Pengelompokan pada Hadits Menggunakan *Naive Bayes Classifier*

Tugas Akhir

diajukan untuk memenuhi salah satu syarat

memperoleh gelar sarjana

dari Program Studi S1 Ilmu Komputasi

Fakultas Informatika

Universitas Telkom

1302144175

Sukmawan Pradika Janusange Santoso

Program Studi Sarjana S1 Ilmu Komputasi

Fakultas Informatika

Universitas Telkom

Bandung

2019

LEMBAR PENGESAHAN

Pengelompokan pada Hadits Menggunakan
Naive Bayes Classifier

Grouping in Hadiths Using the
Naive Bayes Classifier

1302144175
Sukmawan Pradika Janusange Santoso

Tugas akhir ini telah diterima dan disahkan untuk memenuhi sebagian syarat memperoleh gelar pada Program Studi Sarjana S1 Ilmu Komputasi
Fakultas Informatika
Universitas Telkom

Bandung, 26 Maret 2019
Menyetujui

Pembimbing I,

Dr. Kemas Muslim Lhaksmana, S.T., M.ISD
NIP: 13820075

Ketua Program Studi
Sarjana S1 Ilmu Komputasi,

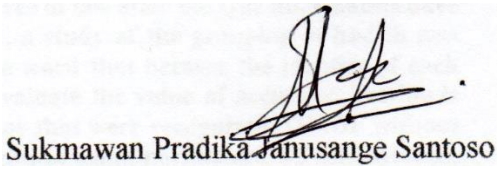
Dr. Deni Saepudin, S.Si., M.Si
NIP: 99750013

LEMBAR PERNYATAAN

Dengan ini saya, Sukmawan Pradika Janusange Santoso, menyatakan sesungguhnya bahwa Tugas Akhir saya dengan judul Pengelompokan pada Hadits Menggunakan *Naive Bayes Classifier* beserta dengan seluruh isinya adalah merupakan hasil karya sendiri, dan saya tidak melakukan penjiplakan yang tidak sesuai dengan etika keilmuan yang belaku dalam masyarakat keilmuan. Saya siap menanggung resiko/sanksi yang diberikan jika di kemudian hari ditemukan pelanggaran terhadap etika keilmuan dalam buku TA atau jika ada klaim dari pihak lain terhadap keaslian karya,

Bandung, 26 Maret 2019

Yang Menyatakan



Sukmawan Pradika Janusange Santoso

Pengelompokan pada Hadits Menggunakan *Naive Bayes Classifier*

Sukmawan Pradika Janusange Santoso[1], Kemas Muslim Lhaksmana[2]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[4]Divisi Digital Service PT Telekomunikasi Indonesia
[1]sukmawan@students.telkomuniversity.ac.id, [2] kemasmuslim@telkomuniversity.ac.id

Abstrak

Hadits merupakan landasan syariat islam dan dijadikan sumber hukum kedua setelah al-Qur'an. Hadits memiliki pokok pembahasan yang bermacam-macam pada setiap bab. Pada tugas akhir ini dilakukan penelitian tentang pengelompokan pada hadits dengan menggunakan TF-IDF untuk menghitung bobot yang diperoleh pada kata yang menjadi identitas pada setiap kategori dan *naïve bayes classifier* untuk memprediksi dataset serta mengevaluasi nilai akurasi. Pengujian dilakukan sebanyak dua kali pengujian dengan skenario yang berbeda. Dari dua skenario diketahui akurasi TF-IDF tanpa normalisasi sebesar 87,97% serta akurasi paling tinggi yaitu pada skenario dengan menggunakan TF-IDF yang dinormalisasi dengan akurasi sebesar 89,29%.

Kata kunci : naive Bayes, klasifikasi

Abstract

Hadith is the basis of Islamic Shari'a and is used as the second source of law after the Qur'an. Hadiths have various subject matter in each chapter. In this final assignment, a study of the grouping of hadith was carried out using TF-IDF to calculate the weight obtained in the word that became the identity of each category and the naïve bayes classifier to predict datasets and evaluate the value of accuracy. Testing is done twice testing with different scenarios. From the two scenarios that were recognized TF-IDF without normalization were 87.97% and also the most efficient in the scenario using normalized TF-IDF with an accuracy of 89.29%.

Keywords: naive Bayes, classification

1. Pendahuluans

Latar Belakang

Hadits adalah merupakan landasan syariat islam dan dijadikan sumber hukum islam yang kedua setelah al-Qur'an [11]. Istilah hadits pada dasarnya berasal dari bahasa Arab yaitu dari kata "Al-hadits" yang artinya adalah perkataan, percakapan atau pun berbicara. Dalam terminologi agama Islam sendiri, dijelaskan bahwa hadits merupakan setiap tulisan yang melaporkan atau pun mencatat seluruh perkataan, perbuatan dan tingkah laku Nabi Muhammad SAW [3][11].

Hadits sejauh ini memiliki bermacam-macam jenis diantaranya, yaitu hadits berdasarkan keutuhan rantai sanad, jumlah penutur, dan tingkat keaslian hadits [3]. Para ulama hadis membagi hadis berdasarkan kualitasnya dalam tiga kategori, yaitu hadits shahih, hadits hasan, hadits dhaif. Pada penelitian ini, hadits yang digunakan adalah berdasarkan tingkat keaslian dengan kategori hadits shahih karena kualitasnya yang lebih tinggi [1].

Pada tugas akhir ini dilakukan penelitian untuk mengelompokkan pokok bahasan dari sekumpulan hadits shahih dengan kategori yang paling umum agar bisa digunakan oleh khalayak yang membutuhkan. Pada jurnal Xhemali, Daniela, Chris J. Hinde, and Roger G. Stone. "Naive Bayes vs. decision trees vs. neural networks in the classification of training web pages." (2009), mengatakan bahwa "Naïve Bayes Classifier memiliki tingkat akurasi yg lebih baik dibanding model classifier lainnya" [5]. Oleh karena itu penulis menggunakan metode naive bayes classifier dengan menggunakan metode pembobotan kata yang paling umum melalui TF-IDF untuk melakukan klasifikasi pada data hadits.

Topik dan Batasannya

Berdasarkan latar belakang, maka rumusan masalah yang akan dibahas dalam penelitian tugas akhir ini adalah sebagai berikut,

1. Bagaimana cara pengelompokan pada hadits menggunakan *naive bayes classifier* ?
2. Bagaimana kinerja sistem berdasarkan akurasi ?
3. Bagaimana hasil performansi dalam mengidentifikasi hadits menggunakan *naive bayes classifier* ?

Adapun batasan masalah di dalam tugas akhir ini, yaitu pada percobaan ini menggunakan data terjemahan bahasa Indonesia yang didapatkan dari website carihadis.com dengan mengambil hadits dari Shahih Bukhari dan pelabelan didapatkan dari website sunnah.com, serta file disusun dengan sedemikian rupa dan menggunakan format *.CSV agar mempermudah pada proses percobaan.

Tujuan

Tujuan dari penelitian pada tugas akhir ini berdasarkan rumusan masalah adalah untuk membangun sistem otomatis yang dapat mengklasifikasikan data teks pada hadits yang sudah diterjemahkan melalui pembelajaran pada sistem dan memudahkan orang-orang dalam mencari hadits sesuai dengan kebutuhannya menggunakan *naive bayes classifier* serta untuk mengetahui kinerja sistem berdasarkan akurasi dan hasil performansi dalam mengidentifikasi hadits yang sudah diterjemakan.

2. Studi Terkait

2.1 Hadits

Hadits pada dasarnya berasal dari bahasa Arab yaitu dari kata "Al-hadits" yang artinya adalah perkataan, percakapan atau pun berbicara. Jika diartikan dari kata dasarnya, maka pengertian hadits adalah setiap tulisan yang berasal dari perkataan atau pun percakapan Rasulullah Muhammad SAW. Dalam terminologi agama Islam sendiri, dijelaskan bahwa hadits merupakan setiap tulisan yang melaporkan atau pun mencatat seluruh perkataan, perbuatan dan tingkah laku Nabi Muhammad SAW. Hadis dijadikan sumber hukum Islam selain al-Qur'an, dalam hal ini kedudukan hadis merupakan sumber hukum kedua setelah al-Qur'an [3][11]. Hadits sejauh ini memiliki bermacam-macam jenis diantaranya, yaitu hadits beradasarkan keutuhan rantai sanad, jumlah penutur, dan tingkat keaslian hadits [3].

2.2 Preprocessing

Preprocessing digunakan untuk mempermudah proses klasifikasi. Tahapan dalam preprocessing adalah sebagai berikut:

1. Remove Punctuation digunakan untuk menghilangkan karakter (tanda baca) yang tidak diperlukan.
2. *Case folding* yang berfungsi untuk membuat teks menjadi huruf kecil.
3. *Tokenization* adalah sebuah proses pemisahan kalimat menjadi kata-kata atau frase.
4. *Stop word removal* digunakan untuk pengecekan apakah termasuk di dalam daftar stoplist seperti kata "karena, yang, dan, atau". Jika kata-kata tersebut masuk kedalam stoplist, maka kata-kata tersebut dihilangkan karena dianggap tidak berkaitan dalam proses klasifikasi.

2.3 Term Frequency – Inverse Document Frequency (TF-IDF)

TF-IDF digunakan untuk menghitung nilai pada setiap kata dalam dokumen melalui *inverse proportion* dari frekuensi kata dalam dokumen tertentu dengan persentase kata yang muncul[4][8]. Untuk mencari TF-IDF, yang pertama dilakukan adalah mencari *term frequency* telebih dahulu dengan rumus:

$$TF(t) = \frac{Count(t)}{\sum_{i=0}^{n} Count(t_i)}$$

Keterangan:

$Count(t)$: Jumlah kemunculan kata t dalam dokumen.

$\sum_{i=0}^{n} Count(t_i)$: Jumlah kemunculan semua kata t dalam dokumen.

Lalu, menentukan *inverse document frequency* dengan menggunakan rumus:

$$IDF(t) = \log\left(\frac{Count(d)}{Count(t,d)}\right)$$

Keterangan:

$Count(d)$: Jumlah dokumen.

$Count(t,d)$: Jumlah dokumen yang memiliki kata t.

Selanjutnya menentukan bobot dengan cara penggabungan dari perhitungan TF dan IDF [4]. Untuk mencari TF-IDF menggunakan rumus:

$$TF\_IDF = TF \cdot IDF$$

2.4 Naive Bayes Classifier

Naïve Bayes Classifier merupakan salah satu metoda machine learning yang memanfaatkan perhitungan probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes[12]. Naive Bayes adalah algoritma pembelajaran yang sering digunakan untuk mengatasi masalah klasifikasi teks. Algoritma naive bayes merupakan sebuah metode klasifiksi yang digunakan untuk memprediksi peluang berdasarkan pengalaman sebelumnya dengan menggunakan metode probabilitas dan statistik[7][12]. Naive Bayes Classifier bekerja sangat baik dibanding dengan model classifier lainnya karena sangat efisien dan mudah diimplementasikan[7]. Hal ini dibuktikan pada jurnal Xhemali, Daniela, Chris J. Hinde, and Roger G. Stone. "Naive Bayes vs. decision trees vs. neural networks in the classification of training web pages." (2009), mengatakan bahwa "Naïve Bayes Classifier memiliki tingkat akurasi yg lebih baik dibanding model classifier lainnya" [5]. Oleh sebab itu, pada tugas akhir ini penulis menggunakan metode naive bayes classifier sebagai metode klasifikasi[6]. Berikut adalah persamaan teorema naive bayes :

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

Keterangan :

X	: Data tes dengan kelas yang belum diidentifikasi
C	: Kategori
P(C|X)	: Probabilitas C berdasarkan kondisi X (posteriori probability)
P(X|C)	: Probabilitas X berdasarkan kondisi C
P(C)	: Probabilitas hipotesis C (prior probability)
P(X)	: Probabilitas X

Ada banyak pekerjaan pada Naive Bayes dan klasifikasi teks. Lewis memberikan ulasan tentang penggunaan Naif Bayes dalam pencarian informasi [Lewis, 1998]. Tidak seperti klasifikasi teks, praktisi pencarian informasi biasanya menganggap independensi antara fitur dan mengabaikan frekuensi kata dan informasi panjang dokumen[9]. *Multinomial Naive Bayes* (MNB) mengimplementasikan algoritma naive bayes untuk data yang terdistribusi secara *multinomially*, dan merupakan salah satu dari dua varian naive bayes yang digunakan dalam klasifikasi teks [5]. Dalam pengklasifikasi MNB setiap dokumen dipandang sebagai kumpulan kata dan urutan kata-kata dianggap tidak relevan[2]. Distribusi ditentukan oleh vektor $\theta_C = (\theta_{C1}, \dots, \theta_{Cn})$ untuk setiap kelas C di mana n adalah jumlah data (dalam klasifikasi teks, ukuran kosakata) dan θ_{Ci} adalah probabilitas P(X|C) dari data i yang muncul pada sampel kelas C.

Parameter θ_C diperkirakan oleh versi smoothed dari kemungkinan maksimum, misal. Penghitungan frekuensi relatif:

$$\theta_{Ci} = \frac{N_{Ci} + \alpha}{N_C + \alpha n}$$

Di mana $N_{Ci} = \sum_{X \epsilon T} X_i$ adalah berapa kali fitur i muncul dalam sampel kelas C pada set pelatihan T, dan $N_C = \sum_{i=1}^{n} N_{Ci}$ adalah jumlah total semua fitur untuk kelas C.

Dengan *smoothing priors* $\alpha \geq 0$ memperhitungkan fitur yang tidak ada dalam sampel pembelajaran dan mencegah probabilitas nol dalam perhitungan selanjutnya. Pengaturan $\alpha = 1$ disebut *laplace smoothing*, sementara $\alpha < 1$ disebut *lidstone smoothing* [10].

3. Sistem yang Dibangun

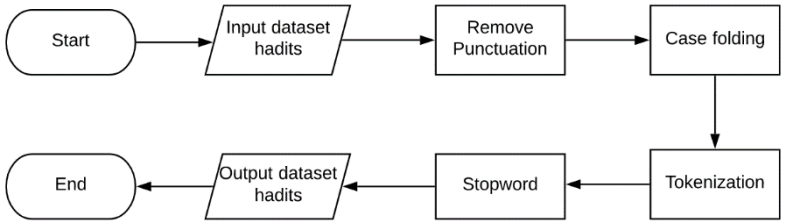
3.1 Pelabelan

Pelabelan yaitu pengambilan data dari website secara manual yang selanjutnya data diberikan label dan dikategorikan sesuai dengan isinya. Pelabelan yang digunakan berasal dari sunnah.com yaitu haji, kiamat, menikah, puasa, shalat, dan zakat. Data berupa hadits shahih Al-Bukhari [6]. Data yang diambil berasal dari terjemahan bahasa Indonesia dengan sumber terpercaya carihadis.com.

3.2 Preprocessing

Pada penelitian ini, preprocessing yang digunakan yaitu berupa remove punctuation, case folding, tokenizing, dan stopwords.

Gambar 1. Diagram preprocessing



Atribut yang digunakan pada data yang diperoleh sebagai penunjang sistem kredibilitas dikembangkan berdasakan penjelasan atribut pada tabel berikut:

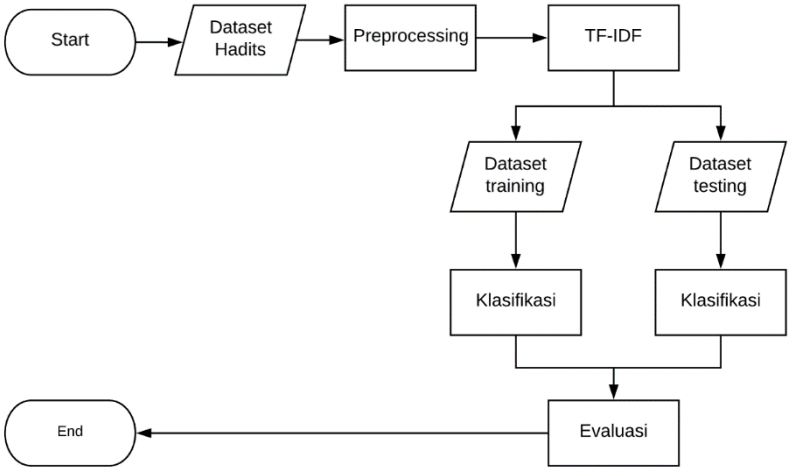
Tabel 2. Atribut pada data yang diperoleh

Atribut	Penjelasan Atribut
Asal Hadits	Asal dari hadits tersebut (hadits shahih siapa)
No. Hadits	Nomor dari hadits sesuai dari sumbernya
Hadits	Isi dari hadits tersebut
Kategori	Pembeda antara satu hadits dengan yang lainnya berdasarkan isinya (terdapat enam kategori)

3.3 Deskripsi Keseluruhan Sistem

Pada tugas akhir ini, yang pertama dilakukan adalah preprocessing data terlebih dahulu. Selanjutnya ketahap TF-IDF untuk mendapatkan bobot. Lalu data dibagi menjadi dua yaitu data training dan data testing. Data yang digunakan berupa data dari carihadis.com. Kemudian data training lanjut ketahap klasifikasi menggunakan algoritma *naive bayes classifier* dengan enam kelas yang sudah ditentukan. Berikutnya data masuk ketahap evaluasi untuk mendapatkan akurasi. Untuk skema diagram sistem dapat dilihat pada gambar dibawah:

Gambar 2. Diagram sistem



4. Evaluasi

4.1 Hasil Pengujian

Pada penelitian ini pengujian dilakukan sebanyak dua kali, pembagian data dilakukan secara random untuk mendapatkan data training dan data testing. Sebelum pembagian data dilakukan, data harus dipreprocessing terlebih dahulu. Hasil dari preprocessing dapat dilihat pada tabel dibawah:

Tabel 2. Hadits sebelum preprocessing

Asal Hadits	No.Hadits	Hadits	Kategori
Shahih Bukhari	7	“Islam dibangun diatas lima (landasan); persak...	Haji
Shahih Bukhari	24	“Aku diperintahkan untuk memerangi manusia hin...	Zakat
Shahih Bukhari	37	“Barangsiapa yang berpuasa karena iman dan men...	Puasa
Shahih Bukhari	57	Ketika Nabi shallallahu’alaihi wasallam berad...	Kiamat
Shahih Bukhari	95	“Ada tiga orang yang akan mendapat pahala dua...	Menikah
Shahih Bukhari	45	“Barangsiapa mengiringi jenazah muslim, karena...	Shalat

Tebel 3. Hadits sesudah preprocessing

Asal Hadits	No.Hadits	Hadits	Kategori
Shahih Bukhari	7	['islam', 'dibangun', 'diatas', 'landasan', 'p...	Haji
Shahih Bukhari	24	['diperintahkan', 'memerangi', 'manusia', 'ber...	Zakat
Shahih Bukhari	37	['berpuasa', 'iman', 'mengharap', 'pahala', 'd...	Puasa
Shahih Bukhari	57	['majelis', 'membicarakan', 'kaum', 'tibatiba'...	Kiamat
Shahih Bukhari	95	['orang', 'pahala', 'kali', 'ahlul', 'kitab', ...	Menikah
Shahih Bukhari	45	['mengiringi', 'jenazah', 'muslim', 'iman', 'm...	Shalat

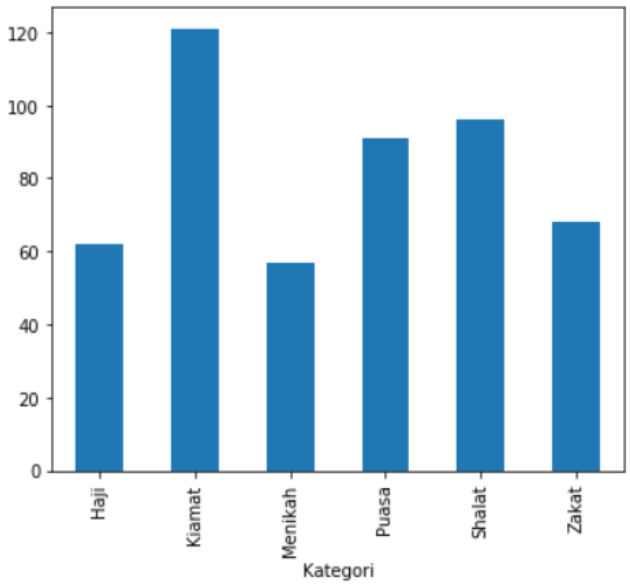
Setelah mendapatkan hasil preprocessing, kategori diberi identitas agar mempermudah proses klasifikasi.

Tabel 4. Identitas kategori

Kategori	Hadits	Kategori_id
Haji	['islam', 'dibangun', 'diatas', 'landasan', 'p...	0
Zakat	['diperintahkan', 'memerangi', 'manusia', 'ber...	1
Puasa	['berpuasa', 'iman', 'mengharap', 'pahala', 'd...	2
Kiamat	['majelis', 'membicarakan', 'kaum', 'tibatiba'...	3
Menikah	['orang', 'pahala', 'kali', 'ahlul', 'kitab', ...	4
Shalat	['mengiringi', 'jenazah', 'muslim', 'iman', 'm...	5

Selanjutnya menampilkan banyaknya data dalam bentuk diagram pada setiap kategori.

Gambar 3. Diagram data



Selanjutnya adalah hasil dari penghitungan akurasi dari naive bayes classifier dengan dua skenario untuk mendapakan perbandingan dengan menggunakan TF-IDF yang dinormalisasi dan tanpa normalisasi serta untuk mendapatkan manakah yang paling baik diantara keduanya yang dapat dilihat pada tabel berikut:

Tabel 6. Hasil akurasi pada dua skenario

TF-IDF (normalisasi)	TF-IDF (non-normalisasi)
89.29%	87.97%

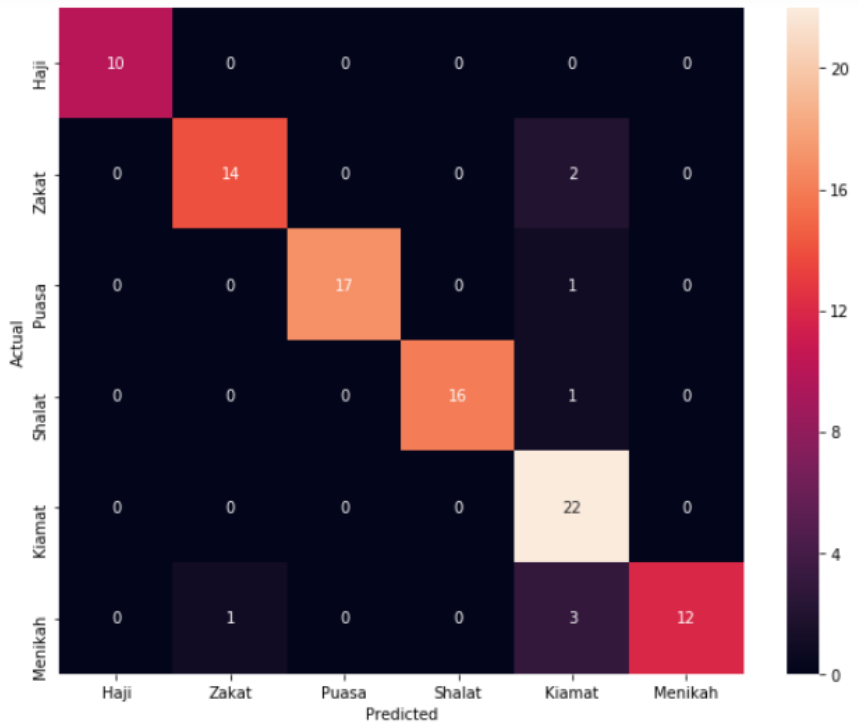
Dilanjutkan dengan hasil yang menunjukkan beberapa kata dan kalimat yang frekuensinya paling banyak (most correlated) pada setiap kategori yang terdapat pada tabel dibawah:

Tabel 5. Hasil most correlated

Kategori	Unigrams	Bigrams
Haji	Umrah Haji	Haji wada Menunaikan haji
Kiamat	Manusia Kiamat	Dihari kiamat Kiamat kelak
Menikah	Menikahi Menikah	Berzina menikah Menikahi wanita
Puasa	Puasa Berpuasa	Orang berpuasa Puasa wishal
Shalat	Adzan Shalat	Melaksanakan shalat Shalat ashar
Zakat	Shadaqah Zakat	Menunaikan zakat Shadakah zakat

Lalu diketahui confusion matrix dari prediksi naive bayes classifier dengan TF-IDF yang dinormalisasikan dengan hasil seperti pada gambar dibawah:

Gambar 4. Hasil confusion matrix



Berikut hasil prescision, recall, dan f-masure yang didapatkan dari confusion matrix.

Tabel 7. Hasil evaluasi

	PRECISION	RECALL	F1-SCORE
HAJI	1.00	1.00	1.00
ZAKAT	0.93	0.88	0.90
PUASA	1.00	0.94	0.97
SHALAT	1.00	0.94	0.97
KIAMAT	0.76	1.00	0.86
MENIKAH	1.00	0.75	0.86
MICRO AVG	0.92	0.92	0.92
MACRO AVG	0.95	0.92	0.94
WEIGTED AVG	0.94	0.92	0.92

4.2 Analisis Hasil Pengujian

Pada anilisis penelitian ini, hasil pengujian dengan melakukan pembobotan TF-IDF memberikan hasil yang baik karena, hal ini disebabkan TF-IDF pada term frekuensi yang sudah dinormalisasi sehingga nilai pembobotan TF-IDF lebih akurat.

5. Kesimpulan

Berdasarkan hasil pengujian, naive bayes memerlukan proses learning dan pengklasifikasian untuk memprediksi probabilitas berdasarkan pengalaman. Lalu pada pengujian ini, didapatkan hasil kinerja dari naive bayes yang sangat baik dengan akurasi terbaik sebesar 89.29%, niave bayes classifier berhasil mendapatkan akurasi dari data Hadits yang telah disusun sedemikian rupa. Pada perbandingan akurasi antara dua skenario, yaitu pada skenario pembobotan menggunakan TF-IDF tanpa normalisasi dan sudah dinormalisasi, dapat disimpulkan bahwa pembobotan TF-IDF harus melalui normalisasi agar mendapatkan hasil akurasi data yang akurat.

Untuk kedepannya (*future work*) diharapkan sistem ini dapat dikembangkan lebih lanjut lagi, baik itu dengan cara menambahkan jumlah data, dan menambahkan kategori yang lebih spesifik lagi.

Daftar Pustaka

[1] Ferdiansyah, H. (2017, Desember 27). Pembagian Hadis Secara Kualitas. Retrieved from Wikihadis: https://wikihadis.id/pembagian-hadis-secara-kualitas/

[2] Frank, E., & Bouckaert, R. R. (2006, September). Naive bayes for text classification with unbalanced classes. In European Conference on Principles of Data Mining and Knowledge Discovery (pp. 503-510). Springer, Berlin, Heidelberg.

[3] Ibrahim, A. (n.d.). Pengertian Hadits dan Jenis-jenis Hadits. Retrieved from Pengertian dan Definisi: https://pengertiandefinisi.com/pengertian-hadits-dan-jenis-jenis-hadits/

[4] INFORMATIKALOGI. (2016, NOVEMBER 12). Pembobotan Kata atau Term Weighting TF-IDF. Retrieved from INFORMATIKALOGI: https://informatikalogi.com/term-weighting-tf-idf/

[5] INFORMATIKALOGI. (2017, April 8). Algoritma Naive Bayes. Retrieved from INFORMATIKALOGI: https://informatikalogi.com/algoritma-naive-bayes/

[6] Jbara, K. (2010). Knowledge Discovery in Al-Hadith Using Text Classification Algorithm.

[7] Kibriya, A. M., Frank, E., Pfahringer, B., & Holmes, G. (2004, December). Multinomial naive bayes for text categorization revisited. In Australasian Joint Conference on Artificial Intelligence (pp. 488-499). Springer, Berlin, Heidelberg.

[8] Ramos, J. (2003). Using TF-IDF to Determine Word Relevance in Document Queries.

[9] Rennie, J. D. (2001). Improving multi-class text classification with naive Bayes.

[10] Scikit-Learn. (n.d.). Multinomial Naive Bayes. Retrieved from scikit-learn: https://scikit-learn.org/stable/modules/naive_bayes.html#multinomial-naive-bayes

[11] Wikipedia. (2019, Februari 28). Hadis. Retrieved from Wikipedia: https://id.wikipedia.org/wiki/Hadis

[12] Ginting, S. L. B., & Trinanda, R. P. (2013). Teknik Data Mining Menggunakan Metode Bayes Classifier untuk Optimalisasi Pencarian pada Aplikasi Perpustakaan (Studi Kasus: Perpustakaan Universitas Pasundan–Bandung). Jurnal Teknologi dan Informasi, 3(2).