

Klasifikasi *Multi Label* pada Hadis Bukhari Terjemahan Bahasa Indonesia Menggunakan *Mutual Information* dan *k-Nearest Neighbor*

Afrian Hanafi¹, Adiwijaya², Widi Astutis³

^{1,2,3}Fakultas Informatika, Universitas Telkom, Bandung
¹afrianhanafi@students.telkomuniversity.ac.id, ²adiwijaya@telkomuniversity.ac.id,
³widiwdu@telkomuniversity.ac.id

Abstrak

Hadis merupakan sumber hukum kedua bagi umat muslim yang setelah Al-Qur'an yang mana berasal dari berbagai bentuk ucapan, tindakan, dan ketetapan Nabi Muhammad SAW atau disebut sebagai sunnah-nya. Untuk mempermudah umat muslim dalam menerapkan ajaran yang ada pada hadis, diperlukan sebuah sistem klasifikasi yang dapat mengkategorikan suatu hadis kedalam suatu kelas ataupun gabungan dari dua diantara ketiga kelas atau disebut dengan klasifikasi *multi label*. Pada pembangunan sistem klasifikasi teks, terdapat berbagai teknik klasifikasi salah satunya yaitu *k-Nearest Neighbor* (KNN). KNN merupakan suatu metode klasifikasi yang sederhana dan efektif untuk klasifikasi teks, namun memiliki kelemahan dalam memproses data dengan dimensi vektor yang tinggi sehingga menyebabkan waktu komputasi menjadi lebih tinggi dan efisiensi dari klasifikasi teks sangat rendah. *Mutual Information* (MI) digunakan sebagai metode *feature selection* untuk mereduksi dimensi vektor karena memiliki kemampuan untuk menunjukkan seberapa berpengaruh suatu fitur dalam melakukan prediksi yang tepat terhadap suatu kelas. Pada penelitian ini *Problem Transformation Method* dengan pendekatan *Binary Relevance* (BR) digunakan agar proses klasifikasi *multi label* dapat dilakukan. Hasil optimum yang didapat pada penelitian ini menunjukkan nilai *hamming loss* sebesar 0.0886 atau sekitar 91.14% data yang terklasifikasi dengan benar dan waktu komputasi selama 595 detik dengan menggunakan MI sebagai *feature selection*, namun tanpa *stemming*.

Kata kunci : klasifikasi *multi-label*, hadis bukhari, *pre-processing*, *k-nearest neighbor*, *mutual information*, *hamming loss*

Abstract

Hadith is the second source of law for Muslims after the Qur'an which comes from various forms of the words, actions and stipulations of the Prophet Muhammad or referred to as his sunnah. In order to make it easier for Muslims to apply the teachings of the hadiths, a classification system is needed that can categorize a hadith into a class or a combination of two of the three classes which called a multi-label classification. In building a text classification system, there are various classification techniques, one of which is k-Nearest Neighbor (KNN). KNN is a simple and effective classification method for text classification, but has a weakness in processing data with high vector dimensions so that the computation time is higher and the efficiency of text classification is very low. Mutual Information (MI) is used as a feature selection method to reduce vector dimensions because it has the ability to show how strong a feature is in making a correct prediction of a class. In this study Problem Transformation Method with the Binary Relevance (BR) approach is used so that the multi label classification process can be accomplished. The optimum results obtained in this study shows the value of hamming loss is 0.0886 or about 91.14% of data were correctly classified and computational time for 595 seconds by using MI as a feature selection, but without stemming.

Keywords : multi-label classification, bukhari's hadith, pre-processing, k-nearest neighbor, mutual information, hamming loss

1. Pendahuluan

1.1 Latar Belakang

Hadis merupakan sumber hukum kedua bagi umat muslim yang setelah Al-Qur'an yang mana berasal dari berbagai bentuk ucapan, tindakan, dan ketetapan Nabi Muhammad SAW atau disebut sebagai sunnah-nya [1][2]. Hadis juga berfungsi sebagai memperjelas dan menegaskan hukum-hukum yang ada di Al-Qur'an. Umumnya hadis tersebut diriwayatkan oleh para ahli hadis, salah satunya yaitu Imam Bukhari. Tujuannya supaya generasi yang akan datang dapat menerapkan nilai-nilai yang diterapkan oleh Nabi Muhammad SAW. Masing-masing hadis memiliki beberapa jenis ajaran yang dapat diterapkan dalam kehidupan umat muslim sehari-hari. Ajaran yang dimaksud seperti anjuran, informasi, hingga larangan-larangan [3]. Namun terdapat kendala dalam mempelajari hadis untuk diterapkan isi dan maknanya di kehidupan sehari-hari apabila umat Islam kesulitan

dalam menggolongkan hadis ke dalam anjuran, larangan, dan informasi maupun gabungan dari kedua diantara ketiga jenis ajaran tersebut. Atas dasar itu, untuk mempermudah umat muslim dalam menerapkan ajaran yang ada pada hadis, diperlukan sebuah sistem klasifikasi yang dapat mengkategorikan dan mengidentifikasi suatu hadis kedalam suatu kelas ataupun gabungan dari dua diantara ketiga kelas atau disebut dengan klasifikasi *multi label*.

Pada pembangunan sistem klasifikasi teks, terdapat berbagai teknik klasifikasi salah satunya yaitu *k-Nearest Neighbor* (KNN). KNN merupakan suatu metode klasifikasi yang sederhana dan efektif untuk klasifikasi teks dengan pendekatan pembelajaran berbasis sampel atau disebut dengan *supervised learning*, yang mana pada tahapan klasifikasi menggunakan seluruh data latih untuk memprediksi label dari data uji. KNN memiliki kelemahan dalam memproses data dengan dimensi vektor yang tinggi sehingga menyebabkan waktu komputasi menjadi lebih tinggi dan efisiensi dari klasifikasi teks sangat rendah [4][5][6]. Salah satu solusi yang dapat dilakukan untuk meningkatkan kinerja dari KNN yaitu dengan mereduksi dimensi vektor [5]. Untuk mereduksi data dengan dimensi vektor yang tinggi perlu dilakukannya *feature selection* untuk menyeleksi kata atau fitur yang dianggap kurang relevan dalam pembentukan suatu model. *Mutual Information* (MI) digunakan sebagai metode *feature selection* karena memiliki kemampuan untuk menunjukkan seberapa berpengaruh suatu fitur dalam melakukan prediksi yang tepat terhadap suatu kelas [7][8].

Berdasarkan penjabaran diatas, maka pada penelitian ini dilakukan klasifikasi *multi label* pada Hadis Bukhari Terjemahan Bahasa Indonesia kedalam suatu kelas ataupun gabungan dari dua diantara ketiga kelas yang ada yaitu anjuran, larangan dan informasi dengan menggunakan KNN sebagai metode klasifikasi dan MI sebagai metode *feature selection*. Namun umumnya klasifikasi teks dengan KNN hanya dapat dilakukan untuk data dengan *single label*. Sehingga *Problem Transformation Method* juga perlu digunakan agar proses klasifikasi *multi label* dapat dilakukan. *Binary Relevance* (BR) merupakan salah satu pendekatan yang ada pada *Problem Transformation Method*. Dasar algoritma dari BR yaitu untuk mendekomposisi permasalahan *multi label* menjadi permasalahan *single label* dan mengintegrasikan hasil klasifikasinya kembali menjadi bentuk *multi label* [9]. Metode evaluasi diperlukan untuk mengetahui hasil performansi dari sistem yang dibangun. Pada penelitian ini, *hamming loss* digunakan untuk mengukur performansi dari hasil klasifikasi *multi label* yang didapatkan.

1.2 Tujuan

Penelitian ini bertujuan untuk membangun sistem klasifikasi *multi label* terhadap data Hadis Bukhari terjemahan Bahasa Indonesia kedalam suatu kelas ataupun gabungan dari dua diantara ketiga kelas yaitu anjuran, larangan, dan informasi menggunakan *K-Nearest Neighbor* dan *Mutual Information* serta melakukan analisis performansi dari sistem klasifikasi yang dibangun menggunakan *hamming loss*.

Tabel 1 Tujuan Penelitian

No	Tujuan	Pengujian	Kesimpulan
1	Membangun sistem klasifikasi menggunakan <i>K-Nearest Neighbor</i> dan <i>Mutual Information</i> pada data Hadis Bukhari terjemahan Bahasa Indonesia.	Sistem melakukan klasifikasi pada data Hadis Bukhari terjemahan Bahasa Indonesia.	Mendapatkan data hadis yang sudah berlabel.
2	Menganalisis performansi dari sistem yang dibangun dalam melakukan klasifikasi.	Melakukan pengujian terhadap penggunaan <i>stemming</i> pada tahap <i>pre-processing</i> , penggunaan <i>Mutual Information</i> sebagai metode <i>feature selection</i> dan pengujian parameter <i>k</i> pada tetangga terdekat atau <i>classifier</i>	Mendapatkan nilai <i>hamming loss</i> minimum dari sistem dalam pengujian yang mana merupakan hasil optimum.

2. Studi Terkait

2.1 Penelitian Terkait

Penelitian mengenai klasifikasi Hadis Bukhari terjemahan Bahasa Indonesia dengan kelas anjuran, larangan dan informasi sudah pernah dilakukan beberapa kali menggunakan pendekatan dan metode yang beragam salah satunya penelitian oleh S. Al Faraby [10] yang berfokus pada permasalahan klasifikasi *single label* dengan menggunakan *Support Vector Machine* dan *Artificial Neural Network*. Pada penelitian ini juga dibangun *classifier* dasar menggunakan metode *rule-based*. Hal tersebut disebabkan karena kelas atau kategori yang dimaksud cukup baru pada saat penelitian dilakukan. Hasil yang didapat pada metode *rule-based classifier* yaitu dengan nilai *f1-score* sebesar 0.69, sedangkan nilai *f1-score* yang optimum dengan pendekatan pembelajaran mesin dihasilkan oleh model *Support Vector Machine* sebesar 0.88 dengan *kernel* linear.

Untuk kasus klasifikasi *multi label* telah dilakukan oleh H. Prasetyo [11] yang mana pada penelitiannya menggunakan *Backpropagation Neural Network* dan *Mutual Information*. Penelitian ini mendapatkan nilai