Klasifikasi Topik Multi Label pada Hadis Bukhari dalam Terjemahan Bahasa Indonesia Menggunakan Random Forest

Adhitia Wiraguna[1], Said Al Faraby, S.T., M.Sc [2], Prof. Dr. Adiwijaya, S.Si, M.Si[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]wiragynaadhitia@students.telkomuniversity.ac.id, [2]saida1faraby@telkomuniversity.ac.id,
[3]adiwijaya@telkomuniversity.ac.id

Abstrak

Hadis merupakan hal yang wajib untuk dipelajari dan diamalkan oleh umat Islam. Terdapat banyak jenis ajaran yang dapat diambil oleh manusia dengan mempelajari hadis. Untuk membantu umat Islam dalam mempelajari hadis, dibutuhkan sistem klasifikasi multi label untuk mengategorikan Hadis Shahih Bukhari terjemahan bahasa Indonesia berdasarkan tiga topik yaitu larangan, anjuran dan informasi. Dalam membangun sistem klasifikasi teks, terdapat berbagai metode klasifikasi yang dapat digunakan, pada penelitian ini menggunakan *Random Forest* (RF). Kesederhanaan algoritma RF dan kemampuan yang baik dalam menghadapi data berdimensi tinggi, membuat RF merupakan metode yang cocok dalam melakukan klasifikasi teks. Namun belum banyak diketahui kemampuan RF untuk klasifikasi *multi label*. Penelitian ini menggunakan metode pendekatan *Problem Transformation* yaitu *Binary Relevance* (BR) dan *Label Powerset* (LP) untuk mengadaptasi RF dalam membangun sistem klasifikasi teks *multi label*. Hasil penelitian menunjukkan bahwa performansi *hamming loss* yang terbaik didapat dari sistem yang menggunakan BR dan tidak menggunakan *stemming* yaitu sebesar 0,0663. Hasil ini menunjukkan bahwa metode BR lebih baik daripada metode LP dalam mengadaptasi algoritma RF untuk melakukan klasifikasi *multi label* terhadap data hadis. Hal ini dikarenakan metode BR menghasilkan model klasifikasi sebanyak jumlah label pada data hadis dan pada sisi lainnya, hasil transformasi data dari penggunaan LP membuat data yang digunakan menjadi *imbalanced*.

Kata kunci : Klasifikasi, hadis, multi label, *random forest*, *problem transformation*

Abstract

Hadith is a mandatory thing to be studied and practiced by Muslims. There are many types of teachings that humans can take by studying the hadith. To assist Muslims in studying the hadith, a multi label classification system is needed to categorize Sahih Bukhari Hadi in Indonesian translation based on three topics, namely prohibition, advice and information. In building a text classification system, there are various classification methods that can be used, in this study using Random Forest (RF). The simplicity of the RF algorithm and good ability to deal with high dimensional data, make RF a suitable method of text classification. But, there is not widely known RF capability for the multi label classification. This study uses the Problem Transformation approach method, namely Binary Relevance (BR) and Label Powerset (LP) to adapt RF in building a multi label classification system. The results showed that the best hamming loss performance obtained from a system that used BR and does not use *stemming* which is equal to 0,0663. These results indicate that the BR method is better than the LP method in adapting the RF algorithm to perform multi label classification of hadith data. This is happened because the BR method produces a classification model of the number of labels in the hadith data and on the other hand, the transformation of data from the use of LP makes the data are imbalanced.

Keywords: Classification, hadith, multi label, random forest, problem transformation

1. Pendahuluan

1.1 Latar Belakang

Hadis yang merupakan dasar Agama Islam tentunya menjadikan hadis sebagai hal yang wajib untuk dipelajari dan diamalkan oleh umat Islam. Di dalam hadis, terdapat beberapa jenis ajaran yang dapat diambil oleh manusia. Beberapa hadis merupakan anjuran untuk umat Islam dalam menjalani kehidupan di dunia. Ada juga hadis yang berisikan larangan-larangan dalam berperilaku sebagai umat Islam. Namun, beberapa hadis ada yang bukan merupakan kedua hal tersebut, yang bisa dikatakan hanyalah sebagai informasi kepada umat Islam. Melihat hal ini, maka hadis dapat dibagi ke dalam tiga topik yaitu anjuran, larangan dan informasi. Untuk mempermudah umat Islam dalam mengamalkan kewajibannya yaitu mempelajari hadis, dibutuhkan sebuah sistem klasifikasi agar dapat mengategorikan hadis berdarkan ketiga topik tersebut. Namun, ternyata secara makna, kategori suatu hadis juga dapat merupakan gabungan antara ketiga kategori tersebut yang berarti suatu hadis dapat memiliki lebih dari satu label. Model klasifikasi dengan model kasus tersebut disebut dengan klasifikasi *multi-label*.

Untuk membangun sistem klasifikasi teks, terdapat berbagai pendekatan klasifikasi yang dapat digunakan, salah satunya adalah *Random Forest* (RF). RF merupakan metode yang baik dalam klasifikasi teks karena

kesederhanaannya dan kemampuan generalisasinya dalam menghadapi data berdimensi tinggi [1]. Kemampuan RF dalam melakukan klasifikasi sudah terkenal baik untuk model klasifikasi *single label* [1] [2], namun belum banyak diketahui kemampuan RF dalam membangun model klasifikasi *multi label*.

Terdapat berbagai pendekatan untuk mengadaptasi permasalahan klasifikasi *multi label*, salah satunya merupakan *Problem Transformation Method*. *Problem Transformation Method* secara umum akan menyelesaikan model permasalahan *multi label* dengan mengubahnya menjadi model permasalahan *single label* dan mengintegrasikan hasilnya kembali ke dalam bentuk *multi label* seperti *Binary Relevance* dan *Label Powerset* [3].

Dengan begitu, permasalahan yang muncul adalah apakah algoritma Random Forest juga merupakan algoritma yang efisien dalam membangun sistem klasifikasi untuk mengategorikan data korpus hadis ke dalam tiga label tersebut dan apakah algoritma Random Forest dapat diadaptasi dengan baik untuk model permasalahan klasifikasi *multi label*. Oleh karena itu, fokus penelitian ini adalah pembangunan sebuah sistem klasifikasi *multi label* teks terhadap data Hadis Shahih Bukhari terjemahan Bahasa Indonesia ke dalam tiga label yaitu anjuran, larangan dan informasi dengan mengimplementasikan algoritma RF dan menganalisis pengaruh jenis metode *Problem Transformation* yang digunakan. Hasil klasifikasi yang didapat juga akan dianalisis untuk mendapatkan gambaran penyebab terjadinya kesalahan model dalam memprediksi kelas dan label dari data. Metode evaluasi yang digunakan dalam penelitian ini adalah *hamming loss*.

1.2 Tujuan

Penelitian yang dilakukan bertujuan untuk membangun sistem yang dapat melakukan klasifikasi topik dengan model klasifikasi *multi label* terhadap data Hadis Shahih Bukhari terjemahan Bahasa Indonesia ke dalam gabungan dari tiga label yaitu anjuran, larangan, dan informasi menggunakan *Random Forest* dan melakukan analisis performa dari sistem klasifikasi yang dibangun menggunakan metode evaluasi *hamming loss*.

Tabel 1. Tujuan dari penelitan yang dilakukan

No	Tujuan	Pengujian	Kesimpulan
1	Membangun sistem klasifikasi menggunakan *Random Forest* untuk melakukan klasifikasi topik pada data hadis terjemahan Bahasa Indonesia.	Sistem melakukan klasifikasi topik pada data hadis terjemahan Bahasa Indonesia.	Mendapatkan data hadis yang sudah berlabel.
2	Menganalisis performa sistem klasifikasi dalam melakukan klasifikasi data teks *multi label*.	Melakukan pengujian terhadap perbedaan penggunaan jumlah trees, jenis *Problem Transfomation* yang digunakan, dan penggunaan *stemming*.	Mendapatkan nilai perfomansi yaitu *hamming loss* minimum dari sistem dalam pengujian.

2. Dasar Teori

2.1 Klasifikasi Hadis

Klasifikasi hadis telah pernah dilakukan dengan menggunakan empat belas kelas seperti pada penelitian [4]. Teks hadis juga telah dapat dikategorisasikan terhadap satu label antara kelas berupa anjuran, larangan dan informasi seperti pada penelitian [4][5]. Tetapi, penelitian-penelitian tersebut masih menggunakan model klasifikasi multi-class dan single-label.

Pada penelitian ini, akan dilakukan klasifikasi topik hadis dalam terjemahan bahasa Indonesia menggunakan tiga kelas berupa anjuran, larangan dan informasi. Model klasifikasi yang digunakan merupakan klasifikasi multi-label. Perbedaan dari penelitian-penelitian sebelumnya adalah dimana suatu hadis dapat dikategorisasikan ke dalam lebih dari satu label.

2.2 Klasifikasi *Multi Label*

Klasifikasi Multi-Label sudah menjadi perhatian para peneliti dalam mengembangkan dan menerapkannya di dunia Machine Learning. Pendekatan untuk menerapkan Multi-label pun sudah banyak diciptakan. Pendekatan dalam metode klasifikasi multi-label dibagi menjadi dua kategori utama, yaitu Problem Transformation Methods dan Algorithm Adaptation Methods [2].

Penelitian ini akan menggunakan metoda Random Forest (RF) yang merupakan pendekatan Problem Transformation Methods. Klasifikasi multi-label pada penelitian ini menggunakan tiga kelas yaitu berupa anjuran, larangan dan informasi.