ISSN 2085-4552

# Implementasi Rocchio's Classification dalam Mengkategorikan Renungan Harian Kristen

Elisabeth Adelia Widjojo, Antonius Rachmat C, R. Gunawan Santosa
Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Kristen Duta Wacana, Yogyakarta
adeliawidjojo@gmail.com, anton@ti.ukdw.ac.id, gunawan@ukdw.ac.id

Diterima 13 Maret 2014
Disetujui 04 Juni 2014

***Abstract*—Nowadays, many Christian institutions are using digital media to save spiritual pictures, musics or videos, even daily devotional articles which is usually printed monthly. Since many daily devotionals are published in the Internet, it will be difficult to find a daily devotional articles with a spesific topic/category. To make it easier, in this research we use Rocchio's classification, which use TF-IDF weighting for classification and centroid calculation in every category to classify daily devotional articles. Every testing article will be matched with the centroid using cosine similariy. As a result, the system accuracy is 73,33% using 20% of feature selection. The highest precision goes to Wisdom category which score is 1 for precision by using 100% feature selection. While the highest recall goes to Motivaton category which score is 1 by using 100% feature selection..**

***Index Terms*—classification, categorization, daily devotional article, Rocchio's classification, centroid, similarity.**

## I. PENDAHULUAN

Banyak lembaga-lembaga kristen termasuk gereja-gereja yang sudah menyimpan data-datanya dalam bentuk digital baik berupa teks, gambar, musik, maupun video. Salah satu data teks yang banyak ditemui adalah teks yang berisi renungan / kotbah yang bisa dijumpai di website-website. Kesulitan yang dihadapi pencari teks renungan adalah sulitnya mencari renungan yang sesuai dengan topik / kategori tertentu yang diinginkan. Beberapa topik yang biasa dicari misalnya topik keselamatan, cinta kasih, tri tunggal, perkawinan, dan lain-lain. Dari kebutuhan tersebut diperlukan suatu sistem yang mampu mengkategorikan renungan secara otomatis berdasarkan suatu kategori tertentu yang disepakati sebelumnya. Dari banyak algoritma klasifikasi, algoritma Rocchio dan Naive Bayes memiliki waktu kompleksitas yang hampir sama, namun Rocchio memiliki tingkat keakuratan yang cukup baik walaupun masih kalah dengan algoritma k-NN. Penulis akan menggunakan algoritma Rocchio's classification dalam penelitian ini.

Masalah yang akan diteliti adalah bagaimana keakuratan Rocchio's classification dalam mengkategorikan renungan harian kristen, serta *precision* dan *recall* untuk masing-masing kategori. Metode yang digunakan dalam penelitian ini adalah studi pustaka yang bertujuan untuk memberikan pengetahuan / teori mengenai hal – hal yang terkait dengan klasifikasi dokumen dan algoritma Rocchio. Studi pustaka dilakukan dengan cara membaca buku, literatur, jurnal dan artikel dari internet yang berhubungan dengan masalah yang dibahas. Kemudian dilanjutkan dengan pengumpulan data dari sumber yang resmi terkait dengan penelitian ini. Langkah terakhir adalah pembuatan sistem yang dilakukan dengan langkah-langkah sebagai berikut : identifikasi permasalahan, perancangan desain aplikasi dan antarmuka, implementasi desain, pengujian sistem dan evaluasi, dan diakhiri dengan pelaporan.

## II. TEXT MINING

*Text mining* merupakan proses pengetahuan yang intensif di mana user berinteraksi dan bekerja dengan sekumpulan dokumen dengan menggunakan beberapa alat analisis [1]. Data teks biasanya berupa sebuah kumpulan dari dokumen tak terstruktur tanpa ada syarat khusus dalam penyusunan dokumennya, sehingga pada text mining diperlukan adanya *preprocessing* dokumen yang nantinya dapat membuat dokumen menjadi lebih terstruktur [2].

### A. Klasifikasi

Menurut Han [3], klasifikasi data dilakukan dalam dua langkah pemrosesan, yaitu tahap pembelajaran (di mana sebuah model klasifikasi dibangun) dan sebuah tahap klasifikasi (di mana model tersebut digunakan untuk memprediksi label kelas dari data yang diberikan). Pada tahap pembelajaran diawali dengan preprocessing dokumen yaitu tokenisasi [2], baru kemudian penghapusan stopword. Setelah itu dilakukan pembobotan TF-IDF setiap token dengan rumus sebagai berikut [4]:

$$w_{ij} = tf_{ij} \, .idf \tag{1}$$

Rumus di atas harus dinormalisasi agar panjang vektornya menjadi 1 dengan cara :

$$w_{ij} = \frac{tf_{ij} \cdot (\log(\frac{N}{n})+1)}{\sqrt{\sum_{k=1}^{t}(tf_{ik})^2 \cdot (\log(\frac{N}{n})+1)^2}} = \frac{tf_{ij}}{\sqrt{\sum_{k=1}^{t}(tf_{ik})^2}} \quad (2)$$

Dimana:

$w_{ij}$ = bobot kata/term $t_j$ terhadap dokumen $d_i$

$tf_{ij}$ = jumlah kemunculan kata/term $t_j$ dalam dokumen $d_i$

idf = nilai pengali dari tf yang ada untuk tiap token, dan akan semakin besar jika suatu token hanya ada di dalam dokumen tertentu saja.

idf = log (N/n) , dimana N adalah jumlah semua dokumen yang ada dalam *database* dan n adalah jumlah dokumen yang mengandung kata/*term* $t_j$

t = jumlah token

Sebelum proses klasifikasi, dilakukan *frequency-based feature selection* untuk memilih n% token-token umum dari keseluruhan token yang ada, token-token inilah yang digunakan dalam proses klasifikasi [5]. Berikut ini adalah algoritma Rocchio untuk klasifikasi teks [5]:

TRAINROCCHIO($\mathbb{C}, \mathbb{D}$)
1 **for each** $c_j \in \mathbb{C}$
2 **do** $D_j \leftarrow \{d : \langle d, c_j \rangle \in \mathbb{D}\}$
3 $\vec{\mu}_j \leftarrow \frac{1}{|D_j|} \sum_{d \in D_j} \vec{v}(d)$
4 **return** $\{\vec{\mu}_1, \ldots, \vec{\mu}_J\}$

APPLY($\{\vec{\mu}(c'), \vec{v}(d)\vec{\mu}_J\}, d$)
1 **return** $\arg\min_j |\vec{\mu}_j - \vec{v}(d)|$

B. Evaluasi Sistem

Untuk mengukur keakuratan sistem digunakan rumus berikut:

$$akurasi\ sistem\ (\%) = \frac{jumlah\ dokumen\ betul}{jumlah\ seluruh\ dokumen\ uji} \cdot 100\% \quad (3)$$

Untuk kelas/kategori dalam jumlah kecil, pengukuran *precision* dan *recall* dianggap lebih optimal [5]. *Precision* dapat diartikan sebagai ketepatan pengukuran, sementara recall dapat diartikan sebagai kelengkapan pengukuran. Istilah lain dari *recall* adalah *sensitivity (true positive rate)*. *Precision* dan *recall* untuk kelas *positive* dapat dirumuskan sebagai berikut [3]:

| | Classified positive | Classified negative |
|---|---|---|
| Actual positive | TP | FN |
| Actual negative | FP | TN |

$$p = \frac{TP}{TP + FP}. \qquad r = \frac{TP}{TP + FN}. \quad (4)$$

Keterangan:

- *True Positives* (TP): merupakan *positive class* (kelas yang ingin dievaluasi) yang terklasifikasikan dengan benar oleh sistem klasifikasi. TP adalah jumlah dari *true positives*.
- *True Negatives* (TN): merupakan *negative class* (kelas selain yang ingin dievaluasi) yang terklasifikasikan dengan benar oleh sistem klasifikasi. TN adalah jumlah dari *true negatives.*
- *False Positives* (FP): merupakan *negative class* yang terklasifikasikan oleh sistem klasifikasi sebagai *positive class*. FP adalah jumlah dari *false positives*.
- *False Negatives* (FN): merupakan *positive class* yang terklasifikasikan oleh sistem klasifikasi sebagai *negative class*. FN adalah jumlah dari *false negatives*.

Selain *precision* dan *recall*, tingkat keakuratan sistem juga dapat dihitung menggunakan *F-measure*, di mana merupakan kombinasi rata-rata harmonic (*weighted harmonic mean*) dari *precision* dan *recall* yang dapat ditulis dengan rumus[5]:

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad (5)$$

untuk bobot β umumnya menggunakan nilai yang seimbang yaitu 1 atau bobot $\alpha = 1/2$. Pada kasus tertentu, bobot β bisa berkisar antara $0 < \beta < 1$ disesuaikan dengan kebutuhan *precision* dan *recall*nya.

## III. HASIL DAN PEMBAHASAN

*1. Feature Selection (FS)* 10%

- Evaluasi Keakuratan Sistem

  Berikut ini adalah tabel hasil pengujian dengan *feature selection* 10%.

Tabel 1. Hasil Pengujian *Feature Selection* 10%

| ID Dokumen | Berkat | Motivator | Iman | Hikmat |
|---|---|---|---|---|
| 61 | T | T | T | T |
| 62 | T | T | T | T |
| 63 | F(Motivator) | T | T | T |
| 64 | T | T | T | F(Berkat) |
| 65 | F(Iman) | T | F(Berkat) | F(Iman) |
| 66 | F(Motivator) | T | T | F(Iman) |
| 67 | F(Iman) | T | T | T |
| 68 | F(Iman) | T | T | T |
| 69 | T | T | T | F(Motivator) |
| 70 | F(Iman) | T | T | F(Berkat) |
| 71 | T | T | T | F(Berkat) |
| 72 | T | T | T | T |
| 73 | F(Motivator) | T | T | T |
| 74 | F(Motivator) | T | T | F(Berkat) |
| 75 | F(Iman) | T | T | T |

Jumlah dokumen benar (T) : 43

Keakuratan sistem:

$$\frac{43}{60} * 100\% = 71{,}67\%$$

- Evaluasi *Precision Recall*

Tabel 2. *Confusion Matrix* dengan FS 10%

| Fakta | Sistem | | | | | |
|---|---|---|---|---|---|---|
| | | Berkat | Motivator | Iman | Hikmat | Total |
| | Berkat | 6 | 4 | 5 | - | 15 |
| | Motivator | - | 15 | - | - | 15 |
| | Iman | 1 | - | 14 | - | 15 |
| | Hikmat | 4 | 1 | 2 | 8 | 15 |
| | Total | 11 | 20 | 21 | 8 | **60** |

Keterangan: *Precision* = P, *Recall* = R, *F-Measure* = F

Kategori Berkat :

P = 6 / (6 + 5) = 0,545

R = 6 / (6 + 9) = 0,4

F = 0,436 / 0,945 = 0,461

Kategori Motivator :

P = 15 / (15 + 5) = 0,75

R = 15 / (15 + 0) = 1

F = 1,5 / 1,75 = 0,857

Kategori Iman :

P = 14 / (14 + 7) = 0,667

R = 14 / (14 + 1) = 0,93

F = 1,24 / 1,597 = 0,776

Kategori Hikmat :

P = 8 / (8 + 0) = 1

R = 8 / (8 + 7) = 0,53

F = 1,06 / 1,53 = 0,693

2. *Feature Selection* 20%

- Evaluasi Keakuratan Sistem

  Berikut ini adalah tabel hasil pengujian dengan *feature selection* 20%.

Tabel 3. Hasil Pengujian *Feature Selection* 20%

| ID Dokumen | Berkat | Motivator | Iman | Hikmat |
|---|---|---|---|---|
| 61 | T | T | T | T |
| 62 | T | T | T | T |
| 63 | F(Motivator) | T | T | T |
| 64 | T | T | T | F(Berkat) |
| 65 | F(Iman) | T | F(Berkat) | F(Iman) |
| 66 | F(Motivator) | T | T | F(Iman) |
| 67 | F(Iman) | T | T | T |
| 68 | F(Iman) | T | T | T |
| 69 | T | T | T | F(Motivator) |
| 70 | F(Iman) | T | T | F(Berkat) |
| 71 | T | T | T | F(Berkat) |
| 72 | T | T | T | T |
| 73 | T | T | T | T |
| 74 | F(Motivator) | T | T | F(Berkat) |
| 75 | F(Iman) | T | T | T |

Jumlah dokumen benar (T) : 44

Keakuratan sistem :

$$\frac{44}{60} * 100\% = 73{,}33\%$$

- Evaluasi *Precision Recall*

Tabel 4. *Confusion Matrix* dengan FS 20%

| Fakta | Sistem | | | | | |
|---|---|---|---|---|---|---|
| | | Berkat | Motivator | Iman | Hikmat | Total |
| | Berkat | 7 | 3 | 5 | - | 15 |
| | Motivator | - | 15 | - | - | 15 |
| | Iman | 1 | - | 14 | - | 15 |
| | Hikmat | 4 | 1 | 2 | 8 | 15 |
| | Total | 12 | 19 | 21 | 8 | **60** |

Keterangan: *Precision* = P, *Recall* = R, *F-Measure* = F

Kategori Berkat :

P = 7 / (7 + 5) = 0,583

R = 7 / (7 + 8) = 0,467

F = 0,545 / 1,05 = 0,519

Kategori Motivator :

P = 15 / (15 + 4) = 0,789

R = 15 / (15 + 0) = 1

F = 1,578 / 1,789 = 0,882

Kategori Iman :

P = 14 / (14 + 7) = 0,667

R = 14 / (14 + 1) = 0,93

F = 1,24 / 1,597 = 0,776

Kategori Hikmat :

P = 8 / (8 + 0) = 1

R = 8 / (8 + 7) = 0,53

F = 1,06 / 1,53 = 0,693

3. *Feature Selection* 30%

- Evaluasi Keakuratan Sistem

  Berikut ini adalah tabel hasil pengujian dengan *feature selection* 30%.

Tabel 5. Hasil Pengujian *Feature Selection* 30%

| ID Dokumen | Berkat | Motivator | Iman | Hikmat |
|---|---|---|---|---|
| 61 | T | T | T | T |
| 62 | T | T | T | T |
| 63 | F(Motivator) | T | T | T |
| 64 | F(Iman) | T | T | F(Berkat) |
| 65 | F(Iman) | T | F(Berkat) | F(Iman) |
| 66 | F(Motivator) | T | T | F(Iman) |
| 67 | F(Iman) | T | T | T |
| 68 | F(Iman) | T | T | T |
| 69 | T | T | T | F(Motivator) |
| 70 | F(Iman) | T | T | F(Berkat) |
| 71 | T | T | T | F(Berkat) |
| 72 | T | T | T | T |
| 73 | T | T | T | T |
| 74 | F(Motivator) | T | T | F(Berkat) |
| 75 | F(Iman) | T | T | T |

Jumlah dokumen benar (T) : 43

Keakuratan sistem :

$$\frac{43}{60} * 100\% = 71{,}67\%$$

- Evaluasi *Precision Recall*

Tabel 6. *Confusion Matrix* dengan FS 30%

| Fakta | Sistem | | | | |
|---|---|---|---|---|---|
| | Berkat | Motivator | Iman | Hikmat | Total |
| Berkat | 6 | 3 | 6 | - | 15 |
| Motivator | - | 15 | - | - | 15 |
| Iman | 1 | - | 14 | - | 15 |
| Hikmat | 4 | 1 | 2 | 8 | 15 |
| Total | 11 | 19 | 22 | 8 | 60 |

Keterangan: *Precision* = P, *Recall* = R, *F-Measure* = F

Kategori Berkat :

P = 6 / (6 + 5) = 0,545

R = 6 / (6 + 9) = 0,4

F = 0,436 / 0,945 = 0,461

Kategori Motivator :

P = 15 / (15 + 4) = 0,789

R = 15 / (15 + 0) = 1

F = 1,578 / 1,789 = 0,882

Kategori Iman :

P = 14 / (14 + 8) = 0,64

R = 14 / (14 + 1) = 0,93

F = 1,19 / 1,57 = 0,758

Kategori Hikmat :

P = 8 / (8 + 0) = 1

R = 8 / (8 + 7) = 0,53

F = 1,06 / 1,53 = 0,693

4. *Feature Selection* 40%

- Evaluasi Keakuratan Sistem

  Berikut ini adalah tabel hasil pengujian dengan *feature selection* 40.

Tabel 7. Hasil Pengujian *Feature Selection* 40%

| ID Dokumen | Berkat | Motivator | Iman | Hikmat |
|---|---|---|---|---|
| 61 | T | T | T | T |
| 62 | T | T | T | T |
| 63 | F(Motivator) | T | T | T |
| 64 | F(Iman) | T | T | F(Berkat) |
| 65 | F(Iman) | T | F(Berkat) | F(Iman) |
| 66 | F(Motivator) | T | T | F(Iman) |
| 67 | F(Iman) | T | T | T |
| 68 | F(Iman) | T | T | T |
| 69 | T | T | T | F(Motivator) |
| 70 | F(Iman) | T | T | F(Berkat) |

| | | | | |
|---|---|---|---|---|
| 71 | T | T | T | F(Berkat) |
| 72 | T | T | T | T |
| 73 | T | T | T | T |
| 74 | F(Motivator) | T | T | F(Berkat) |
| 75 | F(Iman) | T | T | T |

Jumlah dokumen benar (T) : 43

Keakuratan sistem :

$$\frac{43}{60} * 100\% = 71{,}67\%$$

- Evaluasi *Precision Recall*

Tabel 8. *Confusion Matrix* dengan FS 40%

| Fakta | Sistem | | | | | |
|---|---|---|---|---|---|---|
| | | Berkat | Motivator | Iman | Hikmat | Total |
| | Berkat | 6 | 3 | 6 | - | 15 |
| | Motivator | - | 15 | - | - | 15 |
| | Iman | 1 | - | 14 | - | 15 |
| | Hikmat | 4 | 1 | 2 | 8 | 15 |
| | Total | 11 | 19 | 22 | 8 | 60 |

Keterangan: *Precision* = P, *Recall* = R, *F-Measure* = F

Kategori Berkat :

P = 6 / (6 + 5) = 0,545

R = 6 / (6 + 9) = 0,4

F = 0,436 / 0,945 = 0,461

Kategori Motivator :

P = 15 / (15 + 4) = 0,789

R = 15 / (15 + 0) = 1

F = 1,578 / 1,789 = 0,882

Kategori Iman :

P = 14 / (14 + 8) = 0,64

R = 14 / (14 + 1) = 0,93

F = 1,19 / 1,57 = 0,758

Kategori Hikmat :

P = 8 / (8 + 0) = 1

R = 8 / (8 + 7) = 0,53

F = 1,06 / 1,53 = 0,693

5. *Feature Selection* 50%

- Evaluasi Keakuratan Sistem

  Berikut ini adalah tabel hasil pengujian dengan *feature selection* 50%.

Tabel 9. Hasil Pengujian *Feature Selection* 50%

| ID Dokumen | Berkat | Motivator | Iman | Hikmat |
|---|---|---|---|---|
| 61 | T | T | T | T |
| 62 | T | T | T | T |
| 63 | F(Motivator) | T | T | T |
| 64 | F(Iman) | T | T | F(Berkat) |
| 65 | F(Iman) | T | F(Berkat) | F(Iman) |
| 66 | F(Motivator) | T | T | F(Iman) |
| 67 | F(Iman) | T | T | T |
| 68 | F(Iman) | T | T | T |
| 69 | T | T | T | F(Motivator) |
| 70 | F(Iman) | T | T | F(Berkat) |
| 71 | T | T | T | F(Berkat) |
| 72 | T | T | T | T |
| 73 | T | T | T | T |
| 74 | F(Motivator) | T | T | F(Berkat) |
| 75 | F(Iman) | T | T | T |

Jumlah dokumen benar (T) : 43

Keakuratan sistem :

$$\frac{43}{60} * 100\% = 71{,}67\%$$

- Evaluasi *Precision Recall*

Tabel 10. *Confusion Matrix* dengan FS 50%

| Fakta | Sistem | | | | | |
|---|---|---|---|---|---|---|
| | | Berkat | Motivator | Iman | Hikmat | Total |
| | Berkat | 6 | 3 | 6 | - | 15 |
| | Motivator | - | 15 | - | - | 15 |
| | Iman | 1 | - | 14 | - | 15 |
| | Hikmat | 4 | 1 | 2 | 8 | 15 |
| | Total | 11 | 19 | 22 | 8 | 60 |

Keterangan: *Precision* = P, *Recall* = R, *F-Measure* = F

Kategori Berkat :

P = 6 / (6 + 5) = 0,545

R = 6 / (6 + 9) = 0,4

F = 0,436 / 0,945 = 0,461

Kategori Motivator :

P = 15 / (15 + 4) = 0,789

R = 15 / (15 + 0) = 1

F = 1,578 / 1,789 = 0,882

Kategori Iman :

P = 14 / (14 + 8) = 0,64

R = 14 / (14 + 1) = 0,93

F = 1,19 / 1,57 = 0,758

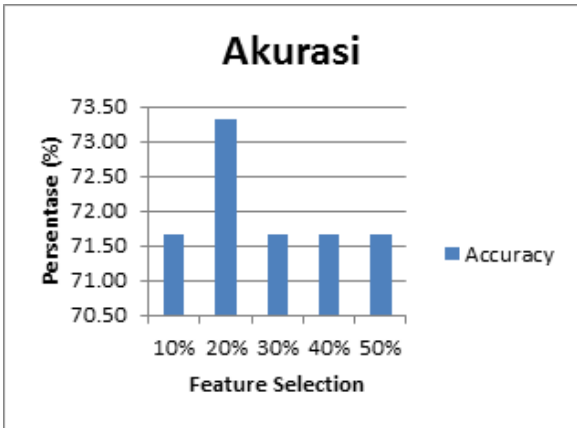
Kategori Hikmat :

P = 8 / (8 + 0) = 1

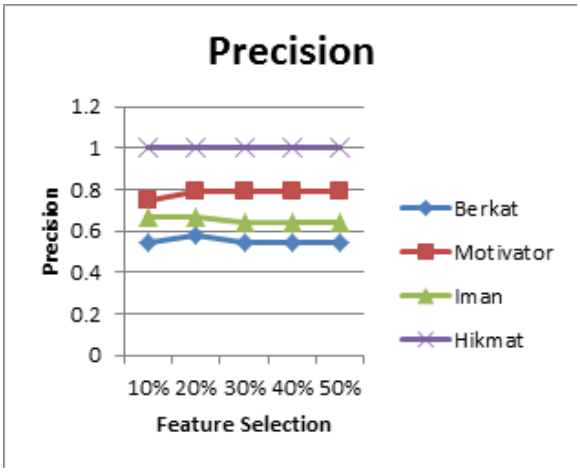
R = 8 / (8 + 7) = 0,53

F = 1,06 / 1,53 = 0,693

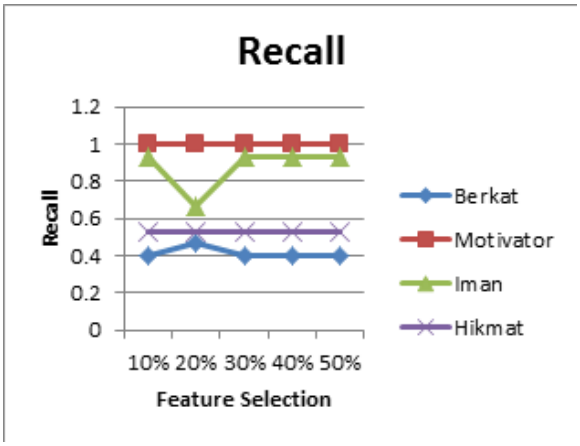
Dari hasil evaluasi 1 sampai dengan 5 dapat disimpulkan dalam grafik 1, 2, dan 3. berikut:




Grafik 1. Hasil Evaluasi Akurasi




Grafik 2. Hasil Evaluasi Precision




Grafik 3. Hasil Evaluasi Recall

6. Evaluasi *Precision Recall* Menurut Sumber Data

Evaluasi ini dibuat berdasarkan hasil pengujian untuk *feature selection* 20%. Hasil pengujian untuk *feature selection* 20% dapat dilihat pada tabel 2. Berikut adalah inisial singkatan untuk setiap sumber data dokumen uji :

Tabel 11. Inisial Sumber Data

| Sumber Data | Inisial |
|---|---|
| Renungan Harian Air Hidup | AH |
| Renungan Harian Bethany (http://www/bethanygraha.org) | BT |
| Renungan Harian Spirit | SP |
| Renungan Harian Online (http://renungan-harian-online.com) | OL |
| Renungan GKPI (http://www.gkpi.or.id/renungan/) | GK |

Berikut ini adalah tabel 12 sumber data dokumen uji. Setiap inisial pada setiap sel mengacu pada tabel 11. Dari ID dokumen 61 sampai dengan ID dokumen 75, terdapat 4 kategori yaitu berkat, motivator, iman, dan hikmat. Setiap kategori memiliki sumber data yang berbeda-beda. Untuk sumber data AH total berjumlah 20, sumber data BT total berjumlah 14, sumber data SP total berjumlah 13, sumber data OL total berjumlah 11, dan sumber data GK total berjumlah 2.

Tabel 12. Sumber Data Dokumen Uji

| ID Dokumen | Berkat | Motivator | Iman | Hikmat |
|---|---|---|---|---|
| 61 | AH | BT | GK | BT |
| 62 | BT | BT | GK | BT |
| 63 | BT | SP | BT | BT |
| 64 | BT | SP | BT | BT |
| 65 | BT | SP | BT | AH |
| 66 | BT | SP | AH | AH |
| 67 | AH | SP | AH | AH |
| 68 | AH | SP | AH | OL |
| 69 | AH | SP | AH | OL |
| 70 | AH | SP | AH | OL |
| 71 | AH | SP | AH | OL |
| 72 | AH | SP | AH | OL |
| 73 | OL | SP | AH | OL |
| 74 | OL | SP | AH | OL |
| 75 | OL | SP | AH | OL |

Berikut ini adalah tabel 13 *confusion matrix* untuk sumber data AH. Total keseluruhan untuk sumber data ini berjumlah 20 dokumen uji. Kategori motivator tidak ada yang diambil dari sumber data AH sehingga tidak dihitung *precision* dan *recall*-nya.

Tabel 13. *Confusion Matrix* Sumber Data AH

| Fakta | Sistem | | | | |
|---|---|---|---|---|---|
| | **Berkat** | **Motivator** | **Iman** | **Hikmat** | **Total** |
| **Berkat** | 4 | - | 3 | - | 7 |
| **Motivator** | - | - | - | - | 0 |
| **Iman** | - | - | 10 | - | 10 |
| **Hikmat** | - | - | 2 | 1 | 3 |
| **Total** | 4 | - | 15 | 1 | **20** |

- Pberkat = 4 / 4 + 0 = 1
- Rberkat = 4 / 4 + 3 = 0,571
- Piman = 10 / 10 + 5 = 0,667
- Riman = 10 / 10 + 0 = 1
- Phikmat = 1 / 1 + 0 = 1
- Rhikmat = 1 / 1 + 2 = 0,333

Berikut ini adalah *confusion matrix* untuk sumber data BT. Total keseluruhan untuk sumber data ini berjumlah 14 dokumen uji.

Tabel 14. *Confusion Matrix* Sumber Data BT

| Fakta | Sistem | | | | |
|---|---|---|---|---|---|
| | **Berkat** | **Motivator** | **Iman** | **Hikmat** | **Total** |
| **Berkat** | 2 | 2 | 1 | - | 5 |
| **Motivator** | - | 2 | - | - | 2 |
| **Iman** | 1 | - | 2 | - | 3 |
| **Hikmat** | 1 | - | - | 3 | 4 |
| **Total** | 4 | 4 | 3 | 3 | **14** |

- Pberkat = 2 / 2 + 2 = 0,5
- Rberkat = 4 / 4 + 3 = 0,571
- Pmotivator = 2 / 2 + 2 = 0,5
- Rmotivator = 2 / 2 + 0 = 1
- Piman = 2 / 2 + 1 = 0,667
- Riman = 2 / 2 + 1 = 0,667
- Phikmat = 3 / 3 + 0 = 1
- Rhikmat = 3 / 3 + 1 = 0,75

Berikut ini adalah confusion matrix untuk sumber data SP. Total keseluruhan untuk sumber data ini berjumlah 13 dokumen uji. Hanya kategori motivator saja yang diambil dari sumber data ini, sehingga precision dan recallnya hanya dihitung untuk kategori motivator.

Tabel 15. *Confusion Matrix* Sumber Data SP

| Fakta | Sistem | | | | |
|---|---|---|---|---|---|
| | **Berkat** | **Motivator** | **Iman** | **Hikmat** | **Total** |
| **Berkat** | - | - | - | - | 0 |
| **Motivator** | - | 13 | - | - | 13 |
| **Iman** | - | - | - | - | 0 |
| **Hikmat** | - | - | - | - | 0 |
| **Total** | - | 13 | - | - | **13** |

- Pmotivator = 13 / 13 + 0 = 1
- Rmotivator = 13 / 13 + 0 = 1

Berikut ini adalah confusion matrix untuk sumber data OL. Total keseluruhan untuk sumber data ini berjumlah 11 dokumen uji. Kategori motivator dan iman tidak ada yang diambil dari sumber data AH sehingga tidak dihitung *precision* dan *recall*-nya.

Tabel 16. *Confusion Matrix* Sumber Data OL

| Fakta | Sistem | | | | |
|---|---|---|---|---|---|
| | **Berkat** | **Motivator** | **Iman** | **Hikmat** | **Total** |
| **Berkat** | 1 | 1 | 1 | - | 3 |
| **Motivator** | - | - | - | - | 0 |
| **Iman** | - | - | - | - | 0 |
| **Hikmat** | 3 | 1 | - | 4 | 8 |
| **Total** | 4 | 2 | 1 | 4 | **11** |

- Pberkat = 1 / 1 + 3 = 0,25
- Rberkat = 1 / 1 + 2 = 0,333
- Phikmat = 4 / 4 + 0 = 1
- Rhikmat = 4 / 4 + 4 = 0,5

Berikut ini adalah *confusion matrix* untuk sumber data GK. Hanya ada 2 renungan yang diambil dari sumber data ini dan kedua renungan tersebut memiliki kategori hikmat, sehingga *precision* dan *recall* untuk sumber data ini hanya dihitung untuk kategori hikmat saja.

Tabel 17. *Confusion Matrix* Sumber Data GK

| Fakta | Sistem | | | | |
|---|---|---|---|---|---|
| | **Berkat** | **Motivator** | **Iman** | **Hikmat** | **Total** |
| **Berkat** | - | - | - | - | 0 |
| **Motivator** | - | - | - | - | 0 |
| **Iman** | - | - | 2 | - | 2 |
| **Hikmat** | - | - | - | - | 0 |
| **Total** | - | - | 2 | - | **2** |

- Piman = 2 / 2 + 0 = 1
- Riman = 2 / 2 + 0 = 1

## IV. KESIMPULAN DAN SARAN

Kesimpulan pada penelitian ini adalah sebagai berikut :

1. Sistem klasifikasi Rocchio memberikan akurasi cukup tinggi untuk *feature selection* 20% yaitu sebesar 73,33%, demikian juga dengan rata-rata precision sebesar 0,76 dan rata-rata recall sebesar 0,73. Dari hasil tersebut dapat diartikan bahwa hasil klasifikasi sistem cukup baik (*fair classification*) [6].
2. Nilai precision tertinggi jatuh pada kategori hikmat dengan nilai precision 1. Sedangkan nilai recall tertinggi jatuh pada kategori motivator dengan nilai recall 1.
3. Peningkatan persentase *feature selection* tidak terlalu mempengaruhi nilai precision dan recall pada setiap kategori.
4. Dari penelitian klasifikasi berdasarkan sumber data, maka sumber data dari Renungan Harian Spirit sangat cocok untuk kategori motivator karena memiliki nilai precision dan recall 1 dari penelitian yang telah dilakukan.

Adapun saran untuk pengembangan penelitian ini adalah sebagai berikut :

1. Diperlukan penggunaan *store procedure* pada bahasa VB.NET untuk mempercepat *preprocessing* data dan mengurangi penggunaan memori.
2. Dapat ditambahkan proses *stemming* dalam bahasa Indonesia untuk lebih meningkatkan akurasi sistem.

## DAFTAR PUSTAKA


[1]. Feldman, R., dan Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge* : Cambridge University Press.

[2]. Weiss, Sholom M., et all. (2005). *Text Mining : Predictive Methods for Analyzing Unstructured Information*. New York : Springer.

[3]. Han, J. & Kamber, M. (2006). *Data Mining : Concepts and Technique 2nd Edition*. San Fransisco : Morgon Kauffman Publishers.

[4]. Intan, Rolly & Defeng, Andrew. (2006). *HARD : Subject Based Search Engine Menggunakan TF-IDF dan Jaccard's Coeffisient*. Diakses pada tanggal 25 Agustus 2012 dari http://puslit.petra.ac.id/files/published/journals/IND/IND060801/IND06080106.pdf

[5]. Manning, Christopher D., et all. (2008). *Introduction to Information Retrieval*. New York : Cambridge University Press.

[6]. Gorunescu, F. (2011). Data Mining Concept Model and Techniques. Berlin: Springer. ISBN 978-3-642-19720-8.