ANALISIS DAN *CLUSTERING* MAKNA AYAT AL-QUR'AN DENGAN *MINI SEARCH ENGINE* BM-25

Dheaz Kelvin Harahap, Daffa Pratama Yudhistira, Azka Muhammad Pinandito, Muhammad Rizky Ramadhan, Reynaldo Arya Budi Trisna.

Abstract: *The Quran contains various knowledge and serves as a guide to life for Muslims. When reading the Quran, Muslims are advised to read the translation of each verse in all surahs that have the same meaning. This is done so that they can understand the Quran as a whole. To facilitate Muslims, this research uses K-Means for clustering and BM-25 as the searching method. The proposed method includes preprocessing the Quranic text, calculating TF-IDF to measure the relevance of terms in the document, clustering using the K-Means algorithm, and using the BM-25 search method as a mini search engine. This methodology is expected to improve the ability to search for and group Quranic verses based on meaning, as well as provide more efficient access to users. In the search for the best searching method, BM-25 was found to have a score of 78.9. Furthermore, BM-25 was able to achieve a score of 97 in searching for the query "tidak" (not) in the Quran. However, the average score for regular queries was only 36.86, and 54.3 for queries with word similarity. This is a comprehensible mini-project on the analysis and grouping of the meanings of Quranic verses using the mini search engine BM-25, which facilitates efficient access to information and enriches the understanding of Quranic studies and the spirituality of individuals who study it. It also inspires further research and the development of more advanced applications in analyzing the meanings of the Quran. The findings and methodology presented in this research can serve as a foundation for other researchers to continue studying in this field, expand the scope of analysis, and improve the performance of the mini search engine BM-25 to provide a deeper understanding of the meanings of Quranic verses for the Muslim community and the academic community.*
Keywords: *Al-Qur'an, Mini Search Engine, TF-IDF, K-means, LDA, Vector Space Model, Fuzzy, BM-25*

Abstrak: Al-Quran berisi berbagai pengetahuan dan menjadi panduan hidup bagi umat Muslim. Ketika membaca Al-Quran, umat Muslim disarankan untuk membaca terjemahan setiap ayat dalam semua surah yang memiliki arti yang sama. Hal ini dilakukan agar mereka dapat memahami Al-Quran secara keseluruhan. Untuk memudahkan umat-umat Muslim, penelitian ini menggunakan *K-Means* sebagi *clustering* dan BM-25 sebagai *searching method.* Metode yang diusulkan mencakup *preprocessing* teks Al-Qur'an, perhitungan TF-IDF untuk mengukur relevansi term dalam dokumen, *clustering* menggunakan algoritma K-Means, dan penggunaan metode pencarian BM-25 *sebagai mini search engine*. Metodologi ini diharapkan dapat meningkatkan kemampuan dalam mencari dan mengelompokkan ayat-ayat Al-Qur'an berdasarkan makna, serta memberikan akses yang lebih efisien kepada pengguna. Dalam pencarian *searching method* terbaik didapatkan BM-25 dengan skor 78,9. Selain itu BM-25 mampu mendapatkan skor 97 dalam mencari *query* di Al-Qur'an pada kata *query "*tidak". Tetapi untuk rata-rata skor pada *regular query* hanya 36,86 dan 54,3 untuk *with word similarity*. Hal ini merupakan suatu *mini project* yang mudah dipahami tentang analisis dan pengelompokan makna ayat-ayat Al-Qur'an menggunakan mini search engine BM-25, yang memfasilitasi akses efisien terhadap informasi dan memperkaya pemahaman studi Al-Qur'an serta spiritualitas individu yang mempelajarinya. Selain itu, menginspirasi penelitian lanjutan dan pengembangan aplikasi yang lebih canggih dalam analisis makna Al-Qur'an. Temuan dan metodologi yang disajikan dalam penelitian ini dapat menjadi pijakan bagi peneliti lain untuk melanjutkan studi dalam bidang ini, memperluas cakupan analisis, dan meningkatkan performa mini search engine BM-25 agar dapat memberikan pemahaman yang lebih mendalam tentang makna ayat-ayat Al-Qur'an bagi umat Islam dan masyarakat akademik.
Kata kunci: Al-Qur'an, *Mini Search Engine,* TF-IDF, K-Means, LDA, Vector Space Model, Fuzzy, BM-25

I. PENDAHULUAN

Al-Quran adalah sebuah kitab suci yang terdiri dari 6236 ayat yang terbagi dalam 114 surat. Kitab Al-Quran digunakan sebagai panduan hidup bagi umat muslim dalam menjalani kehidupan beragama, bermasyarakat, dan bernegara. Al-Qur'an didefinisikan sebagai kalamullah yang berupa mukjizat, diturunkan kepada Nabi Muhammad SAW. Selain dibaca, Al-Quran juga ditulis baik dalam bentuk mushaf atau kaligrafi dan ditafsirkan dengan beragam corak dan pendekatan. Artinya, al-Quran selalu ada dalam keseharian kaum muslim Indonesia. [2]

Masalah agama merupakan salah satu permasalahan yang sulit dalam kehidupan. Untuk mengatasi masalah ini, diperlukan pemahaman yang lebih mendalam tentang agama, dan salah satu sumber ilmu yang penting adalah Al-Quran. Al-Quran mencakup berbagai bidang ilmu pengetahuan, sehingga ilmuwan tertarik untuk melakukan penelitian tentang Al-Quran. Dengan kemajuan teknologi, dapat dibuat sebuah sistem yang dapat membantu masyarakat dalam mempelajari berbagai ilmu yang terdapat dalam Al-Quran. [7]

Menganalisis serta mengelompokkan makna ayat Al-Qur'an dengan *mini search engine* adalah proyek atau penelitian yang dapat membantu masyarakat dalam mempelajari berbagai ilmu yang terdapat dalam Al-Qur'an. Dengan menggunakan teknik analisis dan clustering, ayat-ayat Al-Qur'an akan dikelompokkan berdasarkan makna-makna yang terkandung di dalamnya. *Mini search engine* BM-25 akan digunakan sebagai alat untuk mencari dan mengakses ayat-ayat yang relevan dengan topik yang sedang dipelajari. Hal ini akan memudahkan pemahaman lebih mendalam tentang agama dan memberikan akses yang lebih efisien kepada masyarakat untuk menjalani kehidupan beragama.

Penggabungan antara BM-25 dan *K-Means* memiliki tujuan untuk meningkatkan kemampuan dalam mencari dan mengelompokkan ayat-ayat Al-Qur'an berdasarkan makna. BM-25, sebagai metode pencarian, digunakan untuk mengidentifikasi dan memberikan nilai ke ayat-ayat yang paling relevan dengan *query* pengguna. Sementara itu, *K-Means*, sebagai algoritma *clustering*, digunakan untuk mengelompokkan ayat-ayat yang memiliki makna serupa. Dengan menggabungkan keduanya, pengguna dapat mencari ayat-ayat yang relevan dengan *query* dan pada saat yang sama melihat hubungan antara ayat-ayat tersebut dalam konteks makna yang sama. Hal ini memungkinkan pengguna untuk mendapatkan pemahaman yang lebih holistik dan terstruktur mengenai Al-Qur'an melalui kombinasi fungsi pencarian dan pengelompokan yang optimal.

Berdasarkan pemaparan diatas, metodologi yang diusulkan dalam penelitian ini terdiri dari beberapa tahap. Tahap awal adalah *preprocessing*, di mana dokumen Al-Qur'an akan melalui serangkaian tahapan seperti *case folding, cleaning, tokenization, filtering, stop list, dan stemming* untuk mempersiapkan teks agar lebih mudah diproses. Setelah itu, dilakukan perhitungan *Term Frequency-Inverse Document Frequency* (TF-IDF) untuk mengukur relevansi *term* dalam dokumen. Tahap selanjutnya adalah *clustering* menggunakan algoritma *K-Means*, yang bertujuan untuk mengelompokkan ayat-ayat Al-Qur'an berdasarkan kesamaan makna. Metode *Elbow* digunakan untuk menentukan jumlah kluster terbaik. Selain itu, *centroid distance* digunakan untuk mengukur afinitas antara klaster data. Terakhir, metode pencarian BM-25 digunakan sebagai *mini search engine* untuk mencari dan mengakses ayat-ayat yang relevan dengan query pengguna. Dengan menggunakan metodologi ini, diharapkan penelitian dapat memberikan pemahaman yang lebih mendalam tentang Al-Qur'an dan memudahkan akses masyarakat dalam mempelajari berbagai ilmu yang terkandung dalam Al-Qur'an.

II. DASAR TEORI DAN RELATED WORKS

II.1.1 Al-Quran dan Pencarian Alquran

Al-Qur'an adalah firman Allah SWT yang diturunkan kepada Nabi Muhammad SAW dan membacanya mendapatkan pahala. Jadi pada prinsipnya pengertian al- Qur'an adalah wahyu atau firman Allah SWT untuk menjadi petunjuk atau pedoman bagi manusia yang beriman dan bertaqwa kepada Allah SWT. [2]

Pencarian Al-Qur‟an terdiri dari dua kata yakni kata "Pencarian" dan kata "Al-Qur‟an". Kata Pencarian adalah Pencarian adalah proses mencari atau mencari informasi atau objek yang diinginkan dalam sebuah kumpulan data yang memiliki *type* data yang sama. [13] Dalam konteks teknologi informasi, pencarian sering kali merujuk pada aktivitas mencari data atau informasi menggunakan mesin pencari atau sistem pencarian elektronik. Tujuan dari pencarian adalah untuk menemukan hasil yang relevan atau sesuai dengan kriteria yang diberikan oleh pengguna. Pencarian dapat dilakukan dalam berbagai konteks, termasuk pencarian informasi di internet, pencarian dalam basis data, pencarian dalam dokumen teks, atau pencarian dalam koleksi digital.

II.1.2 *Term Frequency Inverse Document Frequency* (TF-IDF)

TF-IDF digunakan untuk mengekstraksi kata kunci teks yang telah diproses sebelumnya dan mendapatkan bobot kata kunci. Berdasarkan bobot tersebut, metode jarak kosinus digunakan untuk menghitung kesamaan teks dari kata kunci teks dari pesanan kerja pelayanan pelanggan terkait listrik yang berbeda. Hasilnya menunjukkan bahwa dibandingkan dengan metode deteksi berbasis vektor kata dan kalimat yang ditimbang, metode deteksi berbasis model hibrida, metode deteksi berbasis diskriminasi jarak dan fitur kategori yang kuat, nilai pengukuran F1 rata-rata makro lebih tinggi, metode ini membutuhkan waktu yang lebih sedikit dan memiliki kompleksitas yang lebih rendah. [3]

II.1.3. *K-Means*

K-means merupakan salah satu algoritma pengelompokan (*clustering*) yang populer dan efektif. Konsep dasar dari algoritma *K-means* adalah membagi sekumpulan data menjadi beberapa kelompok (*cluster*) berdasarkan kesamaan atribut. Tujuan utama algoritma *K-means* adalah meminimalkan jumlah varian dalam setiap *cluster*. Algoritma ini mengasumsikan bahwa jumlah *cluster* sudah diketahui sebelumnya. Proses *K-means* dimulai dengan menentukan jumlah kelompok (*k*) yang diinginkan

sebelumnya. Kemudian, algoritma K-means akan secara iteratif mencari pusat (*centroid*) untuk setiap kelompok dengan meminimalkan jarak antara data dengan pusat kelompok yang sesuai.[17]

II.1.3. LDA (*Latent Dirichlet Allocation*)

LDA (*Latent Dirichlet Allocation*) digunakan sebagai metode analisis data teks. LDA adalah sebuah model statistik generatif yang digunakan untuk menemukan topik tersembunyi dalam koleksi dokumen. LDA memandang setiap dokumen sebagai kombinasi dari beberapa topik yang tersembunyi. Topik ini merupakan distribusi probabilitas atas kata-kata yang muncul dalam korpus teks. Dengan menggunakan LDA, kita dapat mengidentifikasi topik yang muncul secara inheren dalam dokumen-dokumen tersebut. [4]

Metode LDA akan menemukan kata yang ada pada topik secara semi acak distribusi, menghitung probabilitas topik pada dataset, dan menghitung probabilitas terhadap topik setiap iterasinya. percobaan pemodelan topik sebanyak 5 kali uji iterasi dan jumlah topik yang berbeda-beda dilakukan pada penelitian ini. Setelah percobaan yang disebutkan di atas selesai, data yang dikumpulkan akan dianalisis, dan 1 jumlah topik dengan hasil terbaik akan dipilih dari sekian banyak topik. [11]

II.1.4. *Vector Space Models*

Dalam VSM, setiap dokumen direpresentasikan sebagai vektor dalam ruang dimensi yang mewakili kata-kata atau fitur yang muncul dalam dokumen tersebut. Masing-masing dimensi dalam ruang vektor menggambarkan keberadaan atau frekuensi kata atau fitur tertentu dalam dokumen tersebut. Oleh karena itu, setiap dokumen dalam koleksi teks memiliki representasi vektor yang unik. Dokumen-dokumen dalam bahasa Arab direpresentasikan sebagai vektor dalam ruang fitur yang memungkinkan algoritma klasifikasi untuk mempelajari pola-pola atau hubungan antara fitur-fitur tersebut dan melakukan pengklasifikasian dokumen berdasarkan representasinya dalam ruang vektor. [6]

II.1.5. *Fuzzy K-Means*

Fuzzy K-Means adalah metode pengelompokan data yang memperkenalkan tingkat keanggotaan parsial untuk setiap titik data dalam kluster. Dalam algoritma *K-Means* standar, setiap titik data hanya ditempatkan dalam satu kluster yang memiliki pusat kluster terdekat. Namun, dalam *Fuzzy K-Means*, setiap titik data memiliki keanggotaan *fuzzy* yang mencerminkan tingkat keanggotaan atau probabilitas bahwa titik data tersebut termasuk dalam setiap kluster. Keanggotaan *fuzzy* ini dinyatakan sebagai bilangan real antara 0 hingga 1, di mana nilai 1 menunjukkan tingkat keanggotaan penuh dalam kluster tertentu, nilai 0 menunjukkan tidak ada keanggotaan, dan nilai di antara keduanya menunjukkan tingkat keanggotaan parsial. Dengan adanya keanggotaan *fuzzy*, titik data dapat termasuk secara parsial dalam beberapa kluster, memberikan fleksibilitas yang lebih besar dalam pengelompokan data. [16]

II.1.6. BM-25

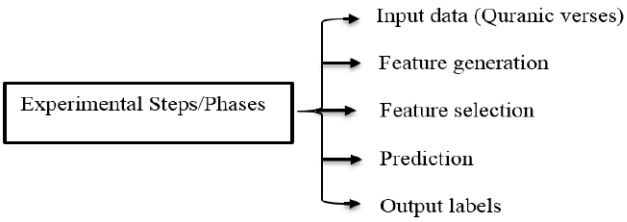
BM25 merupakan perbaikan dari model sebelumnya yang dikenal sebagai TF-IDF (*Term Frequency-Inverse Document Frequency*). Tujuan utama dari algoritma BM25 adalah untuk menghitung relevansi atau skor kecocokan antara sebuah dokumen dengan sebuah *query* atau permintaan pencarian. Dalam BM25, sebuah dokumen direpresentasikan sebagai sekumpulan term atau kata-kata yang terkandung di dalamnya, dan sebuah query juga direpresentasikan sebagai sekumpulan term. Algoritma ini menghitung skor relevansi berdasarkan TF, IDF, dan Term Saturation. [19]

II.2 Related Work

Penelitian Sebelumnya, membahas tentang penggunaan teknik pembelajaran mesin untuk mengklasifikasikan ayat-ayat Al-Qur'an secara otomatis dengan bertujuan untuk mengklasifikasikan ayat-ayat Al-Qur'an ke dalam kelompok tema yang berbeda menggunakan pendekatan berbasis pembelajaran mesin. Penulis menggunakan dataset ayat Al-Qur'an yang telah diberikan label berdasarkan kelompok tema yang telah ditentukan sebelumnya. Tujuannya adalah untuk membangun model klasifikasi yang dapat mengklasifikasikan ayat-ayat Al-Qur'an baru ke dalam kelompok tema yang sesuai.

Metode *K-means clustering* digunakan untuk mengelompokkan ayat-ayat Al-Qur'an berdasarkan fitur-fitur teks yang diekstraksi. Algoritma *K-means* digunakan untuk mengklasifikasikan ayat-ayat baru ke dalam kelompok yang sudah ada. Selain itu, penulis juga menerapkan teknik pembelajaran mesin lainnya, seperti *Support Vector Machine* (SVM), untuk membandingkan kinerja dan akurasi model klasifikasi.

Pada *Paper Extracting Topics from the Holy Quran* (Mohammad Alhawarat), teknik LDA telah dibandingkan dengan teknik pengelompokan *K-means*. Para penulis telah menerapkan kedua teknik LDA dan *K-means* pada kasus ini. Hasil penelitian menunjukkan bahwa LDA memberikan hasil yang lebih baik dibandingkan *K-means* dalam sebagian besar kasus.

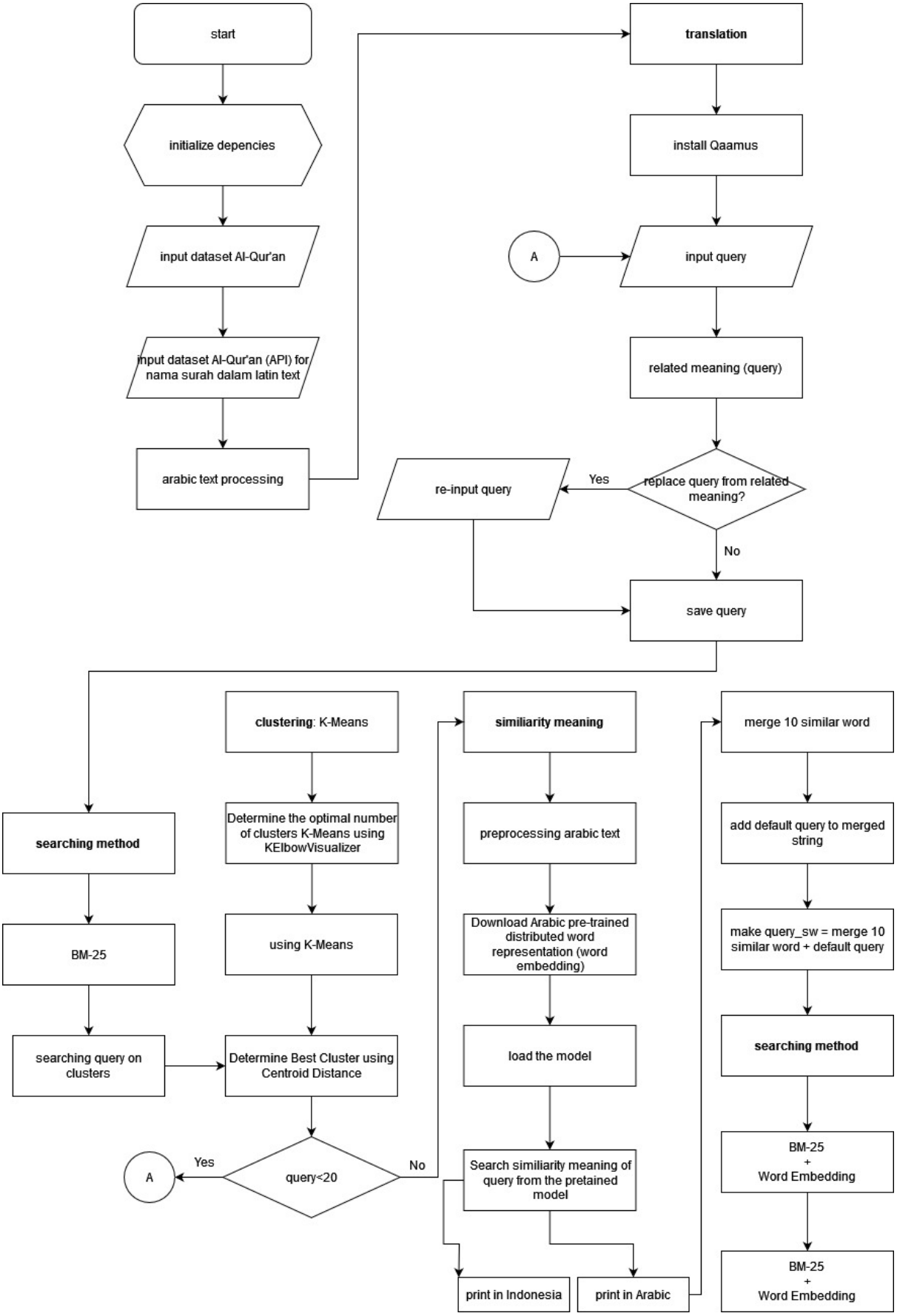


Gambar 1. *Step* Eksperimen

Pada penelitian ini, dilakukan pembelajaran untuk mengekstraksi topik-topik dalam ayat-ayat Al-Qur'an. Proses ekstraksi topik didasarkan pada korpus yang terdiri dari ayat-ayat Al-Qur'an menggunakan metode NLTK, TF-IDF, *Cosine Similarity*, LDA dan BERT. Topik-topik divisualisasikan dan ayat-ayat terkait untuk setiap *query* ditampilkan berdasarkan kata kunci utama topik tersebut. Penulis telah berhasil mengekstraksi dan mengidentifikasi topik-topik yang berhubungan erat di Al-Qur'an. Namun, penulis hanya mengidentifikasi dari kata-kata berdasarkan model yang digunakan sehingga masih kurang jelas menghubungkan kata kunci dari setiap topik dengan ayat-ayat terkait yang sesuai dengan kata kunci topik tersebut. Meskipun demikian, temuan-temuan tersebut dapat membantu dalam mengungkap makna-makna yang lebih dalam dalam Al-Qur'an oleh orang-orang yang ahli dalam studi Al-Qur'an. *Paper* ini bermanfaat untuk sarana media belajar dan memudahkan untuk mencari kata-kata kunci dan *similarity* dari kata yang dicari dalam bahasa indonesia ke bahasa arab sebagai *prediction* dari inputan yang dimasukkan *user* (Gambar 1).

III. METODOLOGI

Penelitian ini akan meneliti tentang analisis dan pengelompokan makna Al-Qur’an dengan *mini search engine* serta membandingkan berbagai metode *clustering* dan pencarian. Gambar 1 merupakan perancangan algoritma penelitian.



Gambar 2. Algoritma Penelitian.

Penelitian ini akan meneliti tentang analisis dan pengelompokan makna Al-Qur'an dengan *mini search engine* serta membandingkan berbagai metode *clustering* dan pencarian. Gambar 1 merupakan perancangan algoritma penelitian.

Tahap preprocessing dilakukan dengan menggabungkan metode *case folding, tokenization, filtering,* dan *stemming*. Setelah itu didapatkan dokumen yang sudah melalui tahap *text processing*. Kemudian, kami membuat tahap komputasi dilakukan clustering menggunakan 3 algoritma berbeda untuk membandingkan 3 metode *clustering* tersebut dan mencari metode yang akan mendapatkan skor lebih besar. Selain itu, kami juga membandingan 3 metode *searching* dan mencari metode yang akan mendapatkan skor lebih besar. Pada penelitian ini, akan mendapatkan kelompok dokumen (ayat) yang relevan berdasarkan *query* yang diberikan oleh pengguna. Pada tahap ini juga dilakukan pengujian.

III.1. Text Preprocessing

Sebelum mengelompokkan dokumen perlu diolah agar lebih mudah diproses dan dapat meningkatkan akurasi yang lebih tinggi, tahap tersebut disebut preprocessing [7]. Adapun tahapan *preprocessing text* yang kami pakai. Pertama, *Case folding* merupakan tahap mengubah semua huruf di semua dokumen yang ada menjadi huruf kecil. Dengan cara ini, semua kata dianggap sama dalam perhitungan. Kedua, *Cleaning* merupakan tahap menghapus karakter yang dianggap tidak terkait dengan informasi seperti tanda baca dan angka [7]. Ketiga, *Tokenization* merupakan proses untuk memotong kalimat di setiap dokumen menjadi kata. [7] Keempat, *Filtering* merupakan tahap pemilihan katakata penting dari hasil tokenized. Kelima, *Stop list* merupakan algoritme yang digunakan dengan konsep menghilangkan kata-kata yang tidak penting. Keenam, *Stemming* merupakan tahapan dimana berbagai kata dengan imbuhan tambahan menjadi kata-kata dasar.

III.2 *Term Frequency-Inverse Document Frequency* (TF-IDF)

Term frequency-inverse document frequency (TF-IDF) merupakan algoritma yang menerapkan dua konsep yaitu menghitung frekuensi suatu *term* dalam suatu dokumen dan menghitung *inverse* frekuensi dokumen yang mengandung *term* tersebut. Setelah menghitung TF-IDF maka perlu dilakukan normalisasi agar nilai TF-IDF tetap berada pada rentang 0 sampai 1. [7]

III.3 *Clustering: K-Means*

Clustering merupakan salah satu metode pembelajaran yang paling penting yang digunakan dalam banyak bidang [10]. Penggunaan algoritma clustering secara cepat menjadi salah satu mekanisme komputasi standar untuk memahami berbagai data. Beragam metode ini telah dikembangkan dalam beberapa dekade terakhir [9]. Tujuan utama dari clustering adalah mencapai kesamaan yang tinggi di dalam klaster dan pada saat yang sama kesamaan yang rendah antar klaster. Berbagai metrik dipertimbangkan untuk kesamaan, yang paling umum adalah jarak [14]. Dalam clustering berbasis jarak, objek yang saling dekat dianggap sebagai klaster.

Algoritma K-Means berdasarkan pembagian adalah jenis algoritma pengelompokan yang memiliki kelebihan dalam kesederhanaan, efisiensi, dan kecepatan. Namun, algoritma ini sangat bergantung pada titik awal dan perbedaan dalam pemilihan sampel awal yang selalu menghasilkan hasil yang berbeda. Selain itu, algoritma ini berdasarkan fungsi target selalu menggunakan metode gradien untuk mendapatkan ekstremum. Arah pencarian dalam metode gradien selalu sejalan dengan arah di mana energi berkurang, yang akan menyebabkan jika titik pusat kelompok awal tidak tepat, maka seluruh algoritma akan mudah terjebak pada titik minimum lokal.

III.4. *Elbow Method*

Metode ini adalah metode visual untuk menguji konsistensi jumlah cluster terbaik dengan membandingkan perbedaan *sum of square error* (SSE) dari setiap *cluster*, perbedaan paling signifikan membentuk sudut siku yang menunjukkan jumlah cluster terbaik. Dalam beberapa studi ini, fokus masih pada optimasi penentuan jumlah *cluster* terbaik dengan metode *elbow*, sementara pemilihan awal pusat *cluster* masih acak. Hal ini memungkinkan jumlah iterasi untuk menempatkan objek dalam cluster berdasarkan pusat cluster baru menjadi lebih banyak sehingga pencapaian kesamaan pola yang terbentuk menjadi lebih lama. [17]

III.5 Centroid Distance

Jarak antara sebuah *instance* dan *cluster centroid* dari semua kelas menunjukkan afinitas antara klaster data, yang merupakan suatu pengetahuan untuk proses konstruksi model dalam fase kolaborasi. *Centroid distance* dapat diukur dengan beberapa metode, seperti jarak *Euclidean*, jarak *Mahalanobis*, atau jarak geodesik dari *manifold*. [8]

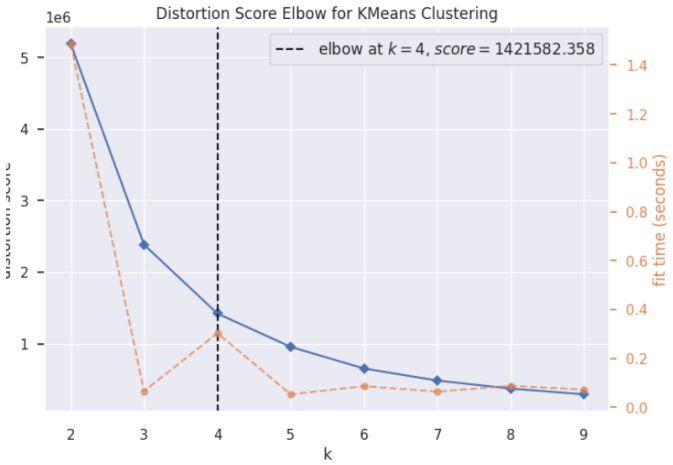
III.6 *Searching Method*: BM-25

Metode pencarian atau *searching method* adalah cara atau strategi yang digunakan untuk menemukan informasi yang relevan atau mencari data yang spesifik dalam suatu konteks. Dalam konteks komputasi, metode pencarian sering digunakan dalam pemrograman atau dalam sistem informasi untuk menemukan data dalam struktur yang lebih besar, seperti array, daftar, atau basis data.

Okapi Best Matching 25 (BM25) adalah fungsi peringkat berbasis *bag-of-words* yang digunakan untuk memperkirakan relevansi suatu dokumen. BM25 juga merupakan suatu model pembelajaran mesin, termasuk model neural yang kuat [12]. BM25 ini merupakan formula terbaik dalam kelas *best match*, dikarenakan formula ini efektif dan memiliki ketepatan dalam mengurutkan dokumen berdasarkan *query* yang dicari.

IV. HASIL DAN PEMBAHASAN

Pada proses pengujian ini, pertama-tama penentuan optimal number k pada K-Means. Optimal number ini nantinya akan menentukan jumlah klaster yang dibuat. Disini kami memakai *KElbowVisualizer* atau *Elbow Method* sebagai penentu jumlah klaster. Kami mendapatkan nilai 4 pada pengujian ini.



Gambar 3. *Elbow Method*

Proses untuk mencari *best cluster* atau klaster terbaik adalah dengan metode *Centroid Distance*. *Centroid Distance* adalah metrik yang digunakan untuk mengukur sejauh mana titik data (*query*) berada dari pusat klaster (*centroid*). Metrik ini berguna untuk menentukan mana yang paling sesuai untuk sebuah *query*. *Query* untuk mencari *best cluster* ada pada tabel 3.

Pada *similiarity word*, kami memakai *word2Vec* sebagai *pretrained* Bahasa Arab dalam Al-Qur'an. Dataset bahasa Arab kami mengambil dari wikiPedia (AraVec) untuk bekerja sama dalam penelitian ini. Selanjutnya dengan bantuan *Gensim* dan *word2Vec* kami dapat memberikan *similarity of word* yaitu makna-makna sama yang akan muncul setelah input *query*.

Dalam menentukan metode yang terbaik untuk penelitian ini, kami tidak langsung mengambil metode saja. Kami melakukan *test* dari tiga metode (*Vector Space Model, Fuzzy,* BM-25) dengan 5 *queries* pada tabel 1.

Tabel 1. *Query* Pengujian untuk Mencari Metode Terbaik

Query	***Translate***
كَذِبٌ	Dusta
فإن	Maka
لا	Tidak atau Jangan
والله	Aku bersumpa
الله	Allah

Pengujian dengan mengambil rata-rata *top 5 highest score relevance* pada setiap *query*. Hasil yang penelitian dapatkan adalah BM-25 dengan skor paling tinggi. Jadi, kami memutuskan untuk memakai BM-25 sebagai metode penelitian.

Tabel 2. Hasil Pengujian untuk Mencari Metode *Searching Terbaik*

Metode	***Score***
Vector Space Model	41
Fuzzy	41,4
BM-25	78,9

Setelah didapatkan *best cluster* dan didapatkan metode *searching* terbaik untuk pengujian ini. Selanjutnya menguji 20 *queries* pada *cluster 3* dengan BM-25. Penelitian ini membaginya menjadi 2 percobaan yaitu dengan *query* biasa dan dengan penggabungan dari kata-kata pada *similarity words* yang telah diproses memakai *word2Vec* sebelumnya.

Tabel 3. *Full Query* Pengujian Penelitian

Query	
Jujur	Tidak atau Jangan
Dia berkata	Mengetahui
Arab	Allah
Kemudian	Mati
Dusta	Lemah
Khianat	Mengatakan
Menyembunyikan	Makan
Bodoh	Minum
Kekal	Bagimu
Maka	Aku bersumpah

Proses selanjutnya, kami mendapatkan hasil setelah melakukan pengujian pada seluruh *query*. Hal-hal yang kami uji adalah rata-rata skor relevansi, benar atau salah, dan tidak mengeluarkan hasil. Untuk pengujian menggunakan *word embedding* kami menambahkan *related* atau *not related* karena adanya ketidak sesuaian pada *query* tetapi memiliki arti atau makna yang sama (berhubungan) dengan *query*. Selain itu, ada *query* yang tidak mengeluarkan output kami tulis dengan *No Result*.

Tabel 4. Hasil Pengujian Penelitian

	Skor Relevansi	**Benar**	**Salah**	***No Result***	***Related***	***Not Related***
Regular query	36,86	65	9	1	-	-
Query with word similarity	54,3	57	21	1	66	12

Pada hasil diatas terlihat jika skor relevansi *query* biasa lebih rendah dari pada dengan *word similarity*. Untuk benar atau salah (*query* ada pada dokumen), *regular query* sedikit lebih besar dari pada *with word similarity*. Hal ini disebabkan kemungkinan representasi vektor kurang akurat ataupun pengaruh variasi data dan pengolahan teks yang menyebabkan *with word similarity* tertinggal dari *regular query*. Adapun *niil result* pada *query* kami yang menurut kami ada kesalahan sistem dalam memproses *query*.

V. DISKUSI

Kami cukup puas dengan hasilnya karena menambah cinta kami kepada Al-Qur'an. Menurut kami, proyek ini menjadi tantangan baru bagi kami. Hasil yang kami dapat cukup bagus karena kami memproses bahasa Arab yang *notabene* bukan bahasa Ibu kami. Selain itu, kami juga berbeda dari kelompok-kelompok lainnya yang menggunakan dokumen dengan teks latin, sedangkan kami sendiri memakai teks Arabic yang sangat susah. Teks Arab sangat susah dimengerti karena sambungan-sambungan antar hurufnya harus diteliti secara hati-hati. Teks Arab lebih rumit karena ada salah sedikit (sambungan, tanda baca, dan *preprocessing text*) langsung memiliki arti yang kacau. Selain itu, dataset yang kami gunakan adalah Al-Qur'an yang mana berisi Bahasa Tuhan. Sedangkan kami memakai *qaamus* atau *dictionary* buatan manusia dan model data wikiPedia. Tatanan Bahasa di dalam Al-Qur'an sendiripun berbeda dengan Bahasa Arab seperti yang dijelaskan di kesimpulan.

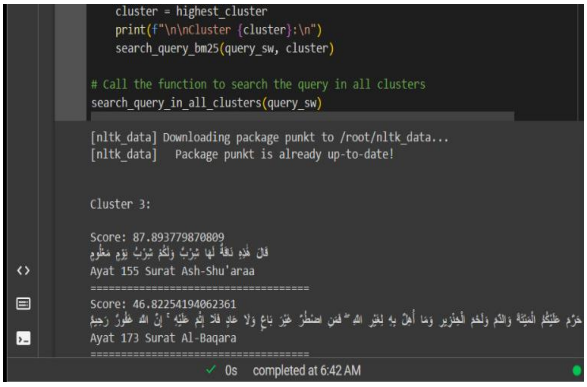
Kekurangan dari penelitian ini yaitu harus menambah jumlah terjemah yang digunakan untuk dapat menerjemahkan secara detail Al-Qur'an dan juga kami menyadari bahwa tidak diperlukan proses *filtering* pada *preprocessing*, dikarenakan keseluruhan makna pada terjemah begitu penting dengan demikian harapannya masyarakat dapat memahahmi

VI. KESIMPULAN

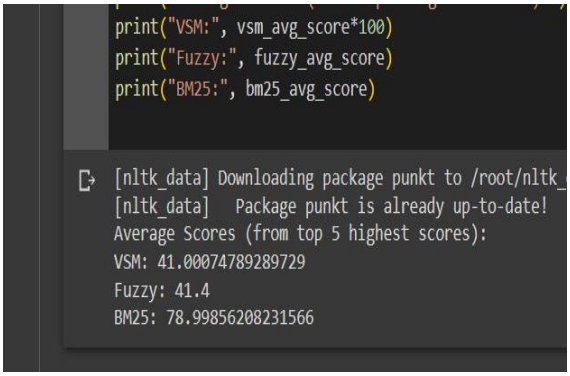
Dalam melakukan analisis dan clustering makna ayat Al-Qur'an dengan mini search engine serta membandingkan berbagai metode, dapat disimpulkan bahwa pendekatan ini memberikan manfaat dalam pemahaman dan penelitian Al-Qur'an. Dengan menggunakan mini search engine, proses pencarian dan indeksasi makna ayat Al-Qur'an dapat dilakukan dengan lebih efisien. Selain itu, melalui analisis makna ayat Al-Qur'an, dapat diidentifikasi pola dan relasi antara ayat-ayat yang memiliki kesamaan topik atau konteks, yang dapat memperkaya pemahaman Al-Qur'an secara keseluruhan.

Berdasarkan hasil pengujian serta analisis pada pengujian pertama, yaitu pengujian menentukan metode *searching*, kami memutuskan untuk memakai BM-25. Setelah serangkaian uji coba, dengan berbagai *query*, kami menemukan bahwa BM-25 mendapatkan skor lebih tinggi dibandingkan 2 metode lain. Untuk *checking process*, kami juga menemukan bahwa BM-25 sesuai dengan *query*. Hal tersebut karena BM-25 menghitung dan mengevaluasi relevansi antara *query* dan dokumen yang ada. Metode ini menggunakan perhitungan bobot yang mempertimbangkan frekuensi kata dalam dokumen dan koleksi dokumen secara keseluruhan. Dengan demikian, BM25 dapat memberikan peringkat yang lebih baik untuk dokumen yang memiliki hubungan yang lebih erat dengan *query*.

Berdasarkan hasil pengujian serta analisis pada pengujian kedua, yaitu pengujian dengan *query* biasa dan dengan penggabungan dari kata-kata pada *similarity words* yang telah diproses memakai *word2Vec* sebelumnya dengan total 20 *queries*. Pada hasil diatas terlihat jika skor relevansi *query* biasa lebih rendah dari pada dengan *word similarity*. Hal itu bisa terjadi karena *with word similarity* menggunakan informasi semantik yang dapat memperkaya *query*, perluasan istilah, analisis semantik lebih dalam dilakukan sehingga skor relevansinya lebih besar. Untuk benar atau salah (*query* ada pada dokumen), *regular query* sedikit lebih besar dari pada *with word similarity*. Hal ini disebabkan kemungkinan representasi vektor kurang akurat ataupun pengaruh variasi data dan pengolahan teks yang menyebabkan *with word similarity* tertinggal dari *regular query*. Walaupun begitu, *with word similarity* memiliki 66 *related* yang berarti ada 9 kata yang tidak persis sama dengan query tetapi memiliki arti yang sama. Adapun *niil result* pada *query* kami yang menurut kami ada kesalahan sistem dalam memproses *query* tersebut yang mana perbedaan antara penulisan *query* dan penulisan Al-Qur'an. Selain itu, kami juga mengambil *pretrained word embedding wikipedia* yang mana model data tersebut tidak sebanding, tidak selengkap, dan tidak sesempurna Al-Qur'an. Sehingga ada sedikit kesalahan dalam memproses data.



Gambar 4a.



Gambar 4b.

Gambar 2 Cuplikan hasil eksperimen : (2a) Pengujian *Searching Query* (2b) Pengujian *Searching Best Method*

Berdasarkan *research* kami, Bahasa Arab dan Bahasa Al-Qur'an tidak memiliki perbedaan, tetapi murni bahasa di Al-Qur'an adalah keistimewaan dari Tuhan YME. kehebatan karya sastra arab yang sangat monumental tidak mampu mengalahkan keindahan bahasa al- Qur'an. Inilah keagungan mukjizat yang dapat dilihat dari tingkatan bahasa yang digunakan dalam al-Qur'an, ia tidak sekedar bahasa Arab biasa, namun bahasa dengan tingkatan yang sangat tinggi. Sehingga menjadikan para ahli bahasa dan para penyair Arab pun berdecak kagum, kekaguman para penyair Arab terhadap bahasa Al-Quran sering terungkap dari lisan mereka. [1]

Dalam penelitian mendatang, dapat dijelajahi penggunaan teknik pengolahan bahasa alami (*Natural Language Processing*) dan pemodelan topik (*Topic Modeling*) yang lebih lanjut untuk menganalisis dan mengelompokkan makna ayat Al-Qur'an. Ini akan memungkinkan identifikasi dan pemahaman yang lebih mendalam terhadap subtopik dan tema yang terkandung dalam Al-Qur'an. Selain itu, penggunaan teknik pembelajaran mendalam (*Deep Learning*) dan jaringan saraf (*Neural Networks*) juga dapat dieksplorasi untuk meningkatkan akurasi dan kekayaan pemahaman hasil *clustering* ayat-ayat Al-Qur'an. Selain itu, nantinya kita dapat memproses Al-Qur'an untuk penerjemahan dan pencarian makna melalui suara untuk membantu orang berkebutuhan khusus.

REFERENSI

[1] Aman, M. (2021) Bahasa Arab dan Bahasa Al-Qur'an. *Tadarus Tarbawy Universitas Muhammadiyah Tangerang*, *3*(1).

[2] Amin, M., & Nurhayat, M. A. (2020). Resepsi Masyarakat terhadap Al-Quran. *Jurnal Ilmu Agama UIN Raden Fatah*, *21*(2), 290–303. https://doi.org/10.19109/jia.v21i2.7423

[3] Du, L., & Hu, C. (2022). Text similarity detection method of power customer service work order based on TFIDF algorithm. *2022 IEEE 5th International Conference on Information Systems and Computer Aided Education (ICISCAE)*, 978–982. https://doi.org/10.1109/ICISCAE55891.2022.9927512

[4] Edison, H., & Carcel, H. (2021). Text data analysis using Latent Dirichlet Allocation: an application to FOMC transcripts. *Applied Economics Letters*, *28*(1), 38–42. https://doi.org/10.1080/13504851.2020.1730748

[5] Ferraro, M. B. (2021). Fuzzy k-Means: history and applications. *Econometrics and Statistics*. https://doi.org/https://doi.org/10.1016/j.ecosta.2021.11.008

[6] Hanandeh, E. S., Abu Awwad, A., & Khassawneh, Y. (2021). Classify Arabic Text using Vector Space Models. *2021 22nd International Arab Conference on Information Technology (ACIT)*, 1–12. https://doi.org/10.1109/ACIT53391.2021.9677134

[7] Havid Albar Purnomo, M., & Abdurrachman Bachtiar, F. (2021). *Pengelompokan Terjemah Al-Quran Departemen Agama menggunakan Metode Fuzzy C-Means* (Vol. 5, Issue 2). http://j-ptiik.ub.ac.id

[8] Kong, D., Bao, Y., & Chen, W. (2020). Collaborative learning based on centroid-distance-vector for wearable devices. *Knowledge-Based Systems*, *194*, 105569. https://doi.org/https://doi.org/10.1016/j.knosys.2020.105569

[9] Li, T., Rezaeipanah, A., & Tag El Din, E. M. (2022). An ensemble agglomerative hierarchical clustering algorithm based on clusters clustering technique and the novel similarity measurement. *Journal of King Saud University - Computer and Information Sciences*, *34*(6, Part B), 3828–3842. https://doi.org/https://doi.org/10.1016/j.jksuci.2022.04.010

[10] Ma, T., Zhang, Z., Guo, L., Wang, X., Qian, Y., & Al-Nabhan, N. (2021). Semi-supervised Selective Clustering Ensemble based on constraint information. *Neurocomputing*, *462*, 412–425. https://doi.org/https://doi.org/10.1016/j.neucom.2021.07.056

[11] Nawang Sari, W. A., & Dwi Purnomo, H. (2022). Topic Modeling Using The Latent Dirichlet Allocation Method On Wikipedia Pandemic Covid-19 Data In Indonesia. *Jurnal Teknik Informatika (Jutif)*, *3*(5), 1223–1230. https://doi.org/10.20884/1.jutif.2022.3.5.321

[12] Nguyen, V., Rybinski, M., Karimi, S., & Xing, Z. (2022). Search like an expert: Reducing expertise disparity using a hybrid neural index for COVID-19 queries. *Journal of Biomedical Informatics*, *127*, 104005. https://doi.org/https://doi.org/10.1016/j.jbi.2022.104005

[13] Saeed, S., Haider, S., & Rajput, Q. (2020). On Finding Similar Verses from the Holy Quran using Word Embeddings. *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*, 1–6. https://doi.org/10.1109/ICETST49965.2020.9080691

[14] Shen, B., Jiang, J., Qian, F., Li, D., Ye, Y., & Ahmadi, G. (2023). Semi-supervised hierarchical ensemble clustering based on an innovative distance metric and constraint information. *Engineering Applications of Artificial Intelligence*, *124*, 106571. https://doi.org/https://doi.org/10.1016/j.engappai.2023.106571

[15] Sinan, M., Leng, J., Shah, K., & Abdeljawad, T. (2023). Advances in numerical simulation with a clustering method based on K–means algorithm and Adams Bashforth scheme for fractional order laser chaotic system. *Alexandria Engineering Journal*, *75*, 165–179. https://doi.org/https://doi.org/10.1016/j.aej.2023.05.080

[16] Tsai, C.-C., Liu, C.-S., & Tai, F.-C. (2021). Adaptive Fuzzy K-Means for Determining Structural Postures of Medical Beds with Multi- Axial Actuators. *2021 International Conference on Fuzzy Theory and Its Applications (IFUZZY)*, 1–6. https://doi.org/10.1109/iFUZZY53132.2021.9605079

[17] Umargono, E., Suseno, J. E., & Gunawan, S. K. V. (2020). K-Means Clustering Optimization Using the Elbow Method and Early Centroid Determination Based on Mean and Median Formula. *Proceedings of the 2nd International Seminar on Science and Technology (ISSTEC 2019)*, 121–129. https://doi.org/10.2991/assehr.k.201010.019

[18] Wang, Z., Zhou, Y., & Li, G. (2020). Anomaly Detection by Using Streaming K-Means and Batch K-Means. *2020 5th IEEE International Conference on Big Data Analytics (ICBDA)*, 11–17. https://doi.org/10.1109/ICBDA49040.2020.9101212

[19] Zhang, Z. (2021). An improved BM25 algorithm for clinical decision support in Precision Medicine based on co-word analysis and Cuckoo Search. *BMC Medical Informatics and Decision Making*, *21*(1), 81. https://doi.org/10.1186/s12911-021-01454-5