dokumen 3, dan dokumen 1 sesuai dengan besaran nilai yang dimiliki masing-masing dokumen. Hal ini menunjukkan bahwa dokumen 4 memiliki kemiripan paling dekat dengan *query*, sehingga dokumen 4 adalah dokumen paling relevan di antara dokumen lainnya.

3.2 Uji Coba dan Evaluasi

Pada tahap uji coba dan evaluasi, sistem perangkingan dokumen diuji dengan metode evaluasi *f-measure*. *F-measure* merupakan metode evaluasi kombinasi dari *precision* dan *recall*. Persamaan 3.6 menunjukkan perhitungan *f-measure*.

$$f - measure = \frac{2.precision.recall}{precision+recall}, \quad (3.6)$$

Precision adalah dokumen yang relevan diantara semua dokumen yang berhasil di-*retrieve* atau dengan kata lain ketepatan informasi yang diminta *user* terhadap hasil jawaban yang diberikan oleh sistem . *Recall* adalah dokumen yang berhasil di-*retrieve* diantara dokumen yang relevan atau dengan kata lain keberhasilan sistem dalam menampilkan ketepatan informasi. Persamaan 3.7 dan 3.8 menunjukkan persamaan dari *precision* dan *recall*.

$$precision = \frac{TP}{TP+FP}, \quad (3.7)$$

$$recall = \frac{TP}{TP+FN}, \quad (3.8)$$

dimana TP adalah *true positive* (dokumen yang dikembalikan dan relevan), FP adalah *false positive* (dokumen yang dikembalikan dan tidak relevan), dan FN adalah *false negative* (dokumen yang tidak dikembalikan dan relevan). Nilai *f-measure* berada pada rentang 0 sampai 1. Nilai *f-measure* untuk sistem temu kembali informasi yang baik adalah dengan memberikan nilai mendekati 1. Dokumen yang akan diuji adalah dokumen hasil *retrieve* oleh sistem perangkingan sebanyak 15 dokumen dari tiap *query*.

Groundtruth yang digunakan pada pengujian penelitian ini berdasarkan data yang dikaji oleh ahli di bidangnya. Data tersebut berdasarkan hasil beberapa keputusan oleh ahli seputar kajian dalam agama Islam. Skenario pengujian yang akan dilakukan adalah dengan membandingkan hasil dokumen yang di-*retrieve* oleh sistem berdasarkan *query* masukan pengguna terhadap hasil dokumen yang telah ditetapkan oleh ahli tersebut dengan *query* yang sama.

BAB 4
HASIL PENELITIAN DAN PEMBAHASAN

Pada bab ini menjelaskan hasil penelitian berupa implementasi metodologi dan pengujian beserta pembahasan. Hasil penelitian meliputi *dataset* yang digunakan, *preprocessing* dokumen, perhitungan kemiripan secara leksikal, perhitungan kemiripan secara semantik dan penggabungan perhitungan kemiripan leksikal dan semantik.

4.1 Lingkungan Uji Coba

Uji coba yang dilakukan pada penelitian ini menggunakan perangkat dengan spesifikasi perangkat lunak sebagai berikut:

e. Sistem operasi *Windows* 10
f. *Processor Intel Core* i5-8250U, *up to* 3.4 GHz
g. Kapasitas memori 8 GB
h. Kapasitas *harddisk* 1 TB

Aplikasi dan *library-library* bahasa pemrograman terkait pengerjaan sistem pada penelitian sebagai berikut:

a. Aplikasi *Microsoft Visual Studio Code*
b. Bahasa pemrograman *Python* 3 versi 3.86. *Library-library* utama meliputi *PyArabic*, *Arabic Light Stemmer*, dan *Arabic-Stopwords*, yang dikhususkan untuk mengolah *dataset* berbahasa Arab.

4.2 *Preprocessing* Dokumen

Dokumen-dokumen yang dijadikan dataset penelitian, sebelum diproses lebih lanjut ke dalam metode penelitian, dilakukan *preprocessing* atau praproses terlebih dahulu. *Preprocessing* yang dilakukan berupa tokenisasi, di mana tokenisasi berupa proses pemecahan dokumen yang jika satu dokumen direpresentasikan ke dalam satu halaman kitab maka pemecahan dokumen berupa pemecahan paragraf menjadi kumpulan kalimat, kemudian kumpulan kalimat menjadi kumpulan kata. Setelah itu, penghapusan *diacritic* atau harokat yang

terdapat pada bahasa Arab. Selanjutnya proses *stopwords removal* yaitu penghapusan kata-kata yang tidak diperlukan atau tidak penting karena jika kata-kata ini dimasukkan maka akan mempengaruhi pembobotan di mana frekuensi kemunculan kata-kata tidak penting tersebut akan muncul lebih banyak daripada kata-kata pentingnya. Proses yang terakhir adalah *stemming* yaitu mengubah suatu kata menjadi kata dasarnya, di mana dalam penelitian ini menggunakan *Arabic Light Stemmer*. Berikut Gambar 4.1 di bawah ini menunjukkan contoh hasil *preprocessing* dari *dataset*. Berdasarkan gambar tersebut bahwa terdapat sejumlah 1.064.008 dokumen yang telah di-*preprocessing* dan akan diproses ke tahap selanjutnya.

	Isi
0	...جزء في خبار شعار تب شيخ امام حافظ بو عبد له مح
1	...سم له رحمن رحيم رب عن سر يا ريم خبر شيخ امام ش
2	...حا قال : قال رسول له صل له على سلم : ستتر من ن
3	...: اخبر بو على بن خلال : خبر جعفر : خبر سلف - 3
4	... ، تجويد بن هب ، روا حجاج بن محمد ، عن بن جريج
...	...
1064004	...خبر حمد بن حسن بن يوب ثن عبد له بن محمد ب - 25
1064005	...خبر بو حفص روق بن عبد كبير سليم بن حمد حب - 26
1064006	...خبر حمد بن محمد بن سن نا بو عبد رحمن نساء ثن ق
1064007	...خبر بو حر بربهار ثن (غير ضح اصل) بن سن (ياض
1064008	... خبر بو كر حمد بن محمد بن حسن دينور نا بو - 27

Gambar 4.1 Contoh Hasil *Preprocessing* Dokumen

4.3 Perhitungan Kemiripan Secara Leksikal

Setelah *dataset* selesai dilakukan *preprocessing*, tahap selanjutnya adalah dengan melakukan perhitungan kemiripan secara leksikal. Pada perhitungan kemiripan ini, dokumen dilakukan pembobotan kata menggunakan TF.IDF. Setelah selesai pembobotan kata, selanjutnya adalah perkalian vektor antara *query* dan tiap dokumen. Dokumen direpresentasikan ke dalam satu halaman kitab (buku). Sehingga diakhir perhitungan menggunakan rumus $LexSim(Q, d_j)$,

tampilan pada sistem adalah isi per halaman kitab. Setiap dokumen masing-masing memiliki nilai $LexSim(Q, d_j)$ yang nantinya akan dimasukkan ke perhitungan penggabungan perhitungan kemiripan leksikal dan semantik. Berikut pada Gambar 4.2 dapat dilihat salah satu contoh hasil dari perhitungan kemiripan secara leksikal. Selain menampilkan tiap dokumen yaitu isi per halaman kitab, isi tersebut juga dilengkapi dengan nomor *database* kitab, kategori, nama kitab, pengarang beserta nomor halamannya. Berdasarkan gambar tersebut terlihat bahwa masing-masing dokumen memiliki nilai kemiripan leksikalnya tersendiri, di mana nilai kemiripan leksikal ini nantinya yang akan digunakan dalam penggabungan perhitungan kemiripan leksikal dan semantik. Nilai kemiripan leksikal dari masing-masing dokumen beragam sesuai dengan tingkat relevansinya terhadap *query* yang digunakan.

	No Database Kitab	Kategori	Nama Kitab	Pengarang	No Halaman	Isi	Leksim
0	10025	الأجزاء الحديثية	التذكرة للحميدي	محمد بن فتوح بن عبد الله بن فتوح بن حميد الأز...	(1/374)	جزء في خبار شعار تب شيخ امام حافظ بو عبد له مج...	0.000000
1	10025	الأجزاء الحديثية	التذكرة للحميدي	محمد بن فتوح بن عبد الله بن فتوح بن حميد الأز...	(1/376)	سم له رحمن رحيم رب عن سر يا ريم خبر شيخ امام ش...	0.000000
2	10025	الأجزاء الحديثية	التذكرة للحميدي	محمد بن فتوح بن عبد الله بن فتوح بن حميد الأز...	(1/377)	حا قال : قال رسول له صل له على سلم : ستتر من ن...	0.000000
3	10025	الأجزاء الحديثية	التذكرة للحميدي	محمد بن فتوح بن عبد الله بن فتوح بن حميد الأز...	(1/378)	اخبر بو على بن خلال : خبر - 3 جعفر : خبر سلف :...	0.000000
4	10025	الأجزاء الحديثية	التذكرة للحميدي	محمد بن فتوح بن عبد الله بن فتوح بن حميد الأز...	(1/379)	تجويد بن هب ، روا حجاج بن محمد ، عن بن جريج ...	0.000000
...	...	...	...	...	...	...	...
1064004	9853	مخطوطات حديثية	أحاديث مسندة في أبواب القضاء - مخطوط	أبو نعيم أحمد بن عبد الله بن أحمد بن إسحاق بن...	(1/17)	خبر حمد بن حسن بن يوب - 25 ثن عبد له بن محمد ب...	0.027867
1064005	9853	مخطوطات حديثية	أحاديث مسندة في أبواب القضاء - مخطوط	أبو نعيم أحمد بن عبد الله بن أحمد بن إسحاق بن...	(1/18)	خبر بو حفص روق بن عبد - 26 كبير سليم بن حمد حب...	0.000000
1064006	9853	مخطوطات حديثية	أحاديث مسندة في أبواب القضاء - مخطوط	أبو نعيم أحمد بن عبد الله بن أحمد بن إسحاق بن...	(1/19)	خبر حمد بن محمد بن سن نا بو عبد رحمن نساء ثن ق...	0.000000
1064007	9853	مخطوطات حديثية	أحاديث مسندة في أبواب القضاء - مخطوط	أبو نعيم أحمد بن عبد الله بن أحمد بن إسحاق بن...	(1/20)	خبر بو حر بربهار ثن (غير ضح اصل) بن سن (ياض...	0.000000
1064008	9853	مخطوطات حديثية	أحاديث مسندة في أبواب القضاء - مخطوط	أبو نعيم أحمد بن عبد الله بن أحمد بن إسحاق بن...	(1/21)	خبر بو كر حمد بن محمد - 27 بن حسن دينور نا بو ...	0.000000

Gambar 4.2 Contoh Hasil Perhitungan Kemiripan Leksikal

4.4 Perhitungan Kemiripan Secara Semantik

Pada perhitungan kemiripan secara semantik, tiap dokumen yang telah dilakukan *preprocessing* dicari kata-kata terdekatnya menggunakan *word embedding*. Dalam penelitian ini, *word embedding* yang digunakan adalah *fasttext*. Setelah didapatkan kata-kata terdekat di tiap dokumen, selanjutnya adalah mengukur kemiripan antara *query* dan kata-kata terdekat tiap dokumen tersebut

menggunakan rumus $SemSim(Q, d_j)$. Setiap dokumen masing-masing memiliki nilai $SemSim(Q, d_j)$ yang nantinya akan dimasukkan ke perhitungan penggabungan perhitungan kemiripan leksikal dan semantik. Berikut pada Gambar 4.3 dapat dilihat salah satu contoh hasil dari perhitungan kemiripan secara semantik. Berdasarkan Gambar 4.3, terlihat bahwa tiap dokumen memiliki nilai kemiripan secara semantiknya masing-masing. Tiap dokumen atau isi dari kitab tersebut juga dilengkapi nomor *database* kitab, kategori, nama kitab, pengarang dan nomor halaman untuk menunjukkan bahwa hasil *retrieve* bisa didapatkan dari kitab-kitab yang berbeda. Dari gambar tersebut terlihat bahwa masing-masing dokumen memiliki nilai kemiripan semantiknya tersendiri. Nilai kemiripan semantik tiap dokumen beragam sesuai dengan tingkat relevansinya terhadap *query* yang digunakan. Nilai kemiripan semantik ini nantinya akan dimasukkan ke dalam penggabungan perhitungan kemiripan leksikal dan semantik.

	No Database Kitab	Kategori	Nama Kitab	Pengarang	No Halaman	Isi	Semsim
0	10025	الأجزاء الحديثية	التذكرة للحميدي	محمد بن فتوح بن عبد الله بن فتوح بن حميد الأز...	(1/374)	جزء في خبار شعار تب شيخ امام حافظ بو عبد له مح...	0.535155
1	10025	الأجزاء الحديثية	التذكرة للحميدي	محمد بن فتوح بن عبد الله بن فتوح بن حميد الأز...	(1/376)	سم له رحمن رحيم رب عن سر يا ريم خبر شيخ امام ش...	0.560256
2	10025	الأجزاء الحديثية	التذكرة للحميدي	محمد بن فتوح بن عبد الله بن فتوح بن حميد الأز...	(1/377)	حا قال : قال رسول له صل له على سلم : ستتر من ن...	0.551559
3	10025	الأجزاء الحديثية	التذكرة للحميدي	محمد بن فتوح بن عبد الله بن فتوح بن حميد الأز...	(1/378)	اخبر بو على بن خلال : خبر - 3 : جعفر : خبر سلف ...	0.569470
4	10025	الأجزاء الحديثية	التذكرة للحميدي	محمد بن فتوح بن عبد الله بن فتوح بن حميد الأز...	(1/379)	تجويد بن هب ، روا حجاج بن ، محمد ، عن بن جريج ...	0.587291
...	...	...	...	...	...	...	...
1064004	9853	مخطوطات حديثية	أحاديث مسندة في أبواب القضاء - مخطوط	أبو نعيم أحمد بن عبد الله بن أحمد بن إسحاق بن...	(1/17)	خبر حمد بن حسن بن يوب - 25 ثن عبد له بن محمد ب...	0.566974
1064005	9853	مخطوطات حديثية	أحاديث مسندة في أبواب القضاء - مخطوط	أبو نعيم أحمد بن عبد الله بن أحمد بن إسحاق بن...	(1/18)	خبر بو حفص روق بن عبد - 26 كبير سليم بن حمد حب...	0.545786
1064006	9853	مخطوطات حديثية	أحاديث مسندة في أبواب القضاء - مخطوط	أبو نعيم أحمد بن عبد الله بن أحمد بن إسحاق بن...	(1/19)	خبر حمد بن محمد بن سن نا بو عبد رحمن نساء ثن ق...	0.557192
1064007	9853	مخطوطات حديثية	أحاديث مسندة في أبواب القضاء - مخطوط	أبو نعيم أحمد بن عبد الله بن أحمد بن إسحاق بن...	(1/20)	خبر بو حر بربهار ثن (غير ضح اصل) بن سن (ياض...	0.566513
1064008	9853	مخطوطات حديثية	أحاديث مسندة في أبواب القضاء - مخطوط	أبو نعيم أحمد بن عبد الله بن أحمد بن إسحاق بن...	(1/21)	خبر بو كر حمد بن محمد - 27 بن حسن دينور نا بو ...	0.553401

Gambar 4.3 Contoh Hasil Perhitungan Kemiripan Semantik

4.5 Penggabungan Perhitungan Kemiripan Leksikal dan Semantik

Pada penggabungan perhitungan kemiripan leksikal dan semantik, tiap dokumen dihitung kemiripannya dengan *query* menggunakan rumus

$LSSim(Q, d_j)$. Perhitungan menggunakan rumus $LSSim(Q, d_j)$ dengan melibatkan nilai alpha. Rumus $LSSim(Q, d_j)$ melibatkan nilai hasil dari rumus $LexSim(Q, d_j)$ dan nilai hasil dari rumus $SemSim(Q, d_j)$. Sebagai contoh penggunaan nilai alpha 0.5 dalam rumus $LSSim(Q, d_j)$ sebagai bagian dari penggabungan perhitungan kemiripan leksikal dan semantik, akan merata-rata antara perhitungan kemiripan leksikal dan perhitungan kemiripan semantik. Berikut Gambar 4.4 menunjukkan salah satu contoh hasil dari penggunaan alpha 0.5 dalam penggabungan perhitungan kemiripan leksikal dan semantik. Pada gambar tersebut terlihat bahwa nilai rata-rata kemiripan leksikal dan kemiripan semantik yang menjadi nilai akhir pada perhitungan $LSSim(Q, d_j)$.

	No Database Kitab	Kategori	Nama Kitab	Pengarang	No Halaman	Isi	Leksim	Semsim	LSSim
817412	32292	فقه عام	موسوعة الفقه الإسلامي	محمد بن إبراهيم بن عبد الله التويجري	(3/183)	قضاء صيام - - 12 الذين جب على صيام ... قضاء اطعام	0.315136	0.786376	0.550756
875503	21751	فقه مالكي	البيان والتحصيل والشرح والتوجيه والتعليل لمسا...	أبو الوليد محمد بن أحمد بن رشد القرطبي	(2/340)	قضاء ومي ، قضي وم يوم اول الذي كان ابتد ...، صيام	0.291431	0.801067	0.546249
815220	32004	فقه عام	مختصر الفقه الإسلامي في ضوء القرآن والسنة	محمد بن إبراهيم بن عبد الله التويجري	(1/634)	قال : لا ، قال : « هل جد ما طعم ين مسك ؟» ... قال	0.312225	0.775363	0.543794
438946	31252	بحوث ومسائل	تحفة الإخوان بأجوبة مهمة تتعلق بأركان الإسلام	عبد العزيز بن عبد الله بن باز	(1/157)	من جب] [رابع صيام] على صيام رمض فضل صيام صي...	0.277439	0.810096	0.543767
439922	31285	بحوث ومسائل	المختصر في فقه العبادات	خالد بن علي بن محمد بن حمود بن علي المشيقح	(1/89)	س : مت قضي من فطر في رمض ؟ ج : جب قضاء قبل حلو...	0.280256	0.792873	0.536565
438961	31252	بحوث ومسائل	تحفة الإخوان بأجوبة مهمة تتعلق بأركان الإسلام	عبد العزيز بن عبد الله بن باز	(1/173)	مرض او سفر لي رمض خر على قضاء قط دون اطعام بعد...	0.246063	0.794274	0.520169
864506	21611	فقه مالكي	التاج والإكليل لمختصر خليل	محمد بن يوسف بن أبي القاسم بن يوسف العبدري ال...	(3/385)	مذهب قول (زمن بيح صوم) لخم : قضاء ... رمض صح في	0.270870	0.768776	0.519823

Gambar 4.4 Contoh Hasil Penggabungan Perhitungan Kemiripan Leksikal dan Semantik

4.6 Contoh Hasil *Retrieve* Dokumen Masing-Masing Metode

Pada subbab ini akan menampilkan contoh hasil retrieve dokumen dari sebuah *query* yang hanya menggunakan kemiripan leksikal, kemiripan semantik, serta gabungan kemiripan leksikal dan semantik. Hasil dokumen tersebut ditunjukkan pada Tabel 4.1 dengan metode kemiripan leksikal, Tabel 4.2 dengan metode kemiripan semantik, Tabel 4.3 dengan metode gabungan kemiripan leksikal dan semantik.

Tabel 4.1 Contoh Hasil *Retrieve* Dokumen Kemiripan Leksikal

Query	Metode	Hasil *Retrieve* Dokumen
إِخْرَاجُ زَكَاةِ ٱلفِطْرِ (membayar zakat fitrah)	Kemiripan Leksikal	فقه عام, موسوعة الفقه الإسلامي, محمد بن إبراهيم بن عبد الله ,32292 التويجري,(5/3),4 - كتاب الزكاة ويشتمل على ما يلي: 1 - الزكاة المفروضة، وتشتمل على ما يلي: 1 - الأموال التي تجب فيها الزكاة وتشمل: 1 - زكاة الذهب والفضة. 2 - زكاة الأوراق النقدية. 3 - زكاة عروض التجارة. 4 - زكاة بهيمة الأنعام. 5 - زكاة الحبوب والثمار. 6 - زكاة الركاز. 7 - زكاة المعادن. 2 - إخراج الزكاة. 3 - آداب إخراج الزكاة. 4 - أهل الزكاة. 2 - زكاة الفطر. . **Terjemahan:** 32292, Yurisprudensi Umum, Encyclopedia of Islamic Fiqh, Muhammad bin Ibrahim bin Abdullah Al-Tuwaijri, (3/5), 4 - Buku zakat dan meliputi: 1 - Zakat yang dikenakan, meliputi: 1 - Dana yang dikeluarkan zakatnya jatuh tempo, termasuk 1 - Zakat emas dan perak. 2 - Zakat uang kertas. 3- Zakat perdagangan barang. 4 - Zakat ternak. 5- Zakat gandum dan buah-buahan. 6 - Zakat bijih. 7 - Zakat mineral. 2 - Membayar zakat. 3 - Etika membayar zakat. 4 - Orang Zakat. 2 - Zakat Fitrah.

Tabel 4.2 Contoh Hasil *Retrieve* Dokumen Kemiripan Semantik

Query	Metode	Hasil *Retrieve* Dokumen
إِخْرَاجُ زَكَاةِ ٱلفِطْرِ (membayar zakat fitrah)	Kemiripan Semantik	بحوث ومسائل أحكام, عبد الله بن صالح ,30926 القصير,(65/1),[الفصل الثاني في مهمات من أحكام زكاة الفطر] [معنى زكاة الفطر] الفصل الثاني في مهمات من أحكام زكاة الفطر 1 - معنى زكاة الفطر. 2 - تاريخ مشروعيتها والدليل عليها. 3 - حكمها. 4 - حكمة مشروعيتها. 5 - على من تجب الفطرة. 6 - أنواع الأطعمة التي تخرج منها زكاة الفطر. 7 - المقدار الواجب في الفطرة. 8 - وقت إخراج الزكاة. 9 - لمن تعطى صدقة الفطر. 10 - إخراج قيمة زكاة الفطر. 11 - نقل زكاة الفطر من بلد الشخص إلى بلد آخر **Terjemahan:** 30926, Riset dan Soal, Abdullah Bin Saleh Al-Qusayr (1/65), [Bab Dua Tentang Tugas Ketentuan Zakat Fitrah] [The Makna Zakat Fitrah] Bab Dua Tugas dari Ketentuan Zakat Fitrah 1 - Pengertian Zakat Fitrah. 2 - Tanggal dan bukti legalitasnya. 3- Aturannya. 4 - Kebijaksanaan legitimasinya. 5- Pada siapa fitrah diperlukan. 6- Jenis makanan yang mengeluarkan zakat fitrah. 7 - Jumlah yang harus dibayar dalam insting. 8 - Saatnya membayar zakat. 9 - Kepada siapa diberikan shadaqat fitrah. 10 - Membayar Zakat Fitrah. 11 - Transfer zakat fitrah dari satu negara ke negara lain.

Tabel 4.3 Contoh Hasil *Retrieve* Dokumen Gabungan Kemiripan Leksikal dan Semantik

Query	Metode	Hasil *Retrieve* Dokumen
إِخْرَاجُ زَكَاةِ ٱلفِطْرِ (membayar zakat fitrah)	Gabungan Kemiripan Leksikal dan Semantik	فقه عام, موسوعة الفقه الإسلامي, محمد بن إبراهيم بن عبد ,32292 الله التويجري,(92/3),4 - يجوز تقديم زكاة الفطر قبل العيد بيوم أو يومين؛ لأن بعض الصحابة رضي الله عنهم كانوا يعطونها قبل الفطر بيوم أو يومين. - حكم إخراج زكاة الفطر بعد خروج وقتها: 1 - زكاة الفطر عبادة من العبادات، ولها وقت يجب أداؤها فيه، ويحرم تأخيرها عن وقتها إلا لعذر. 2 - زكاة الفطر لا تسقط بعد خروج وقتها؛ لأنها حق واجب للفقراء في ذمته، فلا يسقط عنه إلا بالأداء. أما حق الله في التأخير عن وقتها فلا يسقط إلا بالتوبة والاستغفار. - مكان إخراج زكاة الفطر: زكاة المال تُخرج في بلد المال، وزكاة الفطر تُخرج حيثما كان الإنسان، ولا يُعدل عن ذلك إلا لحاجة ومصلحة **Terjemahan:** 32292, Fikih Umum, Encyclopedia of Islamic Fiqih, Muhammad bin Ibrahim bin Abdullah Al-Tuwaijri, (3/92), 4 - Zakat fitrah dapat diberikan satu atau dua hari sebelum Idul Fitri; Karena sebagian sahabat, radhiyallahu 'anhu, biasa memberikannya satu atau dua hari sebelum Idul Fitri. - Hukum mengeluarkan zakat fitrah setelah lewat waktunya: 1 - Zakat fitrah adalah ibadah yang ada waktunya wajib dibayarnya, dan dilarang menunda melebihi waktunya kecuali mengizinkan. 2 - Zakat fitrah tidak dihapuskan setelah waktunya habis. Karena itu adalah hak yang wajib bagi orang miskin, dan tidak dapat dicabut kecuali dengan pembayaran. Adapun hak Allah untuk terlambat pada waktunya tidak akan padam kecuali dengan taubat dan mohon ampun. Di mana membayar zakat fitrah: zakat uang diberikan di negara uang, dan zakat fitrah diberikan di mana pun seseorang berada, dan dia tidak mengubahnya kecuali untuk kebutuhan dan manfaat.

4.7 Uji Coba, Evaluasi dan Analisa Hasil Penelitian

4.7.1 Uji Coba dan Evaluasi Penelitian

Penelitian ini melakukan uji coba terhadap 20 *query* berbahasa Arab. Topik yang digunakan pada *query* beragam, di mana proses penterjemahan *query* ke dalam bahasa Arab membutuhkan pakar yang mampu berbahasa Arab. Hasil *retrieve* dokumen pencarian yang dihasilkan oleh sistem divalidasi oleh dua orang

pakar yang mampu berbahasa Arab serta berkualifikasi sebagai dosen bahasa dan sastra Arab di perguruan tinggi di kota Makassar. Selain itu, penentuan jumlah kitab sebanyak 2900 kitab dengan rumpun ilmu yang beragam juga membutuhkan pakar dalam penggunaannya pada penelitian ini. Setiap kitab/buku memiliki topik bahasannya masing-masing. Namun, penelitian ini terdapat batasan masalah penelitian yaitu dataset kitab-kitab tersebut tidak dilabeli berdasarkan topik bahasan dari setiap kitab/buku tersebut.

Evaluasi terhadap metode usulan dalam penelitian menggunakan *recall*, *precision*, dan *f-measure*. Skenario uji coba terbagi ke dalam tiga bagian yaitu *all query*, *short query*, dan *long query*. *All query* merupakan pengujian yang dilakukan untuk melihat hasil evaluasi terhadap semua *query* secara menyeluruh. Kemudian, *short query* merupakan frekuensi kata dalam *short query* berjumlah kurang dari sama dengan 2 kata. Sedangkan, *long query* merupakan uji coba terhadap frekuensi kata dalam *query* dengan jumlah lebih dari 2 kata.

Seluruh skenario uji coba tersebut dengan membandingkan antara kemiripan leksikal, kemiripan semantik, dan gabungan kemiripan leksikal dan semantik. Nilai alpha yang digunakan pada gabungan kemiripan leksikal dan semantik adalah 0.5. Selain itu, evaluasi juga dilihat dari hasil pencarian sebanyak 5, 10 dan 15 teratas. Seluruh 20 *query* uji coba dapat dilihat pada Tabel 4.4 yang disertai dengan terjemahannya. Namun, pada penelitian *query* yang dijadikan uji coba hanya yang berbahasa Arab saja tanpa terjemahannya.

Skenario uji coba yang dilakukan terhadap *short query* dan *long query* terdapat 12 *query* yang tergolong ke dalam *query* dengan frekuensi kata pendek (*short query*) dan 8 *query* tergolong ke dalam *query* dengan frekuensi kata panjang (*long query*). Uji coba terhadap *short query* dan *long query* dilakukan untuk mengukur performa *query* perangkingan dokumen. Salah satu contoh *short query* misalnya "غَسْلُ ٱلْمَيِّتِ" yang berarti "memandikan jenazah". Contoh *long query* misalnya "صِيَامُ 6 أَيَّامٍ فِى شَوَّال" yang artinya "puasa 6 hari di bulan Syawal". Pembagian *query* secara keseluruhan antara *short query* dan *long query* dapat dilihat pada Tabel 4.5. Sebelumnya telah dijelaskan, *query* merupakan kata kunci penting dalam pencarian suatu dokumen. Jika *query* yang dimasukkan tidak dapat

mewakili keseluruhan dokumen yang akan dicari maka dokumen hasil pencarian yang ditampilkan menjadi kurang relevan.

Tabel 4.4 Seluruh *Query* Uji Coba

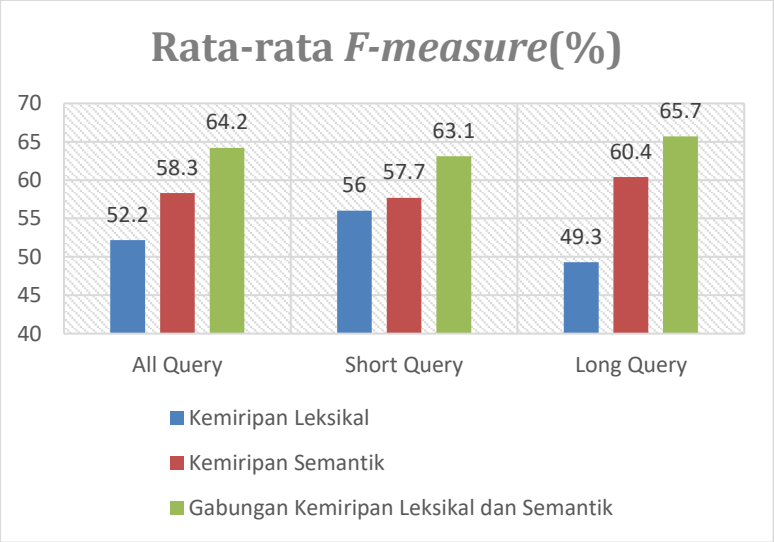
ID Query	*Query*	Terjemahan
Q1	ذِبْحُ ٱلْأَضَاحِي	Menyembelih Hewan Qurban
Q2	صِيَامُ 6 أَيَّامٍ فِى شَوَّال	Puasa 6 Hari di bulan Syawal
Q3	الطَّوَافُ فِى ٱلْحَجِّ	Tawaf Ibadah Haji
Q4	رَمْيُ ٱلْجُمُرَاتِ فِى ٱلْحَجِّ	Lempar Jumroh Ibadah Haji
Q5	المَحَارِمُ لِلْمَرْأَةِ	Mahram Anak Perempuan
Q6	المَحَارِمُ لِلرِّجَالِ	Mahram Anak Laki-Laki
Q7	صَوْمُ التَّطَوُّعِ	Puasa-Puasa Sunah
Q8	مُبْطِلَاتُ الصَّلَاةِ	Hal yang Membatalkan Sholat
Q9	شُرْبُ ٱلْخَمْرِ	Meminum Minuman Keras
Q10	أَكْلُ لَحْمِ ٱلْخِنْزِيْرِ	Memakan Daging Babi
Q11	أَرْكَانُ ٱلْإِيْمَانِ	Rukun Iman
Q12	العَقِيْقَةُ لِلْبَنَاتِ	Aqiqah Anak Perempuan
Q13	صَلَاةِ ٱلْجَنَازَةِ	Sholat Jenazah
Q14	صَلَاةِ التَّهَجُّدِ	Sholat Tahajud
Q15	غَسْلُ ٱلْمَيِّتِ	Memandikan Jenazah
Q16	الصَّلَوَاتُ ٱلْخَمْسُ ٱلْمَفْرُوْضَةُ	Shalat Fardhu Lima Waktu
Q17	إِخْرَاجُ زَكَاةِ ٱلْفِطْرِ	Membayar Zakat Fitrah
Q18	أركان الوضوء	Rukun Wudhu
Q19	قضاء صيام رمضان	Qada Puasa Ramadhan
Q20	الإفطار في الطيارة	Berbuka Puasa di atas Pesawat

Tabel 4.5 *Short Query* dan *Long Query*

Short Query		*Long Query*	
Q1	ذِبْحُ ٱلْأَضَاحِي	Q2	صِيَامُ 6 أَيَّامٍ فِى شَوَّال
Q5	المَحَارِمُ لِلْمَرْأَةِ	Q3	الطَّوَافُ فِى ٱلْحَجِّ
Q6	المَحَارِمُ لِلرِّجَالِ	Q4	رَمْيُ ٱلْجُمُرَاتِ فِى ٱلْحَجِّ
Q7	صَوْمُ التَّطَوُّعِ	Q10	أَكْلُ لَحْمِ ٱلْخِنْزِيْرِ
Q8	مُبْطِلَاتُ الصَّلَاةِ	Q16	الصَّلَوَاتُ ٱلْخَمْسُ ٱلْمَفْرُوْضَةُ
Q9	شُرْبُ ٱلْخَمْرِ	Q17	إِخْرَاجُ زَكَاةِ ٱلْفِطْرِ
Q11	أَرْكَانُ ٱلْإِيْمَانِ	Q19	قضاء صيام رمضان
Q12	العَقِيْقَةُ لِلْبَنَاتِ	Q20	الإفطار في الطيارة
Q13	صَلَاةِ ٱلْجَنَازَةِ		
Q14	صَلَاةِ التَّهَجُّدِ		
Q15	غَسْلُ ٱلْمَيِّتِ		
Q18	أركان الوضوء		

Berdasarkan Gambar 4.5, rata-rata *f-measure* yang dihasilkan dari uji coba *all query*, *short query* dan *long query* metode usulan yaitu penggabungan

perhitungan kemiripan leksikal dan semantik lebih unggul dari metode lainnya yaitu dengan rata-rata *f-measure* sebesar 64.2% pada keseluruhan *query (all query)*, 63.1% dari uji coba *short query*, serta 65.7% pada *long query*. Hal tersebut menunjukkan bahwa metode usulan baik pada frekuensi kata yang sedikit dalam hal ini antara 1 sampai 2 kata ataupun frekuensi kata pada *query* yang berjumlah lebih dari 2 sama-sama mampu menghasilkan pencarian dokumen yang lebih relevan terhadap inputan *query*.



Gambar 4.5 Rata-Rata *F-measure* Uji Coba *Query*

Tabel 4.6 Rata-Rata *Recall* dan *Precision Short Query*

	Recall(%) *Short Query*			*Precision*(%) *Short Query*		
	Top 5	Top 10	Top 15	Top 5	Top 10	Top 15
Kemiripan Leksikal	43.19%	67.06%	80.81%	66.67%	51.67%	42.22%
Kemiripan Semantik	42.70%	74.92%	81.96%	66.67%	58.33%	43.33%
Gabungan Kemiripan Leksikal dan Semantik	53.91%	74.92%	90.64%	83.33%	58.33%	47.78%

Tabel 4.7 Rata-Rata *Recall* dan *Precision Long Query*

	Recall(%) Long Query			*Precision(%) Long Query*		
	Top 5	Top 10	Top 15	Top 5	Top 10	Top 15
Kemiripan Leksikal	35.48%	56.88%	71.98%	62.50%	50.00%	42.50%
Kemiripan Semantik	42.34%	73.85%	86.61%	70.00%	66.25%	50.83%
Gabungan Kemiripan Leksikal dan Semantik	54.29%	74.63%	85.03%	92.50%	66.25%	50.83%

Tabel 4.8 Rata-Rata *Recall* dan *Precision All Query*

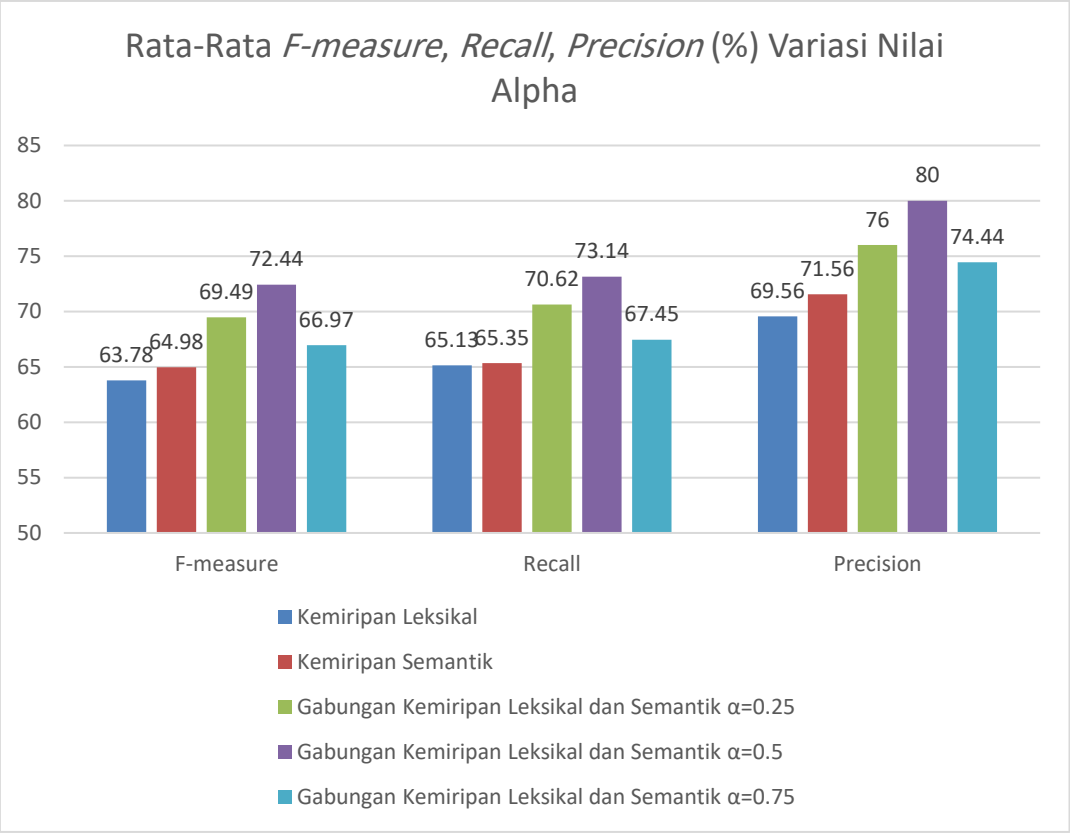
	Recall(%) All Query			*Precision(%) All Query*		
	Top 5	Top 10	Top 15	Top 5	Top 10	Top 15
Kemiripan Leksikal	40.10%	62.99%	77.28%	65.00%	51.00%	42.33%
Kemiripan Semantik	42.56%	74.49%	83.82%	68.00%	61.50%	46.33%
Gabungan Kemiripan Leksikal dan Semantik	54.06%	74.80%	88.40%	87.00%	61.50%	49.00%

Tabel 4.6 menampilkan *recall* dan *precision* terhadap 12 *short query* uji coba. Gabungan kemiripan leksikal dan semantik mendominasi *recall* pada 5 dan 15 teratas sebesar 53.91% dan 90.64%. Tidak hanya pada *recall*, *precision* tertinggi juga didominasi oleh gabungan kemiripan leksikal dan semantik pada hasil pencarian 5 dan 15 teratas yakni 83.33% dan 47.78% namun memiliki *precision* yang sama besar pada hasil pencarian 10 teratas dengan kemiripan semantik yaitu 58.33%. Hal tersebut dikarenakan pada kedua metode, jika dirata-rata me-*retrieved* hasil dokumen relevan yang sama jumlahnya terhadap dokumen yang di-*retrieved*.

Selanjutnya, Tabel 4.7 menampilkan *recall* dan *precision* dari 8 uji coba *long query*. Pada *long query*, *recall* tertinggi sebesar 54.29% dan 74.63% terletak pada hasil pencarian 5 dan 10 teratas dari gabungan kemiripan leksikal dan semantik. Sedangkan pada *precision*, gabungan kemiripan leksikal dan semantik mengungguli metode lainnya sebesar 92.50% pada hasil pencarian 5 teratas, sedangkan pada hasil pencarian 10 dan 15 teratas memiliki *precision* yang sama besar dengan kemiripan semantik. Selain itu Tabel 4.8 menunjukkan hasil *recall* dan *precision* di tiap 5,10, dan 15 pencarian teratas keseluruhan *query*. Berdasarkan tabel tersebut, gabungan kemiripan leksikal dan semantik memiliki nilai tertinggi pada hasil pencarian 5, 10, dan 15 teratas dengan *recall* sebesar 54.06%, 74.80%, dan 88.40% serta *precision* sebesar 87% dan 49% pada hasil pencarian 5 dan 15 teratas. Hasil *recall*, *precision* dan *f-measure* dari masing-masing *query* baik pada *short query* ataupun *long query* dapat dilihat pada lampiran 2, 3 dan 4.

Selain uji coba terhadap *short query* dan *long query*, pengujian juga dilakukan terhadap variasi nilai alpha pada penggabungan perhitungan kemiripan

leksikal dan semantik. Variasi nilai alpha dari metode usulan dibandingkan dengan yang hanya menggunakan kemiripan leksikal saja dan kemiripan semantik saja. Gambar 4.6 menunjukkan hasil perbandingan variasi nilai alpha pada metode usulan dan metode lainnya.



Gambar 4.6 Rata-Rata *F-measure*, *Recall*, *Precision* Variasi Nilai Alpha

Berdasarkan Gambar 4.6, pengujian variasi nilai alpha dilakukan pada 5 *query* yang terdiri dari *short query* dan *long query*. Dari hasil tersebut didapatkan bahwa metode usulan dengan sejumlah variasi nilai alpha mampu mengungguli metode lainnya baik pada *f-measure*, *recall*, dan *precision*. Secara spesifik, metode usulan yaitu gabungan kemiripan leksikal dan semantik dengan nilai alpha 0.5 memiliki hasil evaluasi yang lebih tinggi daripada variasi nilai alpha lainnya yaitu 0.25 dan 0.75. Hal tersebut dikarenakan variasi nilai alpha 0.25 akan memuat porsi nilai kemiripan semantik sebanyak 75% dan kemiripan leksikal sebanyak 25% pada proses perhitungan rumus $LSSim(Q, d_j)$.

Kemudian pada variasi nilai alpha 0.75 memuat porsi nilai kemiripan leksikal sebanyak 75% dan kemiripan semantik sebanyak 25% pada proses perhitungan rumus $LSSim(Q, d_j)$. Sedangkan pada variasi nilai alpha 0.5 baik kemiripan leksikal ataupun kemiripan semantik memiliki porsi yang sama yaitu masing-masing sebesar 50%. Porsi nilai yang lebih besar di salah satu perhitungan kemiripan akan membuat hasil evaluasi cenderung ke salah satu kemiripan dengan dengan porsi lebih besar tersebut. Sebagai contoh pada variasi nilai alpha 0.75 yang memuat porsi nilai kemiripan leksikal lebih banyak dari pada kemiripan semantik akan menghasilkan evaluasi yang cenderung menurun karena porsi nilai kemiripan leksikalnya lebih besar. Sebaliknya juga demikian pada variasi nilai alpha 0.25.

Namun pada variasi nilai alpha 0.5 akan memuat porsi nilai yang seimbang baik pada kemiripan leksikal dan kemiripan semantik, sehingga hasil evaluasi cenderung lebih tinggi karena membagi porsi nilai kemiripan di antara keduanya yang sama besar pada perhitungan rumus $LSSim(Q, d_j)$ yakni penggabungan perhitungan kemiripan leksikal dan semantik.

4.7.2 Analisa Hasil Penelitian

Berdasarkan hasil evaluasi seluruh *query* uji coba, terdapat satu buah *query* yang menghasilkan *recall*, *precision* dan *f-measure* paling rendah di antara lainnya. *Query* tersebut adalah "صَوْمُ التَّطَوُّعِ" yang berarti puasa sunnah. Hasil evaluasi yang rendah tersebut mempengaruhi rata-rata keseluruhan hasil evaluasi *query*. Rendahnya hasil evaluasi *query* tersebut disebabkan karena dimungkinkan tidak ada dokumen yang relevan pada kumpulan dokumen pencarian.

Penelitian ini memiliki sejumlah batasan masalah di antaranya adalah *dataset* yang tidak digolongkan ke dalam topik-topik tertentu (sebagai contoh: topik puasa sunnah) sehingga membuat adanya *query* yang tidak mendapatkan dokumen pencarian yang relevan. Seharusnya agar setiap *query* mendapatkan dokumen pencarian yang relevan, dataset terlebih dahulu digolongkan ke dalam topik yang sama dengan *query*. Sehingga, hasil evaluasi dari setiap *query* dapat optimal.

Selain itu, analisa lainnya yang menyebabkan *query* "صَوْمُ التَّطَوُّعِ" memiliki hasil evaluasi yang rendah dapat disebabkan karena pemilihan kata dalam *query* yang kurang tepat. Meskipun *query* "صَوْمُ التَّطَوُّعِ" dapat diartikan sebagai puasa sunnah, jika dua kata pada *query* tersebut dipecah dan diartikan per masing-masing kata, keterkaitan atau kemiripan antara dua kata tersebut dapat dikatakan cukup jauh kemiripannya. Kata "التَّطَوُّعِ" jika diartikan tersendiri tanpa diikuti kata "صَوْمُ" secara berdampingan dapat berarti "sukarela". Kata "التَّطَوُّعِ" yang berarti "sukarela" ini jika diukur kemiripannya dengan kata "صَوْمُ" yang berarti "puasa" kemiripannya menjadi cukup jauh. Karena kata "التَّطَوُّعِ" diartikan sebagai "sukarela" ketika diartikan tersendiri tanpa digabung dengan kata "صَوْمُ" maka dokumen-dokumen pencarian yang akan dimunculkan adalah yang berkaitan dengan kata "sukarela".

Dokumen-dokumen pencarian yang berkaitan dengan kata "sukarela" yang dimunculkan sebagai hasil pencarian bisa dokumen-dokumen yang berkaitan dengan "hal yang bersifat sukarela", di mana dokumen-dokumen tersebut tidak ada keterkaitan dengan "puasa sunnah". Contoh dokumen yang tidak berkaitan dengan “puasa sunnah” dapat dilihat pada Gambar 4.7. Gambar tersebut menampilkan dokumen yang berkaitan dengan “hal yang bersifat sukarela”. Dokumen tersebut berarti “sedekah secara sukarela”, di mana tidak berkaitan dengan “puasa sunnah”. Padahal dokumen yang seharusnya di-*retrieve* teratas mengenai “puasa sunnah” berada pada hasil retrieve lebih rendah daripada dokumen yang tidak relevan tersebut. Dokumen mengenai “puasa sunnah” tersebut ditunjukkan pada Gambar 4.8.

كتب الألباني, التعليقات الحسان على صحيح ابن حبان وتمييز سقيمه من ,31591 صحيحه، وشاذه من محفوظه, أبو عبد الرحمن محمد ناصر الدين، بن الحاج نوح بن نجاتي بن آدم، الأشقودري الألباني ,(218/5),9 - بَابُ صَدَقَةِ **التَّطَوُّعِ**

Gambar 4.7 Dokumen Tidak Relevan Dari *Query* Puasa Sunnah

فقه مالكي, فقه العبادات على المذهب المالكي, الحاجّة كوكب ,12978
عبيد,(324/1),8- يكره **صوم** ست من شوال لمن كان يُقتدى به، إن صامها
متتابعة ومتصلة بيوم العيد وأظهر صومها لكي لا يعتقد العامة وجوبها، أما إن
اختل شرط من هذه الشروط فلا يكره صومها. رابعاً- **الصوم** المُحرَّم

Gambar 4.8 Dokumen Relevan Dari *Query* Puasa Sunnah

Dari Gambar 4.7 terdapat kata bercetak tebal "**التَّطَوُّعِ**" yang berarti "sukarela", di mana dalam dokumen tersebut hanya terdapat kata tersebut yang sesuai dengan *query*. Sedangkan, pada Gambar 4.8, pada dokumen tersebut hanya terdapat kata "**صوم**" yang berarti "puasa" di mana hanya ada kata tersebut yang sesuai dengan *query*.

Berdasarkan rendahnya hasil evaluasi dari *query* "صَوْمُ التَّطَوُّعِ" dalam me-*retrieve* dokumen pencarian, jika suatu kata kunci pencarian dalam bahasa arab dialihbahasakan ke bahasa lain atau sebaliknya secara tekstual tanpa memperhatikan konteks maka akan mengalami reduksi makna. Selain itu, dokumen kitab seperti hadis memiliki bahasa Arab yang berbeda dari dokumen berbahasa Arab pada umumnya. Bahasa arab yang digunakan dalam dokumen kitab-kitab tersebut menggunakan bahasa Arab klasik. Bahasa Arab klasik ini juga disebut dengan "fusha" yang memiliki arti asli atau murni. Sedangkan pada bahasa Arab pada umumnya disebut amiyyah, yang digunakan dalam kehidupan sehari-hari (Hamsiati, 2019). Tidak hanya itu, morfologi pembentuk kata dalam bahasa Arab sendiri yang kompleks membuat dalam satu makna kata dapat memiliki beragam bentuk serta pengaplikasiannya di beragam tempat atau dokumen berbahasa Arab yang berbeda-beda.

Inilah yang menyebabkan pemilihan kata dalam *query* yang kurang tepat akan menampilkan hasil pencarian dokumen yang kurang relevan khususnya dokumen-dokumen kitab seperti hadis. Meskipun penggabungan perhitungan kemiripan leksikal dan semantik mampu menangkap makna kata yang sama dari bentuk kata berbeda, tetapi jika suatu kata tersebut bukan kata-kata yang dipakai di dalam dokumen seperti hadis maka akan tetap menghasilkan dokumen yang kurang relevan.

Oleh karena itu pencarian dokumen berbahasa Arab khususnya dokumen-dokumen kitab yang berisi kajian Islam dibutuhkan penguasaan bahasa Arab oleh pengguna yang tidak hanya mampu melakukan pencarian menggunakan *query* berbahasa Arab tetapi juga memahami konteks dari dokumen yang akan dicari. Sehingga pemilihan kata dalam *query* yang tepat akan menghasil dokumen yang lebih relevan.

Berdasarkan rata-rata *f-measure* keseluruhan *query*, *short query* dan *long query* yang ditunjukkan pada Gambar 4.5, penggabungan perhitungan kemiripan leksikal dan semantik memberikan hasil yang lebih unggul dalam memberikan informasi yang lebih relevan. Hal ini dikarenakan pada penggabungan perhitungan kemiripan leksikal dan semantik tidak hanya menggunakan pembobotan kata dalam merangking dokumen. Pembobotan kata (kemiripan leksikal) dalam perangkingan dokumen hanya akan me-*retrieve* dokumen yang berisi kata yang secara lafal sama. Hal tersebut akan membuat suatu kata yang memiliki bentuk berbeda namun bermakna sama tidak akan di-*retrieve* sebagai hasil pencarian dokumen. Sebagai contoh kata kata "ذَبَحَ" yang mempunyai arti kata menyembelih. Bentuk lain dari kata tersebut adalah "يَذْبَحُ", dan "نَحَرَ" yang juga punya arti yang sama yaitu menyembelih. Jika hanya melihat dari susunan karakter pembentuk kata dalam hal ini kemiripan leksikal, dua bentuk lain dari kata "ذَبَحَ" tidak akan di-*retrieved* sebagai hasil pencarian dokumen. Oleh karena itu, penggunaan kemiripan semantik bersamaan dengan kemiripan leksikal tidak hanya menangkap kata dengan lafal yang sama tetapi juga akan menangkap makna kata yang sama dari suatu kata meskipun berbeda bentuk penyusun karakternya. Sehingga hasil pencarian dokumen akan menghasilkan lebih banyak dokumen relevan.

BAB 5
KESIMPULAN DAN SARAN

Pada bab ini memberikan penjelasan mengenai beberapa hasil kesimpulan atas penelitian yang dilakukan, serta memberikan saran yang dapat digunakan untuk pengembangan penelitian lebih lanjut.

5.1 Kesimpulan

Penelitian ini menggunakan metode penggabungan perhitungan kemiripan leksikal dan kemiripan semantik untuk merangking dokumen berbahasa Arab, di mana mampu memberikan hasil *retrieved* yang relevan terhadap *query* inputan *user*. Berdasarkan hasil pengujian terhadap 20 *query* berbahasa Arab sebagai bahan untuk inputan *query*, didapatkan beberapa kesimpulan sebagai berikut:

1. Hasil uji coba dan evaluasi yang hanya menggunakan kemiripan leksikal saja didapatkan *f-measure* sebesar 52.2%, 56.0%, 49.3% pada skenario uji coba *all query*, *short query* dan *long query*. Kemiripan leksikal memiliki hasil evaluasi paling rendah di antara kedua metode lainnya. Hal ini disebabkan kemiripan leksikal tidak mampu dalam menangkap makna suatu kata.
2. Hasil uji coba dan evaluasi yang hanya menggunakan kemiripan semantik saja didapatkan *f-measure* sebesar 58.3%, 57.7%, 60.4% pada skenario uji coba *all query*, *short query* dan *long query*. Kemiripan semantik berada pada tingkat yang lebih tinggi dari kemiripan leksikal karena mampu dalam menangkap makna kata. Namun meskipun demikian, kemiripan tersebut hanya mempertimbangkan makna kata saja tanpa mempertimbangkan lafal suatu kata. Hal ini yang menyebabkan kemiripan semantik tergolong masih rendah dalam me-*retrieve* hasil dokumen yang lebih relevan.
3. Hasil uji coba dan evaluasi gabungan kemiripan leksikal dan semantik didapatkan *f-measure* sebesar 64.2%, 63.1%, 65.7% pada skenario uji coba *all query*, *short query* dan *long query*. Gabungan kemiripan leksikal dan semantik memiliki hasil evaluasi yang lebih tinggi daripada kedua metode lainnya

disebabkan karena metode tersebut tidak hanya menggunakan lafal suatu kata (kemiripan secara leksikal) tetapi juga mempertimbangkan makna kata (kemiripan secara semantik) dalam me-*retrieve* dokumen pencarian. Sehingga, hasil dokumen pencarian yang menggunakan metode gabungan kemiripan leksikal dan semantik mampu menghasilkan dokumen relevan lebih banyak daripada kedua metode lainnya.

5.2 Saran

Saran yang dapat diberikan untuk pengembangan penelitian lebih lanjut adalah penggunaan dataset selain bahasa Arab. Selain itu, *dataset* dapat digolongkan terlebih dahulu ke dalam topik-topik tertentu begitu pula dengan *query* agar hasil pencarian dokumen lebih relevan dan hasil evaluasi lebih optimal.

DAFTAR PUSTAKA

Abdi, A., Mariyam, S. and Aliguliyev, R. M. (2018) 'QMOS : Query-based multi-documents opinion-oriented summarization', *Information Processing and Management*. Elsevier, 54(2), pp. 318–338. doi: 10.1016/j.ipm.2017.12.002.

Abriania, G. U. and Yaqin, M. A. (2019) 'Analisis Implementasi Metode Semantic Similarity untuk Pengukuran Kemiripan Makna Antar Kalimat', *ILKOMNIKA: Journal of Computer Science and Applied Informatics*, 1(2), pp. 47–57.

Al-barhamtoshy, H. M. and Jambi, K. M. (2021) 'Arabic Documents Information Retrieval for Printed , Handwritten , and Calligraphy Image', *IEEE Access*, 9, pp. 51242–51257. doi: 10.1109/ACCESS.2021.3066477.

Alghamdi, N. and Assiri, F. (2020) 'A Comparison of fastText Implementations Using Arabic Text Classification', in *Intelligent Systems and Applications Proceedings of the 2019 Intelligent Systems Conference (IntelliSys)*, pp. 306–311.

Alhanjouri, M. (2017) 'Pre Processing Techniques for Arabic Documents Clustering', *International Journal of Engineering and Management Research*, (2), pp. 70–79.

Almarwani, N. and Diab, M. (2017) 'GW QA at SemEval-2017 Task 3 : Question Answer Re-ranking on Arabic Fora', in *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*, pp. 344–348.

Aris, N. (2015) 'Digital Library: Mengenal Al-Maktabah Al-Syamilah', *LIBRARIA: Jurnal Perpustakaan*, 3(2), pp. 166–180. doi: http://dx.doi.org/10.21043/libraria.v3i2.

Bojanowski, P. *et al.* (2017) 'Enriching Word Vectors with Subword Information', *Transactions of the Association for Computational Linguistics*, pp. 135–146. doi: 10.1162/tacl_a_00051.

Bounhas, I. (2019) 'On the Usage of a Classical Arabic Corpus as a Language Resource : Related Research and Key Challenges', *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 18(3), pp. 1–45.

Fauzi, M. A., Arifin, A. Z. and Yuniarti, A. (2017) 'Arabic Book Retrieval using Class and Book Index Based Term Weighting', *International Journal of Electrical and Computer Engineering (IJECE)*, 7(6), pp. 3705–3710. doi: 10.11591/ijece.v7i6.pp3705-3711.

Grave, E. *et al.* (2018) 'Learning Word Vectors for 157 Languages', in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 1–5.

Hajeer, I. S. *et al.* (2017) 'A Hybrid Usage-Based Ranking for Enhancing Arabic Search Engines', *International Journal of Soft Computing*, 12(4), pp. 280–286.

Hamsiati (2019) ‘Pengenalan Morfologi Bahasa Arab bagi Pembelajar Pemula’, *Pusaka Jurnal Khazanah Keagamaan*, 6(1), pp. 111–126.

Holle, K. F. H., Arifin, A. Z. and Purwitasari, D. (2015) ‘Preference Based Term Weighting For Arabic Fiqh Document Ranking’, *Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)*, 8(1), pp. 45–52.

Kaur, H. and Maini, R. (2018) ‘Assessing lexical similarity between short sentences of source code based on granularity’, *International Journal of Information Technology*. Springer Singapore, 1(2), pp. 58–64. doi: 10.1007/s41870-018-0213-1.

Mahdaouy, A. El *et al.* (2018) ‘Improving Arabic information retrieval using word embedding similarities’, *International Journal of Speech Technology*. Springer US. doi: 10.1007/s10772-018-9492-y.

Moatez, E. *et al.* (2017) ‘Semantic Similarity of Arabic Sentences with Word Embeddings’, in *Third Arabic Natural Language Processing Workshop*.

Othman, N., Faiz, R. and Smaili, K. (2019) ‘Enhancing Question Retrieval in Community Question Answering Using Word Embeddings’, in *23rd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*, pp. 485–494. doi: 10.1016/j.procs.2019.09.203.

Sholikah, R. W. *et al.* (2017) ‘Term Weighting based on Positive Impact Factor Query for Arabic Fiqh’, *Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)*, 10(1), pp. 29–35.

Suleiman, D. and Awajan, A. (2018) ‘Comparative Study of Word Embeddings Models and Their Usage in Arabic Language Applications’, *2018 International Arab Conference on Information Technology (ACIT)*. IEEE, pp. 1–7.

Sunilkumar, P. and Shaji, A. P. (2019) ‘A Survey on Semantic Similarity’, in *2019 International Conference on Advances in Computing, Communication and Control (ICAC3)*. IEEE.

LAMPIRAN

Lampiran 1. Rumpun-Rumpun Ilmu *Dataset* Maktabah Syamilah

No.	**Terjemahan**	**Alih Aksara**	**Rumpun Ilmu**	**Jumlah Kitab**
1	Bagian-bagian hadis	*Al-ajza' al-haditsah*	الأجزاء الحديثية	542
2	Nasab-nasab (silsilah-silsilah keturunan)	*Al-ansab*	الأنساب	20
3	Negara-negara, geografi, dan perjalanan-perjalanan	*Al-buldan wa al-jugrafiya wa al-rihlat*	البلدان والجغرافيا والرحلات	65
4	Ilmu tajwid dan ilmu qira'at	*Al-tajwid wa al-qira'at*	التجويد والقراءات	27
5	Dakwah dan keadaan kaum muslimin	*Al-da'wah wa ahwal al-muslimin*	الدعوة وأحوال المسلمين	201
6	Diwan-diwan syair	*Al-dawawin al-syi'riyah*	الدواوين الشعرية	1
7	Politik syar'i dan kekuasaan	*Al-siyasah al-syar'iyah wa al-qadha'*	السياسة الشرعية والقضاء	61
8	Akidah (kepercayaan)	*Al-aqidah*	العقيدة	505
9	Pertanyaan-pertanyaan dan jawaban-jawaban dengan dalil	*Al-'ilal wa al-su'alaat*	العلل والسؤالات	58
10	Sekte-sekte dan cara-cara menjawab mereka	*Al-firaq wa al-rudud*	الفرق والردود	15
11	Ilmu nahwu dan ilmu saraf	*Al-nahw wa al-sharf*	النحو والصرف	108
12	Penelitian-penelitian dan permasalahan-permasalahan	*Buhuts wa masa'il*	بحوث ومسائل	295
13	Ilmu-ilmu yang lain	*'ulum ukhraa*	علوم أخرى	26
14	Ilmu-ilmu hadis	*'ulum hadits*	علوم الحديث	202

15	Ilmu fikih mazhab imam hanbali	*Fiqh hanbaly*	فقه حنبلي	45
16	Ilmu fikih mazhab imam hanafi	*Fiqh hanafy*	فقه حنفي	35
17	Ilmu fikih mazhab imam syafi'i	*Fiqh syafi'y*	فقه شافعي	47
18	Ilmu fikih secara umum	*Fiqh 'am*	فقه عام	45
19	Ilmu fikih mazhab imam malik	*Fiqh maliky*	فقه مالكي	30
20	Indeks-indeks buku dan buku-buku panduan	*Faharis al-kitab wa al-adillah*	فهارس الكتب والأدلة	64
21	Buku-buku ibnu abi dunya	*Kitab ibnu abi al-duniya*	كتب ابن أبي الدنيا	61
22	Buku-buku ibnu qayyim	*Kitab ibnu al-qayyim*	كتب ابن القيم	35
23	Buku-buku islam umum	*Kitab islamiyyah 'ammah*	كتب إسلامية عامة	2
24	Buku-buku albani	*Kitab al-bany*	كتب الألباني	65
25	Buku-buku bahasa	*Kitab al-lughah*	كتب اللغة	46
26	Manuskrip-manuskrip hadis	*Makhthuthat haditsah*	مخطوطات حديثية	299

Lampiran 2. Hasil *recall* masing-masing *short query* dan *long query*

		Recall(%) *Short Query*								
		Kemiripan Leksikal			Kemiripan Semantik			Gabungan Kemiripan Leksikal dan Semantik		
		Top 5	Top 10	Top 15	Top 5	Top 10	Top 15	Top 5	Top 10	Top 15
Q1	ذَبْحُ ٱلأَضَاحِي	33.33%	55.56%	66.67%	44.44%	66.67%	88.89%	55.56%	55.56%	77.78%
Q5	المَحَارِمُ لِلْمَرْأَةِ	40.00%	70.00%	90.00%	40.00%	70.00%	70.00%	50.00%	70.00%	80.00%
Q6	المَحَارِمُ لِلرِّجَالِ	75.00%	100.00%	100.00%	50.00%	75.00%	75.00%	100.00%	100.00%	100.00%
Q7	صَوْمُ التَّطَوُّعِ	25.00%	75.00%	100.00%	25.00%	100.00%	100.00%	25.00%	75.00%	100.00%
Q8	مُبْطِلَاتُ الصَّلَاةِ	50.00%	66.67%	83.33%	66.67%	83.33%	83.33%	66.67%	83.33%	100.00%
Q9	شُرْبُ ٱلخَمْرِ	50.00%	75.00%	75.00%	62.50%	75.00%	75.00%	62.50%	75.00%	87.50%
Q11	أَرْكَانُ ٱلإِيْمَانِ	57.14%	71.43%	71.43%	57.14%	85.71%	85.71%	57.14%	85.71%	85.71%
Q12	العَقِيْقَةُ لِلْبَنَاتِ	30.00%	50.00%	70.00%	30.00%	80.00%	90.00%	40.00%	50.00%	90.00%
Q13	صَلَاةِ ٱلجَنَازَةِ	44.44%	55.56%	66.67%	33.33%	66.67%	77.78%	44.44%	66.67%	77.78%
Q14	صَلَاةِ التَّهَجُّدِ	50.00%	80.00%	90.00%	40.00%	80.00%	80.00%	50.00%	80.00%	100.00%
Q15	غَسْلُ ٱلمَيِّتِ	30.00%	50.00%	90.00%	30.00%	50.00%	80.00%	40.00%	80.00%	100.00%
Q18	أركان الوضوء	33.33%	55.56%	66.67%	33.33%	66.67%	77.78%	55.56%	77.78%	88.89%

		Recall(%) *Long Query*								
		Kemiripan Leksikal			Kemiripan Semantik			Gabungan Kemiripan Leksikal dan Semantik		
		Top 5	Top 10	Top 15	Top 5	Top 10	Top 15	Top 5	Top 10	Top 15
Q2	صِيَامُ 6 أَيَّامٍ فِي شَوَّال	33.33%	44.44%	55.56%	33.33%	77.78%	100.00%	55.56%	66.67%	77.78%
Q3	الطَّوَافُ فِي ٱلْحَجّ	30.77%	69.23%	76.92%	38.46%	76.92%	92.31%	38.46%	69.23%	76.92%
Q4	رَمْيُ ٱلْجُمُرَاتِ فِي ٱلْحَجّ	33.33%	66.67%	66.67%	66.67%	66.67%	100.00%	66.67%	66.67%	66.67%
Q10	أَكْلُ لَحْمِ ٱلْخِنْزِيْرِ	40.00%	60.00%	80.00%	30.00%	70.00%	80.00%	50.00%	80.00%	90.00%
Q16	الصَّلَوَاتُ ٱلْخَمْسُ ٱلْمَفْرُوْضَةُ	44.44%	66.67%	77.78%	44.44%	66.67%	66.67%	55.56%	66.67%	88.89%
Q17	إِخْرَاجُ زَكَاةِ ٱلْفِطْرِ	37.50%	62.50%	100.00%	62.50%	75.00%	75.00%	62.50%	100.00%	100.00%
Q19	قضاء صيام رمضان	20.00%	30.00%	30.00%	30.00%	80.00%	90.00%	50.00%	70.00%	80.00%
Q20	الإفطار في الطيارة	44.44%	55.56%	88.89%	33.33%	77.78%	88.89%	55.56%	77.78%	100.00%

Lampiran 3. Hasil *precision* masing-masing *short query* dan *long query*

		Precision(%) *Short Query*								
		Kemiripan Leksikal			Kemiripan Semantik			Gabungan Kemiripan Leksikal dan Semantik		
		Top 5	Top 10	Top 15	Top 5	Top 10	Top 15	Top 5	Top 10	Top 15
Q1	ذِبْحُ ٱلأَضَاحِي	60.00%	50.00%	40.00%	80.00%	60.00%	53.33%	100.00%	50.00%	46.67%
Q5	المَحَارِمُ لِلْمَرْأَةِ	80.00%	70.00%	60.00%	80.00%	70.00%	46.67%	100.00%	70.00%	53.33%
Q6	المَحَارِمُ لِلرّجَالِ	60.00%	40.00%	26.67%	40.00%	30.00%	20.00%	80.00%	40.00%	26.67%
Q7	صَوْمُ التَّطَوُّعِ	20.00%	30.00%	26.67%	20.00%	40.00%	26.67%	20.00%	30.00%	26.67%
Q8	مُبْطِلَاتُ الصَّلَاةِ	60.00%	40.00%	33.33%	80.00%	50.00%	33.33%	80.00%	50.00%	40.00%
Q9	شُرْبُ ٱلخَمْرِ	80.00%	60.00%	40.00%	100.00%	60.00%	40.00%	100.00%	60.00%	46.67%
Q11	أَرْكَانُ ٱلإِيْمَانِ	80.00%	50.00%	33.33%	80.00%	60.00%	40.00%	80.00%	60.00%	40.00%
Q12	العَقِيْقَةُ لِلْبَنَاتِ	60.00%	50.00%	46.67%	60.00%	80.00%	60.00%	80.00%	50.00%	60.00%
Q13	صَلَاةِ ٱلجَنَازَةِ	80.00%	50.00%	40.00%	60.00%	60.00%	46.67%	80.00%	60.00%	46.67%
Q14	صَلَاةِ التَّهَجُّدِ	100.00%	80.00%	60.00%	80.00%	80.00%	53.33%	100.00%	80.00%	66.67%
Q15	غَسْلُ ٱلمَيِّتِ	60.00%	50.00%	60.00%	60.00%	50.00%	53.33%	80.00%	80.00%	66.67%
Q18	أركان الوضوء	60.00%	50.00%	40.00%	60.00%	60.00%	46.67%	100.00%	70.00%	53.33%

		Precision(%) *Long Query*								
		Kemiripan Leksikal			Kemiripan Semantik			Gabungan Kemiripan Leksikal dan Semantik		
		Top 5	Top 10	Top 15	Top 5	Top 10	Top 15	Top 5	Top 10	Top 15
Q2	صِيَامُ 6 أَيَّامٍ فِى شَوَّال	60.00%	40.00%	33.33%	60.00%	70.00%	60.00%	100.00%	60.00%	46.67%
Q3	الطَّوَافُ فِى ٱلْحَجّ	80.00%	90.00%	66.67%	100.00%	100.00%	80.00%	100.00%	90.00%	66.67%
Q4	رَمْيُ ٱلْجُمَرَاتِ فِى ٱلْحَجّ	20.00%	20.00%	13.33%	40.00%	20.00%	20.00%	40.00%	20.00%	13.33%
Q10	أَكْلُ لَحْمِ ٱلْخِنْزِيْر	80.00%	60.00%	53.33%	60.00%	70.00%	53.33%	100.00%	80.00%	60.00%
Q16	الصَّلَوَاتُ ٱلْخَمْسُ ٱلْمَفْرُوْضَةُ	80.00%	60.00%	46.67%	80.00%	60.00%	40.00%	100.00%	60.00%	53.33%
Q17	إِخْرَاجُ زَكَاةِ ٱلْفِطْرِ	60.00%	50.00%	53.33%	100.00%	60.00%	40.00%	100.00%	80.00%	53.33%
Q19	قضاء صيام رمضان	40.00%	30.00%	20.00%	60.00%	80.00%	60.00%	100.00%	70.00%	53.33%
Q20	الإفطار في الطيارة	80.00%	50.00%	53.33%	60.00%	70.00%	53.33%	100.00%	70.00%	60.00%

Lampiran 4. Hasil *f-measure* masing-masing *short query* dan *long query*

		F-measure(%) *Short Query*								
		Kemiripan Leksikal			Kemiripan Semantik			Gabungan Kemiripan Leksikal dan Semantik		
		Top 5	Top 10	Top 15	Top 5	Top 10	Top 15	Top 5	Top 10	Top 15
Q1	ذَبْحُ ٱلأَضَاحِي	42.86%	52.63%	50.00%	57.14%	63.16%	66.67%	71.43%	52.63%	58.33%
Q5	المَحَارِمُ لِلْمَرْأَةِ	53.33%	70.00%	72.00%	53.33%	70.00%	56.00%	66.67%	70.00%	64.00%
Q6	المَحَارِمُ لِلرّجَالِ	66.67%	57.14%	42.11%	44.44%	42.86%	31.58%	88.89%	57.14%	42.11%
Q7	صَوْمُ التَّطَوُّعِ	22.22%	42.86%	42.11%	22.22%	57.14%	42.11%	22.22%	42.86%	42.11%
Q8	مُبْطِلَاتُ الصَّلَاةِ	54.55%	50.00%	47.62%	72.73%	62.50%	47.62%	72.73%	62.50%	57.14%
Q9	شُرْبُ ٱلخَمْرِ	61.54%	66.67%	52.17%	76.92%	66.67%	52.17%	76.92%	66.67%	60.87%
Q11	أَرْكَانُ ٱلإِيْمَانِ	66.67%	58.82%	45.45%	66.67%	70.59%	54.55%	66.67%	70.59%	54.55%
Q12	العَقِيْقَةُ لِلْبَنَاتِ	40.00%	50.00%	56.00%	40.00%	80.00%	72.00%	53.33%	50.00%	72.00%
Q13	صَلَاةِ ٱلجَنَازَةِ	57.14%	52.63%	50.00%	42.86%	63.16%	58.33%	57.14%	63.16%	58.33%
Q14	صَلَاةِ التَّهَجُّدِ	66.67%	80.00%	72.00%	53.33%	80.00%	64.00%	66.67%	80.00%	80.00%
Q15	غَسْلُ ٱلمَيِّتِ	40.00%	50.00%	72.00%	40.00%	50.00%	64.00%	53.33%	80.00%	80.00%
Q18	أركان الوضوء	42.86%	52.63%	50.00%	42.86%	63.16%	58.33%	71.43%	73.68%	66.67%

		F-measure(%) *Long Query*								
		Kemiripan Leksikal			Kemiripan Semantik			Gabungan Kemiripan Leksikal dan Semantik		
		Top 5	Top 10	Top 15	Top 5	Top 10	Top 15	Top 5	Top 10	Top 15
Q2	صِيَامُ 6 أَيَّامٍ فِى شَوَّال	42.86%	42.11%	41.67%	42.86%	73.68%	75.00%	71.43%	63.16%	58.33%
Q3	الطَّوَافُ فِى ٱلْحَجّ	44.44%	78.26%	71.43%	55.56%	86.96%	85.71%	55.56%	78.26%	71.43%
Q4	رَمْيُ ٱلْجُمُرَاتِ فِى ٱلْحَجّ	25.00%	30.77%	22.22%	50.00%	30.77%	33.33%	50.00%	30.77%	22.22%
Q10	أَكْلُ لَحْمِ ٱلْخِنْزِيْرِ	53.33%	60.00%	64.00%	40.00%	70.00%	64.00%	66.67%	80.00%	72.00%
Q16	الصَّلَوَاتُ ٱلْخَمْسُ ٱلْمَفْرُوْضَةُ	57.14%	63.16%	58.33%	57.14%	63.16%	50.00%	71.43%	63.16%	66.67%
Q17	إِخْرَاجُ زَكَاةِ ٱلْفِطْرِ	46.15%	55.56%	69.57%	76.92%	66.67%	52.17%	76.92%	88.89%	69.57%
Q19	قضاء صيام رمضان	26.67%	30.00%	24.00%	40.00%	80.00%	72.00%	66.67%	70.00%	64.00%
Q20	الإفطار في الطيارة	57.14%	52.63%	66.67%	42.86%	73.68%	66.67%	71.43%	73.68%	75.00%

BIODATA PENULIS

Syadza Anggraini, lahir di Pangkalan Bun pada 11 Agustus 1995. Penulis telah menempuh pendidikan pada TK Kartika XVII-22 Sampit (1999-2001), SD Negeri 1 Mentawa Baru Hulu Sampit (2001-2007), SMP Negeri 1 Sampit (2007-2010), SMA Negeri 1 Sampit (2010-2013), dan S1 Teknik Informatika Universitas Muhammadiyah Malang (2013-2018). Penulis kemudian melanjutkan pendidikan pascasarjana S2 Teknik Informatika Institut Teknologi Sepuluh Nopember Surabaya pada tahun 2018. Selama perkuliahan pascasarjana S2, penulis mengambil bidang minat Komputasi Cerdas dan Visi (KCV). Penulis dapat dihubungi melalui surel sasaanggraini.sa@gmail.com.