

# KLASIFIKASI TEKS MULTI-LABEL PADA HADIS MENGGUNAKAN *MULTINOMIAL NAÏVE BAYES* DENGAN *INFORMATION GAIN* SEBAGAI *FEATURE SELECTION*

A Saskia Nurul Septiani Putri<sup>1</sup>, Said Al Faraby<sup>2</sup>, Mahendra Dwifabri<sup>3</sup>

<sup>1,2,3</sup> Universitas Telkom, Bandung

<sup>1</sup>andisaskiaputri@students.telkomuniversity.ac.id, <sup>2</sup>saidalfaraby@telkomuniversity.ac.id,  
<sup>3</sup>mahendradp@telkomuniversity.ac.id

---

## Abstrak

Hadis merupakan pedoman hidup yang berisi perbuatan ataupun perkataan Nabi Muhammad SAW, sehingga banyak umat muslim yang ingin mempelajarinya. Permasalahan dalam mempelajari hadis adalah jumlah hadis yang sangat banyak dan memiliki kategori untuk setiap hadis. Sehingga dibutuhkan sistem yang mampu mengelompokkan *corpus* pada hadis, supaya hadis mudah dipelajari sesuai dengan kelasnya. Oleh karena itu, dibangun yang mampu sistem mengklasifikasikan hadis ke dalam kelas anjuran, larangan, dan informasi dengan menggunakan kombinasi metode *preprocessing*; pembobotan kata menggunakan TF-IDF; seleksi fitur *Information Gain*; klasifikasi *Multinomial Naïve Bayes*; dan evaluasi *Confusion Matrix*. Hasil penelitian terhadap sistem menunjukkan kinerja terbaik didapatkan menggunakan *preprocessing* tanpa teknik *stopword removal* dan tanpa pengaplikasian seleksi fitur *Information Gain* dengan akurasi 88%. Kombinasi tanpa menggunakan *stopword removal* menghasilkan akurasi yang lebih tinggi karena model dapat memiliki lebih banyak fitur yang dapat diklasifikasikan.

**Kata kunci :** hadis, *information gain*, *multinomial*, *multi label*

---

## Abstract

Hadith is a way of life that contains the actions or words of the Prophet Muhammad SAW, so that many Muslims want to learn it. The problem in studying hadith is that the number of hadiths is very large and has a category for each hadith. So we need a system that is able to classify the corpus of hadith, so that hadith is easy to learn according to its class. Therefore, a system is built that is able to classify hadith into recommendations, prohibitions, and information classes by using a combination of preprocessing methods; word weighting using TF-IDF; Information Gain feature selection; Multinomial Naïve Bayes classification; and evaluation of the Confusion Matrix. The results of the research on the system showed that the best performance was obtained using preprocessing without stopword removal techniques and without the application of Information Gain feature selection with an accuracy of 88%. The combination without using stopword removal results in higher accuracy because the model can have more features that can be classified.

**Keywords:** hadith, *information gain*, *multinomial*, *multi label*

---

## 1. Pendahuluan

Hadis adalah segala sesuatu yang dikaitkan dengan Nabi Muhammad dalam bentuk perkataan, perbuatan dan taqir (sikap diam) [1]. Dengan berkembangnya teknologi dan ilmu pengetahuan, kemajuan juga telah dicapai dalam pengembangan hadits. Buku-buku yang digunakan di atas berasal dari berbagai koleksi hadis yang dikemas dalam bentuk digital. Namun, kebanyakan hadis digital hanya menampilkan terjemahan hadis, tanpa penjelasan untuk setiap hadits, baik itu nasihat, larangan atau informasi. Hal ini memungkinkan umat Islam yang memahami dan mempelajari ilmu hadits salah mengartikan makna hadis. [2].

Pada tahun 2018, dilakukan penelitian [3] dibangun sebuah sistem klasifikasi topik multi-label dan single-label pada hadis *Bukhari* dengan menggunakan *Backpropogation Neural Network* dengan kategori kelas anjuran, larangan, dan informasi. Seleksi fitur yang digunakan adalah IG lalu metode untuk evaluasi menggunakan *Hamming Loss*. Penelitian tersebut mendapatkan akurasi terbaik diangka 88,42% untuk data hadis multi label dan 66,27% untuk data hadis berlabel tunggal [3]. Tahun 2017, pada penelitian [5], kombinasi *Multinomial NB* dan TFIDF digunakan untuk membangun sistem dapat mengelola teks yang terdiri dari 4 kategori, yaitu undangan,

cuti, perintah, dan pinjam ruang dan didapatkan nilai *Accuracy* sebesar 89% [5].

Berdasarkan penelitian yang telah dibahas sebelumnya, kombinasi terhadap pengkategorian *corpus* dari hadis hadis masih kurang untuk terjemahan bahasa Indonesia. Penulis mengusulkan model klasifikasi *multi-label* hadis Bukhari dengan menggunakan metode *Information Gain* (IG) untuk *feature selection* dan metode klasifikasi *Multinomial Naïve Bayes* dengan kategori kelas anjuran, larangan, dan informasi. Metode *information gain* digunakan karena bekerja dengan memberikan nilai pada setiap fitur, dan hanya fitur terpilih yang terkait dengan kategori yang akan digunakan [3], sedangkan metode *Multinomial Naïve* dipilih karena cocok untuk data yang memiliki kosakata yang banyak dan dapat mengurangi tingkat kesalahan [5].

### Topik dan Batasannya

Pada penelitian penulis akan membangun model klasifikasi hadis Bukhari terjemahan bahasa Indonesia. Model yang dibangun menggunakan beberapa proses yaitu *perprocessing*, ekstraksi fitur, seleksi fitur, dan klasifikasi *multinomial naïve bayes*. Dataset yang digunakan berjumlah 7008 dan berasal dari dataset terjemahan hadis Al Bukhari. Kelas yang terdapat pada dataset berjumlah 3 kelas, yaitu anjuran, larangan dan informasi.

### Tujuan

Penelitian ini dilakukan untuk mengetahui *corpus* yang ada pada hadis dengan menggunakan beberapa skenario, yaitu pengaruh teknik *stemming* dan *stopword removal* pada *preprocessing*; nilai *threshold* pada penggunaan *Information Gain*; serta penggunaan data *balance* dan data *imbalace* pada klasifikasi *Multinomial Navive Bayes*.

### Organisasi Tulisan

Bagian selanjutnya pada penelitian yaitu pembahasan mengenai studi terkait pada penelitian, penjelasan mengenai sistem yang dibangun, kemudian evaluasi hasil pengujian, pembahasan mengenai kesimpulan dari penelitian ini, dan terakhir pembahasan saran untuk penelitian selanjutnya.

## 2. Studi Terkait

Pada tahun 2018, dilakukan penelitian [3] dengan dataset terdiri dari hadis shahih Bukhari diterjemahkan ke dalam bahasa indonesia yang telah diberi label sebelumnya. Sistem pertama terdiri dari 1.064 data multi-label dengan 289 di antaranya memiliki lebih dari satu label. Sistem kedua terdiri dari 455 data label tunggal. Penelitian tersebut mengklasifikasikan topik multi label dan label tunggal pada hadis *Bukhari* dengan menggunakan *Backpropogation Neural Network* dengan kategori kelas anjuran, larangan, dan informasi. Algoritma Jaringan Syaraf Tiruan *Backpropagation* dipilih karena dapat melakukan klasifikasi dengan sejumlah besar fitur yang bervariasi. Pada penelitian sebelumnya hasil yang diperoleh menunjukkan bahwa metode Jaringan Syaraf Tiruan *Backpropagation* dapat melakukan klasifikasi teks multi label yang lebih baik dibandingkan dengan metode lainnya. Seleksi fitur yang digunakan adalah IG lalu metode untuk evaluasi menggunakan *Hamming Loss*. Penelitian tersebut mendapatkan akurasi terbaik diangka 88,42% untuk data hadis multi label dan 66,27% untuk data hadis berlabel tunggal [3].

Pada tahun 2018 dilakukan penelitian [4] tentang pengelompokan terjemahan hadis ke dalam tiga kategori yaitu *positive suggestion*, *negative suggestion*, and *information*. Dataset hadis yang digunakan dalam karya ini adalah terjemahan dari bahasa Arab ke Bahasa, dan dikumpulkan dari Koleksi Imam Bukhari yang merupakan bagian dari “Kitab Hadits 9 Imam” yang disediakan oleh aplikasi Lidwa. Pada *preprocessing* dilakukan proses *stemming* dan *stopword removal*, lalu untuk *feature extraction* metode yang digunakan adalah TF-IDF dari *unigram*, *bigram*, dan *trigram*. Pada proses klasifikasi penelitian ini menggunakan metode *Support Vector Machine* (SVM). Dengan menggunakan metode SVM dengan kernel linier mencapai F1Score terbaik sebesar 88% [4].

Tahun 2019, pada penelitian [6] metode pendukung dalam pengujian sistem adalah metode evaluasi *hamming loss* dan metode *k-folds Cross Validation* dimana  $k=5$  sedangkan metode klasifikasi yang digunakan yaitu *Random Forest*. Pada sistem yang dibangun menggunakan RF mendapatkan hasil terbaik dengan menggunakan model pendekatan *multi label BR*, dengan rata-rata *hamming loss* yang didapat yaitu 0,0728992 [6].

### 2.1 Hadis

Hadis merupakan pegangan utama umat muslim setelah Al Qur'an yang menutun umat kepada kehidupan yang lebih baik. Ada beberapa hadis yang telah diriwayatkan kepada umat muslim agar tidak tersesat selama di dunia, diantaranya hadis Bukhari, Muslim, Tirmidzi, Nasa'I, Ibnu Majah, dan Ibnu Daud [8]. Hadis dalam islam memiliki dua fungsi, yaitu sebagai berikut:

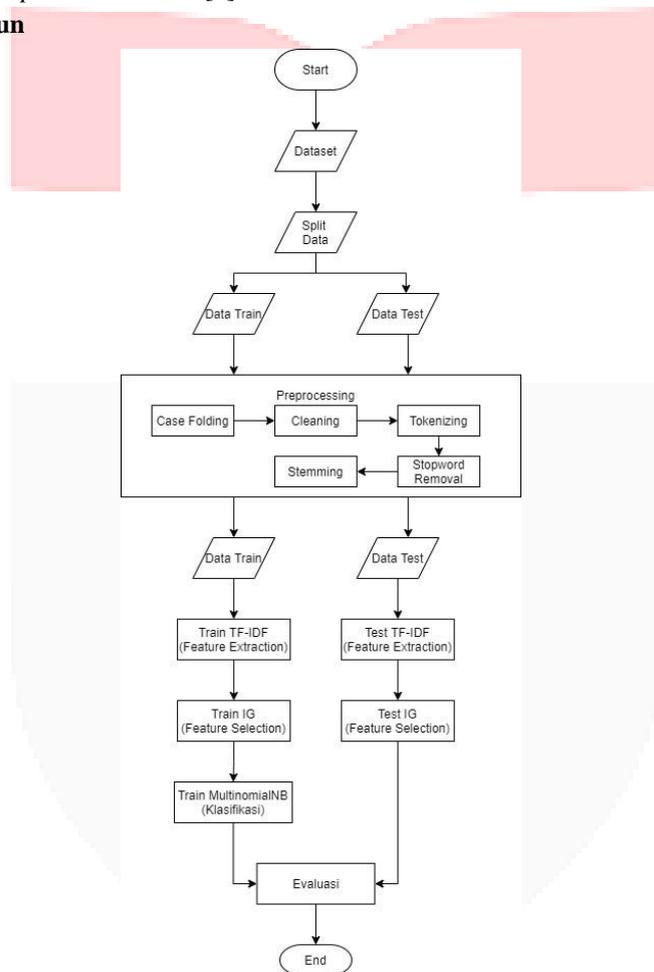
1. Mubayyin Mubayyin adalah fungsi dari Hadis, yang menjelaskan makna global dan universal dari Quran
2. Sumber Hukum Tersendiri Asal muasal hukumnya sendiri adalah fungsi hadis sebagai sumber menjelaskan mengenai sesuatu yang tidak umum [9].

Hadis mencakup hal-hal yang dapat dijadikan sebagai pedoman oleh hamba Allah. Sebagai umat muslim, ada hal yang dilarang sehingga wajib di jauhi dan ada hal yang diperintahkan maka dikerjakan semampu umatnya [18].

## 2.2 Klasifikasi *Multi Label*

Klasifikasi multi-label adalah teknik klasifikasi yang digunakan untuk menentukan label kelas data dengan beberapa label kelas [3]. Terdapat dua pendekatan metode klasifikasi *multi-label*, yaitu *Problem Transformation Methods* dan *Algorithm Adaptation Methods* [8].

## 3. Sistem yang Dibangun



**Gambar 1.** Gambaran Umum Sistem

Dibangun sistem yang dapat mengelompokkan *corpus* pada hadis yang berisi kumpulan data multi-label. Data terdiri dari 3 kategori berupa anjuran, larangan, dan informasi.

### 3.1 Dataset

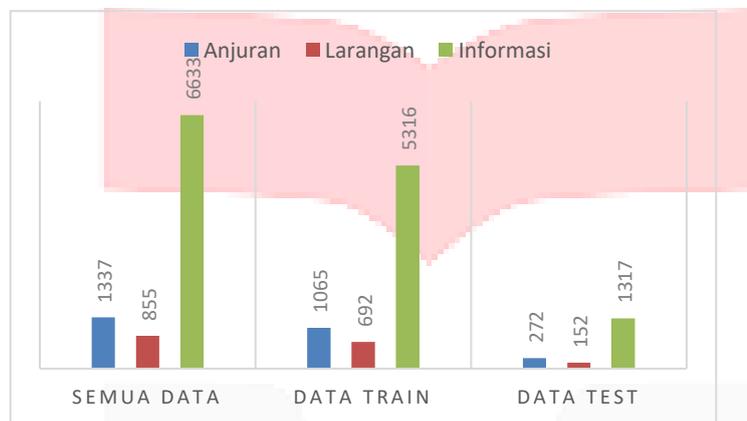
Pada penelitian ini penulis menggunakan dataset hasil terjemahan hadis Shahih Bukhari yang memiliki kategori anjuran, larangan, dan informasi. Contoh dataset hadis ditampilkan pada tabel 1.

**Tabel 1.** Salah satu isi dataset hadis

Hadis	Anjuran	Larangan	Informasi
tidaklah kamu menafkahkan suatu nafkah yang dimaksudkan mengharap wajah Allah kecuali kamu akan diberi pahala termasuk sesuatu yang kamu suapkan ke mulut istrimu.	0	1	1

Dataset hadis memiliki 7008 data dengan persebaran 1337 berlabel anjuran, 844 berlabel larangan, dan 6633 berlabel informasi. Pada dataset ini terdapat 2 jenis pelabelan, yaitu label 1 untuk data yang dianggap termasuk dalam label kelas hadis tersebut dan label 0 untuk data yang bukan termasuk label kelas hadis tersebut. Data asli dari hadis akan dibagi menjadi 2 jenis data seperti pada gambar 2.

Penulis membagi data menjadi 2 bagian yaitu 80% untuk data latih dan 20% untuk data uji dan menggunakan *random state* untuk membuat nilai konsisten pada saat sistem dijalankan.

**Gambar 2.** Pembagian data

### 3.2 Preprocessing

*Preprocessing* merupakan tahap awal klasifikasi untuk menangani permasalahan yang dapat mengganggu hasil klasifikasi, seperti missing value atau format data yang tidak sesuai dengan sistem.

#### 1. Case Folding

*Case folding* merupakan tahap mengkonversi total bacaan dokumen menjadi wujud standar atau huruf kecil.

#### 2. Cleaning

*Cleaning* merupakan proses penghilangan tanda pada bacaan dokumen yang tidak dibutuhkan dalam membangun sistem, seperti “(”, “,”, dan “.”.

#### 3. Tokenizing

*Tokenizing* merupakan proses pembelahan kalimat pada teks dokumen menjadi kata-kata individu.

#### 4. Stopword Removal

*Stopword Removal* merupakan proses menyelapkan perkata yang universal atau yang sangat banyak digunakan. Penelitian ini menggunakan *stopword* bahasa Indonesia dari *Natural Language Toolkit* (NLTK).

#### 5. Stemming

*Stemming* merupakan proses yang mengganti wujud kata bacaan dokumen menjadi kata dasar yang cocok berdasarkan struktur bahasa Indonesia yang baik dan benar. Proses penghilangan imbuhan menggunakan *stemming* Sastrawi.

**Tabel 1.** Tahap *preprocessing*

Proses	Hasil <i>Preprocessing</i>
--------	----------------------------

Data Asli	Ia tidak boleh memakai baju, Imamah (surban yang dililitkan pada kepala), celana panjang, mantel, atau pakaian yang diberi minyak wangi atau za'faran. Jika dia tidak mendapatkan sandal, maka ia boleh mengenakan sepatu dengan memotongnya hingga di bawah mata kaki.
Case Folding	ia tidak boleh memakai baju, imamah (surban yang dililitkan pada kepala), celana panjang, mantel, atau pakaian yang diberi minyak wangi atau za'faran. jika dia tidak mendapatkan sandal, maka ia boleh mengenakan sepatu dengan memotongnya hingga di bawah mata kaki.
Cleaning	ia tidak boleh memakai baju imamah surban yang dililitkan pada kepala celana panjang mantel atau pakaian yang diberi minyak wangi atau zafaran jika dia tidak mendapatkan sandal maka ia boleh mengenakan sepatu dengan memotongnya hingga di bawah mata kaki
Tokenizing	['ia' 'tidak' 'boleh' 'memakai' 'baju' 'imamah' 'surban' 'yang' 'dililitkan' 'pada' 'kepala' 'celana' 'panjang' 'mantel' 'atau' 'pakaian' 'yang' 'diberi' 'minyak' 'wangi' 'atau' 'zafaran' 'jika' 'dia' 'tidak' 'mendapatkan' 'sandal' 'maka' 'ia' 'boleh' 'mengenakan' 'sepatu' 'dengan' 'memotongnya' 'hingga' 'di' 'bawah' 'mata' 'kaki']
Stopword	memakai baju imamah surban dililitkan kepala celana panjang mantel pakaian di beri minyak wangi zafaran mendapatkan sandal mengenakan sepatu memotongnya hingga di bawah mata kaki
Stemming	pakai baju imamah surban dililit kepala celana panjang mantel pakaian diberi minyak wangi zafaran dapat sandal kenakan sepatu potong hingga di bawah mata kaki

Tabel 2 merupakan contoh pengujian pada dokumen untuk setiap tahapan *preprocessing* sehingga pada proses pembangunan model sistem data dapat terklasifikasi dengan baik.

### 3.3 Feature Extraction

Setelah data dibersihkan pada tahap *preprocessing*, selanjutnya akan dikonversin pada fitur ekstraksi atau pembobotan kata dengan menggunakan TF-IDF. [13]. *Term Frequency* yaitu untuk menghitung suatu kata yang diulang dalam sebuah teks. Sedangkan *Inverse Document Frequency* yaitu untuk menghitung probabilitas dalam suatu kata pada teks [14]. Persamaan metode TF-IDF sebagai berikut :

$$TF(d, t) = f(d, t) \quad (1)$$

TF(d,t) pada persamaan 1 merupakan *term frequency*, sedangkan f(d,t) merupakan frekuensi *term t* pada dokumen *frequency*.

$$TF - IDF (d, t) = TF(d, t) * \log \frac{N_d}{df(t)} \quad (2)$$

Untuk persamaan 2,  $N_d$  merupakan jumlah dokumen keseluruhan dan untuk  $df(t)$  merupakan jumlah dokumen yang mengandung *term t*.

Sebagai contoh terdapat 2 dokumen dengan dokumen 1 berisi “Iman memiliki lebih dari enam puluh cabang, dan malu adalah bagian dari iman” dan dokumen 2 berisi “Tinggalkanlah dia, karena sesungguhnya malu adalah bagian dari iman”. Dari kedua dokumen tersebut dihitung bobot setiap pada kalimat seperti tabel.

**Tabel 2.** Contoh pengujian TF-IDF

dokumen	tf		df	IDF ( $\log N/df_i$ )	tf * idf	
	d1	d2			d1	d2
Memiliki	1	0	1	$\log(2/1) = 0.3$	0.3	0

Lebih	1	0	1	$\log(2/1) = 0.3$	0.3	0
Malu	1	1	2	$\log(2/2) = 0$	0	0
Adalah	1	1	2	$\log(2/2) = 0$	0	0
Bagian	1	1	2	$\log(2/2) = 0$	0	0
Tinggalkanlah	0	1	1	$\log(2/1) = 0.3$	0	0.3
Sesungguhnya	0	1	1	$\log(2/1) = 0.3$	0	0.3

Tabel 3 merupakan contoh perhitungan nilai bobot setiap kata menggunakan TFIDF.

### 3.4 Feature Selection

Setelah pembobotan kata pada proses fitur ekstraksi, proses selanjutnya adalah melakukan seleksi fitur. Seleksi fitur merupakan proses pengurangan jumlah fitur sehingga dapat mempercepat proses klasifikasi. Seleksi fitur yang digunakan yaitu *Information Gain* yang dapat mengetahui fitur mana yang mempunyai data paling banyak menurut kelas tertentu. [10]. Nilai dari IG didapatkan berdasarkan penghitungan nilai untuk setiap *uniqueterm* [14].

Nilai IG diperoleh dari hasil pengurangan nilai dari entropi sebelum dilakukan pemisahan dengan nilai dari entropi setelah dilakukan pemisahan. *Information Gain* memiliki nilai dengan rentang 0 sampai dengan 1, yang memiliki arti semakin tinggi nilai IG dari atribut tersebut maka akan tinggi tingkat keterkaitan atribut terhadap data. Setelah dilakukan perhitungan nilai IG untuk setiap atribut, lalu dilakukan penyeleksian terhadap data terkait dengan menggunakan nilai *threshold* telah ditentukan. Pada penelitian ini, nilai *threshold* yang digunakan yaitu 0.1, 0.2, 0.3 dan 0.4. Data yang memenuhi kriteria akan digunakan sebagai atribut untuk pembangunan model sistem. Rumus *entropy* dapat dirumuskan sebagai berikut,

$$Entropy(Y) = -p_{(+)} \log_2 p_{(+)} - p_{(-)} \log_2 p_{(-)} \quad (3)$$

dimana Y merupakan data sample yang digunakan untuk training pada IG,  $p_{(+)}$  merupakan probabilitas atribut memiliki nilai positif, sedangkan  $p_{(-)}$  merupakan probabilitas atribut memiliki nilai negatif. Selanjutnya, perhitungan nilai *gain* dengan menggunakan nilai *entropy* pada persamaan 4.

$$Gain(Y, X) = Entropy(Y) - \sum_{v \in \text{Values}(X)} \frac{|Y_v|}{|Y|} \times Entropy(Y_v) \quad (4)$$

Y adalah kelas dan X adalah atribut. Nilai  $|Y_v|$  menunjukkan jumlah data dari kelas v, dimana v merupakan nilai dari atribut yang terdiri dari kelas positif dan kelas negatif.  $|Y|$  menunjukkan jumlah keseluruhan data, sedangkan  $Entropy(Y_v)$  merupakan *entropy* untuk sampel dengan nilai v.

**Tabel 3.** Top 10 fitur IG setiap label

Term Anjuran	Nilai IG Anjuran	Term Larangan	Nilai IG Larangan	Term Informasi	Nilai IG Informasi
Sumah	1.000000	Jangan	1.000000	Jual	1.000000
Hendak	0.994066	Larang	0.581705	Larang	0.853241
Wanita	0.987488	Reda	0.187917	Jangan	0.827488
Mendatangi	0.946902	Kalimat	0.163379	Ikhlas	0.808723
Falaa	0.938558	Diperintahkan	0.156043	Rekaat	0.792118
Shaumnya	0.934265	Pintupintunya	0.150273	Wahai	0.787679
Lahmi	0.898768	Dahar	0.148934	Nashara	0.780293
Rizki	0.898116	Tetua	0.148467	Mangsa	0.733188
Orantuanya	0.897698	Salat	0.147660	Menjalani	0.712881
asysyabiy	0.895035	Tabit	0.145279	malaikat	0.706726

Tabel 4 merupakan top 10 nilai IG pada *term* setiap label pada dokumen dataset yang akan digunakan pada penelitian ini.

### 3.5 Multinomial Naïve Bayes

Hasil dari proses perhitungan pada seleksi fitur *Information Gain* akan menjadi inputan pada tahap klasifikasi. Model Multinomial mampu mempertimbangkan frekuensi setiap kata yang muncul dalam dokumen [13]. Model *Multinomial NB* secara umum dapat digambarkan dengan probabilitas dari dokumen  $d$  menjadi bagian dari kelas  $c$  [15]. Dataset yang digunakan berupa multi-label, tetapi *Multinomial NB* hanya mampu mengklasifikasikan multiclass sehingga dibangun sistem untuk setiap label. Pada penelitian ini, penggunaan Multinomial pada data multi-label dilakukan dengan membangun 3 model, yaitu model untuk label anjuran, model untuk label larangan, dan model untuk label informasi. Berikut rumus yang digunakan,

$$P(c|d) = P(c) \prod_{1 \leq k \leq n_d} P(t_k|c) \quad (5)$$

dimana  $P(t_k | c)$  merupakan peluang kemunculan suatu term  $t_k$  (term dokumen  $d$ ) dalam dokumen kelas  $c$  sedangkan  $P(c)$  merupakan prior peluang dari suatu dokumen yang ada pada kelas  $c$ . Berikut rumus untuk menentukan peluang dari suatu kelas:

$$P(c) = \frac{\text{Jumlah suatu kelas dokumen kelas latih}}{\text{Jumlah total dokumen kelas yang ada}} \quad (7)$$

Sedangkan kondisi probabilitas  $P(t|c)$  dapat dihitung menggunakan jumlah kemunculan dari  $t$  pada dokumen data training, termasuk beberapa kemunculan suatu *term* pada suatu dokumen. Berikut merupakan formula dari perhitungan tersebut

$$P(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{c't'}} \quad (8)$$

Tujuan utama pada klasifikasi teks yaitu mencari kelas terbaik pada dokumen. Klasifikasi *Multinomial Naïve Bayes* menggunakan *maximum a posteriori* (MAP) untuk mencari kelas terbaik.

$$c_{map} = \underset{c \in C}{\operatorname{argmax}} P(c|d) = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{1 \leq k \leq n_d} P(t_k|c) \quad (9)$$

Pada *equation 9*, ada banyak probabilitas yang muncul berulang kali, hal tersebut dapat menyebabkan *floating point underflow*. Untuk menghindari hal tersebut maka dilakukan proses penjumlahan suatu peluang kata yang muncul menggunakan logaritma  $\log(x \cdot y) = \log(x) + \log(y)$ . Pada *Multinomial Naïve Bayes* untuk mencari MAP adalah sebagai berikut:

$$c_{map} = \underset{c \in C}{\operatorname{argmax}} [\log P + \sum_{1 \leq k \leq n_d} \log P(t_k|c)] \quad (10)$$

### 3.6 Evaluasi

Proses akhir dalam melakukan klasifikasi adalah evaluasi sistem. Pada penelitian penulis mengevaluasi model dengan menggunakan *confusion matrix* untuk melihat kinerja mode klasifikasi [16]. Pada teknik evaluasi *confusion matrix* dilakukan perhitungan akurasi, *precision*, *recall*, dan *F1-Score* [17]. Akurasi digunakan untuk mengetahui seberapa akurat model dalam mengklasifikasikan dokumen dengan benar. Pengevaluasian hasil klasifikasi dapat dilihat pada tabel 7.

**Tabel 4.** Tabel *Confusion Matrix*

Nilai Prediksi	Nilai Aktual	
	Positif	Negatif
Positif	TP	FP
Negatif	FN	TN

TP : Hasil klasifikasi benar sesuai dengan keputusan sistem (True Positive)  
FP : Hasil klasifikasi salah menurut sistem (False Positive)



TN : Hasil klasifikasi yang memang bukan merupakan anggota kategori tersebut (True Negative)  
 FN : Dokumen yang seharusnya termasuk dalam hasil klasifikasi tapi tidak menjadi bagian anggota kategori tersebut (False Negative).

Hasil dari *confusion matrix* akan digunakan untuk evaluasi sistem pada *precision*, *recall*, dan F1-Score. Rumus yang digunakan untuk ketiga metode evaluasi tersebut ditunjukkan pada formula 11, 12, dan 13.

$$Precision = \frac{TP}{TP + FN} \quad (11)$$

$$Recall = \frac{TP}{TP + FP} \quad (12)$$

$$F - Measure = \frac{2PR}{Precision + Recall} \quad (13)$$

Hasil akurasi dari evaluasi akan digunakan sebagai acuan untuk performansi pada setiap model yang dibangun. Jika dataset memiliki jumlah data FN dan FP yang *symmetric* maka nilai akurasi akan digunakan sebagai acuan, sedangkan jika tidak *symmetric* maka F1 Score yang akan menjadi acuan. Untuk *precision* akan dijadikan sebagai acuan jika lebih menginginkan terjadinya TP dan *recall* jika memilih FP lebih baik dari pada FN.

#### 4. Evaluasi

Pada penelitian ini dilakukan 3 skenario pengujian berupa skenario pertama yaitu pengujian terhadap *preprocessing* dengan melihat pengaruh proses *stemming* dan *stopword removal*. Skenario kedua yaitu feature selection untuk melihat pengaruh setiap variasi threshold dengan menggunakan *information gain*. Sedangkan skenario ketiga yaitu membandingkan dua implementasi dari *Naive Bayes Classifier* yaitu *Multinomial* dan *Gaussian*.

##### 4.1 Analisis Pengaruh Penggunaan *Stemming* dan *Stopword Removal*

Pada skenario pertama dilakukan proses pengujian terhadap teknik *preprocessing* dengan kombinasi menggunakan *stemming* dan *stopword removal*; tanpa menggunakan *stemming*; tanpa menggunakan *stopword removal*; dan tanpa menggunakan *stemming* dan *stopword removal*. Proses klasifikasi menggunakan TFIDF sebagai fitur ekstraksi, *Information Gain* dengan *threshold* 0.1 sebagai fitur seleksi, *Multinomial Naive Bayes* sebagai implementasi klasifikasi, dan dilakukan evaluasi dengan menggunakan *Confusion Matrix*.

**Tabel 5.** Tabel hasil pengujian *stemming* dan *stopword* data test

Stopword Removal	Stemming	Performansi
		Akurasi
Y	Y	80%
Y	-	81%
-	Y	85%
-	-	82%

Berdasarkan hasil pengujian terhadap skenario pertama, kombinasi teknik *preprocessing* dengan kinerja sistem terbaik yaitu pada percobaan tanpa menggunakan *stopword removal* dengan hasil akurasi 85% dengan nilai F1-Score 71%, serta didapatkan jumlah data yang salah terklasifikasikan pada label anjuran sebanyak 354 data, label larangan sebanyak 100 data, dan label informasi sebanyak 187 data. Pengujian menggunakan proses *stemming* dan *stopword removal* memiliki penurunan hasil klasifikasi sebanyak 5%, hal tersebut diakibatkan karena dalam menggunakan kedua proses tersebut lebih banyak lagi data yang salah terklasifikasi dalam pembangunan model. Pada percobaan lain dengan menggunakan kombinasi *stopword removal* dan *stemming* didapatkan hasil akurasi 80% serta didapatkan jumlah data yang salah terklasifikasikan pada label anjuran sebanyak 254 data, label larangan sebanyak 254 data, dan label informasi sebanyak 214 data. Percobaan tanpa

menggunakan *stemming* didapatkan hasil akurasi 81% serta didapatkan jumlah data yang salah terklasifikasikan pada label anjuran sebanyak 350 data, label larangan sebanyak 271 data, dan label informasi sebanyak 196 data. Percobaan tanpa menggunakan keduanya yaitu tanpa *stopword removal* dan tanpa *stemming* didapatkan hasil akurasi 82% serta didapatkan jumlah data yang salah terklasifikasikan pada label anjuran sebanyak 334 data, label larangan sebanyak 265 data, dan label informasi sebanyak 143 data. Ketiga percobaan tersebut menghasilkan akurasi yang lebih rendah dibanding tanpa menggunakan *stopword removal*.

**Tabel 7.** Tabel perbandingan label sebenarnya dengan label hasil prediksi

Proses	Data Hadis	Label Sebenarnya		
		Anjuran	Larangan	Informasi
Data asli	Janganlah kalian berdusta terhadapku (atas namaku), karena barangsiapa berduasta terhadapku dia akan masuk neraka.	0	1	1
Data Asli	Barangsiapa berwudhu hendaklah mengeluarkan (air dari hidung), dan barangsiapa beristinja' dengan batu hendaklah dengan bilangan ganjil.	1	0	1
		Label Prediksi		
		Anjuran	Larangan	Informasi
Full Prerocessing	dusta hadap (namaku), barangsiapa dusta hadap masuk neraka.	0	0	1
Tanpa Stopword	janngan dusta hadap (namaku), karena barangsiapa dusta hadap dia akan masuk neraka.	0	1	1
Dengan Stemming	siapa wudhu hendak keluar (air dari hidung), dan siapa beristinja' dengan batu hendak dengan bilangan ganjil	0	0	1
Tanpa Stemming	Barangsiapa berwudhu hendaklah mengeluarkan (air dari hidung), barangsiapa beristinja' dengan batu hendaklah bilangan ganjil	1	0	1

Berdasarkan perbandingan yang ditampilkan pada tabel 9, dengan penghilangan kata-kata yang dianggap tidak penting dapat mengubah makna dari setiap kalimat sehingga memengaruhi hasil dari setiap prediksi.

#### 4.2 Analisis Pengaruh Penggunaan Fitur Seleksi

Pada skenario kedua, penulis melakukan pengujian terhadap pengaruh fitur seleksi *Information Gain* dalam mencari akurasi terbaik dengan menggunakan metode *preprocessing*, fitur ekstraksi TFIDF, implementasi klasifikasi *Multinomial Naïve Bayes*, dan akan dievaluasi menggunakan *confusion Matrix*. Berdasarkan hasil terbaik pada skenario pertama, tahapan yang akan digunakan untuk *preprocessing* yaitu, *case folding*, *data cleaning*, *tokenizing*, dan *stemming* (tanpa *stopword removal*). Percobaan pertama dilakukan dengan tanpa menggunakan seleksi fitur *information gain* dan didapatkan hasil pada tabel 8.

**Tabel 6.** Tabel hasil pengujian tanpa menggunakan fitur seleksi

Data	Akurasi
Data Train	91%
Data Test	83%

Berdasarkan hasil tabel 8, terjadi *overfitting* dimana pengujian terhadap data train mendapatkan akurasi yang lebih baik dibanding pengujian pada data test. Hal tersebut kemungkinan terjadi karena jumlah yang digunakan pada data test terlalu sedikit dan *overfitting* menunjukkan bahwa model sistem yang dibangun hanya berfokus pada data train, sehingga pada data test tidak memberikan hasil yang cukup baik. Untuk mengatasi masalah

*overfitting* maka dilakukan pengujian dengan menggunakan seleksi fitur *information gain*. *Information Gain* melakukan penyeleksian dengan menggunakan *threshol*. Tabel 9 merupakan hasil pengujian dengan menggunakan seleksi fitur.

**Tabel 7.** Hasil pengujian *threshol*

Threshol	Akurasi
0.1	85%
0.2	83%
0.3	81%
0.4	79%

Berdasarkan hasil tabel 9, hasil pengujian dengan menggunakan *threshol* 0.1 mendapatkan hasil akurasi sistem terbaik dengan 85% dengan jumlah data yang salah terklasifikasi yaitu 354 data pada label anjuran, 100 data pada label larangan, dan 187. Hal tersebut mengalami penurunan 6.2% dibanding jumlah data yang salah terklasifikasi pada pengujian tanpa IG dengan 499 data pada label anjuran, 182 pada label larangan, dan 157 pada label informasi. Hal ini membuktikan, fitur yang memiliki nilai IG diatas 0.1 cukup untuk mewakili keseluruhan fitur dalam mempresentasikan suatu kelas. Penggunaan *threshol* 0.1 mendapatkan hasil yang lebih baik dibandingkan pengujian terhadap *threshol* yang lain. Hal tersebut ditunjukkan pada label 10.

**Tabel 8.** Jumlah data salah terklasifikasi label anjuran

Label	0.1	0.2	0.3	0.4	0.1 $\cap$ 0.2	0.1 $\cap$ 0.3	0.1 $\cap$ 0.2 $\cap$ 0.3 $\cap$ 0.4
Anjuran	354	394	414	433	299	323	194

Hasil tabel 10 menunjukkan bahwa jumlah data yang salah terklasifikasi pada *threshol* 0.1 lebih sedikit dibanding penggunaan *threshol* 0.2, 0.3, dan 0.4. Beberapa data yang salah pada *threshol* yang lebih tinggi dapat diperbaiki dengan menggunakan *threshol* 0.1, namun meskipun mendapatkan hasil akurasi yang lebih tinggi beberapa data juga tidak mampu untuk diperbaiki. Setiap *threshol* memiliki kinerja yang berbeda-beda untuk setiap data. Hal tersebut dapat dibuktikan pada tabel 11.

**Tabel 11.** Contoh data dan hasil pengklasifikasiannya

Data	Keterangan
Dari Muhammad, hamba Allah dan Rasul-Nya untuk Heraclius. Penguasa Romawi, Keselamatan bagi siapa yang mengikuti petunjuk.	Dapat diklasifikasikan dengan baik pada setiap pengujian <i>threshol</i> .
Semua perbuatan tergantung niatnya, dan (balasan) bagi tiap-tiap orang (tergantung) apa yang diniatkan.	Tidak dapat diklasifikasikan dengan baik pada setiap pengujian <i>threshol</i>
Iman memiliki lebih dari enam puluh cabang, dan malu adalah bagian dari iman.	Mampu diklasifikasikan dengan baik dengan 0.1 tapi tidak dengan 0.2
Terkadang datang kepadaku seperti suara gemerincing lonceng dan cara ini yang paling berat buatku, lalu terhenti sehingga aku dapat mengerti apa yang disampaikan.	Mampu diklasifikasikan dengan baik pada 0.2 namun tidak pada 0.1

Berdasarkan hasil terhadap seluruh pengujian, dapat disimpulkan dengan menggunakan seleksi fitur dapat mengatasi masalah *overfitting* dan hasil pengujian kinerja sistem didapatkan dengan menggunakan *threshol* 0.1.

#### 4.3 Analisis Perbandingan Penggunaan Data *Balance* dan Data *Imbalance*

Skenario ketiga dilakukan dengan melakukan pengujian terhadap data *balance* dan data *imbalance*. Data yang menjadi imputan untuk kedua klasifikasi sebelumnya telah dilakukan pembersihan data dengan *preprocessing* (tanpa *stopword removal*), pembobotan kata dengan TF-IDF, dan seleksi fitur atau atribut menggunakan *Information Gain*. Data diuji menggunakan klasifikasi *Multinomial Naïve Bayes* dan dievaluasi menggunakan *confusion matrix*. Hasil pengujian ditampilkan pada tabel 12.

**Tabel 9.** Tabel pengujian data *balance* dan data *imbalance*

Data	Jumlah data						Performansi (Treshold 0.1)	
	Anjuran		Larangan		Informasi		Akurasi	F1-Score
	1	0	1	0	1	0		
Balance	4535	4535	4908	4908	5316	5316	85%	71%
Imbalance	1065	4535	692	4908	5316	284	88%	47%

Pengujian pertama dilakukan dengan menggunakan jumlah data asli yang dikategorikan sebagai data *imbalance*. Data tersebut dikatakan *imbalance* karena persebaran data antara yang benar merupakan kelas tersebut “1” dengan yang bukan merupakan bagian kelas tersebut “0” memiliki perbedaan jumlah yang signifikan. Data tidak seimbang mengakibatkan prediksi klasifikasi yang tidak seimbang. Pada penelitian ini, hasil akurasi yang didapatkan dengan data tidak seimbang yaitu 88%. Karena pada *confusion matrix* nilai *false positif* dan *false negatif* tidak *symmetric* maka *F1-Score* digunakan sebagai acuan dan nilai *F1-Score* yang didapatkan dengan menggunakan data tidak seimbang yaitu 47%.

Pengujian kedua dilakukan dengan menggunakan data seimbang yaitu data yang memiliki jumlah label sama semua. Data yang digunakan merupakan data asli yang telah diproses dengan melakukan teknik *oversampling*. Pada proses ini teknik *oversampling* yang digunakan yaitu SMOTE (*Sintetis Minoritas Sampling Teknik*). Berdasarkan hasil pengujian, akurasi yang dihasilkan dengan menggunakan data seimbang atau *balance* 85% dimana hasil tersebut lebih rendah dibandingkan dengan menggunakan data tidak seimbang. Akan tetapi, karena nilai *false positif* dan *false negatif* yang tidak *symmetric* maka yang menjadi acuan yaitu *F1-Score* yang memiliki hasil lebih tinggi dibanding menggunakan data tidak seimbang yaitu 71%.

## 5. Kesimpulan

Berdasarkan hasil pengujian dari beberapa skenario terhadap klasifikasi teks hadis bukhari, kinerja sistem terbaik didapatkan dengan kombinasi preprocessing tanpa teknik *stopword removal*, pembobotan kata TFIDF, seleksi fitur IG dan klasifikasi *Multinomial* NB dengan hasil akurasi 85%. Pengujian menggunakan proses tersebut mendapatkan hasil kinerja sistem terbaik karena tidak terjadinya *overfitting* dan menggunakan data yang seimbang, sehingga sistem dapat dievaluasi dengan baik. Penggunaan teknik *stopword removal* mendapatkan hasil yang kurang baik karena pada proses *stopword removal* dilakukan penghilangan kata yang umum atau kata yang dianggap tidak penting sehingga dapat menghilangkan informasi lengkap dari kalimat. Dengan melakukan perbandingan dengan penelitian [3], hasil akurasi yang didapatkan lebih rendah dengan menggunakan *stopword removal* dan tanpa menggunakan *stemming*.

Adapun saran yang dapat diterapkan pada penelitian selanjutnya, yaitu melakukan pengurangan atau penambahan kata yang dianggap tidak relevan pada teknik *stopword removal* sehingga dapat membantu meningkatkan hasil kinerja sistem yang lebih baik.

## Referensi

- [1] Zein, Muhammad Ma'shum. *Ulumul Hadits dan Musthalah Hadits*. (2008).
- [2] R. Melita, V. Amrizal, H. B. Suseno, and T. Dirjam. Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Syarah Umdatil Ahkam). *Jurnal Teknik Informatika*, vol.2, no.1, 2018.
- [3] M. Y. Abu Bakar, Adiwijaya, and S. Al Faraby. Multi-Label Topic Classification of Hadith of Bukhari (Indonesian Language Translation) Using Information Gain and Backpropagation Neural Network. *In Proceedings of the 2018 International Conference on Asian Language Processing*, IALP 2018, 2019, pp. 344–350.
- [4] S. Al Faraby, E. R. Rachmawati Jasin, A. Kusumaningrum, and Adiwijaya. Classification of hadith into positive suggestion, negative suggestion, and information, *in Proceedings of the 2017 International Conference on Data and Information Science*, ICDIS 2017, doi :10.1088/1742-6596/971/1/01204.

- [5] D. H. Kalokasari, Dr. I. M. Shofim, and A. H. Setyaningrum. Implementasi Algoritma Multinomial Naive Bayes Classifier Pada Sistem Klasifikasi Surat Keluar. *Jurnal Teknik Informatika*, vol.10, no.2, 2017, doi: 10.150408.
- [6] A. Wiraguna, S. Al Faraby, and Adiwijaya. Klasifikasi Topik Multi Label pada Hadis Bukhari dalam Terjemahan Bahasa Indonesia Menggunakan Random Forest. *Vol.6, no.1*, pp. 2335-9365, 2019.
- [7] H. Fauzan, Adiwijaya, and S. Al Faraby. Pengklasifikasian Topik Hadits Terjemahan Bahasa Indonesia Menggunakan Latent Semantic Indexing dan Support Vector Machine. *Vol.5, no.3*, pp. 2335-9365, 2018.
- [8] E. Riviera, R. Jasin, and S. Al-faraby. Klasifikasi Anjuran, Larangan dan Informasi pada Hadis Sahih Al-Bukhari berdasarkan Model Unigram menggunakan Artificial Neural Network (ANN). *Vol. 4, no.3*, pp. 4683–4694, 2017.
- [9] D. T. M. Syair Audi Liri Sacra, Said Al Faraby. Bukhari Menggunakan Naive Bayes Classifier Classification of Suggestion, Prohibition and Information of Shahih Bukhari Hadith Using Naive Bayes Classifier. *Vol. 4, no. 3*, pp. 4794–4802, 2017.
- [10] S. H. A. Aini, Y. A. Sari, and A. Arwan. Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naive Bayes. *J. Pengemb. Teknol. Inf. dan Ilmu Komputer*; Vol 2 No 9, vol. 2, no. 9, pp. 2546–2554, Feb. 2018.
- [11] P. D. Nugraha, S. Al Faraby, and Adiwijaya. Klasifikasi Dokumen Menggunakan Metode k-Nearest Neighbor (kNN) dengan Information Gain. *Vol.5, no.1*, pp. 2335-9365, 2018.
- [12] F. Handayani, F. Setio. Implementasi Algoritma Naive Bayes Classifier Dalam Pengklasifikasian Teks Otomatis Pengaduan Dan Pelaporan Masyarakat Melalui Layanan Call Center 110. *Jurnal Teknik Elektro Unnes*, vol. 7, no. 1, 2015, pp. 19-24, doi: 10.15294/jte.v7i1.8585.
- [13] A. Rahma, Wiranto, and A. Doewes. New Classification Using Multinomial Naive Bayes. *ITSMART: Jurnal Teknologi dan Informasi*, vol.6, no.1, pp. 2301-7201, 2017.
- [14] Xu Y, Jones G, Li J T, Wang B, dan Sun C M. 2007. A Study on Mutual Information-based Feature Selection for Text Categorization. *Journal of Computational Information Systems* 3:3 1007-1012
- [15] UP, O. e. (2009, April 1). Naive Bayes text classification. Retrieved from nlp stanford: <https://nlp.stanford.edu/IR-book/html/htmledition/naive.bayes-text-classification-1.html>
- [16] M. A. Ulfa, B. Irmawati and A. Y. Husodo. Twitter Sentiment Analysis using Naive Bayes Classifier with Mutual Information Feature Selection. *J-COSINE*, vol. 2, no. 2, pp. 106-111, 2018.
- [17] A. Rahman, W. and A. Doewes. Online News Classification Using Multinomial Naive Bayes. *ITSMART: Jurnal Ilmiah Teknologi dan Informasi*, vol. 6, no. 1, pp. 32-38, 2017.
- [18] H. A. Ibnu. *Ensiklopedia Anjuran dan Larangan*. Kudus: Pustaka As Sunnah.