Klasifikasi Multi Label Pada Hadis Bukhari Terjemahan Bahasa Indonesia menggunakan Random Forest, Mutual Information, dan Chi-Square

Hadiyan Nadzri Harish[1], Said Al Faraby[2], Mahendra Dwifebri[3]

[1,2,3] Universitas Telkom, Bandung

[1]Hnadzriaji@student.telkomuniversity.ac.id, [2]saidalfaraby@telkomuniversity.ac.id, [3]mahendradp@telkomuniversity.ac.id

Abstrak

Hadis merupakan sumber hukum kedua bagi umat Islam setelah Al-Qur'an. Oleh karena itu, umat Islam dianjurkan untuk mengetahui dan mengamalkannya. Umumnya Hadis dikoleksi oleh beberapa imam besar, salah satunya adalah koleksi Hadis Imam Bukhari. Dalam Hadis terdapat beberapa kategori jenis ajaran, seperti jenis ajaran Hadis yang mengandung anjuran, larangan, dan informasi. Untuk mengenali karakteristik Hadis terjemahan Bahasa Indonesia berdasarkan kategorinya, pada penelitian ini akan dibangun sebuah sistem yang mampu menggelompokkan Hadis kedalam tiga kategori yaitu anjuran, larangan, dan informasi. Dalam mengelompokkannya diperlukan sebuah sistem klasifikasi. Pada penelitian ini, berbagai metode klasifikasi dapat digunakan salah satunya adalah Random Forest. Random forest merupakan metode klasifikasi yang memiliki kemampuan menggeneralisasi suatu data berdimensi tinggi. Random Forest ini dipercaya dapat menyelesaikan proses klasifikasi dengan hasil yang akurat, namun memiliki kelemahan yaitu terjadinya *overfitting* ketika menghadapi jenis data dengan jumlah fitur yang banyak. Dalam penelitian ini seleksi fitur yang akan digunakan yaitu Chi-Square. Metode seleksi fitur dapat membantu proses penyeleksian dari sekumpulan fitur asli dengan tujuan menyisihkan fitur-fitur yang tidak relevan terhadap masing-masing kelas. Nilai akurasi optimum yang dihasilkan dari beberapa pengujian yang dilakukan menunjukkan nilai akurasi sebesar 91,7% data terklasifikasi dengan benar menggunakan Chi-Square sebagai fitur seleksi dan tanpa proses stemming.

Kata kunci: klasifikasi multi-label, hadis bukhari, Random Forest, Mutual Information, Chi-square

Abstract

Hadith is the second source of law for Muslims after the Qur'an. Therefore, Muslims are encouraged to know and practice it. Generally, Hadith is collected by several high priests, one of which is the Hadith collection of Imam Bukhari. There are several categories of teachings in the Hadith, such as the types of containing suggestions, prohibitions, and information. To identify the characteristics of Hadith translated into Indonesian by category, this research will build a system capable of classifying Hadith into three categories, namely recommendations, prohibitions, and information. In classifying Hadith, a classification system is needed. In developing this system, various classification methods can use, one of which is Random Forest. Random forest is a classification method that can generalize high-dimensional data. This Random Forest is believed to complete the classification process with accurate results but has a weakness, namely overfitting when dealing with data types with many features. In this study, the feature selection that will use is Chi-Square. The feature selection method can help select features from the original feature set to eliminate are features not relevant to each class. Optimum results obtained from several tests in this study showed an accuracy value of 91.7% of data classified correctly using Chi-Square as a selection feature, and without stemming process.

Keywords : Multi-label classification, bukhari's hadith, Random forest, Mutual Information, Chi-Square.

1. Pendahuluan

Hadis bagi umat islam merupakan sumber hukum kedua setelah Al-Qur'an yang dimana mengandung ajaran dari berbagai sunnah yang dilakukan Nabi Muhammad SAW. Setiap hadis terdiri dari dua bagian yaitu *Sanad* yang berartikan kumpulan nama para pembilang hadis yang menunjukkan keaslian dari hadis tersebut dan *Matan* yang berartikan penjelasan dari hadis [10]. Umumnya hadis dikumpulkan oleh beberapa imam besar, dan hadis Bukhari merupakan kumpulan hadis yang disusun oleh salah satu imam besar yaitu Imam Bukhari [2]. Maka dari itu, umat Islam dianjurkan untuk mengetahui dan mengamalkannya dengan mengenali hadis dari kumpulan buku hadis yang tersedia dalam versi cetak hingga versi digital. Di dalam hadis, terdapat beberapa jenis ajaran yang dapat diambil seperti jenis ajaran yang mengandung anjuran, larangan, dan informasi [1][13].

Atas dasar tersebut, dalam penelitian ini akan membangun suatu sistem yang dimana dapat mengelompokkan Hadis terjemahan Bahasa indonesia berdasarkan karateristik dari masing-masing kategori Hadis yaitu anjuran, larangan, dan informasi. Dalam pembangunannya, diperlukan sistem klasifikasi teks yang dapat mengelompokkan dan mengidentifikasi suatu hadis kedalam beberapa kategori. Proses klasifikasi ini dapat disebut dengan klasifikasi *multi label*. Untuk itu berbagai pendekatan klasifikasi yang dapat digunakan, salah satunya metode *Random Forest* (RF). *Random forest* merupakan metode klasifikasi yang memiliki performansi paling baik dalam melakukan klasifikasi teks. Hal ini disebabkan metode *Random Forest* memiliki kemampuan menggeneralisasi suatu data berdimensi tinggi[11]. *Random forest* memiliki kelemahan yaitu Ketika menghadapi jenis data yang mengandung jumlah kelas fitur yang sangat banyak[7][11]. Hal tersebut dapat menyebabkan metode RF memiliki nilai bias untuk banyak kelas data, sehingga dapat membuat terjadinya *overfitting* [11]. Untuk mencegah hal tersebut, dapat dilakukan dengan mereduksi dimensi vektor[2]. Proses yang dapat dilakukan yaitu proses seleksi fitur dengan tujuan menyeleksi kata atau fitur yang dianggap kurang berpengaruh dalam pembentukan suatu model untuk tahap klasifikasi[2][7]. Metode *Chi-Square* digunakan sebagai metode seleksi fitur karena memiliki kemampuan untuk memilih fitur-fitur yang berpengaruh dalam menghasilkan prediksi yang tepat dalam suatu kelas[1][2].

Topik dan Batasannya

Dalam penelitian akan dilakukan perancangan sistem klasifikasi *multi label* pada dataset hadis yang berjumlah 7000 data. Dataset hadis terlebih dahulu sudah dilakukan labeling dan divalidasi oleh pihak ulama. Sebelum dilakukan klasifikasi, dataset terlebih dahulu akan dilakukan *preprocessing* dan dilakukan *stemming* dengan bantuan library *sastrawi*.

Tujuan

Berdasarkan penjelasan dari latar belakang, maka akan dilakukan penelitian dengan klasifikasi *multi label* pada Hadis Bukhari terjemahan Bahasa Indonesia menggunakan *Random Forest* sebagai metode klasifikasi serta *Chi-Square* sebagai metode seleksi fitur. Setelah itu, Pada penelitian akan dilakukan pengujian dengan mengidentifikasi pengaruh dari penggunaan seleksi fitur, pengujian pengaruh penggunaan data *stemming* & tanpa *stemming*, dan pengujian pengaruh terhadap pengaturan nilai *threshold* yang digunakan untuk mendapatkan hasil optimum dari proses klasifikasi.

2. Studi Terkait

Penelitian yang mengklasifikasikan Hadis Bukhari terjemahan Bahasa Indonesia kedalam tiga kelas yaitu kelas larangan, anjuran, dan informasi telah dilakukan dengan menggunakan berbagai metode. Salah satunya adalah penelitian yang dilakukan oleh A. Hanafi [1] dengan menambahkan metode *feature selection* yang dimana penelitiannya menggunakan *K – Nearest Neighbor* sebagai metode klasifikasi dan *Mutual Information* sebagai *feature selection*. Hasil yang di dapat dari penelitian tersebut mendapatkan nilai optimum dengan akurasi sebesar 91.14% tanpa stemming dengan waktu proses klasifikasi sebesar 595 detik.

Zahra Putri Agusta [11] melakukan penelitian kasus lain dengan memanfaatkan metode klasifikasi *Random Forest* dan *Balanced Random Forest*. Pada penelitian ini menunjukkan hasil akurasi optimum yang diperoleh dengan nilai aktual *positif rate* sebesar 93,42% menggunakan *Random Forest* dan 98,51% saat dilakukan klasifikasi dengan metode *Balanced Random Forest*. Penelitian yang dilakukan oleh Syair Audi Liri Sacra [9] melakukan klasifikasi teks hadis bersifat *single-label* dengan menggunakan Metode *Naïve Bayes* sebagai metode klasifikasi dan *Chi-Square* sebagai metode seleksi

fitur. Hasil penelitian ini mendapatkan nilai optimum dengan akurasi 94,62% berdasarkan *F-Measure* dan *tanpa stemming*.

Penelitian selanjutnya dilakukan oleh Muhammad Yuslan Abu Bakar [10] yang dimana melakukan penelitian terhadap klasifikasi *multi label* pada hadis menggunakan metode *Backpropagation Neural Network* dengan metode *feature selection* yang digunakan yaitu *information gain*. Dari hasil penelitian yang dilakukan dengan bantuan *feature selection Information gain*, nilai optimum yang didapatkan sebesar 88.42% data multi label terklasifikasi dengan benar dan 65,275% single label terklasifikasi dengan benar.

Klasifikasi Teks Hadis

Klasifikasi hadis telah dilakukan percobaan dengan dikategorikan ke beberapa *multi label* antara kelas berupa anjuran, larangan, dan informasi seperti pada penelitian sebelumnya[2][12]. Berdasarkan dari beberapa penelitian yang ada, berbagai macam proses klasifikasi teks dapat dilakukan dengan kebutuhan dan kepentingan yang berbeda-beda. Tetapi secara umum proses klasifikasi teks dapat dilakukan dengan enam tahapan yaitu pengumpulan dataset, proses pembersihan data, ekstrasi fitur, seleksi fitur, membangun model klasifikasi dan melakukan proses evaluasi performansi [1].

Klasifikasi Multi Label

Klasifikasi *multi label* merupakan permasalahan yang sering dijumpai dalam proses klasifikasi teks. Data *multi label* banyak dijumpai dengan terkelompoknya masing-masing teks dokumen kedalam dua kelas atau lebih. Dalam proses klasifikasi *multi label* ini, masing-masing dokumen yang ada didalam data latih memiliki satu set label dengan tujuan untuk memprediksi dokumen yang belum diketahui labelnya.

Karena hal tersebut permasalahan pada proses klasifikasi *multi label* dapat diatasi dengan algoritma pendekatan *problem transformation method*, salah satunya yaitu *Binary Relevance*. Algoritma *Binary Releveance* ini bekerja dengan mendekomposisikan data *multi label* kedalam bentuk *single label* dengan tujuan mempermudah proses klasifikasi [1].

Random Forest

Random forest merupakan algoritma *Ensemble Learning* yang menggunakan dan membangun struktur tree dalam prosesnya. Dalam peng-implementasiannya akan dibangun sebuah *decision tree* dengan memilih atau mengambil data secara random. Untuk mengelompokkan kelas pada suatu data, *Random Forest* memanfaatkan sistem voting dari hasil decision tree yang dibangun. Kinerja *Random Forest* di adaptasi dari *decision tree*, dengan setiap tree yang dikembangkan dari sample bootstrap berdasarkan data latih. Dari *decision tree* yang dibangun, Random Forest kemudian akan melakukan prediksi untuk masing-masing *decision tree.* Setelah itu dari sekian banyak hasil prediksi, metode *Random Forest* akan melakukan voting untuk menentukan model mana yang memiliki performansi yang baik [3][11]. Tahap-tahap yang dilakukan dalam proses klasifikasi *Random Forest* secara garis besar sebagai berikut:

1) Menentukan jumlah decision tree
2) Melakukan pengecekan apakah jumlah *decision tree* sudah sesuai dengan aturan yang ditetapkan
3) Jika sudah, maka proses klasifikasi selesai dan masuk kedalam tahap pembangunan *decision tree*
4) Melakukan voting terhadap prediksi yang dibangun dari model yang dihasil dari pembangunan *decision tree.*

Secara umum teknis yang diterapkan oleh *Random Forest* akan membuat hasil prediksi menjadi lebih efisien. Selain itu *Random Forest* memiliki kelebihan yang lain yaitu parameter yang dapat diubah dan tidak terlalu sensitif terhadap data bersifat *outlier* [3].

Mutual Information

Mutual Information (MI) adalah metode seleksi fitur yang sudah digunakan secara umum untuk melakukan penyeleksian fitur. *Mutual Information* melakukan perhitungan dengan mengukur jumlah informasi yang ada pada fitur dan mengetahui serta menetapkan fitur tersebut sebagai nilai tertinggi [8]. Dengan hasil pengukuran tersebut dapat diketahui fitur- fitur yang memiliki pengaruh dalam melakukan

proses klasifikasi yang tepat. Perhitungan pada metode *Mutual Information* memiliki rumus yang ditunjukkan pada persamaan (1).

$$I(X,Y) = \sum_{uc\in\{1,0\}} \sum_{uc\in\{1,0\}} P(X=uc, Y=ud) \log^2 \frac{P(X=uc,\ Y=ud)}{P(X=uc)P(Y=ud)} \quad (1)$$

Variabel X pada persamaan (1) merupakan variabel acak yang memiliki *uc* (nilai dokumen mengandung term c) sama dengan 1 dan *uc* (dokumen tidak mengandung term c) sama dengan 0. Untuk variabel Y pada persamaan (1) merupakan variabel acak dengan nilai *ud* (dokumen berada di kelas d) sama dengan 1 dan *ud* (dokumen tidak berada di kelas d) sama dengan 0. Dengan hal itu, Persamaan (1) dapat dijabarkan dengan jelasnya seperti pada persamaan (2).

$$I(X, Y) = \frac{D11}{D} log_2 \frac{D.D11}{D1.D1} + \frac{D01}{D} log_2 \frac{D.D01}{D0.D1} + \frac{D10}{D} log_2 \frac{D.D10}{D1.D0} + \frac{D00}{D} log_2 \frac{D.D00}{D0.D0} \quad (2)$$

Keterangan:
D = Total dokumen yang terdapat *uc* dan *ud* atau (D = D00 + D01 + D10 + D11).
D1. = Total dokumen yang terdapat *uc* atau (D1. = D10 + D11).
D1 = Total dokumen yang terdapat *uc* atau (D1 = D01 + D11).
D0. = Total dokumen yang tidak terdapat *uc* atau (D0. = D01 + D00).
D0 = Total dokumen yang tidak terdapat *ed* atau (D0 = D10 + D00).

Chi-Square

Chi-square adalah metode seleksi fitur yang dalam perhitungannya memanfaatkan distribusi statistika dengan mengukur nilai ketergantungan antara term dan kategori. *Chi-square* menggunakan uji independensi antara dua kejadian yaitu kejadian kemunculan kata dan kejadian kemunculan kelas. Dalam perhitungannya bila nilai *Chi-square* < 0,108, maka dapat dikatakan hubungan antara dua variabel tersebut tidak ada [5]. *Chi-Square* dalam proses perhitungannya menggunakan teori statistika untuk melakukan uji independensi sebuah kata dengan kategori. Proses perhitungan statistika terjadi dengan menghitung kemunculan dari fitur dan kategori. Setelah itu setiap nilai term yang dihasilkan akan dilakukan pengurutan dari yang tertinggi [8]. Persamaan *Chi-Square* ditunjukkan pada persamaan 3.

$$X^2(X_i, Y_j) = \frac{S(KN-LM)^2}{(K+M)(L+N)(K+L)(M+N)} \quad (3)$$

Keterangan :

S = Total dokumen dalam training set
K = Total dokumen dalam y_j yang didalamnya terdapat x_i
L = Total dokumen yang tidak termasuk y_j, tapi didalamnya terdapat x_i
M = Total dokumen dalam y_j yang didalamnya tidak terdapat x_i
N = Total dokumen yang tidak termasuk y_j dan didalamnya tidak terdapat x_i

Dalam penggunaan *Chi-Square* ini dipilih guna melihat korelasi fitur terhadap kategori.

ReliefF

ReliefF merupakan metode algoritma pemilihan atribut yang berbasis pada instan atau record. Pemilihan atribut dalam algoritma ini dilakukan dengan melakukan perhitungan skor relevansi pada pembobotan untuk setiap fitur/term yang terpilih secara acak dengan memperhatikan perhitungan dari kemunculan fitur yang ternominasi sebagai *near hit* (fitur tetangga terdekat dari fitur/term yang terpilih pada kelas yang sama) dan *near miss* (fitur tetangga terdekat dari fitur/term yang terpilih pada kelas yang berbeda) [15]. Algoritma dari metode seleksi *ReliefF* dapat dilihat sebagai berikut.
Algoritma *ReliefF*:

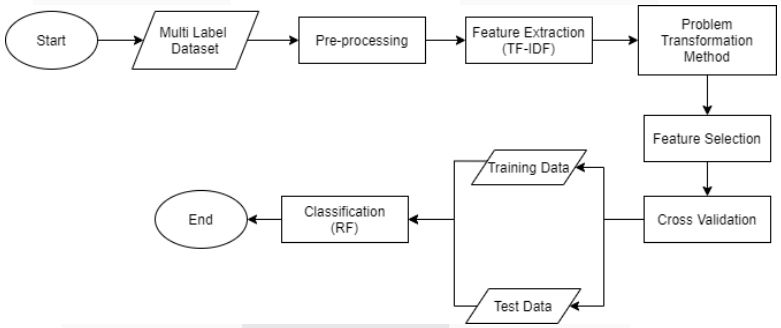
1. *Set all weights W[A] :=0,0;*
2. *For I:=1 to m do begin*
3. *Randomly select an instance R_i;*
4. *Find k nearest hits H_j;*
5. *For each class C ≠ class (R) do*
6. *Find k nearest misses $M_j(C)$;*
7. *For A= 1 to #attributtes do*
8. $$\sum C \neq class(R_i) \left[\left(\frac{P(C)}{1-P(cl(R_i))}\right)\sum_{j=1}^{k} diff(A,R_i,M_j)(C))\right]/(m.k);$$
9. *end;*

Dalam metode perhitungan *ReliefF* semakin tinggi kemunculan *nearhit* (fitur tetangga terdekat dari fitur/term yang terpilih pada kelas yang sama) pada fitur/term yang dipilih secara acak, maka kemungkinan besar fitur tersebut merupakan fitur yang memiki releveansi besar pada kelas tersebut. Jika kemunculan *near miss* (fitur tetangga terdekat dari fitur/term yang terpilih pada kelas yang berbeda) semakin tingi pada pada fitur/term yang dipilih secara acak, maka kemungkinan besar fitur tersebut tersisih dalam proses perhitungan [14][15].

3. Metodologi

System Overview

Dalam penelitian ini akan membangun suatu sistem yang dapat melakukan klasifikasi pada Hadis Bukhari Terjemahan Bahasa Indonesia dengan dataset bersifat *multi-label* menggunakan *Random Forest* dan metode seleksi fitur yaitu *Chi-Square.* Dalam peneltian ini akan menggunakan 2 seleksi fitur lainnya yaitu *Mutual information* dan *ReliefF* dengan tujuan untuk membandingkan kinerja dari seleksi fitur yang digunakan dalam penelitian ini. Gambar 1 menunjukkan alur kerja sistem yang akan dirancang secara umum.



Gambar 1 Flowchart Sistem

Dataset

Dataset yang digunakan dalam penelitian ini adalah data Hadis Bukhari terjemahan Bahasa Indonesia sebanyak 7000 Hadis yang dimana setiap datanya sudah dilakukan labeling dan divalidasi oleh Ulama. Dataset ini bersifat *multi label* yang terdiri dari 3 kelas, yang pertama Hadis dengan kelas anjuran. Hadis anjuran merupakan sunnah yang dilakukan atau diucapkan oleh Nabi Muhammad SAW yang dimana *Matan*(isi) dari hadis tersebut menggambarkan suatu ajaran yang sepatutnya dilakukan oleh umat muslim. Yang kedua, Hadis dengan kelas larangan merupakan sunnah yang dilakukan atau diucapkan oleh Nabi Muhammad SAW yang dimana *Matan*(isi) dari hadis tersebut menggambarkan suatu ajaran yang seharusnya tidak dilakukan oleh umat muslim. Dan yang ketiga, Hadis dengan kelas informasi merupakan sunnah yang dilakukan atau diucapkan oleh Nabi Muhammad SAW yang dimana *Matan*(isi) dari hadis tersebut menjelaskan hukum dari semua ajaran yang dilakukan oleh Rasulullah SAW dan juga cara beribadah.

Data hadis dapat termasuk lebih dari satu kelas, oleh karena itu dalam penelitian ini akan dilakukan proses klasifikasi *multi-label.* Klasifikasi *multi-label* pada umumnya merupakan klasifikasi data yang dimana tiap dokumen yang ada pada data latih memiliki satu set label dengan tujuan untuk

memprediksi dokumen yang belum diketahui kelasnya. Representasi dataset pada penelitian ini dapat dilihat pada Tabel 1.

Tabel 1. Representasi Dataset

Hadis	Anjuran	Larangan	Informasi
Sembelihlah tidak apa apa kemudian datang orang lain dan berkata aku tidak menyadari ternyata ketika berkurban aku belum melempar jumrah. Nabi bersabda lemparlah dan tidak apa apa dan tidaklah Nabi ditanya tentang sesuatu perkara sebelum dan sesudahnya, kecuali beliau menjawab lakukanlah dan tidak apa apa.	1	0	1

Data Preprocessing

Sebelum masuk kedalam tahap klasifikasi, pada dataset yang digunakan akan dilakukan *preprocessing*. Tahap *Preprocessing* ini dilakukan dengan tujuan menghasilkan data yang bersih untuk tahap klasifikasi sehingga menjadi lebih optimal dan dapat meningkatkan kinerja sistem dalam mengelompokkan dokumen berdasarkan kelasnya. Proses *preprocessing* yang dilakukan yaitu *noise removal* dengan tujuan menghapus angka, tanda baca, dan karakter lainnya. Selanjutnya *case folding* dengan tujuan membuat huruf kapital yang ada didalam data menjadi huruf tidak kapital, setelah itu *tokenizing*, dan yang terakhir *stopword removal* dengan tujuan menghilangkan berbagai kata yang dapat dikatakan kurang berpengaruh.

Feature Extraction

Setelah tahap *preprocessing* selesai, tahap selanjutnya akan masuk kedalam tahap ekstraksi fitur. *Term Frequency-Inverse Document Frequency* (TF-IDF) merupakan metode ekstraksi fitur yang digunakan dalam penelitian ini. TF-IDF berasal dari term pembobotan yang diperoleh dari data yang sudah melewati tahap *preprocessing*. Proses TF-IDF ini dilakukan dengan cara memberikan score berdasarkan frekuensi term yang terdapat dalam dokumen. Rumus perhitungan dengan metode TF-IDF dapat dilihat pada persamaan (4).

$$W_{ij} = tf_{ij}\, x\, log\, (\frac{D}{dfi}) \quad (4)$$

Sebelum masuk kedalam tahap seleksi fitur, agar mempermudah proses klasifikasi *multi label* pada penelitian ini akan dilakukan proses *problem transformation method*. Algoritma *problem transformation method* yang digunakan *adalah Binary Relevance (BR).* Algoritma *Binary Releveance* ini bekerja dengan mendekomposisikan data yang masih berbentuk *multi label* menjadi *single label* dengan memecah 3 kelas dalam penelitian ini yang dimana terdiri dari kelas larangan, anjuran, dan infomasi. Setelah itu proses klasifikasi dilakukan pada masing-masing kelas.

Feature Selection

Tahap selanjutnya adalah tahap seleksi fitur. Seleksi fitur merupakan tahapan yang dilakukan untuk memilih fitur yang berpengaruh besar dan menghilangkan fitur yang kurang berpengaruh atau kurang relevan dalam membangun model klasifikasi. Seleksi fitur yang digunakan dalam penelitian ini adalah *Chi-Square*. Dalam pengujian akan menggunakan *Mutual Information* dan *ReliefF* sebagai metode pembanding kinerja dari seleksi fitur yang digunakan dalam penelitian ini *(Chi-Square)*. Proses pengujian seleksi fitur dalam penelitian dilakukan secara terpisah dengan melakukan skenario pengujian penggunaan seleksi fitur *Chi-Square* terlebih dahulu setelah itu dilanjutkan dengan skenario pengujian dengan menggunakan seleksi fitur *Mutual Information* dan *ReliefF* pada area kerja yang berbeda.

Skenario pertama dalam proses seleksi fitur ini akan dimulai dengan metode *Chi-Square*. Proses penyeleksian fitur pada *Chi-Square* dilakukan dengan menyeleksi seluruh kata pada masing-masing kategori. Kata yang terpilih merupakan kata yang memiliki *score Chi-Square* ≥ 0,108. Kata yang berada

dibawah *score* tersebut dianggap tidak memenuhi syarat *Chi-Square* sehingga kata tersebut tidak dimasukkan kedalam proses *learning*. Untuk melakukan perhitungan *Chi-Square* dapat menggunakan rumus yang ditunjukkan pada persamaan (3). Setelah melakukan perhitungan menggunakan metode *Chi-Square*, didapatkan masing-masing fitur kata yang diurutkan dari nilai yang paling tinggi. Tabel 2 menunjukkan tiga fitur terbaik pada masing-masing kelas berdasarkan nilai *Chi-Square*.

Tabel 2. 3 Fitur Terbaik Pada Masing-Masing Kelas Bedasarkan Nilai Chi-Square

Larangan		Anjuran		Informasi	
Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
janganlah	0,47	hendaklah	0,28	melarang	0,21
melarang	0,4	kalian	0,18	beli	0,15
jangan	0,25	maka	0,16	makanlah	0,14

Pada skenario kedua pengujian seleksi fitur dilakukan secara terpisah dari pengujian sebelumnya dengan menggunakan metode *Mutual Information*. Proses perhitungan dilakukan dengan menghitung jumlah informasi yang ada didalam fitur pada masing-masing kelas. Perhitungan *Mutual Information* dapat dilakukan dengan menggunakan rumus yang ditunjukkan pada persamaan (1) & (2). Setelah melakukan perhitungan menggunakan *Mutual Information*, didapatkan berbagai fitur/kata yang memiliki pengaruh tinggi terhadap masing-masing kelas. Tabel 3 menunjukkan tiga fitur terbaik pada masing-masing kelas berdasarkan nilai *MI*.

Tabel 3. 3 Fitur Terbaik Pada Masing-Masing Kelas Bedasarkan Nilai Mutual Information

Larangan		Anjuran		Informasi	
Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
janganlah	0,06	hendaklah	0,03	melarang	0,0105
melarang	0,04	kalian	0,014	beli	0,0055
jangan	0,018	maka	0,012	janganlah	0,0052

Pada skenario ketiga pengujian seleksi fitur dilakukan secara terpisah dari pengujian sebelumnya dengan menggunakan metode *ReliefF*. Proses perhitungan pada metode seleksi fitur *ReliefF* dilakukan dengan menghitung skor relevansi pada pembobotan untuk setiap fitur/term yang terpilih secara acak dengan memperhatikan perhitungan dari kemunculan fitur yang ternominasi sebagai *near hit* (fitur tetangga terdekat dari fitur/term yang terpilih pada kelas yang sama) dan *near miss* (fitur tetangga terdekat dari fitur/term yang terpilih pada kelas yang berbeda). Setelah melakukan perhitungan, algoritma dari seleksi fitur *ReliefF* akan mengecek fitur/*term* yang terpilih. Jika fitur tersebut memiliki *near hit* yang lebih tinggi, maka fitur tersebut ditetapkan sebagai fitur yang memiliki pengaruh dalam kelas tersebut. Tabel 4 menunjukkan tiga fitur terbaik pada masing-masing kelas berdasarkan nilai *ReliefF*.

Tabel 4. 3 Fitur Terbaik Pada Masing-Masing Kelas Bedasarkan Nilai ReliefF

Larangan		Anjuran		Informasi	
Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
larang	0,052	shalatlah	0,0087	larang	0,0109
jual	0,012	perintah	0,0086	jual	0,0085
haram	0,0061	puasa	0,0068	makan	0,0071

Klasifikasi dengan Random Forest

Setelah melakukan proses seleksi fitur, tahap selanjutnya yaitu melakukan klasifikasi dengan metode *Random Forest*. Sebelum melakukan proses klasifikasi, akan dilakukan proses pembagian proporsi data untuk *cross-validation* dalam uji evaluasi proses klasifikasi. Nilai proporsi data yang digunakan sebesar 70% untuk data latih dan 30% untuk data uji. Setelah itu dilakukan proses klasifikasi Random Forest dengan memasukkan 70% data latih untuk membangun sejumlah *decision tree* yang telah ditentukan dengan aturan tertentu. Hasil prediksi yang di dapat adalah hasil proses voting dari pembelajaran berdasarkan *decision tree*. Pada penelitian ini model Random Forest dibangun dengan

membuat banyak *decision tree* dan *Random Forest* kemudian akan melakukan prediksi untuk masing-masing *decision tree*. Dari sekian banyak hasil prediksi, metode *Random Forest* akan melakukan voting untuk menentukan model mana yang memiliki performansi yang baik. Setelah mendapatkan model pohon terbaik, pada proses evaluasi menggunakan 30% data uji akan dilakukan dengan model pohon yang telah didapatkan guna melihat hasil kinerja dari proses klasifikasi.

4. Hasil Pengujian

Penelitian ini melakukan tiga skenario pengujian untuk mencari nilai akurasi paling optimal yang dapat dihasilkan sistem. Pengujian pertama dilakukan untuk mengetahui pengaruh penggunaan fitur seleksi dalam menghasilkan nilai akurasi yang optimal. Dalam pengujian pertama akan dilakukan dua skenario yang pertama pengujian untuk melihat pengaruh dari penggunaan seleksi fitur yaitu *Chi-Square*. Selanjutnya akan dilakukan pengujian menggunakan *Mutual Information* dan *ReliefF* untuk membandingkan performansi metode seleksi fitur yang digunakan dan menentukan metode mana yang dapat menghasilkan akurasi yang optimal. Pengujian kedua adalah mengetahui pengaruh terhadap nilai akurasi dengan menggunakan data *stemming* dan *non-stemming*. Terakhir, pengujian untuk mengetahui pengaruh pengaturan nilai threshold terhadap nilai akurasi. Dalam proses pengujian, metode klasifikasi yang digunakan adalah *Random Forest.*

Pengujian Pengaruh Terhadap Penggunaan Chi-Square sebagai Seleksi Fitur

Pengujian pertama dilakukan untuk mengetahui pengaruh penggunaan fitur seleksi dalam menghasilkan nilai akurasi yang optimal. Dalam pengujian pertama akan dilakukan dua skenario. Yang pertama, pengujian untuk melihat pengaruh dari penggunaan seleksi fitur yaitu *Chi-Square.* Selanjutnya pada skenario kedua dalam pengujian pertama, akan dilakukan pengujian menggunakan *Mutual Information* (MI) dan *ReliefF* untuk membandingkan performansi metode seleksi fitur yang digunakan dan menentukan metode mana yang dapat menghasilkan akurasi yang optimal. Dalam pengujian ini, data yang digunakan telah melalui tahap *pre-processing* tanpa *stemming*. Setelah itu akan dilakukan proses ekstraksi fitur oleh TF-IDF dan klasifikasi menggunakan Random Forest. Pada pengujian pertama, nilai threshold yang digunakan dalam seleksi fitur yang digunakan sebesar 0.1 dan menggunakan nilai *cross-validation* sebesar 70% untuk data latih & 30% untuk data uji. Hasil pengujian pengaruh terhadap penggunaan seleksi fitur ditunjukkan pada tabel 5.

Tabel 5. Hasil Pengujian Pengaruh Terhadap Penggunaan Seleksi Fitur

Skenario Pengujian	Nilai Akurasi
Tanpa Menggunakan Fitur Seleksi	76%
Menggunakan Fitur Seleksi Chi-Square	**91,7%**
Menggunakan Fitur Seleksi MI	91,3%
Menggunakan Fitur Seleksi ReliefF	88,24%

Berdasarkan tabel 5 hasil pengujian pertama pada skenario pertama dengan membandingkan penggunaan seleksi fitur dan tanpa seleksi fitur, menunjukkan proses klasifikasi yang mendapatkan nilai akurasi optimum dengan menggunakan seleksi fitur yaitu *Chi-square*. Nilai akurasi yang diperoleh pada percobaan menggunakan fitur seleksi adalah 91,7% data diklasifikasikan dengan benar. Sedangkan pada percobaan tanpa menggunakan fitur seleksi, nilai akurasinya adalah 76% data terklasifikasi dengan benar. Berdasarkan nilai akurasi yang dihasilkan, penggunaan *Chi-square* sebagai metode seleksi fitur memiliki nilai yang lebih unggul. Hal ini terjadi karena metode *Chi-Square* dapat memilih fitur yang berpengaruh besar dan menghilangkan fitur yang kurang berpengaruh atau kurang relevan dalam membangun model klasifikasi. Oleh karena itu, penggunaan seleksi fitur dapat meningkatkan kemampuan mendeskripsikan suatu golongan dalam proses klasifikasi hadis.

Dari hasil pengujian pertama pada skenario kedua yaitu menguji performansi dari masing-masing seleksi fitur, berdasarkan tabel 5 hasil percobaan diperoleh nilai akurasi terbaik menggunakan metode seleksi fitur yaitu *Chi-Square*. Nilai akurasi yang diperoleh pada percobaan menggunakan seleksi fitur *Chi-Square* adalah 91,7% data terklasifikasi dengan benar. Sedangkan pada percobaan menggunakan metode seleksi fitur *Mutual Information* didapatkan nilai akurasi 91,3% data terklasifikasi dengan benar. Dan pada percobaan menggunakan seleksi fitur *ReliefF* didapatkan nilai akurasi 88,24%

data terklasifikasi dengan benar. Berdasarkan nilai akurasi yang dihasilkan, metode seleksi fitur *Chi-Square* memiliki hasil yang lebih unggul dibandingkan dengan seleksi fitur *Mutual Information* dan *ReliefF*. Hal ini disebabkan metode *Chi-Square* dalam proses perhitungannya menerapkan teori statistik dalam melakukan uji independensi suatu istilah atau kata dengan kategorinya. Proses perhitungan statistik terjadi dengan menghitung kemunculan dari fitur dan kemunculan dari kategori pada masing-masing kelas.

Pengujian Pengaruh Terhadap Penggunaan Data Stemming dan non-Stemming

Pada pengujian kedua dilakukan menggunakan data *stemming* dan tidak *stemming* dengan tujuan untuk mengetahui penggunaan data yang dapat menghasilkan nilai akurasi yang optimum. *Stemming* sendiri merupakan proses mengubah kata dalam kalimat menjadi ke bentuk kata dasarnya. Dalam melakukan proses *stemming* pada data hadis, penulis menggunakan *library* Sastrawi. Dalam pengujian kedua, proses ekstraksi fitur dilakukan menggunakan TF-IDF dan klasifikasi menggunakan *Random Forest.* Pada pengujian ini, nilai *threshold* yang akan digunakan pada seleksi fitur adalah 0.1 dan menggunakan nilai *cross-validation* sebesar 70% untuk data latih & 30% untuk data uji. Hasil pengujian dapat dilihat pada tabel 6 merupakan hasil pengujian pengaruh terhadap penggunaan data *stemming* dan Tanpa *stemming*.

Tabel 6. Hasil Pengujian Pengaruh Terhadap Penggunaan Data Stemming dan Tanpa Stemming

Skenario Pengujian	Nilai Akurasi
Menggunakan Data Stemming	89%
Menggunakan Data Tanpa Stemming	**91,7%**

Berdasarkan tabel 6, pengujian mendapatkan nilai akurasi terbaik dengan menggunakan data tanpa *stemming*. Nilai akurasi yang dihasilkan tanpa *stemming* sebesar 91,7% data multilabel terklasifikasi dengan benar. Sedangkan dengan menggunakan data yang sudah dilakukan *stemming*, nilai akurasi yang didapatkan sebesar 89% data multilabel terklasifikasi dengan benar. Berdasarkan nilai akurasi yang dihasilkan, proses *stemming* tidak bisa dilakukan pada setiap kata dalam hadis. Hal ini karena proses *stemming* setiap kata dalam hadis akan diubah ke bentuk dasarnya. Sehingga dapat membuat makna kalimat dalam hadis tersebut berubah. Contoh pada kalimat "menunaikan zakat", akan berubah menjadi "tunai zakat" jika dilakukan proses stemming. Menghilangkan kata imbuhan "me" dan "-kan" pada kata menunaikan dapat merubah makna dalam kalimat tersebut yang semula bermakna anjuran akan berubah menjadi suatu kalimat yang bersifat informasi.

Pengujian Pengaruh Terhadap Penggunaan Nilai Threshold

Pada pengujian ini dilakukan dengan tujuan untuk mengetahui penggunaan nilai *threshold* pada *feature selection* yang dapat menghasilkan nilai akurasi optimum dengan menggunakan *Chi-Square* sebagai metode *feature selection*. Data yang digunakan dalam pengujian merupakan data yang sudah dilakukan *preprocessing* dan tanpa *stemming*, menggunakan nilai *cross validation* dengan nilai 70% untuk data latih dan 30% untuk data uji, menggunakan TF-IDF sebagai metode ekstraksi fitur, dan menggunakan *Random Forest* untuk metode klasifikasi. Pada pengujian ini dilakukan dilakukan 3 tahap, yang pertama dengan nilai threshold 0.09, yang kedua dengan nilai threshold 0.1, dan yang ketiga dengan nilai threshold 0.2. Hasil pengujian pengaruh terhadap penggunaan nilai *threshold* dapat dilihat pada tabel 7.

Tabel 7. Hasil Pengujian Pengaruh Terhadap Penggunaan Nilai Threshold

Skenario Pengujian	Nilai Akurasi
Menggunakan Nilai Threshold 0.09	91%
Menggunakan Nilai Threshold 0.1	**91,7%**
Menggunakan Nilai Threshold 0.2	91,5%

Berdasarkan tabel 7, hasil dari pengujian dilakukan terhadap penggunaan nilai *threshold* pada *feature selection* menggunakan metode klasifikasi *Random Forest*, nilai akurasi yang diperoleh pada tahap pengujian pertama dengan nilai *threshold* 0.09 menghasilkan nilai akurasi sebesar 91% data terklasifikasi dengan benar. Pada pengujian tahap kedua dengan nilai *threshold* 0.1 menghasilkan nilai

akurasi sebesar 91,7% data terklasifikasi dengan benar. Dan pada pengujian tahap ketiga dengan nilai *threshold* 0.2 menghasilkan nilai akurasi sebesar 91,5% data terklasifikasi dengan benar.

Dari nilai akurasi yang diperoleh, pengujian yang menghasilkan nilai akurasi optimum dilakukan dengan menggunakan nilai *threshold* sebesar 0.1. Dalam pengujian ini, jika nilai *threshold* semakin tinggi maka akurasi yang didapatkan akan bertambah. Tetapi, jika penggunaan nilai *threshold* telah mencapai nilai kinerja optimum, maka akurasi yang dihasilkan pada proses selanjutnya condong menurun. Hal ini disebabkan jika nilai *threshold* yang digunakan semakin besar pada tahap seleksi fitur, maka dapat dikatakan kemungkinan terdapat fitur/*term* lainnya yang dianggap memiliki pengaruh terhadap penggambaran suatu kelas tidak masuk atau tidak digunakan dalam proses seleksi fitur. Hal tersebut dapat mempengaruhi proses klasifikasi yang dimana kemampuan dalam menggambarkan suatu kelas pada hadis menurun.

5. Kesimpulan

Berdasarkan hasil dari analisis dan pengujian yang dilakukan terhadap beberapa skenario yang telah disusun, maka dapat disimpulkan bahwa penggunaan *feature selection*, nilai *cross validation*, dan pengaturan nilai *threshold* dapat mempengaruhi proses klasifikasi. Nilai akurasi optimal yang didapatkan dari proses penelitian ini sebesar 91,7% data terklasifikasi dengan benar. Berbagai parameter dan metode telah digunakan dalam penelitian yang dilakukan yaitu penggunaan *Chi-Square* sebagai metode *feature selection*, menggunakan TF-IDF sebagai ekstraksi fiturnya, menggunakan data yang sudah dilakukan preprocessing tanpa stemming, menggunakan nilai *threshold* sebesar 0.1 pada *feature selection*, dan menggunakan nilai *cross validation* sebesar 70% untuk data latih dan 30% untuk data uji pada proses klasifikasi. Proses *stemming* pada tahap *preprocessing* tidak menghasilkan performansi yang baik dibandingkan dengan data yang di-*preprocessing* tanpa *stemming*. Karena proses *stemming* pada masing-masing kata dalam hadis akan mengubah kata kedalam bentuk dasarnya. Sehingga dapat membuat makna kalimat dalam hadis tersebut berubah.

Sedangkan hasil penggunaan dari metode seleksi fitur *Chi-Square* menghasilkan performansi yang lebih unggul dibandingkan dengan *Mutual Information* dan *ReliefF*. Hal ini disebabkan metode *Chi-Square* dalam proses perhitungannya menerapkan teori statistika untuk melakukan uji independensi sebuah *term* atau kata dengan kategori. Proses perhitungan statistika terjadi dengan menghitung kemunculan dari fitur dan kemunculan dari kategori pada masing-masing kelas. Setelah itu dilanjutkan dengan melakukan pengurutan nilai *term* dari yang tertinggi.

Adapun saran untuk penelitian kedepannya yaitu menambah dokumen hadis yang sebelumnya telah dilakukan labeling oleh ahli hadis dan pesebaran data. Hal ini dilakukan agar dapat mengatasi persebaran data yang tidak merata pada dataset dan menambah variasi data dengan tujuan untuk meningkatkan referensi sistem dalam melakukan proses pengelompokkan hadis berdasarkan fitur/term pada masing-masing kelas Hadis. Selain itu perlu dilakukan tahap *pre-processing* lebih baik lagi, Karena berdasarkan pengamatan yang dilakukan terhadap dataset yang digunakan terdapat beberapa *typo* di dalam kata hadis yang perlu diperbaiki. Hal ini dilakukan agar dapat menghasilkan data yang lebih bersih dalam proses klasifikasi.

REFERENSI

[1] A. Hanafi, W. Astuti. 2020. Multi Label Classification in Bukhari Hadith Indonesian Translation Using Mutual Information and k-Nearest Neighbor. Sisfokom Journal Volume 09 Nomor 03 PP 357 – 364.

[2] Purbolaksono, M.D., Reskyadita, F.D., A.A., 2020. Suryani, and AF Huda,"Indonesian text classification using back propagation and sastrawi stemming analysis with information gain for selection feature,". Int. J. Adv. Sci. Eng. Inf. Technol, 10(1), p.234.

[3] I. Annapoorani, Karthikrajan, Sentihinathan, B. Shanmugam, D. Goyal, R. Samikanmu. 2020. Deep Learning Applications and Intelligent Decision Making in Enginering. India: IGI Global.

[4] P. Riky Sutriadi, A Julio. 2017. Twitter sentiment analysis with Naïve Bayes Classification using Mutual Information Feature selection and Inverse Document Frequency. Computer Science Journal.

[5] Abraham, Ranjit. 2009.Effective Discreatization and Hybrid Feature Selection Using Naïve Bayesian Classifier for Medical Data Mining. Dr. MGR University. Chennai, India

[6] Syair Audi L.S., Said Al – Faraby, Danang Triantoro M. 2017. Classification of advice, Prohibition, and Information on Sahih Bukhari Hadith using Naïve Bayes Classifier. Informatika Journal

[7] Oktanisa, I., & Supianto, A. A. 2018. Comparison of Classification Techniques in Data Mining for Direct Marketing Banks. *Journal of Information Technology and Computer Science*, *5*(5), 567-576.

[8] J. Ling, I. P. Eka N. Kencana, T. Bagus Oka. 2014. Sentiment Analysis Using the nave Bayes classifier method with chi square feature selection. E-Journal of Matematika, 3(3).

[9] Sacra, S., Faraby, S., & Triantoro, D. 2017. Classification of advice, Prohibitions, and Information on Sahih Bukhari Hadith Using Naive Bayes Classifier. *eProceedings of Engineering*, *4*(3).

[10] M. Y. Abu Bakar. 2018. Multi-Label Topic Classification of Hadith of Bukhari (Indonesian Language Translation) Using Information Gain and Backpropagation Neural Network". 2018 International Conference on Asian Language Processing (IALP), 2018, pp. 344-350, doi: 10.1109/IALP.2018.8629263.

[11] Agusta, Z.P.2019. Modified balanced random forest for improving imbalanced data prediction. International Journal of Advances in Intelligent Informatics, 5(1), pp.58-65

[12] Al-Faraby, S., Jasin, E.R.R. and Kusumaningrum, A., 2018, March. Classification of Hadith into positive suggestion, negative suggestion, and information. In Journal of Physics: Conference Series (Vol. 971, No. 1, p. 012046). IOP Publishing.

[13] Ponilan, I.R., Bijaksana, M.A. and Raharusun, A.S., 2019, March. Search relevant retrieval on indonesian translation hadith document using query expansion and smoothing probabilistic model. In Journal of Physics: Conference Series (Vol. 1192, No. 1, p. 012032). IOP Publishing.

[14] A. Fadli and muhamamad imron rosadi, "klasifikasi penyakit jantung koroner menggunakan seleksi fitur dan support vector machine", *explorit*, vol. 10, no. 2, pp. 32-40, May 2021.

[15] Fitriani, Irma & Basuki, Setio & Minarno, Agus. (2020). Seleksi Fitur Relieff Pada Klasifikasi Malware Android Menggunakan Support Vector Machine(SVM). Jurnal Repositor. 2. 1529. 10.22219/repositor.v2i11.901.