REKAYASA
Journal of Science and Technology
https://journal.trunojoyo.ac.id/rekayasa

REKAYASA, 13(2): 2020 p.172-180
ISSN: 0216-9495 (Print)
ISSN: 2502-5325 (Online)

Perbandingan Metode *Term Weighting* terhadap Hasil Klasifikasi Teks pada Dataset Terjemahan Kitab Hadis

Ana Tsalitsatun Ni'mah [1*)], Agus Zainal Arifin [2)]
[1,2)] Teknik Informatika, Institut Teknologi Sepuluh Nopember
[1)] anatsalits@gmail.com, [2)] agusza@cs.its.ac.id

DOI: https://doi.org/10.21107/rekayasa.v13i2.6412

Comparison of a term weighting method for the text classification in Indonesian hadith

ABSTRACT

Hadith is the second source of reference for Islam after the Qur'an. Currently, hadith text is researched in the field of technology for capturing the values of technology knowledge. With the research of the Book of Hadith, retrieval of information from the hadith certainly requires the representation of text into vectors to optimize automatic classification. The classification of the hadith is needed to be able to group the contents of the hadith into several categories. There are several categories in certain Hadiths that are the same as other Hadiths. Shows that there are certain documents of the hadith that have the same topic as other Hadiths. Therefore, a term weighting method is needed that can choose which words should have high or low weights in the Hadith Book space to optimize the classification results in the Hadith Books. This study proposes a comparison of several term weighting methods, namely: Term Frequency Inverse Document Frequency (TF-IDF*), Term Frequency Inverse Document Frequency Inverse Class Frequency* (TF-IDF-ICF), *Term Frequency Inverse Document Frequency Inverse Class Space Density Frequency* (TF-IDF-ICS$_\delta$F) *and Term Frequency Inverse Document Frequency Inverse Class Space Density Frequency Inverse Hadith Space Density Frequency* (TF-IDF-ICS$_\delta$F-IHS$_\delta$F). *This research compares the term weighting results to the 9 Hadith Book Translation dataset applied to the Naive Bayes classification engine and SVM. 9 Books of Hadith are used, namely: Sahih Bukhari, Sahih Muslim, Abu Dawud, at-Turmudzi, an-Nasa'i, Ibn Majah, Ahmad, Malik, and Darimi. The trial results show that the classification results using the* TF-IDF-ICS$_\delta$F-IHS$_\delta$F *term weighting method outperformed another term weighting, namely getting a Precession* of 90%, *Recall* of 93%, F1-*Score* of 92%, *and Accuracy of* 83%.
Keywords: *Term Weighting*, *Classification*, *Hadith*, TF-IDF, TF-IDF-ICF, TF-IDF-ICS$_\delta$F, TF-IDF-ICS$_\delta$F-IHS$_\delta$F.

ABSTRAK

Hadis adalah sumber rujukan agama Islam kedua setelah Al-Qur'an. Teks Hadis saat ini diteliti dalam bidang teknologi untuk dapat ditangkap nilai-nilai yang terkandung di dalamnya secara pegetahuan teknologi. Dengan adanya penelitian terhadap Kitab Hadis, pengambilan informasi dari Hadis tentunya membutuhkan representasi teks ke dalam vektor untuk mengoptimalkan klasifikasi otomatis. Klasifikasi Hadis diperlukan untuk dapat mengelompokkan isi Hadis menjadi beberapa kategori. Ada beberapa kategori dalam Kitab Hadis tertentu yang sama dengan Kitab Hadis lainnya. Ini menunjukkan bahwa ada beberapa dokumen Kitab Hadis tertentu yang memiliki topik yang sama dengan Kitab Hadis lain. Oleh karena itu, diperlukan metode *term weighting* yang dapat memilih kata mana yang harus memiliki bobot tinggi atau rendah dalam ruang Kitab Hadis untuk optimalisasi hasil klasifikasi dalam Kitab-kitab Hadis. Penelitian ini mengusulkan sebuah perbandingan beberapa metode *term weighting,* yaitu: *Term Frequency Inverse Document Frequency* (TF-IDF), *Term Frequency Inverse Document Frequency Inverse Class Frequency* (TF-IDF-ICF), *Term Frequency Inverse Document Frequency Inverse Class Space Density Frequency* (TF-IDF-ICS$_\delta$F), dan *Term Frequency Inverse Document Frequency Inverse Class Space Density Frequency Inverse Hadith Space Density Frequency* (TF-IDF-ICS$_\delta$F-IHS$_\delta$F). Penelitian ini melakukan perbandingan hasil *term weighting* terhadap dataset Terjemahan 9 Kitab Hadis yang diterapkan pada mesin klasifikasi Naive Bayes dan SVM. 9 Kitab Hadis yang digunakan, yaitu: Sahih Bukhari, Sahih Muslim, Abu Dawud, at-Turmudzi, an-Nasa'i, Ibnu Majah, Ahmad, Malik, dan Darimi. Hasil uji coba menunjukkan bahwa hasil klasifikasi menggunakan metode *term weighting* TF-IDF-ICS$_\delta$F-IHS$_\delta$F mengungguli *term weighting* lainnya, yaitu mendapatkan Precission sebesar 90%, Recall sebesar 93%, F1-Score sebesar 92%, dan Accuracy sebesar 83%.
Kata Kunci: *Term Weighting*, Klasifikasi, Hadis, TF-IDF, TF-IDF-ICF, TF-IDF-ICS$_\delta$F, TF-IDF-ICS$_\delta$F-IHS$_\delta$F.

PENDAHULUAN

Hadis menurut bahasa memiliki arti *al-khabar* (berita), yaitu sesuatu yang dipercakapkan dan dipindahkan dari seseorang kepada orang lain (Arifin, 2013). Hadis didefinisikan sebagai ucapan, perbuatan, atau penetapan yang disandarkan kepada Nabi Muhammad SAW (Saloot et al., 2016). Hadis dalam KBBI disebut juga sebagai sunnah yaitu perkataan (sabda), perbuatan, ketetapan, dan persetujuan dari Nabi Muhammad SAW yang dijadikan landasan syariat Islam. Hadis dijadikan sumber hukum Islam selain Al-Qur'an, dalam hal ini kedudukan hadis merupakan sumber hukum kedua setelah Al-Quran.

Article History:
Received: January, 12th 2020; **Accepted**: May, 10th 2020
REKAYASA ISSN: 2502-5325 has been Accredited by Ristekdikti (Arjuna) Decree: No. 23/E/KPT/2019 August 8th, 2019 effective until 2023

Cite this as:
Ni'mah, A.T., & Arifin, A.Z. (2020). Perbandingan Metode *Term Weighting* terhadap Hasil Klasifikasi Teks pada Dataset Terjemahan Kitab Hadis. Rekayasa, 13(2), 172-180. https://doi.org/10.21107/rekayasa.v13i2.6412

Kitab Hadis adalah kumpulan beberapa dokumen hadis yang disusun oleh *mudawwin* atau *mukharrij* (ahli hadis). Kitab Hadis yang paling populer adalah Kitab Hadis Sunni yang terdiri dari 9 Kitab oleh 9 Imam, yaitu (Azmi, Al-Qabbany, & Hussain, 2019): Shahih Muslim disusun oleh Muslim (204-262 H), Shahih Bukhari disusun oleh Bukhari (194-256 H), Sunan at-Turmudzi disusun oleh At-Turmudzi (209-279 H), Sunan Abu Dawud disusun oleh Abu Dawud (202-275 H), Sunan Ibnu Majah disusun oleh Ibnu Majah (209-273), Muwatta Malik disusun oleh Imam Malik (93-179 H), Sunan an-Nasa'i disusun oleh an-Nasa'i (215-303 H), Sunan Darimi disusun oleh Ad-Darimi (181-255 H), dan Musnad Ahmad disusun oleh Imam Ahmad bin Hambal (164-241 H). Setiap Kitab dikategorikan berdasarkan topik yang tertuang di dalam Hadis tersebut (Rostam & Malim, 2019). Ada beberapa kategori dalam beberapa Kitab Hadis yang sama (Shalat, Iman, Ilmu, Zakat, Puasa, Thaharah, dll). Hal tersebut menandakan ada beberapa hadis pada Kitab Hadis tertentu memiliki topik yang sama dengan Kitab Hadis lainnya.

Teks Hadis saat ini telah diteliti dalam bidang teknologi untuk dapat ditangkap nilai-nilai yang terkandung di dalam teksnya secara pegetahuan teknologi dengan *Natural Language Processing* (NLP). Salah satu keuntungan dari menerapkan NLP terhadap teks Hadis adalah implementasi dari sistem cerdas yang dapat menjawab pertanyaan apa pun dengan data dari Hadis, dan dapat membantu masyarakat Muslim dan non-Muslim, untuk memahami Hadis (Saloot et al., 2016). Dengan adanya penelitian terhadap Kitab-kitab Hadis tersebut, pengambilan informasi dari Hadis tentunya membutuhkan representasi teks ke dalam vektor untuk mengoptimalkan tugas mesin klasifikasi otomatis. Secara umum, representasi teks ke vektor dapat diklasifikasikan menjadi dua, yaitu: pengindeksan dan *term weighting* (Ren & Sohrab, 2013). Dalam *Vector Space Model* (VSM), konten teks direpresentasikan sebagai vektor dalam ruang term. *Term weighting* merupakan proses penghitungan bobot tiap *term* yang dicari pada setiap dokumen sehingga dapat diketahui ketersediaan dan kemiripan suatu *term* di dalam dokumen (Sabbah et al., 2017). *Term weighting* adalah tingkat kepentingan term t_i dalam dokumen d_j.

Term weighting yang paling populer adalah *Term Frequency Inverse Document Frequency* (TF-IDF) (Ren & Sohrab, 2013). TF-IDF melakukan pembobotan *term* pada tiap dokumen dengan memperhatikan kelangkaan *term* pada keseluruhan dokumen. Hasil akhir dari TF-IDF adalah bobot *term* pada tiap-tiap dokumen. TF-IDF memberikan bobot tinggi pada *term* yang jarang muncul pada seluruh dokumen. TF-IDF ini memiliki kekurangan yaitu menghilangan informasi kategori pada tiap dokumen sehingga dilakukanlah penelitian untuk mengembangkan TF-IDF yaitu *Term Frequenxy Inverse Document Frequency Inverse Class Frequency* (TF-IDF-ICF) yang lebih memperhatikan informasi kategori di dalam melakukan pembobotan *term*. TF-IDF-ICF juga melakukan pembobotan *term* pada tiap dokumen seperti halnya TF-IDF dengan hasil bobot kata pada tiap dokumen pula. Perbedaannya pada perhitungan ICF, jika IDF hanya memperhitungkan kelangkaan kata pada keseluruhan dokumen, ICF juga menghitung kelangkaan kata pada keseluruhan kelas. Namun, TF-IDF-ICF tidak memperhatikan kepadatan dokumen tiap kelas terhadap munculnya term. Sehingga dikembangkan kembali dengan pendekatan term weighting yaitu *Term Frequency Inverse Document Frequency Inverse Class Space Density Frequency* (TF-IDF-ICS$_\delta$F) yang lebih memperhatikan kepadatan dokumen pada ruang kelas terhadap munculnya term. Pada kasus Kitab Hadis, dibutuhkan pula pembobotan yang dapat menentukan bobot tinggi dan rendah untuk kepadatan ruang Kitab Hadis. *Term weighting* yang digunakan untuk mengukur kepadatan ruang Kitab Hadis adalah *Term Frequency Inverse Document Frequency Inverse Class Space Density Frequency Inverse Hadith Space Density Frequency* (TF-IDF-ICS$_\delta$F- IHS$_\delta$F).

Skema *term weighting* memainkan peran penting dalam klasifikasi teks. Dibutuhkan suatu metode *term weighting* yang dapat menghasilkan lebih banyak *term* yang kaya informasi dan menetapkan nilai bobot *term* yang sesuai untuk persyaratan klasifikasi teks. Klasifikasi Hadis diperlukan untuk dapat mengelompokkan isi dari hadis tersebut ke

dalam beberapa kategori. Pada beberapa Kitab Hadis memiliki beberapa kategori yang sama sehingga ada beberapa hadis antar Kitab hadis yang memiliki topik yang sama, oleh sebab itu diperlukan sebuah metode *term weighting* yang dapat menyeleksi kata mana saja yang harus memiliki bobot tinggi atau rendah pada ruang Kitab Hadis untuk optimasi hasil klasifikasi pada kitab-kitab hadis.

Penelitian ini mengusulkan sebuah perbandingan beberapa metode *term weighting* yang diterapkan untuk Klasifikasi Terjemahan 9 Kitab Hadis.

METODE

Beberapa metode *term weighting* yang akan dibandingkan dan beberapa metode klasifikasi yang digunakan sebagai pembanding di dalam proses uji coba dan evaluasi. Bab yang akan dijelaskan antara lain, yaitu: *Preprocessing*, TF-IDF, TF-IDF-ICF, TF-IDF-$ICS_{\delta}F$, TF-IDF-$ICS_{\delta}F$- $IHS_{\delta}F$, *Naive Bayes*, dan *Support Vector Machine* (SVM).

Preprocessing

Pemrosesan Teks (*Text Preprocessing*) adalah suatu proses pengubahan bentuk data yang belum terstruktur menjadi data yang terstruktur sesuai dengan kebutuhan untuk proses yang lebih lanjut (Dogan & Uysal, 2019). *Text Preprocessing* umumnya terdiri dari case *folding, tokenization, filtering* dan *stemming* (Deposit, Shi, & Jianping, 2018). Masing-masing proses memiliki manfaat terhadap dokumen yang diolah (Uysal & Gunal, 2014).

Case folding adalah konversi bentuk string menjadi lowercase (Uysal & Gunal, 2014). Pada case folding bermanfaat dalam proses selanjutnya pada stemming karena pada stemming saat penghapusan imbuhan, dia mengenali karakter huruf besar dan huruf kecil.

Filtering adalah pemfilteran kata-kata yang tidak mengandung makna atau biasa disebut *stopword*, pada proses *filtering* kata yang termasuk dalam *stopword* dihilangkan (Tala, 2003). Langkah ini melakukan penghapusan stopword yang bermanfaat saat proses pembobotan. Karena semakin banyak kata yang tidak mengandung makna yang ikut dalam proses pembobotan, maka akan semakin terganggu hasil bobot yang mewakili setiap dokumen.

Stemming adalah proses yang menyediakan pemetaan variasi kata morfologis yang berbeda ke dalam kata dasar / umum mereka (Tala, 2003). *Stemming* untuk bahasa Indonesia telah mengalami banyak perkembangan. Pertama kali pengembangan stemming bahasa Indonesia dilakukan oleh Adriani, M dan Nazief, B. pada tahun 1996 dengan diberi nama *Confix Stripping Algorithm* (Adriani, M., Nazief, B., Asian, J., Tahaghoghi, S. M. M., and Williams, 2007). Algoritma tersebut menggunakan tata bahasa Indonesia di dalam proses penghapusan imbuhan katanya, dengan melakukan pengecekan kebenaran kata dasar pada Kamus Besar Bahasa Indonesia. Selanjutnya algoritma *Confix Stripping* dikembangkan kembali oleh Arifin, Z.A. dan Setiono, A.N. pada tahun 2002 (Adriani, M., Nazief, B., Asian, J., Tahaghoghi, S. M. M., and Williams, 2007). Algoritma tersebut melakukan pendekatan dengan menghilangkan prefiks terlebih dahulu kemudian suffiks. Algoritma selanjutnya pada pengembangan *stemming* bahasa Indonesia adalah *Porter Indonesian* (Tala, 2003). Algoritma ini melakukan proses *stemming* dengan mengadaptasi algoritma *Porter Stemming* pada bahasa Inggris. *Porter stemming* pada bahasa Inggris melakukan penghapusan imbuhan secara sederhana tanpa menggunakan pengecekan terhadap kamus kata dasar. Proses tersebut diadaptasikan terhadap bahasa Indonesia. Algoritma pengembangan stemming bahasa Indonesia adalah *Enhanced Confix Stripping* (ECS) (Mahendra, 2007). Kemudian algoritma tersebut mengalami pengembangan pada penelitian (Andita Dwiyoga Tahitoe, 2010).

TF-IDF

Term Frequency Inverse Document Frequency (TF-IDF) adalah salah satu metode pembobotan *statistical* (Kim & Gil, 2019). TF-IDF mengkalikan *term frequency* (TF) sebagai penghitung *frequency term* dalam sebuah dokumen dengan *inverse document frequency* (IDF) sebagai nilai keinformatifan sebuah *term* (kelangkaannya pada keseluruhan dokumen). Pengertian *Term frequency* (TF) adalah setiap kata diasumsikan memiliki kepentingan yang

proporsional terhadap jumlah kemunculan kata pada dokumen. *Inverse Document Frequency* (IDF) memperhatikan kemunculan term pada kumpulan dokumen (Dogan & Uysal, 2019). Fungsi IDF memberikan skor terendah untuk term yang muncul dalam banyak dokumen dalam ruang dokumen $D = d_1, d_2, d_3, \dots, d_n$ (Ren & Sohrab, 2013). Latar belakang pembobotan ini adalah *term* yang jarang muncul pada kumpulan dokumen sangat bernilai. Kepentingan tiap *term* diasumsikan memilki proporsi yang berkebalikan dengan jumlah dokumen yang mengandung term. *Term frequency* (TF) adalah metode sederhana dari pembobotan kata. *Term frequency* memperhatikan kemunculan *term* di dalam dokumen. Bobot tinggi dalam TF-IDF dicapai oleh frekuensi *term* tinggi (dalam dokumen yang diberikan) dan frekuensi dokumen rendah dari *term* dalam seluruh kumpulan dokumen, karena itu bobot cenderung menyaring istilah umum. Ketika sebuah istilah muncul di lebih banyak dokumen, rasio di dalam hasil logaritma mendekati 1, mendekatkan IDF dan TF-IDF ke 0. TF-IDF divisualisasikan dalam Persamaan (1) (Ren & Sohrab, 2013).

$$W_{TF*IDF(t_i,d_j)} = tf_{t_i,d_j} \times \left(1 + \log \frac{D}{df_{(t_i)}}\right) \quad (1)$$

dimana, $W_{TF*IDF(t_i,d_j)}$ adalah bobot *term i* pada dokumen *j*. tf_{t_i,d_j}adalah jumlah *term i* di dalam dokumen *j*. *D* adalah jumlah seluruh dokumen. $df_{(t_i)}$ adalah jumlah seluruh dokumen yang mengandung *term i*.

TF-IDF-ICF

TF-IDF tidak memperhatikan persebaran *term* pada keragaman kelas, hanya berbasis pada dokumen (Ren & Sohrab, 2013). Kemudian dikembangkanlah TF-IDF menjadi *Term Frequency Inverse Document Frequency Inverse Class Frequency* (TF-IDF-ICF.) Metode *Inverse Class Frequency* (ICF) diadopsi dari metode IDF yaitu dengan menggunakan *inverse* perbandingan jumlah kelas dengan jumlah kelas yang mengandung istilah (Yang, Cai, Leung, Lau, & Li, 2019). Dalam hal pengindeksan berorientasi kelas, subset dokumen dari ruang dokumen $D = d_1, d_2, d_3, \dots, d_n$ dialokasikan ke kelas tertentu. Sehingga semakin sering istilah muncul dalam dokumen yang ada dalam kelas tersebut, maka bobot istilah semakin mendekati 0. Fungsi ICF memberikan skor terendah term yang muncul di beberapa kelas di ruang kelas $C = C_1, C_2, C_3, \dots, C_n$. Oleh karena itu, representasi numerik dari suatu term adalah produk dari *Term Frequency* (parameter lokal), IDF (parameter global), dan ICF (parameter global kategori). Persamaan dari TF-IDF-ICF dapat dilihat pada Persamaan (2) (Ren & Sohrab, 2013)

$$W_{TF*IDF*ICF(t_i,d_j,c_k)} = tf_{t_i,d_j} \times \left(1 + \log \frac{D}{df_{(t_i)}}\right) \times \left(1 + \log \frac{C}{cf_{(t_i)}}\right) \quad (2)$$

dimana, $W_{TF*IDF*ICF(t_i,d_j,c_k)}$adalah bobot *term i* pada dokumen *j*. tf_{t_i,d_j} adalah jumlah *term i* di dalam dokumen *j*. *D* adalah jumlah seluruh dokumen. $df_{(t_i)}$ adalah jumlah seluruh dokumen yang mengandung *term i*. *C* adalah jumlah seluruh kelas. $cf_{(t_i)}$adalah jumlah kelas yang mengandung *term i*.

TF-IDF-ICS$_\delta$F

Skema *term weighting* TF-IDF dan TF-IDF-ICF menekankan pada term yang jarang muncul, yang mendukung tingginya bobot *term* yang muncul hanya dalam beberapa dokumen dikalikan dengan fungsi IDF dan yang hanya muncul dalam beberapa kelas dikalikan dengan fungsi ICF. Pada ICS$_\delta$F menghitung kepadatan dokumen pada ruang kategori berdasarkan setiap *term*, *Inverse Class Space Density Frequency* (ICS$_\delta$F) dikalikan dengan TF-IDF untuk menghasilkan TF-IDF-ICS$_\delta$F. Karena fungsi ICF memberikan skor terendah untuk istilah-istilah yang muncul di beberapa kelas tanpa memperhatikan tentang ruang kelas, perhitungan ICS$_\delta$F kemudian diusulkan. Metode term weighting ini untuk meningkatkan kinerja klasifikasi. Dokumen identik ini yang dikaitkan dengan istilah tertentu mungkin merupakan sub-bagian dari kategori tertentu c_k. Karena itu, penting untuk mengeksplorasi karakteristik kemunculan istilah dalam ruang dokumen $D = d_1, d_2, d_3, \dots, d_n$ dan kelas ruang. Dalam pengindeksan berorientasi kelas, subset dokumen dari ruang dokumen global dialokasikan ke kelas tertentu c_k (*k = 1, 2, ..., m*) sesuai dengan topik mereka. Karena itu,

ruang kelas didefinisikan sebagai $C = \{(d_{11}, d_{12}, d_{13}, \dots, d_{1n})$ *anggota* C_1, $(d_{11}, d_{12}, d_{13}, \dots, d_{1n})$ *anggota* C_2 , $(d_{m1}, d_{m2}, d_{m3}, \dots, d_{mn})$ *anggota* C_m } (Ren & Sohrab, 2013). TF-IDF-ICS$_\delta$F diawali dengan menghitung kepadatan kelasnya (C$_\delta$) yaitu menghitung dokumen yang mengandung term pada kategori tertentu (c_k), dengan Persamaan (3) (Ren & Sohrab, 2013).

$$C_\delta(t_i) = \frac{n_{c_k}(t_i)}{N_{c_k}} \quad (3)$$

dimana, $C_\delta(t_i)$ adalah kepadatan kelas terhadap *term i*. $n_{c_k}(t_i)$ adalah jumlah dokumen di dalam kelas c_k yang mengandung *term i*. N_{c_k} adalah jumlah keseluruhan dokumen di dalam kelas c_k.

Kemudian dilanjutkan dengan menghitung kepadatan ruang kelas, yaitu jumlah dari kepadatan keseluruhan kelas yang ada (CS_δ), dengan persamaan (4) (fuji ren).

$$CS_\delta(t_i) = \sum_{c_k} C_\delta(t_i) \quad (4)$$

dimana, $CS_\delta(t_i)$adalah kepadatan ruang kelas terhadap *term i*. $C_\delta(t_i)$ adalah kepadatan kelas terhadap *term i*. c_k adalah kelas (*k = 1, 2, ..., m*). Kemudian hasil dari kepadatan ruang kelas ($CS_\delta(t_i)$) dilakukan inverse sesuai dengan konsep pada TF-IDF-ICF sebelumnya, dengan persamaan (5) (fuji ren).

$$ICS_\delta F(t_i) = log\left(\frac{C}{CS_\delta(t_i)}\right) \quad (5)$$

dimana, $ICS_\delta F(t_i)$ adalah *inverse* kepadatan ruang kelas terhadap *term i*. *C* adalah jumlah keseluruhan kelas. $CS_\delta(t_i)$ adalah kepadatan ruang kelas terhadap *term i*. Langkah selanjutnya adalah melakukan perkalian hasil *inverse* kepadatan ruang kelas terhadap *term i* ($ICS_\delta F(t_i)$) dengan TF-IDF, seperti pada persamaan (6) (fuji ren).

$$W_{TF*IDF*ICS_\delta F(t_i,d_j,c_k)} = tf_{t_i,d_j} \times \left(1 + \log\frac{D}{df_{(t_i)}}\right) \times \left(1 + \log\frac{C}{CS_\delta(t_i)}\right) \quad (6)$$

dimana, $W_{TF*IDF*ICS_\delta F(t_i,d_j,c_k)}$ adalah bobot *term i* pada dokumen *j* di dalam kelas *k*. tf_{t_i,d_j} adalah jumlah *term i* di dalam dokumen *j*. *D* adalah jumlah seluruh dokumen. $df_{(t_i)}$ adalah jumlah dokumen yang mengandung *term i*. *C* adalah jumlah seluruh Kelas. $CS_\delta(t_i)$ adalah kepadatan ruang kelas terhadap *term i*.

TF-IDF-ICS$_\delta$F-IHS$_\delta$F

Term Frequency Inverse Document Frequency Inverse Class Space Density Frequency Inverse Hadith Space Density Frequency (TF-IDF-ICS$_\delta$F-IHS$_\delta$F) adalah metode *term weighting* pengembangan dari TF-IDF-ICS$_\delta$F. Jika ICS$_\delta$F memperhatikan kepadatan ruang kelas terhadap suatu term, maka IHS$_\delta$F lebih memperhatikan kepadatan ruang Hadis terhadap suatu term. Kepadatan ruang Hadis dihitung untuk mengetahui seberapa besar bobot term jika dihitung pula kelangkaan term tersebut dari keseluruhan Hadis. Semakin term tersebut jarang muncul pada banyak Hadis, maka term tersebut memiliki nilai invers yang tinggi. Kemudian Proses pertama dalam perhitungan IHS$_\delta$F adalah menghitung kepadatan kelas terlebih dahulu seperti pada Persamaan (3). Selanjutnya, hasil dari kepadatan kelas terhadap *term i* dijumlahkan untuk mendapatkan kepadatan Hadis terhadap *term i* seperti pada persamaan (7).

$$H_\delta(t_i) = \sum_{c_k}^{h_l} C_\delta(t_i) \quad (7)$$

dimana, $H_\delta(t_i)$ adalah kepadatan Hadis terhadap *term i*. $C_\delta(t_i)$ adalah kepadatan kelas terhadap *term i*. c_k adalah kelas (*k = 1, 2, ..., m*). h_l adalah hadis (*l = 1, 2, ..., 9*). Kepadatan kelas (C$_\delta$) dengan kepadatan Hadis (H$_\delta$) sangatlah berbeda, C$_\delta$ menghitung seberapa besar kepadatan kemunculan term pada kelas tertentu dengan melakukan perbandingan jumlah dokumen yang memiliki term pada kelas tertentu dibanding jumlah dokumen keseluruhan pada kelas tertentu sedangkan H$_\delta$ menghitung seberapa besar kepadatan kemunculan term pada Hadis tertentu dengan melakukan penjumlahan kepadatan kelas sebanyak kelas yang ada pada Hadis tertentu. Kepadatan hadis terhadap *term i* dijumlahkan untuk mendapatkan kepadatan ruang hadis terhadap *term i* seperti pada persamaan (8).

$$HS_\delta(t_i) = \sum_{h_l} H_\delta(t_i) \quad (8)$$

dimana, $HS_\delta(t_i)$adalah kepadatan ruang hadis terhadap *term i*. $H_\delta(t_i)$ adalah kepadatan Hadis terhadap *term i*. h_l adalah hadis (*l = 1, 2, ..., 9*). Kepadatan ruang kelas (CS$_\delta$) berbeda

dengan kepadatan ruang Hadis (HS_δ), CS_δ menghitung seberapa besar kepadatan kemunculan term pada keseluruhan kelas sedangkan HS_δ menghitung seberapa besar kepadatan kemunculan term pada keseluruhan Hadis. Selanjutnya hasil kepadatan ruang hadis terhadap *term i* dilakukan inverse untuk mengetahui tingkat kelangkaan *term* terhadap ruang hadis seperti pada persamaan (9).

$$IHS_\delta F(t_i) = log\left(\frac{H}{HS_\delta(t_i)}\right) \quad (9)$$

$IHS_\delta F(t_i)$ adalah *inverse* kepadatan ruang hadis terhadap *term i*. *H* adalah jumlah keseluruhan hadis. $HS_\delta(t_i)$ adalah kepadatan ruang hadis terhadap *term i*. Selanjutnya hasil inverse dilakukan perkalian dengan persamaan (6) untuk mengetahui bobot *term* yang memperhatikan kepadatan ruang kelas dan juga kepadatan ruang hadis seperti pada persamaan (10).

$$W_{TF*IDF*ICS_\delta F*IHS_\delta F(t_i,d_j,c_k,h_l)} = tf_{t_i,d_j} \times \left(1 + \log\frac{D}{df_{(t_i)}}\right) \times \left(1 + \log\frac{C}{CS_\delta(t_i)}\right) \times \left(1 + log\left(\frac{H}{HS_\delta(t_i)}\right)\right) \quad (10)$$

dimana, $W_{TF*IDF*ICS_\delta F*IHS_\delta F(t_i,d_j,c_k,h_l)}$ adalah bobot *term i* pada dokumen *j* di dalam kelas *k* pada hadis *l*. tf_{t_i,d_j} adalah jumlah *term i* di dalam dokumen *j*. *D* adalah jumlah seluruh dokumen. $df_{(t_i)}$ adalah jumlah dokumen yang mengandung *term i*. *C* adalah jumlah seluruh Kelas. $CS_\delta(t_i)$ adalah kepadatan ruang kelas terhadap *term i*. *H* adalah jumlah keseluruhan hadis. $HS_\delta(t_i)$ adalah kepadatan ruang hadis terhadap *term i*. Hasil pembobotan term dengan memperhatikan kepadatan ruang kelas dan kepadatan ruang hadis ini memberikan nilai yang lebih tinggi terhadap term yang langka pada kelas dan hadis, dan memberikan nilai yang rendah pada term yang jarang muncul. Inverse kepadatan ruang kelas dan ruang hadis ini tetap dikalikan dengan banyaknya term tersebut muncul pada tiap dokumen, sehingga jika term tersebut muncul sangat banyak di dalam suatu dokumen maka bobot term tersebut tetap tinggi. Karena umumnya topik dari sebuah dokumen lebih sering disebutkan di dalam dokumen tersebut. Nilai inverse kepadatan ruang kelas dan ruang hadis ini digunakan untuk menyeleksi term yang terlalu banyak muncul di keseluruhan kelas dan keseluruhan hadis.

Naive Bayes

Naive Bayes adalah salah satu algoritma klasifikasi formal tertua, dan banyak digunakan algoritma klasifikasi. Dalam model Bayesian, asumsi didasarkan pada probabilitas sebelum dan sesudah. Menemukan probabilitas dari jenis dokumen tertentu, d_j anggota C hanya dapat didasarkan pada t_i observasi (Azalia, Bijaksana, & Huda, 2019) .

Support Vector Machine (SVM)

Pada pendekatan machine learning, Support Vector Machine (SVM) dianggap sebagai salah satu algoritma yang paling kuat dan akurat (Ren & Sohrab, 2013). Berbeda dengan strategi jaringan saraf yang hanya mencari hyperplane pemisah antara kelas, SVM mencoba menemukan hyperplane terbaik di ruang input (Yusup, Bijaksana, & Huda, 2019).

HASIL DAN PEMBAHASAN

Bagian ini akan memaparkan dataset yang digunakan, skenario uji coba, dan hasil uji coba.

Dataset

Pengumpulan data dilakukan dengan cara melakukan crawl menggunakan aplikasi crawler yang terdapat pada website https://www.octoparse.com/. Proses pengambilan dilakukan dengan menganalisa beberapa alamat website yang akan dijaring kumpulan Hadisnya dari website https://tafsirq.com/. Hadis yang digunakan merupakan hadis dari 9 Imam, yaitu:

- Shahih Bukhari : https://tafsirq.com/hadits/bukhari
- Shahih Muslim : https://tafsirq.com/hadits/muslim
- Sunan Abu Dawud : https://tafsirq.com/hadits/abu-daud
- Sunan at-Turmudzi : https://tafsirq.com/hadits/tirmidzi
- Sunan an-Nasa'i : https://tafsirq.com/hadits/nasai

- Sunan Ibnu Majah : https://tafsirq.com/hadits/ibnu-majah
- Musnad Ahmad : https://tafsirq.com/hadits/ahmad
- Muwatta Malik : https://tafsirq.com/hadits/malik
- Sunan Darimi : https://tafsirq.com/hadits/darimi

Masing-masing jumlah Hadis yang digunakan pada setiap kitab, yaitu: Sunan Abu Dawud sebanyak 4419 hadis, Musnad Ahmad sebanyak 4305 hadis, Shahih Bukhari sebanyak 6638 hadis, Sunan Darimi sebanyak 2949 hadis, Sunan Ibnu Majah sebanyak 4285 hadis, Muwatta Malik sebanyak 1587 hadis, Shahih Muslim sebanyak 4930 hadis, Sunan an-Nasa'i sebanyak 5364 hadis, dan Sunan at-Turmudzi sebanyak 3625 hadis. Perincian jumlah hadis yang digunakan dapat dilihat pada Tabel 1. Total dokumen yang diolah di dalam penelitian ini adalah 38102 dokumen dengan 370 kategori dan 9 Kitab Hadis.

Sebuah dokumen di dalam Hadis, memiliki 2 macam isi yaitu: Sanad dan Matan (Azmi et al., 2019). Sanad adalah riwayat, Matan adalah isi inti atau redaksi dari sebuah dokumen Hadis. Penelitian ini hanya menggunakan Matan saja.

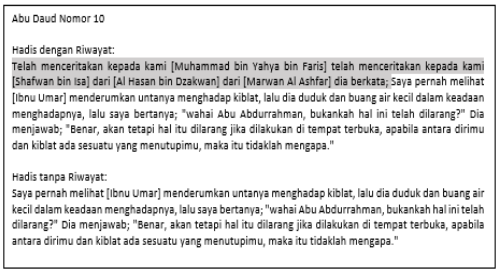
Cara memperoleh Matan adalah dengan melakukan penghapusan manual Sanad dari setiap dokumen Hadis. Contoh dokumen Hadis asli dengan disertai Sanad dan dokumen Hadis yang telah dihapus Sanadnya hanya berisi Matan saja dapat dilihat pada Gambar 1.

Pengujian Kinerja

Uji coba adalah tahapan di mana melakukan pengujian terhadap kesiapan sistem. Uji coba pada penelitian ini menggunakan *Confusion matrix*. *Confusion matrix* merupakan salah satu metode yang dapat digunakan untuk mengukur kinerja suatu metode klasifikasi.

Tabel 1. Rincian Jumlah Dokumen dalam Kitab Hadis

Hadis	Jumlah Dokumen	Jumlah Kategori
Bukhari	6638	77 kategori
Muslim	4930	56 kategori
Dawud	4419	35 kategori
Turmudzi	3625	49 kategori
Nasa'i	5364	51 kategori
Ibnu Majah	4285	32 kategori
Ahmad	4305	14 kategori
Malik	1587	32 kategori
Darimi	2949	24 kategori


Abu Daud Nomor 10

Hadis dengan Riwayat:
Telah menceritakan kepada kami [Muhammad bin Yahya bin Faris] telah menceritakan kepada kami [Shafwan bin Isa] dari [Al Hasan bin Dzakwan] dari [Marwan Al Ashfar] dia berkata; Saya pernah melihat [Ibnu Umar] menderumkan untanya menghadap kiblat, lalu dia duduk dan buang air kecil dalam keadaan menghadapnya, lalu saya bertanya; "wahai Abu Abdurrahman, bukankah hal ini telah dilarang?" Dia menjawab; "Benar, akan tetapi hal itu dilarang jika dilakukan di tempat terbuka, apabila antara dirimu dan kiblat ada sesuatu yang menutupimu, maka itu tidaklah mengapa."

Hadis tanpa Riwayat:
Saya pernah melihat [Ibnu Umar] menderumkan untanya menghadap kiblat, lalu dia duduk dan buang air kecil dalam keadaan menghadapnya, lalu saya bertanya; "wahai Abu Abdurrahman, bukankah hal ini telah dilarang?" Dia menjawab; "Benar, akan tetapi hal itu dilarang jika dilakukan di tempat terbuka, apabila antara dirimu dan kiblat ada sesuatu yang menutupimu, maka itu tidaklah mengapa."

Gambar 1. Contoh Isi Dokumen di dalam Kitab Hadis

Tabel 2. *Confusion Matrix*

Kelas	Terklasifikasi Positif	Terklasifikasi Negatif
Positif	TP (*True Positive*)	FN (*False Negative*)
Negatif	FP (*False Positive*)	TN (*True Negative*)

Keakuratan sistem diukur dengan confusion matrix, terdapat 4 (empat) pengukuran hasil klasifikasi. Keempat pengukuran tersebut adalah False Positive (FP), False Negative (FN), True Positive (TP), dan True Negative (TN). Nilai True Negative (TN) adalah data negatif yang terdeteksi sebagai data negatif.
False Positive (FP) adalah nilai dari data negatif yang terdeteksi menjadi data positif. True Positive (TP) adalah data positif yang terdeteksi menjadi data positif. False Negative (FN) adalah kebalikan dari True Positive yaitu data positif yang terdeteksi menjadi data negatif. Tabel Confusion Matrix dapat dilihat pada Tabel 2.

Skenario Uji Coba

Uji coba adalah tahapan di mana melakukan pengujian terhadap kesiapan sistem. Uji coba pada penelitian ini menggunakan Confusion matrix. Confusion matrix merupakan salah satu metode yang dapat digunakan untuk mengukur kinerja suatu metode klasifikasi. Skenario uji coba terdiri dari 2, yaitu:

1. Perbandingan Hasil Rata-rata Precission, Recall, F1-Score, dan Accuracy pada classifier Naive Bayes
2. Perbandingan Hasil Rata-rata Precission, Recall, F1-Score, dan Accuracy pada classifier SVM

Hasil Uji Coba

Hasil uji coba sesuai dengan skenario uji coba pada bab sebelumnya dipaparkan pada Tabel 3 dan 4. Tabel tersebut merangkum seberapa besar perbandingan hasil klasifikasi dari 4 *term weighting* yang digunakan di dalam penelitian ini. Dari hasil rangkuman terebut disebutkan bahwa metode TF-IDF dengan mesin klasifikasi Naive Bayes mendapatkan hasil yaitu *Precission* sebesar 82%, *Recall* sebesar 72%, *F1-Score* sebesar 75%, dan *Accuracy* sebesar 81%. TF-IDF-ICF dengan mesin klasifikasi Naive Bayes mendapatkan hasil yaitu *Precission* sebesar 90%, *Recall* sebesar 93%, *F1-Score* sebesar 91%, dan *Accuracy* sebesar 83%. TF-IDF-ICS$_\delta$F dengan mesin klasifikasi Naive Bayes mendapatkan hasil.

Precission sebesar 91%, *Recall* sebesar 86%, *F1-Score* sebesar 88%, dan *Accuracy* sebesar 93%. TF-IDF-ICS$_\delta$F-IHS$_\delta$F dengan mesin klasifikasi Naive Bayes mendapatkan hasil yaitu *Precission* sebesar 96%, *Recall* sebesar 93%, *F1-Score* sebesar 94%, dan *Accuracy* sebesar 95%. TF-IDF dengan mesin klasifikasi SVM mendapatkan hasil yaitu *Precission* sebesar 80%, *Recall* sebesar 65%, *F1-Score* sebesar 70%, dan *Accuracy* sebesar 76%. TF-IDF-ICF dengan mesin klasifikasi Naive Bayes mendapatkan hasil yaitu *Precission* sebesar 82%, *Recall* sebesar 70%, *F1-Score* sebesar 74%, dan *Accuracy* sebesar 82%. TF-IDF-ICS$_\delta$F dengan mesin klasifikasi Naive Bayes mendapatkan hasil yaitu *Precission* sebesar 88%, *Recall* sebesar 80%, *F1-Score* sebesar 83%, dan *Accuracy* sebesar 88%. TF-IDF-ICS$_\delta$F-IHS$_\delta$F dengan mesin klasifikasi Naive Bayes mendapatkan hasil yaitu *Precission* sebesar 90%, *Recall* sebesar 93%, *F1-Score* sebesar 92%, dan *Accuracy* sebesar 83%. Dari hasil tersebut dapat terlihat bahwa hasil terbaik

Tabel 3. Hasil Uji Coba pada Mesin Klasifikasi Naive Bayes

Confusion Matrix	**TF-IDF**	**TF-IDF-ICF**	**TF-IDF-ICS$_\delta$F**	**TF-IDF-ICS$_\delta$F-IHS$_\delta$F**
Precission	0,82	0,90	0,91	**0,96**
Recall	0,72	**0,93**	0,86	**0,93**
F1-Score	0,75	0,91	0,88	**0,94**
Accuracy	0,81	0,83	0,93	**0,95**

Tabel 4. Hasil Uji Coba pada Mesin Klasifikasi Support Vector Machine

Confusion Matrix	**TF-IDF**	**TF-IDF-ICF**	**TF-IDF-ICS$_\delta$F**	**TF-IDF-ICS$_\delta$F-IHS$_\delta$F**
Precission	0,80	0,82	0,88	**0,90**
Recall	0,65	0,70	0,80	**0,93**
F1-Score	0,70	0,74	0,83	**0,92**
Accuracy	0,76	0,82	**0,88**	0,83

didapatkan dari *term weighting* TF-IDF-ICS$_\delta$F-IHS$_\delta$F untuk keseluruhan mesin klasifikasi. TF-IDF-ICS$_\delta$F-IHS$_\delta$F mendapatkan hasil terbaik pada mesin klasifikasi Naive Bayes.

KESIMPULAN

Kesimpulan dari penelitian ini adalah metode *term weighting* TF-IDF-ICS$_\delta$F-IHS$_\delta$F telah mampu memberikan hasil terbaik dan mengungguli *term weighting* lain di dalam penerapannya terhadap mesin klasifikasi dengan dataset terjemahan 9 Kitab Hadis. Hasil yang diperoleh pada mesin klasifikasi Naive Bayes adalah *Precission* sebesar 90%, *Recall* sebesar 93%, *F1-Score* sebesar 92%, dan *Accuracy* sebesar 83%.

Saran untuk penelitian selanjutnya adalah melakukan penelitian terhadap perbandingan dengan lebih banyak *term weighting,* mesin klasifikasi, dan penerapan terhadap dataset selain Kitab Hadis.

DAFTAR PUSTAKA

Adriani, M., Nazief, B., Asian, J., Tahaghoghi, S. M. M., and Williams, H. E. (2007). Stemming Indonesian. *ACM J. Educ. Resour. Comput. 6, 4, 38*(December), 307–314. https://doi.org/10.1145/1316457.1316459

Andita Dwiyoga Tahitoe, D. P. (2010). Implementasi Modifikasi Enhanced Confix Stripping Stemmer Untuk Bahasa Indonesia Dengan Metode Corpus Based Stemming. *Jurnal Ilmiah*, 1–15.

Arifin, Z. (2013). Studi Kitab Hadis. In *Srudi*. https://doi.org/10.1364/OL.36.000127

Azalia, F. Y., Bijaksana, M. A., & Huda, A. F. (2019). Name indexing in Indonesian translation of hadith using named entity recognition with naïve bayes classifier. *Procedia Computer Science*, *157*, 142–149. https://doi.org/10.1016/j.procs.2019.08.151

Azmi, A. M., Al-Qabbany, A. O., & Hussain, A. (2019). Computational and natural language processing based studies of hadith literature: a survey. *Artificial Intelligence Review*, *52*(2), 1369–1414. https://doi.org/10.1007/s10462-019-09692-w

Deposit, C., Shi, L. I., & Jianping, C. (2018). Prospecting Information Extraction by Text Mining Based on Convolutional Neural Networks — A Case Study of the Lala. *IEEE Access*, *6*, 52286–52297. https://doi.org/10.1109/ACCESS.2018.2870203

Dogan, T., & Uysal, A. K. (2019). Improved inverse gravity moment term weighting for text classification. *Expert Systems with Applications*, *130*, 45–59. https://doi.org/10.1016/j.eswa.2019.04.015

Kim, S. W., & Gil, J. M. (2019). Research paper classification systems based on TF - IDF and LDA schemes. *Human-Centric Computing and Information Sciences*. https://doi.org/10.1186/s13673-019-0192-7

Mahendra. (2007). Enhanced Confix Stripping Stemmer And Ants Algoritm For Classifying News Document in Representation of Textual. *Technology*, (April), 149–158.

Ren, F., & Sohrab, M. G. (2013). *Class-indexing-based term weighting for automatic text classification*. *236*, 109–125.

Rostam, N. A. P., & Malim, N. H. A. H. (2019). Text categorisation in Quran and Hadith: Overcoming the interrelation challenges using machine learning and term weighting. *Journal of King Saud University - Computer and Information Sciences*, (xxxx). https://doi.org/10.1016/j.jksuci.2019.03.007

Sabbah, T., Selamat, A., Selamat, M. H., Al-Anzi, F. S., Viedma, E. H., Krejcar, O., & Fujita, H. (2017). Modified frequency-based term weighting schemes for text classification. *Applied Soft Computing Journal*, *58*, 193–206. https://doi.org/10.1016/j.asoc.2017.04.069

Saloot, M. A., Idris, N., Mahmud, R., Ja'afar, S., Thorleuchter, D., & Gani, A. (2016). Hadith data mining and classification: a comparative analysis. *Artificial Intelligence Review*, *46*(1), 113–128. https://doi.org/10.1007/s10462-016-9458-x

Tala, F. Z. (2003). A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia. *M.Sc. Thesis, Appendix D*, *pp*, 39–46.

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing and Management*, *50*(1), 104–112. https://doi.org/10.1016/j.ipm.2013.08.006

Yang, K., Cai, Y., Leung, H. Fung, Lau, R. Y. K., & Li, Q. (2019). ITWF: A framework to apply term weighting schemes in topic model. *Neurocomputing*, *350*, 248–260. https://doi.org/10.1016/j.neucom.2019.02.048

Yusup, F. A., Bijaksana, M. A., & Huda, A. F. (2019). Narrator's name recognition with support vector machine for indexing Indonesian hadith translations. *Procedia Computer Science*, *157*, 191–198. https://doi.org/10.1016/j.procs.2019.08.157