Klasifikasi Multi-Label Ayat-Ayat Al-Qur'an Menggunakan Random Forest dan Word Centrality

Rizky Aria Mu'allim[1], Kemas M. Lhaksmana[2]

[1,2]Fakultas Informatika, Universitas Telkom, Bandung
[1]rizkyariamuallim@students.telkomuniversity.ac.id, [2]kemasmuslim@telkomuniversity.ac.id

Abstrak

Penelitian ini memanfaatkan teknologi untuk analisis otomatis topik dalam ayat Al-Qur'an, mengembangkan cakupan analisis dengan klasifikasi ke dalam 15 kategori, termasuk satu 'tidak berlabel'. Fokus penelitian meliputi perbandingan efektivitas antara Random Forest, SVM, dan Naïve Bayes dalam sistem klasifikasi topik ayat Al-Qur'an, dengan Word Centrality sebagai fitur. Tahapan pra-pemrosesan seperti tokenisasi dan penghapusan stopword diterapkan, bersama dengan metode TF-IDF dan TW-IDF. Hasil menunjukkan bahwa Random Forest mencatat skor Hamming Loss terendah dalam skenario TW-IDF, namun hasil TF-IDF dalam skenario menggunakan stopword tidak lebih baik dibandingkan dengan SVM, berturut-turut adalah 0.949 dan 0.0927. Pengujian tanpa penghapusan stopword juga menunjukkan keunggulan relatif hasil hamming loss Random Forest dalam beberapa skenario. Hasil penelitian ini mengindikasikan bahwa penerapan word centrality sebagai metode ekstraksi fitur dalam klasifikasi ayat-ayat Al-Qur'an berpengaruh pada penurunan nilai Hamming Loss.

Kata kunci : klasifikasi multi-label, al-qur'an , word centrality, svm, naïve bayes, random forest.

Abstract

This research utilizes technology for automatic analysis of topics in verses of the Qur'an, expanding the scope of analysis with classification into 15 categories, including one 'unlabeled'. The focus of the research includes a comparison of the effectiveness between Random Forest, SVM, and Naïve Bayes in the Al-Qur'an verse topic classification system, with Word Centrality as a feature. Pre-processing stages such as tokenization and stopword removal are applied, along with TF-IDF and TW-IDF methods. The results show that Random Forest recorded the lowest Hamming Loss score in the TW-IDF scenario, but the TF-IDF results in the scenario using stopwords were no better than SVM, respectively 0.949 and 0.0927. Tests without stopword removal also show the relative superiority of Random Forest's hamming loss results in several scenarios. The results of this research indicate that the application of word centrality as a feature extraction method in the classification of Al-Qur'an verses has an effect on reducing the Hamming Loss value.

Keywords: multi-label classification, al-qur'an , word centrality, svm, naïve bayes, random forest.

1. Pendahuluan

Latar Belakang

Al-Qur'an, kitab suci utama umat Islam, merupakan kalam Allah SWT yang diturunkan kepada Nabi Muhammad SAW. Terbagi menjadi 114 surah dan 30 juz dengan total 6.236 ayat, Al-Qur'an menyajikan prinsip kehidupan, hubungan manusia dengan Allah, serta moral dan etika [1]. Struktur dan makna kata dalam Al-Qur'an penting untuk memahami pesan-pesannya. Pemrosesan bahasa alami (NLP) memberikan potensi besar dalam analisis teks-teks agama, termasuk Al-Qur'an, memudahkan klasifikasi ayat berdasarkan topik.

Perkembangan NLP, terutama dalam klasifikasi multi-label, relevan untuk studi Al-Qur'an. Al-Qur'an sering menyajikan ayat dengan makna multi-topik, menunjukkan kebutuhan untuk klasifikasi multi-label [2]. Penelitian ini memanfaatkan Random Forest dan Word Centrality untuk klasifikasi multi-label ayat Al-Qur'an. Random Forest dipilih karena kemampuannya mengolah data kompleks [3], sedangkan Word Centrality membantu mengidentifikasi kata kunci dalam menentukan topik ayat.

Penelitian yang berkaitan dengan masalah ini [4] telah dilaksanakan dengan memanfaatkan dua algoritma klasifikasi, yaitu Naive Bayes dan Support Vector Machine (SVM). Dalam penelitian tersebut, delapan topik telah dijadikan fokus, serta tiga metode pengukuran sentralitas yang digunakan, yaitu *Degree*, *Betweenness*, dan *Closeness*. Penelitian tersebut melaksanakan dua jenis percobaan, yakni dengan penerapan penghapusan kata berhenti dan tanpa penghapusan kata berhenti. Hasil dari penelitian tersebut menunjukkan bahwa SVM memiliki kinerja yang lebih baik dibandingkan Naive Bayes, dengan nilai kehilangan hamming sebesar 0.15 dan 0.21, pada kondisi dengan penerapan penghapusan kata berhenti.

Dari penelitian tersebut, penulis melakukan penelitian dengan menambahkan algoritma lain untuk membandingkan dengan penelitian sebelumnya dan menambahkan topiknya menjadi lima belas. Proses penelitian ini melibatkan pra-pemrosesan dataset Al-Qur'an, termasuk tokenisasi dan penghapusan *stopword*, untuk memastikan analisis yang akurat. Metode ekstraksi fitur seperti TF-IDF dan TW-IDF, digunakan untuk menonjolkan frekuensi dan pentingnya kata dalam konteks teks. Tujuannya adalah mengembangkan metode efisien

dan sistematis dalam memahami dan menginterpretasikan ayat-ayat Al-Qur'an, memberikan pemahaman yang lebih mendalam tentang teks suci ini.

Topik dan Batasannya

Penelitian ini bertujuan mengukur sentralitas kata dalam ayat Al-Qur'an dan membandingkan kinerja Random Forest dengan model klasifikasi lain. Dataset yg digunakan mencakup 15 topik, termasuk 'tidak berlabel', dengan pra-pemrosesan melalui tokenisasi dan penghapusan stopword. Teknik ekstraksi fitur yang digunakan adalah TF-IDF dan TW-IDF, diterapkan pada data dengan dan tanpa stopword. Proses klasifikasi menggunakan SVM, Naive Bayes, dan Random Forest. Digunakan metode Hamming Loss dalam mengukur performansi setiap metode klasifikasi.

Tujuan

Penelitian ini difokuskan pada perbandingan kinerja antara model klasifikasi Random Forest dan dua model lain, yaitu SVM dan Naïve Bayes. Selain itu, penelitian ini mengusung pembangunan sistem klasifikasi topik untuk ayat Al-Qur'an dengan memasukkan pengukuran Word Centrality sebagai teknik ekstraksi fitur. Penelitian ini juga bertujuan untuk mengevaluasi efektivitas sistem ini dalam dua kondisi pengujian yang berbeda, satu menggunakan penghapusan *stopword* dan satu lagi tanpa menggunakan penghapusan *stopword*.

Organisasi Tulisan

Artikel ini terorganisasi dalam beberapa bagian: Pendahuluan diuraikan pada bagian pertama. Bagian kedua membahas studi yang relevan. Desain sistem dan metodologi penelitian dijelaskan dalam bagian ketiga. Analisis dari hasil penelitian disajikan pada bagian keempat, dan kesimpulan penelitian disimpulkan pada bagian kelima.

2. Studi Terkait

Penelitian terdahulu [4] telah menerapkan Word Centrality pada klasifikasi ayat Al-Qur'an menggunakan Graph of Word (GoW), baik dengan maupun tanpa penghapusan stopword. Hasilnya menunjukkan 'Allah' sebagai kata paling sentral dengan GoW dan penghapusan stopword, menggunakan SVM dan Naïve Bayes untuk peningkatan kinerja yang signifikan.

Terdapat juga penelitian [5] yang melakukan studi tentang pendeteksian SMS Spam menggunakan metode Graph Centrality. Mereka menerapkan teknik Degree Centrality, Eccentricity Centrality, dan Closeness Centrality, dan mengintegrasikannya dengan algoritma Random Forest, SVM, dan Naïve Bayes. Hasil penelitian ini menunjukkan adanya peningkatan yang berarti dalam klasifikasi teks, khususnya dalam konteks deteksi SMS Spam, dengan menggunakan Degree Centrality.

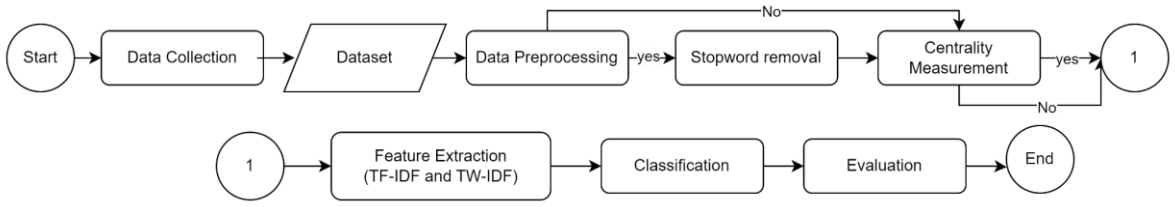
Penelitian terkait [6] menggunakan Word Centrality untuk klasifikasi 10 topik ayat Al-Qur'an dalam terjemahan Inggris. Teknologi ini otomatisasi pemberian topik pada ayat-ayat, membandingkan SVM, Naïve Bayes, KNN, dan Decision Tree. Hasil menunjukkan 'Allah' sebagai kata paling sentral, khususnya saat menghapus stopword.

Penelitian [7] mengkategorikan teks Hadith Sahih Al-Bukhari dengan Random Forest, mengatasi kelemahan overfitting Decision Tree. Dengan 1650 hadits terbagi menjadi tiga kategori, metode ini mencapai F1-Score 90% menggunakan 100 bootstrap tree, dinilai dengan K-Fold Cross Validation.

Dalam penelitian Tugas Akhir ini, penulis mengadopsi beberapa teknik yang telah diterapkan dalam studi-studi sebelumnya. Salah satu pendekatan yang diterapkan oleh penulis adalah analisis sentralitas kata dengan menggunakan metode Graph of Word (GoW). Hasil dari analisis sentralitas kata tersebut selanjutnya diintegrasikan sebagai fitur tambahan dalam proses klasifikasi yang dijalankan oleh beberapa model klasifikasi, termasuk algoritma Support Vector Machine (SVM), Naïve Bayes, dan Random Forest.

3. Sistem yang Dibangun

Dalam Tugas Akhir ini, akan dikembangkan sebuah sistem yang mengikuti alur proses sebagaimana digambarkan dalam Flowchart pada Gambar 1.



Gambar 1. *Flowchart* perancangan sistem

Pengumpulan Data

Penelitian ini menggunakan dataset gabungan. Teks ayat Al-Qur'an dalam bahasa Arab diambil dari Tanzil.net, sementara label untuk klasifikasi ayat diperoleh dari Dataverse. Dataset ini mengklasifikasikan ayat ke dalam 15 topik berbeda seperti kepercayaan, sains, etika, hukum, dan lain-lain[8]. Ayat Al-Qur'an didapat dari website Tanzil.net, dengan total 6236 ayat Al-Qur'an [9]. Pada tabel 1 merupakan contoh data yang diperoleh, yaitu surah Al-Ikhlas. Berikut ini adalah contoh data dari Surah Al-Ikhlas dalam kumpulan data tersebut. 15 Topik yg digunakan pada penelitian penulis, ada pada tabel 2

Tabel 1. Contoh data Surah Al-Ikhlas pada dataset

Verse	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
بسم الله الرحمن الرحيم قل هو الله أحد	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
الله الصمد	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
لم يلد ولم يولد	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ولم يكن له كفوا أحد	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Tabel 2. Topik pada *dataset* yang digunakan

No	Topic	No	Topic
1	Arkanul Islam	9	Akhlak
2	Iman	10	Rules and Community relations
3	Al-Qur'an	11	Matters relating to the law
4	Science	12	Country and Society
5	Amal	13	Agriculture and Trade
6	Dakwah	14	History and Story
7	Jihad	15	Religions
8	Human and Community relations	16	Non-Labeled

Preprocessing

Pra-pemrosesan data dimulai dengan tokenisasi, memecah kalimat menjadi kata atau token, diikuti oleh penghapusan stopword untuk mengeliminasi kata-kata sering muncul namun tidak esensial. Proses ini diilustrasikan pada Tabel 3 menggunakan contoh surah Al-Ikhlas.

Tabel 3. Penerapan preprocessing pada surah Al-Ikhlas

Tahap	Sebelum	Sesudah
Tokenization	بسم الله الرحمن الرحيم قل هو الله أحد	'بسم', 'الله', 'الرحمن', 'الرحيم', 'قل', 'هو', 'الله', 'أحد'
Stopword	'بسم', 'الله', 'الرحمن', 'الرحيم', 'قل', 'هو', 'الله', 'أحد'	'بسم', 'الله', 'الرحمن', 'الرحيم', 'قل', 'الله', 'أحد'

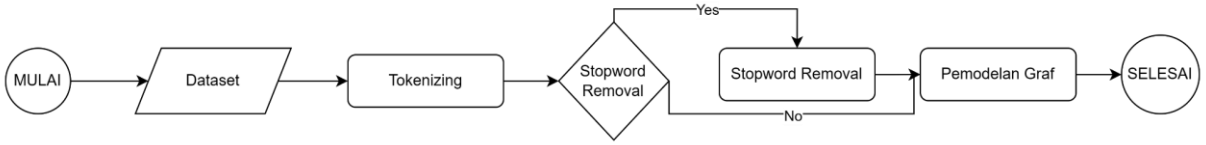
Pengukuran Word Centrality

Word Centrality, digunakan untuk identifikasi kata kunci, menerapkan prinsip sentralitas jaringan sosial dalam Graph of Word (GoW) [10]. Ini efektif untuk menilai pentingnya kata dalam dokumen, mengubah teks menjadi graf untuk analisis mendalam.

GoW dibuat dengan merepresentasikan setiap kata dalam dokumen sebagai node dan menghubungkan setiap kata tersebut dengan edge. Pada ilustrasi yang diberikan dalam Gambar 2, terlihat contoh GoW pada Surah Al-Fatihah yang belum melalui tahap penghapusan *Stopword*

Gambar 2. Penerapan GoW pada surah Al-Fatihah

Dalam penelitian ini, penulis mengembangkan dua versi model Graph of Word (GoW): satu dengan proses penghapusan *stopword* dan satu lagi tanpa penghapusan *stopword*. Sebagai langkah awal, teks data harus melewati proses pembersihan dan tokenisasi. Untuk model GoW pertama, dilakukan penghapusan *stopword* pada teks yang telah diolah sebelumnya. Alur proses pembuatan GoW dapat dilihat pada Gambar 3 berikut.



Gambar 3. Alur pembuatan GoW

Setelah mengembangkan model Graph of Word (GoW), penelitian ini melanjutkan dengan analisis word centrality menggunakan degree centrality, closeness centrality, dan betweenness centrality. Metode ini digunakan untuk mengidentifikasi kata kunci dalam Al-Qur'an, dengan mengilustrasikannya pada model GoW dari Surah Al-Fatihah, di mana bobot pada edge menunjukkan frekuensi koneksi antarkata.

Tahap selanjutnya dalam penelitian ini adalah pengukuran sentralitas dalam model Graph of Word, menggunakan metode Degree Centrality, Closeness Centrality, dan Betweenness Centrality. Pengukuran ini bertujuan untuk menentukan node kunci dalam graf, dengan nilai sentralitas yang dipengaruhi oleh jumlah koneksi setiap verteks (Degree PageRank) [11].

Degree Centrality dihitung berdasarkan jumlah koneksi *edge* yang dimiliki oleh sebuah node dalam suatu jaringan [12]. Dalam model Graph of Word (GoW) yang digunakan, node Vi akan terhubung dengan node Vj jika kata yang diwakili oleh Vj muncul di dekat kata yang diwakili oleh Vi. Rumus perhitungan Degree Centrality dijelaskan dalam persamaan (1).

$$C_{Deg}(Vi) = \frac{|N(Vi)|}{|V|-1} \quad (1)$$

$C_{Deg}(Vi)$ merupakan pengukuran sentralitas Degree Centrality dari sebuah node Vi pada GoW. $N(Vi)$ adalah himpunan node yang terhubung dengan *node* Vi, dan V adalah jumlah total *node* dalam GoW.

Betweenness Centrality mengukur frekuensi sebuah node berada di jalur terpendek antara dua node lain dalam graf [13]. Node dengan nilai betweenness centrality yang tinggi berarti lebih sering terletak di jalur terpendek. Perhitungan Betweenness Centrality berdasarkan total jalur terpendek yang melalui node tersebut dijelaskan dalam persamaan (2).

$$C_{Bet}(i) = \sum_{j<k} \frac{g_{jk}(i)}{g_{jk}} \quad (2)$$

Nilai Betweenness Centrality untuk node i, $C_{Bet}(i)$, dihitung dengan membandingkan jumlah jalur terpendek $Bet\ (i)$ yang melewati node i dengan jumlah total jalur terpendek antara semua pasangan node j dan k. g_{jk} merepresentasikan total jalur terpendek antara node j dan k, sementara $g_{jk}(i)$ menunjukkan jumlah jalur terpendek melalui node i.

Closeness Centrality menghitung rata-rata jarak terpendek dari node ke semua node lain dalam graf, menunjukkan posisi sentral dengan jarak rata-rata yang kecil [13]. Metode ini bersama dengan Degree PageRank dan Betweenness Closeness Eccentricity, digunakan untuk menilai sentralitas verteks.

$$C_{Close} = \frac{N-1}{\sum d(i,j)} \quad (3)$$

C_{Close} adalah nilai closeness centrality dari node i, sedangkan $d(i,j)$ adalah nilai jarak terpendek dari node i ke rj, dan N adalah jumlah keseluruhan node dalam graf

Ekstraksi Fitur

Dalam proses ekstraksi fitur, teknik seperti Vector Space Model, Bag-of-Words, dan TF-IDF vital dalam mengubah teks menjadi format numerik untuk efisiensi pengolahan oleh algoritma pembelajaran mesin. Model Bag-of-Words, yang menginterpretasikan dokumen sebagai koleksi kata, berperan penting dalam mempersiapkan fitur teks untuk model machine learning [14].

TF-IDF, yang menggabungkan term frequency (tf) dan inverse document frequency (idf), adalah teknik ekstraksi fitur penting dalam pemrosesan bahasa alami. Term frequency menunjukkan frekuensi kata dalam dokumen, sementara inverse document frequency menilai keunikannya di seluruh korpus [15]. Kombinasi ini membantu dalam menilai relevansi kata dalam konteks korpus. TF-IDF dihitung menggunakan persamaan (4).

$$W_{t,d} = tf_{t,d} \times \log(\frac{D}{df}) \quad (4)$$

Dimana count(t,d) adalah jumlah kemunculan term t pada dokumen d, dfi adalah jumlah dokumen yang mengandung term i, N adalah jumlah keseluruhan dokumen, dan wt,d adalah nilai pembobotan TF-IDF.

Dalam penelitian ini, dibandingkan tiga teknik pembobotan kata yang berbeda untuk proses ekstraksi fitur: TF-IDF dan TW-IDF. Penelitian [4] menggunakan TW-IDF sebagai metode ekstraksi fitur. Implementasi TW-IDF ini mirip dengan metode TF-IDF, yang dijelaskan dalam persamaan (4), (5), dan (6). Nilai word centrality dari persamaan (1), (2), dan (3) digunakan sebagai Term Weight (TW), dan pada graf tak berarah, TW-IDF dihitung menggunakan persamaan (5).

$$w_{t,d} = tw_t \times \log(\frac{D}{df}) \quad (5)$$

Dimana twt,d merupakan nilai centrality pada kata t di dalam dokumen d

Klasifikasi

Sistem dalam penelitian ini dikembangkan dengan tiga metode klasifikasi untuk mengevaluasi dampak penggunaan word centrality. Untuk klasifikasi data yang terdiri dari 15 topik yang ditentukan selama pengumpulan data, termasuk satu topik yang dikategorikan sebagai 'tidak berlabel', tiga metode yang diadopsi adalah Support Vector Machine (SVM), Naïve Bayes, dan Random Forest.

Support Vector Machine (SVM) adalah algoritma pembelajaran mesin yang digunakan untuk klasifikasi dan regresi. Dalam klasifikasi teks (Text Classification), SVM umumnya digunakan untuk klasifikasi biner (Binary Classification), tetapi selain itu SVM juga dapat digunakan untuk klasifikasi dengan lebih dari dua target klasifikasi (multi-class classification). SVM secara dasarnya adalah sebuah algoritma klasifikasi linear biner yang menggunakan batas keputusan tunggal untuk memisahkan dua kelas [16].

Naïve Bayes, metode klasifikasi yang berdasarkan Teorema Bayes, memprediksi kelas sampel dari probabilitas setiap kelas, mengolah dokumen sebagai vektor fitur biner [17]. Ini merupakan metode pembelajaran terarah yang efektif untuk klasifikasi multilabel [18].

Random Forest, pengembangan dari Decision Tree, mengatasi overfitting dengan menggunakan CART berbasis Gini Index [7]. Metode ini membangun banyak CART yang melakukan voting untuk menentukan kategori dominan, menggunakan bootstrapping sebagai 'bootstrap tree'.

Metode Evaluasi

Penelitian ini menggunakan Hamming Loss sebagai metode evaluasi, yang efektif untuk mengukur kinerja sistem klasifikasi multi-label. Hamming Loss menghitung tingkat keakuratan prediksi dengan menilai jumlah kesalahan prediksi dan label yang tidak teridentifikasi. Ini merupakan metrik penting dalam klasifikasi multi-label, yang mengatasi pembelajaran di mana setiap instance bisa memiliki label berganda, dan dihitung menggunakan persamaan (6) [19].

$$Hamming\ Loss = \frac{1}{NL}\sum_{i=1}^{N}\sum_{j=1}^{L} \times \left[\widehat{y_j^{(i)}} \neq y_j^{(i)}\right] \quad (6)$$

N merupakan total jumlah dokumen, L merupakan total class, $\hat{y}j\ (i)$ merupakan jumlah class prediksi, $yj\ (i)$ merupakan jumlah class sesungguhnya, $[\hat{y}\ j\ (i) \neq yj\ (i)\ ]$ merupakan total kesalahan prediksi dari total keseluruhan prediksi.

4. Evaluasi

Penelitian ini melibatkan dua skenario evaluasi. Pertama, mengevaluasi dampak penghapusan stopword pada klasifikasi, dengan membandingkan hasil klasifikasi dengan dan tanpa penghapusan stopword. Kedua, membandingkan tiga metode klasifikasi, metode ekstraksi fitur, dan efek dari pengukuran word centrality. Performansi diukur dengan menggunakan metode Hamming Loss untuk semua skema perbandingan.

4.1 *Word Centrality* pada Ayat-ayat Al-Qur'an

Pengukuran word centrality pada Al-Qur'an telah dilakukan dengan dua versi model Graph of Word: satu dengan penghapusan *stopword* dan satu tanpa. Hasil pengukuran dapat dilihat pada Tabel 4 dan 5, yang menampilkan lima kata dengan nilai centrality tertinggi untuk setiap metode pengukuran, baik dengan maupun tanpa penggunaan *stopword* removal.

Dalam model GoW yang menerapkan penghapusan *stopword* (Tabel 5), kata 'الله' muncul sebagai kata paling sentral dengan nilai centrality tertinggi dalam *degree*, *closeness*, dan *betweenness centrality*. Sedangkan model GoW tanpa penghapusan *stopword* (4) menunjukkan bahwa kata-kata yang muncul adalah *stopword*, dengan kata 'من' memiliki nilai centrality tertinggi. Dari hasil tersebut, menunjukan penghapusan stopword dapat memberikan informasi kata yang memiliki pengaruh paling tinggi dalam kita suci Al-Qur'an.

Tabel 4. Hasil 5 pengukuran masing-masing centrality dengan skenario tanpa stopword removal

Degree Centrality			Closeness Centrality			Betweenness Centrality		
No	Kata	Centrality	No	Kata	Centrality	No	Kata	Centrality
1	من	0.131230	1	من	0.486024	1	من	0.191937
2	الله	0.083743	2	الله	0.468976	2	الله	0.109177
3	في	0.066523	3	ما	0.443924	3	في	0.081350
4	ما	0.060066	4	في	0.440269	4	ما	0.060689
5	إن	0.054483	5	إن	0.436982	5	إن	0.055568

4.2 Klasifikasi Topik pada Ayat-ayayt Al-Qur'an

Penelitian ini menerapkan tiga metode pembelajaran mesin (Random Forest, SVM, Naïve Bayes) dan membandingkan empat metode ekstraksi fitur (TF-IDF, TW IDF Degree, TW-IDF Closeness, TW- IDF Betweenness) dalam sistem klasifikasi. Proses ini melibatkan pengamatan terhadap pengaruh penghapusan *stopword* dan diukur dengan metode *hamming loss*. Hasilnya, yang terlihat pada Tabel 6 dan 7, menunjukkan sedikit peningkatan kinerja *hamming loss* dengan penggunaan *stopword removal*.

Pada Tabel 6, terlihat bahwa penggunaan TW-IDF dalam ekstraksi fitur meningkatkan kinerja sistem klasifikasi untuk Random Forest. Dengan dan tanpa menggunakan stopword pada klasifikasi *random forest*, memberikan hasil yang tidak terlalu jauh. Lalu, implementasi TW-IDF tidak memberikan peningkatan signifikan pada kinerja sistem untuk SVM dan Naïve Bayes. Namun, metode TW-IDF Betweenness pada tabel 6, dalam klasifikasi Naïve Bayes menghasilkan skor hamming loss yang lebih baik dibandingkan dengan hasil yang ditunjukkan pada Tabel 7.

Lebih lanjut, Tabel 7 menunjukkan bahwa tanpa penerapan penghapusan *stopword*, beberapa metode klasifikasi memberikan hasil yang lebih baik. Ini termasuk hasil dari TF-IDF, TW-IDF Closeness, dan TW-IDF Betweeness pada SVM serta TW-IDF Closenesspada Naïve Bayes. Namun, dari hasil pada tabel 6 dan 7, hasil hamming loss untuk klasifikasi *random forest* tidak dapat lebih baik daripada hasil hamming loss untuk klasifikasi *svm* dalam skenario TF-IDF.

Tabel 5. Hasil 5 pengukuran masing-masing centrality dengan skenario menggunakan stopword removal

Degree Centrality			Closeness Centrality			Betweenness Centrality		
No	Kata	Centrality	No	Kata	Centrality	No	Kata	Centrality
1	الله	0.126055	1	الله	0.446627	1	الله	0.362296
2	قال	0.037192	2	قال	0.382243	2	قال	0.075379
3	قل	0.027242	3	قالوا	0.373609	3	الأرض	0.046631

4	الأر ض	0.026899	4	قل	0.371855	4	قل	0.044191
5	قالوا	0.024223	5	الأرض	0.366681	5	قالوا	0.041477

Tabel 6. Hasil klasifikasi pada pengujian TF-IDF dan TW-IDF, dengan *stopword removal menggunakan hamming loss*

Metode	TF-IDF	TW-IDF Deg	TW-IDF Close	TW-IDF Bet
Random Forest	**0.0949**	0.0974	0.0966	0.0969
SVM	0.0927	0.1009	**0.0926**	0.1027
Naïve Bayes	0.0981	0.1023	**0.0977**	0.1034

Tabel 7 . Hasil klasifikasi pada pengujian TF-IDF dan TW-IDF, tanpa *stopword removal menggunakan hamming loss*

Metode	TF-IDF	TW-IDF Deg	TW-IDF Close	TW-IDF Bet
Random Forest	**0.0956**	0.0957	0.0959	0.0964
SVM	**0.0925**	0.1096	0.1034	0.1101
Naïve Bayes	**0.0991**	0.1079	0.0994	0.1082

5. Kesimpulan

Penelitian ini menunjukkan bahwa dalam konteks klasifikasi multi-label ayat-ayat Al-Qur'an dengan menggunakan Word Centrality, metode Random Forest tidak menunjukkan kinerja yang lebih unggul dibandingkan SVM dalam skenario ekstraksi fitur TF-IDF, dengan skor Hamming Loss berturut-turut adalah 0.949 dan 0.0927. Namun, ketika penghapusan stopword diterapkan, Random Forest mencatat skor Hamming Loss yang lebih rendah pada TW-IDF Degree sebesar 0.0974, Closeness 0.0966, dan Betweenness 0.0969, menunjukkan keunggulannya dalam beberapa skenario tertentu. Temuan ini mengindikasikan bahwa penerapan Random Forest bersama Word Centrality membentuk strategi yang efektif dalam menganalisis teks Al-Qur'an, khususnya untuk pengukuran TW-IDF, baik dengan maupun tanpa penghapusan stopword.

Untuk penelitian di masa mendatang, disarankan untuk mengadopsi pendekatan teknologi pemrosesan bahasa alami (Natural Language Processing, NLP) yang canggih, seperti model bahasa neural seperti BERT, khususnya dalam menganalisis hubungan kontekstual antara ayat-ayat Al-Qur'an. Tujuan utama dari pendekatan ini adalah untuk mengeksplorasi dan memahami bagaimana ayat-ayat yang berbeda dalam Al-Qur'an saling berinteraksi dan terhubung, baik secara semantik maupun tematis.

Daftar Pustaka

[1] Mukjizat al-Quran dan as-sunnah tentang IPTEK. (1997). Indonesia: Gema Insani Press.

[2] Khonsa Izzaty, A. M., Mubarok, M. S., Huda, N. S., & Adiwijaya. (2018). A Multi-Label Classification on Topics of Quranic Verses in English Translation Using Tree Augmented Naïve Bayes. 2018 6th International Conference on Information and Communication Technology (ICoICT). doi:10.1109/icoict.2018.8528802

[3] Jitsakul, Watchreewan; Meesad, Phayung; Sodsee, Sunantha (2017). [IEEE 2017 2nd International Conference on Information Technology (INCIT) - Nakhonpathom, Thailand (2017.11.2-2017.11.3)] 2017 2nd International Conference on Information Technology (INCIT) - Enhancing comment Feedback classification using text classifiers with word centrality measures. , (), 1–5. doi:10.1109/INCIT.2017.8257879

[4] Yulianto, Ferdian, Kemas Muslim Lhaksmana, and Danang Triantoro Murdiansyah. "Classifying Quranic Verse Topics using Word Centrality Measure." Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi) 5.3 (2021): 594-601.

[5] Ishtiaq, A., Islam, M. A., Azhar Iqbal, M., Aleem, M., & Ahmed, U. (2019). Graph Centrality Based Spam SMS Detection. 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST). doi:10.1109/ibcast.2019.8667174

[6] Achmad Salim Aiman, Kemas Muslim Lhaksmana, & Jondri. (2022). Topic Classification of Quranic Verses in English Translation Using Word Centrality Measurement. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 6(5), 803 - 809. https://doi.org/10.29207/resti.v6i5.4358

[7] Afianto, M. F., Adiwijaya, & Al-Faraby, S. (2018). Text Categorization on Hadith Sahih Al-Bukhari using Random Forest. Journal of Physics: Conference Series, 971, 012037. doi:10.1088/1742-6596/971/1/012037

[8] Adiwijaya, Said Al Faraby, and Mohamad Syahrul Mubarok. 2021. Indonesian Translation of the Holy Quran (Multi-label), Dataverse Telkom University, DOI : https://doi.org/10.34820/FK2/XQCNPN

[9] H. Zarrabi-Zadeh, "Tanzil Documents," 2007. http://tanzil.net/docs/home.

[10] Huanhuan Liu, ; Wanggen Wan, ; Jing Lu, ; Wenhui Li, ; Xiaoqing Yu, (2013). [Institution of Engineering and Technology IET International Conference on Smart and Sustainable City 2013 (ICSSC 2013) - Shanghai, China (19-20 Aug. 2013)] IET International Conference on Smart and Sustainable City 2013 (ICSSC 2013) - Centrality study and empirical analysis of microblog network. , (), 304–308. doi:10.1049/cp.2013.2010

[11] Vega-Oliveros, Didier A.; Gomes, Pedro Spoljaric; E. Milios, Evangelos; Berton, Lilian (2019). A multi-centrality index for graph-based keyword extraction. Information Processing & Management, 56(6), 102063–. doi:10.1016/j.ipm.2019.102063

[12] Howlader, P., & Sudeep, K. S. (2016). Degree centrality, eigenvector centrality and the relation between them in Twitter. 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). doi:10.1109/rteict.2016.7807909

[13] Nakajima, Kazuki; Iwasaki, Kenta; Matsumura, Toshiki; Shudo, Kazuyuki (2018). [IEEE 2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom) - Melbourne, Australia (2018.12.11-2018.12.13)] 2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom) - Estimating Top-k Betweenness Centrality Nodes in Online Social Networks. , (), 1128–1135. doi:10.1109/BDCloud.2018.00167

[14] Deng, Xuelian; Li, Yuqing; Weng, Jian; Zhang, Jilian (2018). Feature selection for text classification: A review. Multimedia Tools and Applications, (), –. doi:10.1007/s11042-018-6083-5

[15] Yahav, I., Shehory, O., & Schwartz, D. (2018). Comments Mining With TF-IDF: The Inherent Bias and Its Removal. IEEE Transactions on Knowledge and Data Engineering, 1–1. doi:10.1109/tkde.2018.2840127

[16] Sheykhmousa, M., Mahdianpari, M., Ghanbari, H., Mohammadimanesh, F., Ghamisi, P., & Homayouni, S. (2020). Support Vector Machine vs. Random Forest for Remote Sensing Image Classification: A Meta-analysis and systematic review. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 1–1. doi:10.1109/jstars.2020.3026724

[17] Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, & Sung Hyon Myaeng. (2006). Some Effective Techniques for Naive Bayes Text Classification. IEEE Transactions on Knowledge and Data Engineering, 18(11), 1457–1466. doi:10.1109/tkde.2006.180

[18] Pane, R. A., Mubarok, M. S., Huda, N. S., & Adiwijaya. (2018). A Multi-Lable Classification on Topics of Quranic Verses in English Translation Using Multinomial Naive Bayes. 2018 6th International Conference on Information and Communication Technology (ICoICT). doi:10.1109/icoict.2018.8528777

[19] Wu, Guoqiang & Zhu, Jun. (2020). Multi-label classification: do Hamming loss and subset accuracy really conflict with each other?.