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