Kategorisasi Teks pada Hadits Sahih Al-Bukhari menggunakan *Random Forest*

Text Categorization on Hadith Sahih Al-Bukhari using Random Forest

Muhammad Fauzan Afianto[1], Prof. Dr. Adiwijaya, S.Si, M.Si. .[2], Said Al Faraby, S.T., M.Sc.[3]
[1,2,3]Prodi S1 Teknik Informatika, Fakultas Informatika, Universitas Telkom
[1]**fauzanafianto@telkomuniversity.ac.id**, [2] **adiwijaya@telkomuniversity.ac.id**,
[3]**saidalfaraby@telkomuniversity.ac.id**

Abstrak

Al-Hadits merupakan kumpulan dari sabda, perbuatan, ketetapan, dan persetujuan Rasulullah *Shallallahu 'Alaihi wa Salam* yang merupakan sumber hukum Islam kedua setelah Al-Qur'an. Sebagai dasar agama Islam, Muslim wajib mempelajari, menghafalkan, dan mengamalkan Al-Quran dan Al-Hadits. Satu dari imam besar sekaligus orang yang meriwayatkan Al-Hadits adalah Imam Bukhari. Beliau menghabiskan waktu selama 16 tahun dalam meriwayatkan Al-Hadits yang jumlahnya sebanyak 2602 Hadits tanpa perulangan dan lebih dari 7000 jika dengan perulangan. Kategorisasi teks otomatis merupakan sebuah kegiatan membangun perangkat lunak yang mampu mengklasifikasikan teks dokumen atau *Hypertext* ke dalam kategori atau kode subjek yang sudah ditentukan sebelumnya. Algoritma yang akan digunakan adalah *Random Forest* yang merupakan perkembangan dari *Decision Tree*. Dalam penelitian tugas akhir ini, penulis memutuskan untuk membuat sebuah sistem yang mampu mengkategorisasikan teks dokumen yang memuat Hadits yang diriwayatkan oleh imam Bukhari berdasarkan kategori anjuran, larangan, dan informasi. Adapun dalam metode evaluasinya, perhitungan *K-Fold Cross Validation* dengan *F1-Score* yang didapat sebesar 90%.

Kata kunci : ***Kategorisasi Teks Dokumen, Hadits Sahih Al-Bukhari, Random Forest, K-fold cross validation, F1-score.***

Abstract

Al-Hadith is a collection of words, deeds, provisions, and approvals of Rasulullah Shallallahu 'Alaihi wa Salam that becomes the second fundamental laws of Islam after Al-Qur'an. As a fundamental of Islam, Muslims must learn, memorize, and practice Al-Qur'an and Al-Hadith. One of venerable Imam which was also the narrator of Al-Hadith is Imam Bukhari. He spent over 16 years to compile about 2602 Hadith (without repetition) and over 7000 Hadith with repetition. Automatic text categorization is a task of developing software tools that able to classify text of hypertext document under pre-defined categories or subject code. The algorithm that would be used is Random Forest, which is a development from Decision Tree. In this final project research, the author decided to make a system that able to categorize text document that contains Hadith that narrated by Imam Bukhari under several categories such as suggestion, prohibition, and information. As for the evaluation method, K-fold cross validation with F1-Score will be used and the result is 90%.

Keywords* : *Automatic Document Categorization, Hadith Sahih Al-Bukhari, Random Forest, K-fold cross validation, F1-score.

1. Pendahuluan

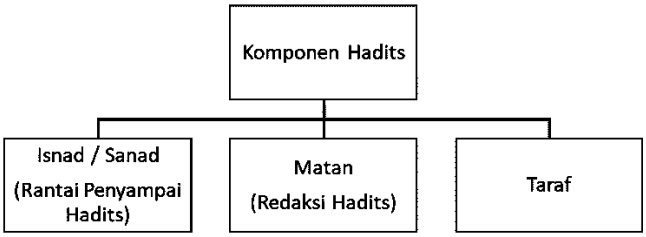
Kategorisasi teks otomatis adalah kemampuan sebuah sistem untuk mengklasifikasikan teks dokumen atau *Hypertext* ke dalam kategori atau kode subjek yang sudah ditentukan sebelumnya secara otomatis. Belakangan ini, minat akan kategorisasi teks otomatis melonjak dikarekanan semakin banyaknya teks dokumen dalam bentuk digital dan kebutuhan akan mengorganisasi teks dokumen yang banyak itu agar lebih mudah digunakan[9]. Dalam kategorisasi teks otomatis, berbagai pendekatan *Machine Learning* telah diimplementasikan selama bertahun-tahun misalnya dengan algoritma *Decision Tree Learning*, *Bayesian Learning*, *Nearest Neighbour*, dan *Artificial Neural Networks*. Diantara itu, *Decision Tree Learning* memiliki kedudukan yang unggul sebagai *Rule Based Classifier*. Namun algoritma ini memiliki kelemahan yaitu seringnya terjadi *Overfitting*.

Dari situ, penulis memutuskan untuk membuat sebuah sistem yang mampu mengkategorisasikan teks dokumen yang memuat hadis sahih Al-Bukhari berdasarkan kategori anjuran, larangan, dan informasi sebagai pendekatan yang banyak dilakukan dengan harapan penelitian ini memudahkan umat Islam dalam mengenal dan mengamalkan agamanya. Algoritma yang digunakan adalah *Random Forest* yang merupakan perkembangan dari algoritma *Decision Tree*, *Random Forest* memiliki mekanisme *Bootstrapping* yang dapat memperbaiki masalah *Overfitting* pada algoritma *Decision Tree*. Adapun dalam evaluasinya mekanisme perhitungan *K-Fold Cross Validation* dengan F1-Score digunakan.

2. Studi Literatur

2.1. Hadits Sahih Al-Bukhari

Hadis dikumpulkan, dikategorisasi, dan diverifikasi oleh beberapa imam besar. Abu Abdillah Muhammad bin Ismail bin Ibrahim bin al-Mughirah bin Bardizbah al-Ju'fi al-Bukhari atau yang lebih dikenal dengan sebutan Imam Bukhari adalah satu dari perawi hadis sekaligus imam besar yang meriwayatkan hadis. Beliau menghabiskan waktu selama 16 tahun dalam meriwayatkan hadis yang jumlahnya sebanyak 2602 hadis tanpa perulangan dan lebih dari 7000 jika dengan perulangan[9]. Mengacu pada[11] hadis terdiri dari komponen-komponen yang digambarkan pada Gambar 1.



Gambar 1 Komponen Hadits

Berikut penjabaranya :

- **Sanad** terletak pada bagian awal hadis, berisi rantai penyampai hadis. Sanad juga menentukan kualitas keaslian atau autentikasi dari sebuah hadis.
- **Matan** terletak setelah sanad, berisi isi dari hadis itu sendiri. Dalam penelitian ini, bagian matan-lah yang digunakan sebagai data teks yang dikategorisasikan.
- **Rawy** atau rawi adalah yang menyampaikan atau menuliskan dalam suatu kitab apa-apa yang pernah didengar dan diterimanya (hadis) dari seseorang (gurunya). Bentuk jamak dari rawy adalah ruwah dan kata kerja dari menyampaikan hadis dinamakan me-rawy (meriwayatkan) hadis [10].

2.2. Term Weighting : TF-IDF

Term Weighting berfungsi untuk menghasilkan nilai bobot kata yang merupakan suatu indikator untuk mengetahui tingkat kepentingan kata dalam dokumen[1]. Salah satu metode perhitungan bobot dalam suatu kata (Term Weighting) adalah *Term Frequency - Inverse Document Frequency* (TF-IDF)[4].

$$Word_{xy} = TermFrequency_{xy} * log \frac{SumDocument}{TermFrequencyDocument_{xy}}$$

1. $Word_{xy}$ = Bobot kata $TermFrequency_y$ terhadap dokumen $SumDocument_x$
2. $TermFrequency_{xy}$ = Jumlah kemunculan kata $TermFrequency_y$ terhadap dokumen $SumDocument_x$
3. $SumDocument$ = Jumlah dokumen pada data
4. $TermFrequencyDocument_{xy}$ = jumlah kemunculan $TermFrequency_{xy}$ dalam dokumen $SumDocument$

2.3. Random Forest

Random Forest adalah perkembangan dari *Decision Tree* yang juga salah satu algoritma *Ensemble Learning*. *Ensemble Learning* sendiri dikenal sebagai sebuah proses *Learning* yang menggabungkan hasil dari beberapa klasifier lain lalu dalam menentukan kelas suatu masukan, yang dilakukan adalah *Voting* dari semua klasifier yang dimuatnya.

Adapun dalam praktiknya, *Random Forest* dapat mengatasi kelemahan dari *Decision Tree* yang berupa *Overfitting*. Ini dikarenakan dalam melatih modelnya, *Random Forest* membangun banyak *Decision Tree* dengan mekanisme *Bootstrap*. Merujuk ke salah satu algoritma dasar *Random Forest* yaitu *Classification and Regression Tree* (CART). Salah satu teknik Impurity yang dapat digunakan untuk membangun CART ini yaitu *Gini Index*[3]. *Gini Index* sendiri merupakan ukuran statistik yang menyatakan persebaran dari suatu data. Ketika membangun model, *Gini Index* mengukur tingkat ketidaksamaan antara setiap atribut terhadap kelas. Semakin besar *Gini Index* yang didapatkan suatu atribut terhadap kelas maka atribut tersebut dianggap penting.

2.4. Confusion Matrix

Pada penelitian tugas akhir ini dilakukan pengukuran performansi menggunakan *Scoring* pada *Confusion Matrix* yaitu *Precision*, *Recall* dan *F1-Score*. Pada dasarnya, *Confusion Matrix* dilihat dari perspektif setiap kelas secara biner. jika kelas yang di kategorisasikan ada 3 maka nilai *Precision* dan *Recall* mengikuti jumlah kelasnya. Berikut adalah penjabaranya:

1. *Precision*, merupakan jumlah dari kategori yang diprediksi benar dan terbukti benar. *Precision* digunakan untuk mengukur seberapa tepatkah prediksi suatu sistem dalam menyatakan prediksinya.

$$Precision = \frac{\sum True\ Positive}{\sum True\ Positive + \sum False\ Positive}$$

2. *Recall* merupakan jumlah dari kategori yang memang benar dan berhasil di prediksi benar. *Recall* digunakan untuk mengukur seberapa banyak kategori memang benar yang berhasil di prediksi benar.

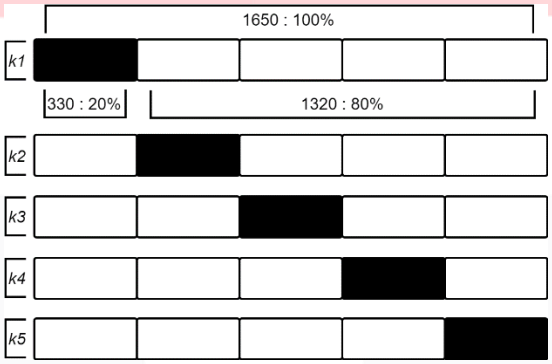
$$Recall = \frac{\sum True\ Positive}{\sum True\ Positive + \sum False\ Negative}$$

3. *F1-Score* atau *Harmonic Mean* merupakan gabungan antara *Precision* dan *Recall*. *F1-Score* digunakan untuk mengukur kombinasi nilai yang telah dihasilkan dari *Precision* dan *Recall* sehingga menjadi satu nilai pengukuran.

$$F1Score = \frac{Pecision * Recall}{Pecision + Recall}$$

2.5. K-Fold Cross Validation

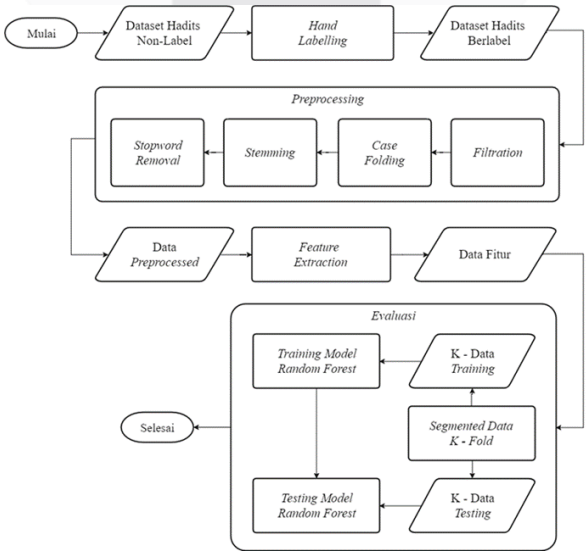
Dikarenakan *Dataset* pada penelitian tugas akhir tidak dipisahkan oleh ahli. Maka *K-Fold Cross Validation* digunakan untuk menguji performansi sistem. Pada metode ini, data akan dipisahkan menjadi *K - Fold* yaitu data-1, data-2, ..., data-k. Penentuan jumlah k dapat diatur secara manual. Misalnya pada Gambar 2.



Gambar 2 Segmentasi K-Fold

3. Perancangan Sistem

Sistem yang dibangun adalah kategorisasi dokumen hadis Al-Bukhari dengan data masukan yang berupa matan hadis. Dikarenakan data yang digunakan belum memiliki label, maka dilakukan proses *Hand Labelling* terlebih dahulu Pada proses Akuisisi Data, dilanjutkan *Feature Extraction* yang memuat *Term Weighting*. Adapun pembentukan model *Training* serta model *Testing* digabung pada tahap *Cross Validation*. Di sana, model klasifikasi dengan algoritma *Random Forest* dibangun dan langsung diuji menggunakan metode *K-Fold*. Luaran yang dihasilkan sudah berbentuk *F1-Score*. Secara garis besar, Gambar 3 akan menunjukkan mekanisme sistem ini.



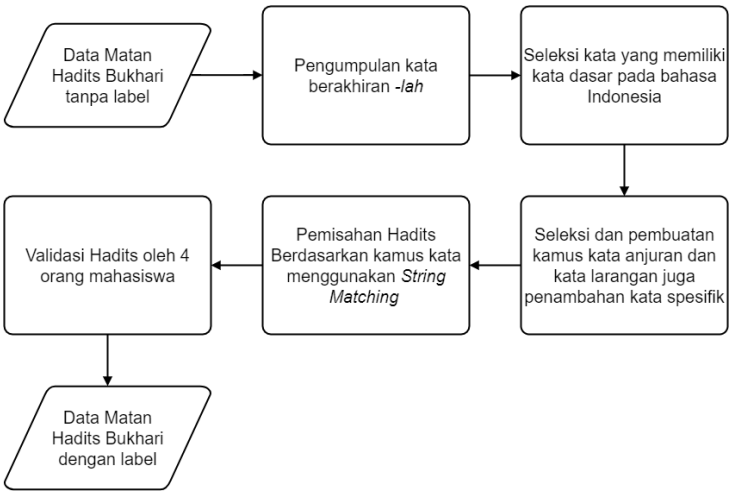
Gambar 3 Flowchart Rancangan Sistem

3.1. Akuisisi Data

Dataset yang digunakan adalah bagian matan pada hadis sahih Al-Bukhari yang diambil dari perangkat lunak desktop "Kitab hadis 9 Imam" dari Lembaga Ilmu Dakwah dan Publikasi Sarana Keagamaan (LIDWA) dengan jumlah 7008 hadis. Dataset hadis ini berbahasa Indonesia dengan atribut.

1. Nomor hadis
2. Nomor Bab
3. Matan hadis
4. Kelas hadis (setelah Hand Labelling)

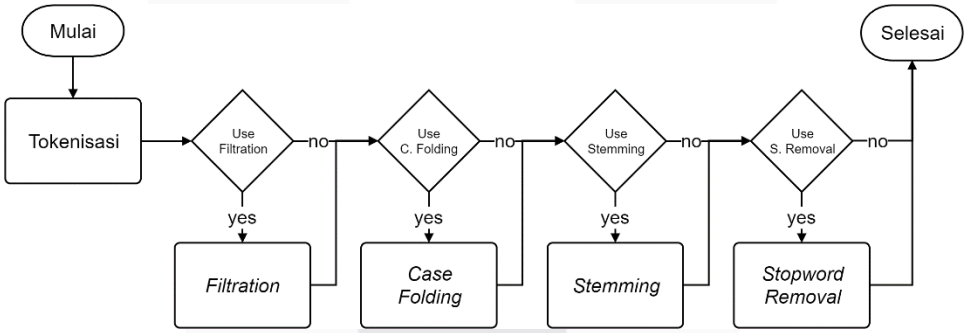
Dikarenakan belum adanya kelas pada hadis, maka *Hand Labelling* dilakukan untuk membagi hadis menjadi 3 kelas yaitu kelas Anjuran, Larangan, dan Informasi. Mekanisme *Hand Labelling* terdiri dari beberapa tahap seperti pada Gambar 4.



Gambar 4 Hand Labelling

3.2. Preprocessing

Preprocessing dilakukan untuk mereduksi variansi dari kata. Menurut beberapa sumber tertulis, variansi pada kata juga termasuk ke dalam *Noise*. *Preprocessing* memuat 4 subproses yang akan dijabarkan pada Gambar 5.

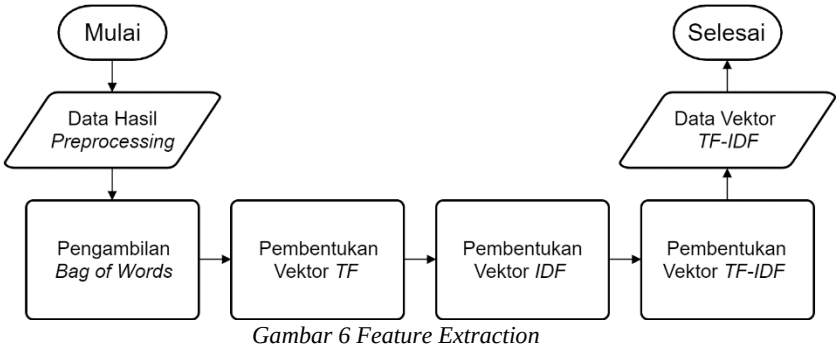


Gambar 5 Preprocessing

Mengacu pada Gambar 5 Setiap subproses dapat dilakukan atau tidak bergantung pada skenario yang akan dijabarkan pada skenario pengujian.

3.3. Feature Extraction

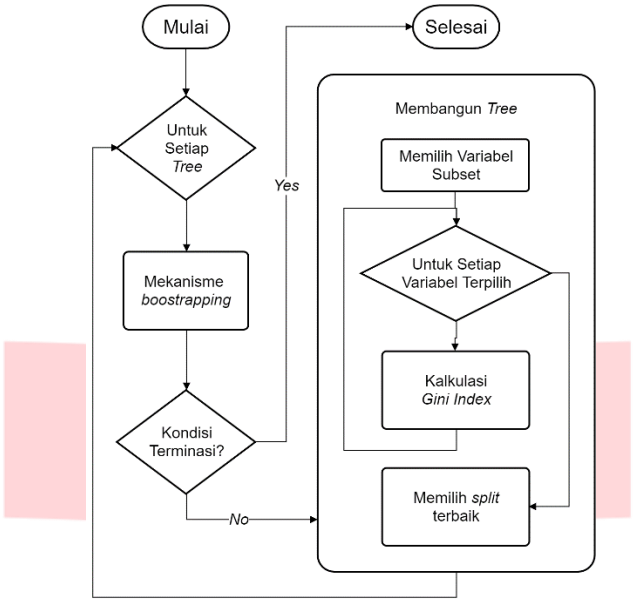
Pada tahap ini, data hasil *Preprocessing* diolah menjadi representasi vektor agar dapat digunakan untuk masukan klasifier. TF-IDF sendiri terbagi atas 3 tahap utama yaitu perhitungan TF, IDF, dan TF-IDF. Digambarkan pada Gambar 6.



Gambar 6 Feature Extraction

3.4. Classification

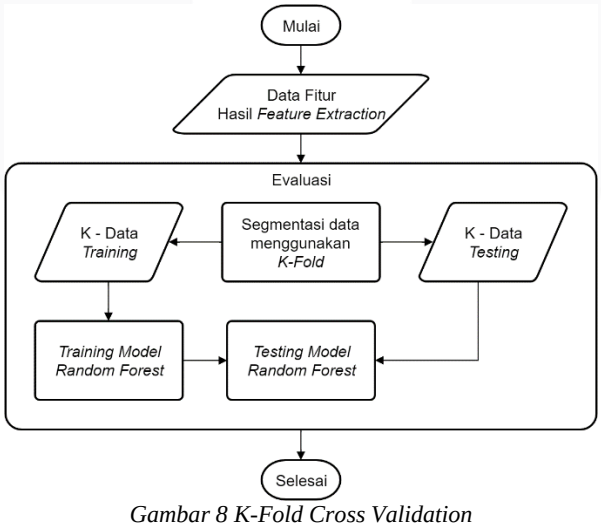
Sederhananya, ini adalah kumpulan dari *Decision Tree* yang dibuat dengan menggunakan mekanisme *Bootstrapping*. Setelah itu maka data masukan akan diklasifikasikan oleh semua *Tree* yang sudah dibuat tadi dan hasil kelas akhirnya merupakan *Voting* dari semua *Decision Tree* itu. Seperti pada Gambar 7.



Gambar 7 Random Forest

3.5. Evaluation

K-Fold Cross Validation sendiri sejatinya membagi data berdasarkan nilai K yang sebelumnya telah ditentukan. Adapun dalam menentukan jumlah K, proporsi data *Training* dan *Testing* yang akan dihasilkan haruslah diperhatikan. Seperti yang digambarkan pada Gambar 2 Maka dapat disimpulkan bahwa nilai K yang berjumlah 5 sesuai dengan proporsi data *Training* : *Testing* (80 : 20). Mekanisme dari *Splitting* sendiri digambarkan pada Gambar 8.



Gambar 8 K-Fold Cross Validation

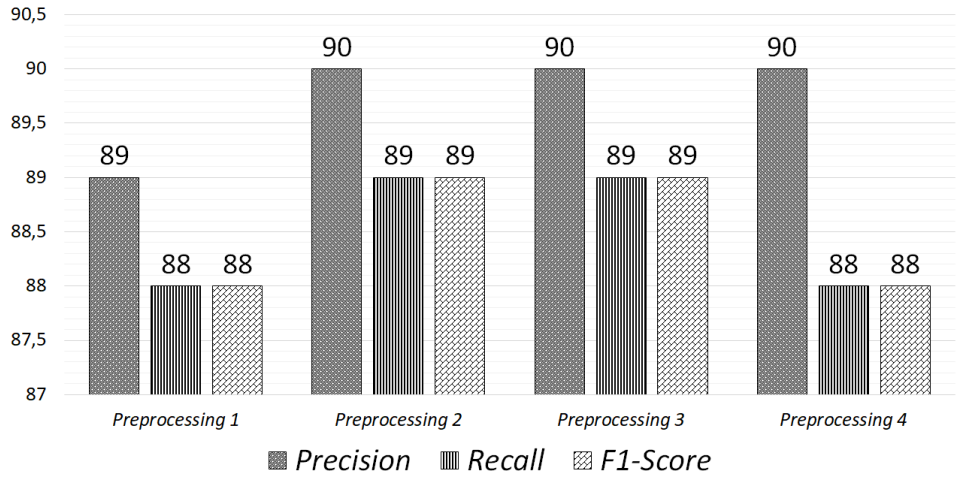
4. Skenario Pengujian

Berikut adalah beberapa skenario pengujian yang dilakukan. Adapun detail dari skenario ini memuat observasi parameter dengan rentang tertentu terhadap performansi sistem, pada setiap tahap observasinya parameter yang dinilai menghasilkan performansi terbaik pada sistem digunakan untuk tahap selanjutnya. Berikut penjabaranya:

1. Mengetahui pengaruh *Preprocessing* terhadap performansi sistem untuk mendapatkan model terbaik.
2. Mengetahui pengaruh jenis *Feature Extraction* terhadap performansi sistem untuk mendapatkan model terbaik.
3. Mengetahui pengaruh jumlah *Tree* yang dibuat dengan mekanisme Bootsrapping pada klasifier *Random Forest* terhadap performansi sistem untuk mendapatkan model terbaik.
4. Mengetahui bahwa *Random Forest* lebih baik dari *Decision Tree* dalam mengkategorisasikan dokumen hadis sahih Al-Bukhari dengan kategori anjuran, larangan, dan informasi.

4.1. Skenario 1

Skenario pengujian I bertujuan untuk mengamati pengaruh perubahan parameter pada tahap *Preprocessing* terhadap performansi sistem. Skenario dilakukan dengan membuat beberapa kombinasi dari subproses pada tahap *Preprocessing*.

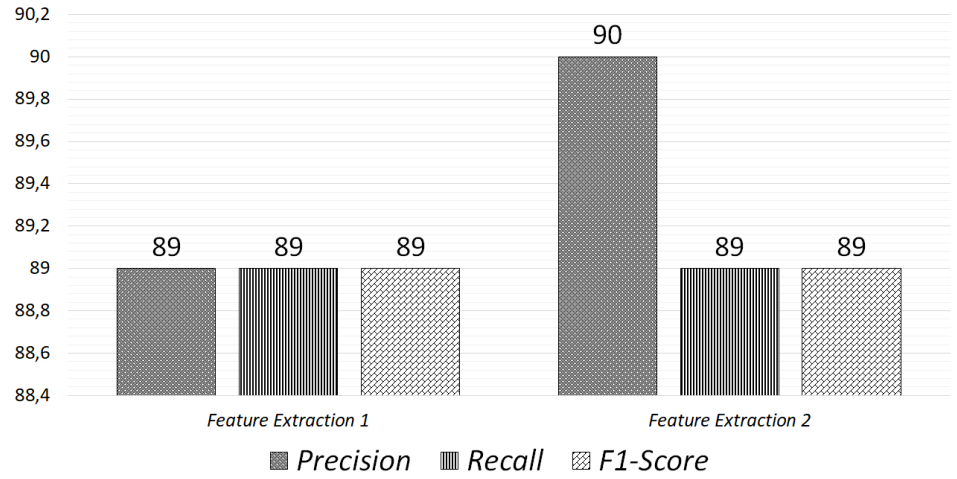


Gambar 9 Hasil Skenario Preprocessing

Berdasarkan Gambar 9 performansi terbaik dihasilkan pada skenario *Preprocessing* 3 dan 2 yang keduanya hampir melibatkan semua mekanisme *Preprocessing* yaitu *Filtration*, *Case Folding*, *Stemming*, kecuali pada skenario *Preprocessing* 2 yang tidak melibatkan *Stopword Removal*. Dapat disimpulkan juga bahwa *Stemming* memberikan peningkatan sebesar 1% pada nilai *F-1 Score* terhadap *Precision* dan *Recall*, ini ditunjukkan pada skenario *Preprocessing* 2 dan 3 yang menggunakan mekanisme *Stemming* sedangkan skenario *Preprocessing* 1 dan 4 yang tidak menggunakan *Stemming*. Mekanisme *Stopword Removal* sama sekali tidak memberikan peningkatan pada performansi ini ditunjukan dengan samanya nilai *F-1 Score* pada skenario *Preprocessing* 1 dan 4 yaitu 88% juga skenario *Preprocessing* 2 dan 3 yaitu 89%.

4.2. Skenario 2

Skenario pengujian II bertujuan untung mengamati pengaruh perubahan parameter pada tahap *Feature Extraction* terhadap performansi sistem. metode yang dapat digunakan dalam membobotkan *Term* ada dua yaitu hanya menggunakan *Term Frequency* dan satu lagi memperhitungkan signifikansi kata terhadap dokumen dengan *Term Frequency-Inverse Document Frequency*. Seperti pada Gambar 10.

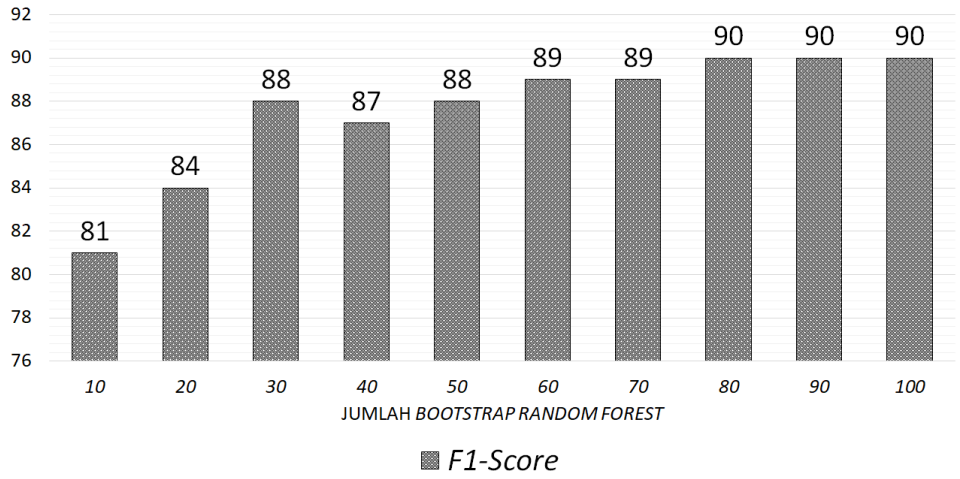


Gambar 10 Hasil Skenario Feature Extraction

Berdasarkan Gambar 10 maka dapat dilihat performansi *F1-Score* yaang dimiliki sama yaitu 89%. Kendati demikian, skenario *Feature Extraction 2* yang menggunakan TF-IDF memiliki nilai *Precision* yang lebih tinggi yaitu 90% berbeda 1% dari skenario *Feature Extraction* 1.

4.3. Skenario 3

Skenario pengujian III bertujuan untuk mengamati pengaruh perubahan parameter yang ada pada klasifier terhadap performansi sistem. Klasifier yang digunakan adalah *Random Forest* di mana parameter yang akan diobservasi adalah jumlah *Tree* yang dibuat melalui mekanisme *Bootstrapping*.

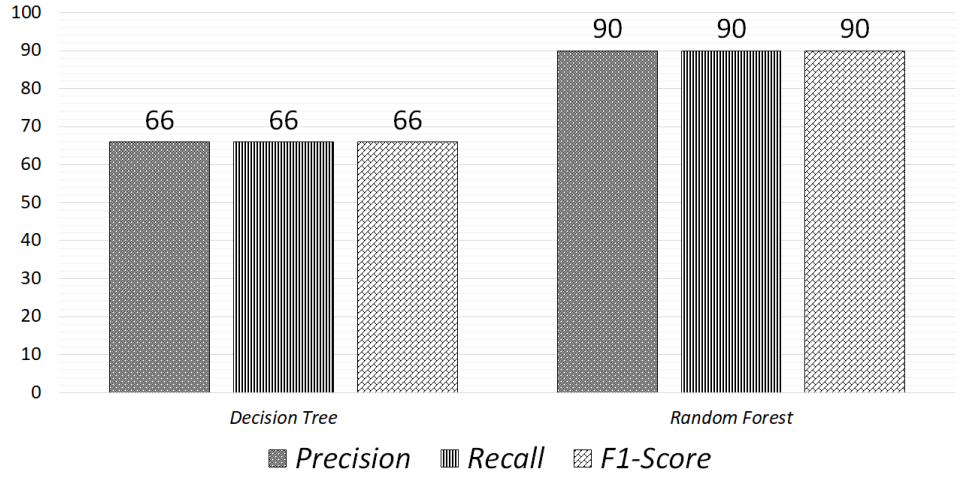


Gambar 11 Hasil Skenario Bootsrapping

Berdasar pada Gambar 11 maka dapat disimpulkan bahwa jumlah *Tree* yang dinilai cukup optimal pada *Random Forest* dimulai dari jumlah 80 Tree ke atas. Dengan jumlah *F1-Score* yang stagnan di 90%.

4.4. Skenario 4

Skenario pengujian IV ini bertujuan untuk menjawab seberapa baik *Decision Tree* dengan parameter yang dinilai optimal dalam melakukan kategorisasi dokumen, disamping itu diuji juga seberapa terbuktikah *Random Forest* dalam mengatasi masalah *Overfitting* yang dimiliki oleh *Decision Tree*.



Gambar 12 Hasil Skenario Perbandingan Random Forest dengan Decision Tree

Berdasarkan Gambar 12 maka dapat disimpulkan bahwa *Random Forest* terbukti memperbaiki kesalahan yang dimiliki oleh *Decision Tree* dalam melakukan kategorisasi dokumen hadis sahih Al-Bukhari. Ini dibuktikan dengan nilai *F-1 Score* yang didapat yaitu 90% sedangkan untuk *Decision Tree* hanya 66%.

5. Kesimpulan

Berdasar dari hasil pengujian dan analisis dengan setiap sekanrio yang telah dirumuskan, maka kesimpulan pada penelitian tugas akhir ini sebagai berikut:

1. Kategorisasi dokumen hadis sahih Al-Bukhari dengan data sampling 1650 Hadits dengan persebaran data 550 *Record* untuk setiap kelasnya menggunakan *Random Forest* dengan mekanisme Preprocessing berupa *Stemming*, *Case Folding*, dan *Filtration*, dengan *Feature Extraction* berupa *TF-IDF* dan jumlah *Bootsrapping Tree* 100 dan mekanisme evaluasi *K-Fold Cross Validation* dengan jumlah K sebanyak 5. Menghasilkan nilai *F1-Score* sebesar 90%.
2. Perubahan *Feature Extraction* baik itu hanya TF maupun TF-IDF tidak memberikan pengaruh yang signifikan pada performansi sistem dengan *F1-Score*, yaitu hanya peningkatan Precision sebesar 1% pada TF-IDF dengan *F1-Score* dari 89% menjadi 90%.
3. Mekanisme *Preprocessing* berperan terutama untuk *Stemming*, *Case Folding*, dan *Filtration* yang menghasilkan *F1-Score* terbesar 89% lainhalnya dengan *Stopword Removal* sama sekali tidak mempengaruhi performansi sistem dengan *F1-Score*.
4. *Random Forest* terbukti dapat mengatasi masalah pada *Decision Tree* yaitu *Overfitting* dengan perbedaan nilai *F1-Score* yang cenderung signifikan dibandingkan skenario yang lain. Yaitu 66% untuk *Decision Tree* dan 90% untuk *Random Forest*.

6. Saran

Berdasar dari hasil kesimpulan penelitian ini, maka saran yang dapat diberikan dari pada penelitian tugas akhir ini sebagai berikut:

1. Akan lebih baik jika kelas dilabeli oleh ahli, baik dari segi keilmuan maupun dari segi persebaran data.
2. *Feature Selection* seperti *Chi Square* dapat diimplementasikan.
3. Dataset yang digunakan diganti dengan yang lebih bersih dari *typo* dan ketidak konsistenan istilah.

Daftar Pustaka

[1] Adiwijaya. Aplikasi Matriks dan Ruang Vektor. Graha Ilmu, 2014.

[2] Adiwijaya. Matematika Diskrit dan Aplikasinya. Alfabeta, 2016.

[3] Arifin, A. H. R. Z., Mubarok, M. S., and Adiwijaya, A. Learning struktur bayesian networks menggunakan novel modified binary differential evolution pada klasifikasi data. In Indonesia Symposium on Computing (IndoSC) 2016 (2016).

[4] Aziz, R. A., Mubarok, M. S., and Adiwijaya, A. Klasifikasi topik pada lirik lagu dengan metode multinomial naive bayes. In Indonesia Symposium on Computing (IndoSC) 2016 (2016).

[5] Harrag, F., El-Qawasmah, E., and Al-Salman, A. M. S. Stemming as a feature reduction technique for arabic text categorization. In Programming and Systems (ISPS), 2011 10th International Symposium on (2011), IEEE, pp. 128–133.

[6] Harrag, F., El-Qawasmeh, E., and Pichappan, P. Improving arabic text categorization using decision trees. In Networked Digital Technologies, 2009. NDT'09. First International Conference on (2009), IEEE, pp. 110–115.

[7] Mubarok, M. S., Adiwijaya, and Aldhi, M. D. Aspect-based sentiment analysis to review products using na¨ıve bayes. In AIP Conference Proceedings (2017), vol. 1867, AIP Publishing, p. 020060.

[8] Najeeb, M. M. Towards innovative system for hadith isnad processing. Int J Comput Trends Technol 18, 6 (2014), 257–259.

[9] Naji Al-Kabi, M., Kanaan, G., Al-Shalabi, R., Al-Sinjilawi, S. I., and Al-Mustafa, R. S. Al-hadith text classifier. Journal of Applied Sciences 5 (2005), 584–587.

[10] Rahman, D. F. Ikhrisar Musthalahu'l-Hadits. PT Al-Ma'arif, 1985.

[11] Saloot, M. A., Idris, N., Mahmud, R., Ja'afar, S., Thorleuchter, D., and Gani, A. Hadith data mining and classification: a comparative analysis. Artificial Intelligence Review 46, 1 (2016), 113–128.