Identifikasi Ayat-Ayat Al-Quran yang Sulit Menggunakan *Random Forest* dan SVM

Satya Rayyis Baruna[1], Kemas Muslim Lhaksmana[2]
[1,2]Fakultas Informatika, Universitas Telkom, Bandung
[1]rayyis@students.telkomuniversity.ac.id,
[2]kemasmuslim@telkomuniversity.ac.id

1. Pendahuluan

Latar Belakang

Diyakini oleh lebih dari satu miliar umat muslim, Al-Quran memegang kedudukan fundamental dalam Islam sebagai wahyu ilahi yang diturunkan kepada Nabi Muhammad melalui perantara Malaikat Jibril. Kitab suci ini terdiri dari 114 surah dengan jumlah ayat bervariasi di setiap surahnya, menghasilkan total kata berkisar antara 77.277 hingga 77.934 kata [1]. Al-Quran berisi kata-kata yang cukup mudah untuk dibaca, namun juga memiliki sejumlah kata yang sering dianggap sulit oleh sebagian orang, sehingga dibutuhkan pendekatan efektif untuk mengidentifikasi dan mengategorikan tingkat kesulitan ayat-ayatnya. Kekayaan linguistik dan kedalaman makna Al-Quran menjadi tantangan tersendiri dalam pemahaman, terutama bagi yang tidak menguasai bahasa Arab atau memiliki keterbatasan dalam pendidikan agama.

Dalam penelitian [2], ditemukan sejumlah faktor utama yang memengaruhi keterbacaan teks Arab, yakni panjang kalimat, panjang kata, frekuensi kata, kompleksitas sintaksis, serta jumlah kata unik dalam teks. Khususnya panjang kalimat dan kata dianggap sangat signifikan karena berpengaruh langsung terhadap beban kognitif pembaca. Selain itu, frekuensi kata dan dan kompleksitas struktur kalimat juga turut menambah tingkat kesulitan dalam memahami teks. Hal ini menunjukan bahwa semakin panjang dan kompleks suatu teks, semakin rendah tingkat keterbacaannya dan semakin sulit dipahami oleh pembaca. Penelitian lain [3] juga menjelaskan bahwa panjang teks memiliki pengaruh yang signifikan terhadap evaluasi keterbacaan, di mana teks yang lebih panjang cenderung menghasilkan skor keterbacaan yang lebih rendah.

Sebagai upaya mendukung pembelajaran membaca Al-Qur'an bagi umat Muslim, penelitian ini akan mengidentifikasi penggunaan metode *Random Forest* dan SVM dalam proses klasifikasi ayat-ayat Al-Qur'an yang sulit, guna meningkatkan pemahaman terhadap ayat-ayat tersebut. Penelitian [4] menunjukan bahwa metode *Random Forest* berhasil diterapkan dalam eksperimen yang membandingkan tujuh jenis fitur linguistik pada enam korpus teks dalam bahasa Inggris dan Rusia. Hasil penelitian menunjukkan bahwa model *Random Forest* memberikan hasil yang kompetitif dalam mengklasifikasikan keterbacaan teks. Dalam beberapa kasus, model ini menunjukkan peningkatan performa ketika diperkaya dengan fitur linguistik tertentu, seperti fitur morfologi dan sintaksis. Penelitian [5] juga menunjukan bahwa metode *Random Forest* memiliki keunggulan yang signifikan dalam mengklasifikasikan teks terjemahan ayat-ayat Al-Qur'an. Hasil dalam penelitian ini menunjukkan bahwa *Random Forest* mencapai akurasi sebesar 66,37%, yang lebih tinggi dibandingkan dengan algoritma SVM yang hanya mencapai 50,56%. Dalam penelitian lain [6], SVM digunakan untuk mengklasifikasi ayat-ayat Al-Quran yang sulit dan membuktikan secara efektif dengan akurasi sebesar 95,17%.

Penelitian ini akan berfokus pada pengklasifikasian ayat-ayat yang sulit dalam Al-Quran, berbeda dengan penelitian-penelitian sebelumnya yang berfokus pada keterbacaan dan klasifikasi teks secara umum. Upaya ini bertujuan untuk menentukan metode paling efektif dalam mengklasifikasikan ayat-ayat yang sulit dalam Al-Quran. Selain itu, hasilnya diharapkan dapat memberikan kontribusi berharga bagi pengembangan bidang *Natural Language Processing* dan kajian Al-Quran, serta menjadi dasar bagi penciptaan alat atau aplikasi yang mempermudah pembacaan Al-Quran secara lebih efisien.

Topik dan Batasannya

Penelitian ini berfokus pada pengembangan model klasifikasi untuk mengidentifikasi ayat-ayat Al-Qur'an yang cenderung sulit dipahami, terutama bagi pembaca tanpa dasar kuat bahasa Arab atau ilmu tafsir. Kerumitan bahasa dan kedalaman makna dalam Al-Qur'an sering menjadi penghalang dalam proses pembelajaran dan pemahaman, terutama bagi kalangan pemula. Oleh karena itu, penelitian ini menghadirkan pendekatan berbasis teknologi melalui metode machine learning yang dapat membantu mengidentifikasi ayat-ayat mana yang relatif lebih sulit secara linguistik.

Penelitian ini mengombinasikan *Natural Language Processing* (NLP) dengan teknik klasifikasi *supervised learning*, menggunakan algoritma *Random Forest* sebagai metode utama dan *Support Vector*

Machine sebagai pembanding. Fitur-fitur linguistik dari ayat-ayat Al-Qur'an, seperti panjang kalimat, kompleksitas kosakata, dan bobot kata (melalui TF-IDF), menjadi masukan (*input*) dari model ini. Hasil (*output*) yang diharapkan adalah kategori tingkat kesulitan berupa label "mudah" atau "sulit" untuk setiap ayat.

Dalam penelitian ini, penulis membatasi ruang lingkup pada analisis teks ayat-ayat Al-Qur'an dalam bahasa Arab tanpa melibatkan tafsir atau terjemahan. Dataset yang digunakan merupakan kumpulan 625 ayat yang sudah dilabeli berdasarkan tingkat kesulitan oleh sumber terdahulu. Pelabelan ini tidak diperoleh melalui survei atau pendapat langsung dari pembaca baru, sehingga penelitian tidak melibatkan proses pengumpulan opini publik tambahan. Pemrosesan teks dibatasi pada langkah-langkah dasar seperti *tokenization* dan *stopword removal*, tanpa melakukan *stemming* dan *lemmatization*.

Analisis kesulitan ayat dilakukan hanya berdasarkan ciri-ciri linguistik seperti panjang kalimat atau kompleksitas kosakata, tanpa memperhitungkan konteks teologis atau sejarah yang mungkin juga memengaruhi tingkat pemahaman. Dengan pembatasan ini, diharapkan penelitian tetap fokus, realistis, dan memberikan kontribusi awal dalam mengembangkan sistem klasifikasi kesulitan ayat Al-Qur'an secara otomatis.

Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, penelitian ini dirumuskan untuk menjawab pertanyaan-pertanyaan berikut:

1. Bagaimana membangun model klasifikasi otomatis yang mampu mengidentifikasi tingkat kesulitan ayat-ayat Al-Qur'an menggunakan algoritma *Random Forest* dan teknik *Natural Language Processing* (NLP)?
2. Bagaimana pemanfaatan fitur linguistik, seperti panjang kalimat, kompleksitas kosakata, dan bobot kata berbasis TF-IDF, dapat membantu proses klasifikasi ayat ke dalam kategori *mudah* atau *sulit*?
3. Bagaimana kinerja model klasifikasi yang dibangun ketika dievaluasi menggunakan metrik *accuracy, precision, recall,* dan *F1-score* ?

Tujuan

Penelitian ini memiliki tujuan sebagai berikut :

1. Membangun model klasifikasi otomatis untuk mengidentifikasi tingkat kesulitan ayat-ayat Al-Qur'an menggunakan algoritma *Random Forest* dan teknik *Natural Language Processing* (NLP).
2. Memanfaatkan fitur linguistik seperti panjang kalimat, kompleksitas kosakata, dan bobot kata berbasis TF-IDF untuk mengelompokkan ayat ke dalam kategori mudah atau sulit.
3. Melakukan evaluasi performa model menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*.

Organisasi Tulisan

Tugas akhir ini dimulai dengan pendahuluan yang menyajikan latar belakang pentingnya Al-Quran dalam Islam dan tantangan memahami ayat-ayat kompleks, serta motivasi pengembangan sistem klasifikasi otomatis dengan *Random Forest* dan SVM untuk mengidentifikasi tingkat kesulitan ayat. Dilanjutkan dengan kajian literatur yang menganalisis penelitian sebelumnya tentang klasifikasi teks Al-Quran menggunakan berbagai metode machine learning. Kemudian bagian sistem yang dibangun menjelaskan dataset Al-Quran, proses preprocessing data, ekstraksi fitur linguistik seperti panjang kalimat dan bobot TF-IDF, membagi data menjadi dua subset utama, serta implementasi algoritma *Random Forest*. Diikuti bagian evaluasi yang menyajikan performa model dalam mengklasifikasikan ayat-ayat berdasarkan tingkat kesulitan dengan metrik *accuracy*, *precision*, *recall*, dan *F1-score*, serta analisis mendalam tentang fitur yang paling berpengaruh. Diakhiri dengan kesimpulan yang merangkum temuan utama serta saran untuk pengembangan penelitian di masa depan, dengan materi pendukung seperti detail dataset dan kode program disertakan dalam lampiran.