

## KLASIFIKASI MULTI LABEL AYAT AL-QURAN MENGUNAKAN MAJORITY VOTING

Bagus Hariyadi<sup>1</sup>, Kemas Muslim Lhaksana<sup>2</sup>

<sup>1,2</sup>Fakultas Informatika, Universitas Telkom, Bandung

<sup>1</sup>bagushariyadi@students.telkomuniversity.ac.id, <sup>2</sup>pembimbing1@telkomuniversity.ac.id,

---

### Abstrak

Kitab suci Al-Quran adalah panduan dan pedoman bagi umat muslim. Al-Quran membahas banyak hal di dalamnya seperti arkanul islam, amal, sains, dan lain-lain. Dengan demikian, untuk mempermudah proses pemahaman ayat suci Al-Quran diperlukan klasifikasi topik dari masing-masing ayat. Model yang dibuat dalam penelitian ini adalah model klasifikasi multi-label yang dibangun dengan metode majority voting. Metode ini dipilih karena secara umum memberikan performa model yang lebih baik dibandingkan hanya menggunakan satu model dasar saja. Setelah itu, performa model akan diukur menggunakan dua jenis ukuran yaitu hamming loss dan mikro F1-score. Berdasarkan hasil pengujian dan eksperimen yang dilakukan performa terbaik dari majority voting diraih ketika menggunakan kombinasi model Random Forest, SVM, dan Adaptive Boosting. Kombinasi model tersebut meraih hamming loss sebesar 0.0878 dan mikro F1-Score sebesar 0.5143. Selain itu, metode majority voting dibuktikan dapat meningkatkan performa dibandingkan dengan penggunaan model secara individu.

**Kata kunci :** Al-Quran, topik, majority voting

---

### Abstract

The holy book of Al-Quran is a guide and manual for Muslims. Al-Quran covers various topics such as the iman, amal, science, and more. Therefore, to facilitate the understanding of the verses of Al-Quran, it is important to categorize the topics of each verse. The model created in this study is a multi-label classification model developed using the majority voting method. This method was chosen because it generally provides better model performance than just one basic model. Subsequently, the model's performance is measured using two types of measures hamming loss and micro F1-score. Based on the experiments, the best performance is achieved using a combination of Random Forest, SVM, and Adaptive Boosting models. This combination achieved 0.5143 for the micro F1-Score and 0.0878 for the hamming loss. Additionally, the majority voting method is proven to improve performance compared to the models used individually.

**Keywords:** Al-Quran, topic, majority voting

---

## 1. Pendahuluan

### Latar Belakang

Kitab suci Al-Quran merupakan kitab yang diturunkan kepada Nabi Muhammad SAW. Salah satu keistimewaan Al-Quran adalah mukjizat yang akan tetap relevan selamanya. Hal ini dapat dilihat pada aspek bahasa arab yang sempurna, tidak mungkin untuk ditiru, dan bukti nyata Al-Quran dalam penemuan ilmiah pada saat ini [1]. Al-Quran terdiri dari 114 surah dan 6236 ayat [2]. Ayat-ayat tersebut bisa berisi ajakan berbuat kebaikan, larangan, aturan bermasyarakat, dan lain-lain. Terlebih lagi, satu ayat dapat memiliki lebih dari satu kategori sehingga klasifikasi yang harus dilakukan adalah klasifikasi multi-label [3].

Secara umum untuk menyelesaikan masalah tersebut perlu digunakan supervised learning jenis klasifikasi perlu digunakan agar dapat melakukan klasifikasi secara otomatis. Akan tetapi, penggunaan supervised learning secara tradisional tidak mampu untuk mengkategorikan data yang bersifat multi-label seperti data ayat Al-Quran [4]. Melihat hal tersebut perlu digunakan paradigma multi-label learning untuk menjadi solusi atas masalah klasifikasi multi-label. Perbedaan pendekatan supervised learning dan multi-label learning terletak pada data yang digunakan untuk pelatihannya. Supervised learning memiliki satu kategori untuk satu observasi, sedangkan multi-label learning satu observasi dapat memiliki lebih dari satu kategori [5]. Setelah mengetahui pendekatan yang akan digunakan langkah selanjutnya adalah melakukan pemodelan menggunakan model pembelajaran mesin.

Model pembelajaran mesin yang digunakan dalam penelitian ini adalah model berjenis ensemble learning. Ensemble learning adalah metode pemodelan yang menggabungkan dua atau lebih model untuk meraih performa yang lebih baik daripada performa model ketika digunakan secara individu [6]. Metode ensemble yang digunakan dalam penelitian ini adalah Majority Voting dengan model dasar Random Forest, SVM, dan logistic regression. Model dasar tersebut digunakan karena popularitas dan performa yang baik dalam melakukan klasifikasi teks [7].

## Topik dan Batasannya

Permasalahan yang diangkat dalam penelitian ini adalah mencari kombinasi model dalam pembangunan model majority voting. Studi kasus yang digunakan dalam penelitian ini adalah melakukan klasifikasi multi-label untuk topik dari ayat Al-Quran dalam bahasa arab. Model-model dasar yang digunakan dalam penelitian ini adalah Random Forest, Support Vector Machine, Logistic Regression, Adaptive Boosting, dan Gradient Boosting. Dari permasalahan tersebut, hal yang akan dijelaskan dalam penelitian ini adalah pertama, bagaimana menentukan kombinasi model yang paling optimal dalam melakukan klasifikasi multi label dari topik ayat Al-Quran dan kedua adalah bagaimana cara mengukur performa model klasifikasi multi label dalam melakukan klasifikasi topik ayat Al-Quran.

## Tujuan

Tujuan dari penelitian ini adalah pertama, memaparkan performa dari model majority voting untuk klasifikasi multi-label dalam hal klasifikasi topik dari ayat Al-Quran ke dalam 15 kategori. Kedua, menunjukkan kombinasi model terbaik berdasarkan ukuran hamming loss dan mikro F1-score untuk melakukan klasifikasi topik. Ketiga, menunjukkan peningkatan performa dari metode majority voting dibandingkan dengan performa model individu.

## Organisasi Tulisan

Bagian pertama dari tulisan ini adalah penjelasan terkait dengan latar belakang, batasan topik, serta tujuan diadakannya penelitian. Selanjutnya, bagian kedua adalah bagian studi terkait yang memaparkan beberapa penelitian yang pernah dilakukan sebelumnya dan selaras dengan penelitian ini. Bagian ketiga menjelaskan sistem yang dibangun dan digunakan dalam penelitian ini. Bagian keempat berisi evaluasi dari sistem yang dibangun pada bagian tiga. Terakhir bagian kelima memaparkan kesimpulan berdasarkan hasil yang didapat pada penelitian ini.

## 2. Studi Terkait

Penelitian yang dilakukan oleh Al-Kabi pada tahun 2021 [1] yang mengkategorikan ayat Al-Quran menjadi salah satu dari 14 topik. Dataset yang digunakan dalam penelitian ini adalah data ayat Al-Quran dengan bahasa asalnya yakni bahasa Arab. Algoritma yang digunakan adalah empat jenis algoritma pembelajaran mesin yaitu K-Nearest Neighbors (KNN), Support Vector Machine, Naïve Bayes, dan Decision Tree. Kemudian, model dilatih menggunakan dataset yang sudah dibersihkan menggunakan teknik pembersihan dan ekstraksi fitur untuk data teks. Hasil dari penelitian ini adalah metode Naïve Bayes memberikan performa paling baik dengan skor akurasi sebesar 99.9% mengalahkan algoritma yang lain.

Kemudian penelitian tentang klasifikasi topik pada ayat Al-Quran di tahun 2021 oleh yulianto [8]. Penelitian ini menggunakan dataset Al-Quran dalam bahasa Arab seperti penelitian Al-Kabi [1]. Akan tetapi, penelitian yang dilakukan yulianto berfokus pada aspek degree centrality dan penggunaan ekstraksi fitur yang bervariasi dalam eksperimennya. Model yang digunakan dalam eksperimen ini adalah model SVM dan Naïve Bayes dan dievaluasi menggunakan Hamming Loss. Hasil dari penelitian ini menunjukkan model SVM dengan ekstraksi fitur TF-IDF memiliki hasil performa yang lebih baik dengan hamming loss sebesar 0.154.

Penelitian serupa dengan penelitian yulianto dilakukan Aiman pada tahun 2022 [9] yang melakukan klasifikasi topik Al-Quran menggunakan degree centrality. Penelitian ini serupa karena penelitian ini menggunakan metode yang sama, tetapi menggunakan dataset yang berbeda yakni dataset Al-Quran terjemahan bahasa inggris. Terdapat perbedaan pada model yang digunakan, penelitian ini menambahkan model KNN dan Decision Tree disamping SVM dan Naïve Bayes. Penelitian ini membuktikan model SVM dan ekstraksi fitur TF-IDF adalah model dengan performa terbaik untuk kasus klasifikasi topik Al-Quran dengan hamming loss sebesar 0.111.

Pemilihan metode yang berbeda dilakukan oleh Mohamed [10] menggunakan ensemble learning pada salah satu surah Al-Quran yakni Al-Baqarah dalam bahasa arab. Surah tersebut dipilih karena memiliki jumlah ayat yang paling banyak yaitu 286 ayat. Kemudian, ayat tersebut diklasifikasikan menggunakan 393 topik, meskipun memiliki distribusi yang tidak seimbang. Metode yang digunakan adalah ensemble voting dengan model dasar SVM, logistic regression, dan random forest. Hasil penelitian ini membuktikan bahwa metode voting classifier memberikan hasil yang lebih baik daripada model dasarnya. Ensemble voting melebihi performa dari ketiga model dasar sebesar 4.6% dari segi performa akurasi.

Penelitian yang menggunakan dataset Al-Quran terjemahan bahasa inggris juga dilakukan oleh Chairulfikri tahun 2022 [11]. Penelitian ini berfokus pada klasifikasi multi-label ayat Al-Quran menjadi 15 kategori sama seperti penelitian [12]. Penelitian Chairulfikri menggunakan model berbasis naïve bayes yang divalidasi menggunakan metode cross-validation. Setelah itu mengimplementasikan metode ensemble majority voting, kemudian dievaluasi menggunakan ukuran hamming loss. Penelitian ini meraih performa terbaik dengan menggabungkan multinomial naïve bayes dan bernouli naïve bayes dengan hamming loss sebesar 0.1167. Penelitian ini meraih performa yang lebih rendah dibandingkan dengan penelitian yang dilakukan oleh Nurfikri