Analisis Name Matching untuk Nama Arab Menggunakan Metode N-gram dan Jaccard Similarity

Muhammad Rizki Chairulloh[1], Moch. Arif Bijaksana[2], Bambang Ari Wahyudi[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]rizkim47@gmail.com, [2]arifbijaksana@gmail.com,
[3]bambangari@gmail.con

Abstrak

Dalam ilmu Rijalul Hadis dijelaskan tentang sejarah ringkas para rawi hadis dan riwayat hidupnya, baik dari generasi sahabat, tabi'in maupun tabi'it tabi'in. Dari pengertian tersebut, kedudukan ilmu ini sangat penting, sebab nilai suatu hadis sangat dipengaruhi oleh karakter dan perilaku serta biografi perawi itu sendiri. sebagai contoh nama Muhammad dengan Muhamad, itu adalah nama yang sama meskipun dengan ejaan yang berbeda. Sehingga perlu adanya penelitian untuk menentukan kecocokan nama meskipun dengan ejaan yang berbeda. Pencocokan nama pada penelitian ini menggunakan metode *n-gram* untuk memecah nama menjadi bagian *substring* kemudian dihitung nilai kecocokannya dengan metode *jaccard similarity* dengan nilai *treshold* yang diberikan sebesar >= 0.7 Selain itu, dilakukan perhitungan untuk menilai kinerja dari metode yang digunakan yaitu *n-gram* dan *jaccard similarity* dengan menghitung nilai *precision, recall, f-measure* dan akurasi. Penilaian kinerja ini didapatkan dengan membandingkan hasil yang diberikan oleh sistem dengan *gold standart* yang telah dibuat dan diverifikasi oleh ahlinya. Dari pengujian yang telah dilakukan rata-rata akurasi yang didapatkan sebesar 0.85714286. ini berarti menunjukan sistem yang dibuat sudah baik.

Kata kunci : ***n-gram, jaccard similarity, precision, recall, f-measure***

Abstract

In the science of Rijalul Hadith it is explained about the concise history of the hadith narrators and their biographies, both from the generation of friends, tabi'in and tabi'it tabi'in. From this understanding, the position of science is very important, because the value of a hadith is strongly influenced by the character and behavior and the biography of the narrator itself. as an example of Muhammad's name with Muhammad, that is the same name even though with a different spelling. So there needs to be research to determine the name match even with different spellings. Name matching in this study uses the n-gram method to break the name into a substring, then the suitability value is calculated with the jaccard similarity method with the given threshold value of >= 0.7. In addition, a calculation is performed to assess the performance of the method used is n-gram and jaccard similarity by calculating the values of precision, recall, f-measure and accuracy. This performance assessment is obtained by comparing the results provided by the system with gold standards that have been created and verified by experts. From the tests that have been done, the average accuracy obtained is 0.85714286. this means showing the system is already good.

Keywords: ***n-gram, jaccard similarity, precision, recall, f-measure***

1. Pendahuluan

Latar Belakang

Ilmu Rijalul Hadis ialah ilmu untuk mengetahui para perawi hadis dalam kapasitasnya sebagai perawi. Dalam ilmu Rijalul Hadis dijelaskan tentang sejarah ringkas para rawi hadis dan riwayat hidupnya, baik dari generasi sahabat, tabi'in maupun tabi'it tabi'in. Dari pengertian tersebut, kedudukan ilmu ini sangat penting, sebab nilai suatu hadis sangat dipengaruhi oleh karakter dan perilaku serta biografi perawi itu sendiri [6]. Nama-nama perawi hadis bisa kita ketahui di dalam hadis Shohih Bukhori maupun hadis yang lainnya. Nama perawi biasanya ditulis dalam bahasa *Arab*, ketika nama perawi dalam tulisan bahasa *Arab* ditulis ke dalam tulisan latin, ejaan penulisannya bisa berbeda-beda. Dari ejaan yang berbeda-beda itu bisa dikatakan itu nama perawi yang sama atau bisa dikatakan itu nama perawi yang berbeda, sehingga bisa mempengaruhi kedudukan suatu hadis. Sebagai contoh nama "Muhammad" dengan "Mochammad". Penelitian ini dibuat untuk mencocokan nama perawi yang diinputkan dengan nama perawi yang berada di terjemahan Hadist Shohih Bukhori. Dalam penelitian ini menggunakan metode *n-gram* dan

untuk mengukur tingkat nilai kecocokan dengan menggunakan metode Jaccard Similarity. Pada proses N-gram nama dipecah menjadi bentuk *bigram* , sebagai contoh nama Muhammad di pecah menjadi bentuk Bigram menjadi Mu, uh, ha, am, mm, ma, ad. Setelah itu diukur nilai tingkat kecocokannya menggunakan koefisien kesamaan *jaccard similarity*. Pengukuran kinerja sistem ini mencari nilai Precision, Recall dan F-measure. Apabila nilai dari ketiga penilaian ini menunjukan hasil yang tinggi, menandakan bahwa sistem memiliki kemampuan yang baik.

Topik dan Batasannya

Adapun perumusan masalah sesuai dengan latar belakang diatas adalah untuk mengetahui bahwa nama yang dicari itu sama meskipun ejaannya berbeda dengan nama yang dicocokan dan mencari nama yang mempunyai nilai kecocokan yang tinggi. Pada penelitian ini mempunyai batasan masalah yaitu inputan *query* berupa satu kata nama, dataset yang digunakan berupa Hadist Shohih Bukori no 1-100.

Tujuan

Adapun tujuan dari penelitian ini adalah mengetahui bahwa nama yang dicari tersebut sama dengan nama yang dicocokan dengan menggunakan metode *n-gram* dan *jaccard similarity*. Menghitung seberapa mirip nama tersebut dan mengetahui nilai kualitas sistem pencarian nama yang dicocokan dengan menggunakan *precision*, *recall*, *f1-score* dan akurasi.

2. Studi Terkait

2.1 Hadis

Hadis adalah sesuatu hal tentang nabi yang berisi tentang semua yang dilakukan atau berkaitan oleh nabi dapat berupa segala hal tentang nabi atau bisa juga tentang kehidupan nabi sebelum atau sesudah menjadi nabi. Hadist dapat berupa sesuatu yang menjadi referensi yang kemudian diterapkan, dilakukan atau dituliskan kembali [1].

2.2 Pedoman Transliterasi Aksara Arab ke Latin

Panduan alih aksara dari huruf *Arab* ke huruf latin dalam ejaan bahasa Indonesia diatur dalam surat Keputusan Bersama Menteri Agama dan Menteri P dan K nomor 158 tahun 1987 - Nomor:0543 b/u/1987. Pedoman ini disusun untuk menunjukan perbedaan, supaya perbedaan tersebut dapat dipahami. Banyaknya variasi dalam penulisan kata dari bahasa *Arab*, hendaknya kita mengutamakan kata populer dalam penggunaanya. Didalam lampiran pada penelitian ini akan dilampirkan pedomoan transliterasi aksara arab ke latin sesuai dengan surat Keputusan Bersama Menteri Agama dan Menteri P dan K nomor 158 tahun 1987 - Nomor:0543 b/u/1987 [7].

2.3 Name Matching

Name matching ini bukan hanya merupakan komponen di proses *"identity matching"* (Pencocokan Identitas). Dua data yang ditunjukkan dapat berbeda namun ditunjukkan bahwa nama itu sama yaitu antara "Barack" dan "Husein". Nama keluarga dapat juga menjadikan bahwa kedua orang tersebut merupakan orang yang sama contohnya dari keluarga yang sama yaitu "Obama". Dalam perbandingan nama belum tentu cukup untuk membuktikan identitas itu sendiri [5]. *Name matching* adalah salah satu unsur yang merupakan *"identity matching"* (pencocokan identitas) yang berdampak langsung pada pergeseran yang terjadi ke arah *linguistic matching* yang menjadikan teks berfokus ke cara nama yang dicocokkan. *linguistic matching* adalah bidang yang relatif baru berdasarkan pengetahuan yang diambil dari ilmu linguistik. Setiap aturan *linguistic matching* (pencocokan linguistik) memiliki dua fungsi yaitu fungsi yang cocok dan fungsi analitis. Contohnya adalah peraturan yang diperlukan untuk mencocokan dua nama yaitu "Yeltsin" dan "Jelzin". Nama juga hanya memiliki satu identitas yang mungkin juga bisa dicocokkan dengan pencariaannya [5].

2.4 N-Gram

Metode N-gram didasarkan pada ide membagi *string* karakter ke dalam rantai kecil atau disebut dengan *substring*, dengan panjang yang telah ditentukan sebelumnya. Seringkali, rantai kecil atau *substring* diatur dengan panjang tiga karakter dan dikenal sebagai *3-gram* atau *trigrams*. Sebagai contoh nama Muhammad dapat dibagi menjadi 6 *trigram* yaitu Muh, uha, ham, amm, mma dan mad. Dalam kasus lain, nama dapat dibagi menjadi rantai kecil atau *substring* menjadi rantai dua karakter (*bigrams*) atau empat karakter (*tetagrams*). sistem umum untuk metode *n-gram* adalah mendeteksi bahasa dalam teks yang lebih panjang. *n-gram* juga digunakan dalam konteks pencocokan nama, yang tujuannya adalah untuk menentukan jumlah *n-gram* konkuren dalam sepasang

nama tertentu. Jika proporsinya mencukupi dari *n-gram* adakah sama, maka nama-nama itu dianggap cocok. Dalam contoh nama-nama fonetis yang mirip Muhammad dan Muhamad, itu bisa dijadikan *bigrams, trigrams* dan *tetagram*. Jumlah n-gram concurring antara dua nama akan lebih tinggi menggunakan *bigrams* dan lebih rendah jika menggunakan *tetagrams*. [3] Bisa dilihat pada gambar dibawah ini

Tetagrams					
Muhammad	Muha	uham	hamm	amma	mmad
Muhamad			hama	amad	

Trigrams						
Muhammad	Muh	uha	ham	amm	mma	mad
Muhamad				ama		

Bigrams							
Muhammad	Mu	uh	ha	am	mm	ma	ad
Muhamad							

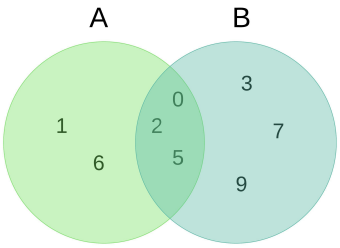
Gambar 1. contoh tetagrams, trigram dan bigrams

2.5 Jaccard Similarity

Jaccard Coeficient adalah salah satu metode yang dipakai untuk menghitung similarity antara dua object (items). Kesamaan Jaccard menggunakan ukuran pembagian dari kedua objek A dan B. [4]

$$JS(A,B) = \frac{|A \cap B|}{|A \cap B| + |A \triangle B|} = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

Sebagai contoh ada dua set A = 0, 1, 2, 5, 6 dan B = 0, 2, 3, 5, 7, 9. Seberapa mirip set A dan Set B ?
Kesamaan Jaccard dapat didefinisikan di bawah ini.



Gambar 2. Perhitungan *Jaccard Similarity*

$$JS(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|0,2,5|}{|0,1,2,3,5,6,7,9|} = \frac{3}{8} = 0.375 \tag{2}$$

2.6 Matching Quality

Pengukuran kinerja sistem dapat dilakukan dengan mengetahui nilai Precision, Recall dan F-measure. Apabila nilai dari ketiga penilaian ini menunjukan hasil yang tinggi, menandakan bahwa sistem memiliki kemampuan yang baik. Ada kategori yang akan sesuai dengan kecocokan dari dua nama. TP (*True Positive*) adalah pasangan nama yang benar-benar keduanya mirip. FP (*False Positive*) adalah pasangan nama yang telah diklasifikasikan tetapi sebenarnya tidak cocok, ini juga bisa disebut dengan kecocokan palsu. TN (*True Negative*) adalah pasangan nama yang telah diklasifikasikan tidak cocok dan memang benar-benar tidak cocok. FN (*False Negative*) adalah pasangan nama yang telah diklasifikasikan cocok tetapi sebenarnya tidak cocok. Dari kategori tersebut akan dihitung nilai [2]

- *Precision*, merupakan tingkat ketepatan antara informasi yang dimiliki oleh pengguna dengan hasil jawaban dari sistem. Maka perhitungan *precision* dinotasikan pada persamaan dibawah ini.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

- *Recall*, merupakan tingkat keberhasilan sistem yang dapat menemukan kembali sebuah informasi. Maka perhitungan *recall* dinotasikan pada persamaan dibawah ini.

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

- *F-measure*, ukuran yang menggabungkan *precision* dan recall, Maka perhitungan *f-measure* dinotasikan pada persamaan dibawah ini

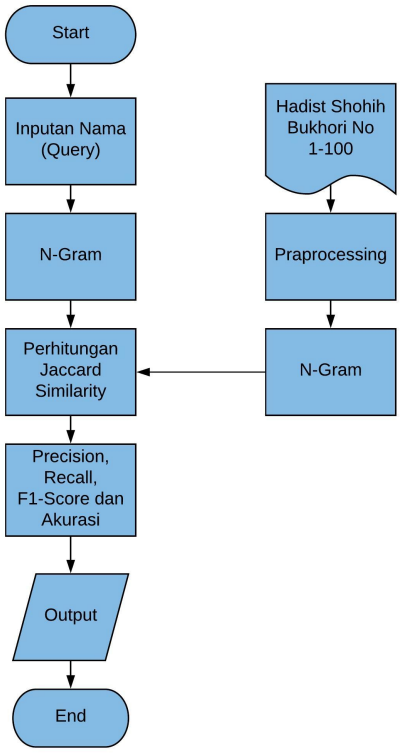
$$F\text{-}measure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5}$$

Dengan memanfaatkan TP (*true positive*), FP (*false positive*), FN (*false negative*), TN (*true negative*) yang ada pada *precision and recall*, nilai perhitungan akurasi bisa didapatkan. Adapun cara menghitung akurasi [**?**] yaitu

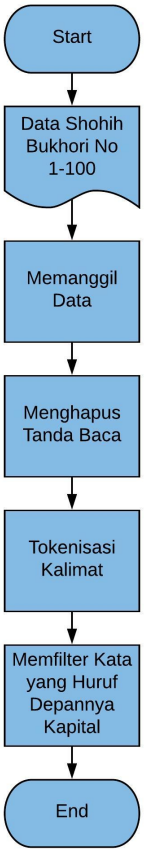
$$Acurracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

3. Sistem yang Dibangun

Dalam penilitian ini akan dibangun sebuah sistem yang dapat memenuhi tujuan dari penelitian tugas akhir ini. Penelitian ini menggunakan dataset Hadist Shohih Bukhori No 1-100 yang sudah diterjemahkan ke dalam bahasa indonesia yang sudah disimpan dalam bentuk *file* yang tiap *file*-nya berisi satu nomor Hadist Shohih Bukhori. Pada awalnya, sistem akan meminta inputan dari *user* yang merupakan nama orang. Selanjutnya inputan dari *user* diproses ke dalam metode *n-gram*. Pada proses *praprocessing* sistem memanggil *file* dataset, kemudian sistem menghapus tanda baca dan memisah-misah kalimat menjadi perkata kemudian diambil kata yang mempunyai huruf depannya kapital. Selanjutnya, hasil kata yang didapatkan dari proses *praprocessing* akan masuk ke dalam proses *n-gram*. Setelah inputan dan hasil kata dari proses *praprocessing* sudah dalam berbentuk *n-gram (bigrams)* akan dilanjutkan ke dalam proses perhitungan kecocokan dengan metode *jaccard similarity*. Selanjutnya dilakukan perhitungan nilai *precission*, *recall*, *f1-score* dan *accurancy*. Hasil *output* yang dihasilkan berupa nama yang mirip dengan inputan dari *user*, nilai *precission*, *recall*, *f-measure* dan akurasi. Berikut dibawah ini adalah gambaran umum sistem.



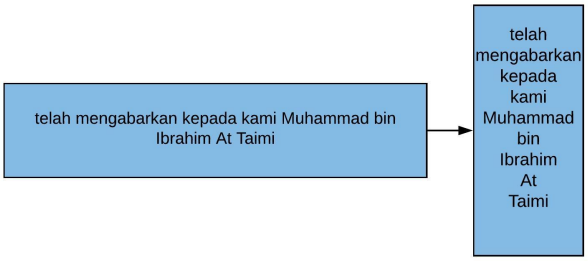
Gambar 3. Gambaran Umum Sistem



Gambar 4. Proses Praprocessing

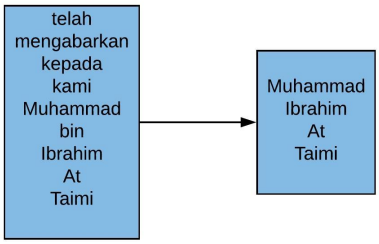
Proses *preprocessing* data yang lebih detail akan dijelaskan dibawah ini.

1. Data Hadist Shohih Bukhori No 1-100 dimasukan ke dalam *file* .txt secara manual. Setiap satu *file* .txt berisi satu nomor Hadist Shohih Bukhori.
2. Memanggil data yaitu data yang sudah berbentuk *file* .txt dipanggil oleh sistem
3. Menghapus tanda baca yang ada pada Hadist Shohih Bukhori.
   - Contoh : telah mengabarkan kepada kami Muhammad bin Ibrahim At Taimi, bahwa dia pernah mendengar Alqamah bin Waqash Al Laitsi berkata; saya pernah mendengar Umar bin Al Khaththab
   - Hasil : telah mengabarkan kepada kami Muhammad bin Ibrahim At Taimi bahwa dia pernah mendengar Alqamah bin Waqash Al Laitsi berkata saya pernah mendengar Umar bin Al Khaththab
4. Memisahkan kalimat menjadi sebuah kata seperti dibawah ini



Gambar 5. Memisah menjadi kata

5. Memfilter kata yang depannya menggunakan huruf kapital seperti dibawah ini



Gambar 6. Mengambil kata yang huruf depan kapital

6. Mencocokan dengan nama inputan (*query*) dengan metode *n-gram* dan *jaccard similarity* yaitu dengan inputan Muhammad.
   - Kedua nama akan masuk ke dalam proses *n-gram*, yang digunakan adalah bentuk *bigram*. hasilnya pada gambar dibawah ini.

Bigrams							
Muhammad	Mu	uh	ha	am	mm	ma	ad
Muhamad							

Gambar 7. Bentuk *bigrams*

   - Selanjutnya masuk ke dalam proses perhitungan *jaccard similarity* dengan perhitungan dibawah ini.

$$JS(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|Mu, uhs, ha, am, ad|}{|Mu, uh, ha, am, mm, ma, ad|} = \frac{6}{7} = 0.857142857 \quad (7)$$

7. Mencari *precision, recall, f1-score* dan akurasi dari banyaknya kemunculan kecocokan yang *thresholdnya* >= 0.7 sampai dengan 1 dengan nama yang ada di *gold standart*

4. Evaluasi

4.1 Skenario dan Hasil Pengujian

Pengujian yang dilakukan dalam penelitian ini bertujuan untuk mengetahui nilai kecocokan dari dua nama. Program akan akan mencocokan hasil *query input* dari *user* terhadap dataset yang berupa Hadis Shohih Bukhori No 1-100. *Input* maupun *output* sistem berupa pasangan nama perawi arab dari transliterasi latin. Penelitian ini menggunakan metode *n-gram* dan *jaccard similarity*. Diharapakan dengan pengujian ini akan diperoleh hasil dari *precision, recall, f1-score* dan akurasi. Tahapan skenario pengujian akan dijelaskan sebagai berikut.

1. Inputan nama (*query*) didapatkan dari kuisioner yang telah diisi oleh responden
2. Sistem akan mengeluarkan hasil dari inputan nama (*query*) berupa nama yang ada didalam dataset Hadis Shohih Bukhori No 1-100
3. Diberikan nilai *threshold*, dimana jika nilai kemiripan nama yang dikeluarkan memiliki nilai >= 0.7 maka dianggap memiliki kemiripan dan jika nilai kemiripan < 0.7 maka dianggap tidak memiliki kemiripan
4. Didapatkan nama dari hasil keluaran sistem yang berdasarkan pada *threshold* yang telah ditentukan
5. Sistem akan menampilkan nilai kemiripan dan jumlah kemunculan nama, jumlah kemunculan nama di sistem akan dicocokan dengan jumlah yang ada di *gold standart*. *Gold Standart* yang telah dibuat sudah diverifikasi oleh DR H Agus Suyadi Raharusun Lc, M.Ag selaku dosen Fakultas Ushuluddin, UIN Sunan Gunung Djati Bandung.
6. Sistem menampilkan nilai *precision, recall, f1-score* dan akurasi

Dari skenario pengujian didapatkan hasil pengujian seperti gambar dibawah ini.

No	Nama Arab	Query	Goldstandart	Jumlah di Gold Standart	Jumlah di Sistem	Jaccard Similarity	Precision	Recall	F1-Score	Akurasi
1	أَبُوْ	Abu	Abu	145	145	1	1	1	1	1
2		Abuu			0	0	0	0	0	0
3		Abū			0	0	0	0	0	0
4	عَائِشَة	Aisyah	Aisyah	13	13	1	1	1	1	1
5		Aaisyah			0	0	0	0	0	0
6		Aissyah			13	0.833333333	1	1	1	1
7	قَتَادَة	Qotadah	Qotadah	9	9	1	1	1	1	1
8		Qotaadah			9	0.857142857	1	1	1	1
9		Qotaadata			9	0.75	1	1	1	1
10	مُحَمَّد	Muhammad	Muhammad	52	52	1	1	1	1	1
11		Muhammadu			52	0.875	1	1	1	1
12		Muhamad			52	0.857142857	1	1	1	1
13	عُمَر	Umar	Umar	21	21	1	1	1	1	1
14		`Umar			21	0.75	1	1	1	1
15		Umaro			21	0.75	1	1	1	1
16	بِلَال	Bilal	Bilal	4	4	1	1	1	1	1
17		Bilaal			4	0.8	1	1	1	1
18		Bilall			4	0.8	1	1	1	1
19	عَبْد الله	Abdullah	Abdullah	82	82	1	1	1	1	1
20		Abdullahi			82	0.875	1	1	1	1
21		Abdullaah			82	0.875	1	1	1	1
21	Rata-rata					0.762981859	0.857142857	0.857142857	0.857142857	0.85714286

Gambar 8. Hasil Pengujian *threshold* 0.7

4.2 Analisis Hasil Pengujian

Hasil pengujian ini menunjukan adanya hasil nilai kecocokan yang beragam dengan *threshold* yang telah diberikan yaitu >= 0.7 sampai dengan 1. Pada gambar 9 hasil pengujian diatas nama inputan Abuu dicocokan dengan nama Abu yang ada di *gold standrt*, sistem tidak menampilkan hasilnya karena nilai kecocokan kurang dari *threshold* yang sudah ditentukan. Kemudian untuk inputan Abu dicocokan dengan nama Abu yang ada di *threshold* menapilkan hasil nilai kecocokan yang tinggi yaitu 1 dan jumah kemunculan di sistem sebanyak 145 sama dengan jumlah yang ada di *gold standart*. Kemudian untuk pengujian dengan inputan Bilaal dengan nama Bilal yang ada di *gold standrt* yaitu 0.8 dan jumlah kemunculan disistem sebanyak 4 sama dengan di *gold standrt*. Dari pengujian diatas dengan *query* yang didapatkan dari responden yang berbeda-beda, sistem ini menghasilkan nilai rata-rata kecocokan sebesar 0.762981859, nilai rata-rata *precision* sebesar 0.857142857, nilai rata *recall* sebesar 0.857142857, nilai rata-rata *f1-score* sebesar 0.857142857 dan nilai rata-rata akurasi sebesar 0.857142857. Dari nilai yang didapatkan tersebut menunjukan bahwa sistem ini sudah cukup baik karena nilai yang didapatkan tinggi.

5. Kesimpulan

Kesimpulan yang bisa diambil dari hasil pengujian dan analisis diatas adalah mencari nama sesuai dengan varian nama (*query*) yang didapatkan dari responden, bahwa sistem sudah mendapatkan nilai *precision, recall, f1-score* dan akurasi yang tinggi, rata-rata akurasi sebesar 0.857142857. ini menunjukan bahwa sistem sudah baik. Saran untuk pengembangan selanjutnya adalah perlu adanya pembuatan korpus nama periwayat hadis untuk meningkatkan akurasi dari program ini.

Daftar Pustaka

[1] S. M. Al-Qaththan. *Pengantar Studi Ilmu Hadits*. Pustaka Al Kautsar, 2012.

[2] M. ALIFIKRI. *INDONESIAN NAME MATCHING USING MACHINE LEARNING SUPERVISED APPROACH*. Telkom University, 2017.

[3] V. M. Bertrand Lisbach. *Linguistic Identity Matching*. www.springer.com, 2013.

[4] U. o. U. Jeff M. Phillips. *Data Mining*. Springer, 2015.

[5] B. Lisbach and V. Meyer. *Linguistic identity matching*. Springer, 2013.

[6] M. D. Online. *Definisi dari Rijalul Hadis Atau Rawi Hadis*. http://www.jejakislam.com/2017/03/definisi-dari-rijalul-hadits-atau-rawi-hadits.html, Diakses 8 Agustus 2018.

[7] Wikipedia. *Pedoman alih aksara Arab ke Latin*. https://id.wikipedia.org/wiki/, Diakses 5 Agustus 2018.