Akreditasi Ristekdikti, No:
30/E/KPT/2019 (Sinta 4)
DOI: https://doi.org/10.31294/p.v21i2

Paradigma – Jurnal Informatika dan Komputer,
Vol. 22, No. 2 September 2020
P-ISSN 1410-5063, E-ISSN: 2579-3500

Analysis Name Entity Disambiguation Using Mining Evidence Method

Adelya Astari[1], Moch Arif Bijaksana[2], Arie Ardiyanti Suryani[3]

[1]Telkom University Bandung
e-mail: aadelyaaastarii@student.telkomuniversity.ac.id

[2]Telkom University Bandung
e-mail: arifbijaksana@telkomuniversity.ac.id

[3]Telkom University Bandung
e-mail: ardiyanti @telkomuniversity.ac.id

Abstract *- Hadith is the second guideline and source of Islamic teachings after the Qur'an. One of the most Saheeh hadith is the book of Saheeh al-Bukhaari. Hadith Sahih Bukhari has a chain of narrators, hadith numbers, and contents of different contents. This tradition also has science that discusses the history of the narrators of the hadith called the Science of Rijalul Hadith. In the Sahih Bukhari hadith there are the names of the narrators of the hadith who have the same name, causing obligation between names. That makes it difficult for many ordinary people to understand these ambiguous names because it is not yet known whether the two names are the same person or not. So, it raises the problem of a name ambiguation for ordinary people who cannot distinguish whether the name of the narrator is the same person or not. To solve these problems, a solution is built, namely the disambiguation of names to eliminate the ambiguity of the name by checking the name, hadith number, narrators chain, content topics, circles, countries, and companions of the Prophet that are seen from the 3 last names before the Prophet based on the chain of narrators. Also, the solution is assisted by using a method Mining Evidence with several other approaches, i.e. Association label documents, word association labels, context similarity, cosine similarity, and word2vec to obtain all similarity values between name entities. After the similarity values are obtained, the data are grouped using the Clustering algorithm. This system is expected to be able to produce a good system performance with a confusion matrix based on value precision, recall, and accuracy.*

Keywords: Disambiguation, Entity Name, Mining Evidence, Sahih Bukhari, Similarity

INTRODUCTION

Hadith is the second guideline and source of Islamic teachings after the Qur'an. One of the most authentic hadith books is the book of *Sahih Bukhari* because it requires liqa 'or meeting between the narrator and his teacher. *Sanad* is a series of narratives that deliver death to the Prophet Muhammad. Hadith Sahih Bukhari has *knowledge* that discusses the history of narrators or narrators of traditions called *Ilmu Rijalul Hadith*(Chairulloh et al., n.d.). This Hadith has a chain of narrators, hadith number, and contents of different contents. In the Sahih Bukhari hadith there are the names of the narrators of the hadith who have the same name, giving rise to name ambiguation. It makes many ordinary people who find it difficult to understand with an ambiguous name. Because it is not yet known whether the two names are the same person or not. For example, there is the name Naif 'in Hadith number 2338 and Hadith number 2360(Dia et al., n.d.). Both names are the same name if only seen from the name. However, if examined further, not necessarily the two names are the same person. Therefore, a solution was built, namely *disambiguation* of names to eliminate ambiguity in names by checking the names, hadith numbers, narrators' chains, circles, countries, and companions of the prophet that can be seen from the 3 last names before the prophet based on the narrator's chain. Therefore, the right method as a solution is *Mining Evidence* which consists of several approaches, namely *the association label document, word association label, context Similarity, and word vector*(Farnham & Rowland, 1968) to obtain all the similarity values between name entities. After all the context similarity values are obtained, the data are grouped using the *Clustering algorithm*(Gupitasari, 2019). The *clustering algorithm* is used to group data that have a similarity in character or similarity between name entities. This system is expected to produce a good system performance based on the value *of precision, recall, and accuracy*(Hoffart et al., 2011).

Based on the background description that has been presented, some problems in this final assignment can be formulated such as how to identify the named entity, eliminates the obligation of the *named entity*, and input the dataset in the form of a set of named entity sets taken from the Sahih Bukhari Hadith

Naskah diterima: 2 Juni 2020, direvisi: 20 Juli 2020, disetujui: 31 Agustus 2020

where the output is the name extraction data has been manually labeled and clustering after getting the *similarity* value. Of all the problems that have been made, the limitation on this final project is in the form of name entities that are already available for the extraction of names, data taken from *Sahih Bukhari* and *Wikipedia,* and only uses *Indonesian* and *English*.

The goal to be achieved in this thesis is to build a *dataset* in the form of the extraction of names from the Hadith Sahih Bukhari. Then, the data is eliminated by using the *Mining Evidence* method and *Cosine Similarity*. In addition to knowing the performance of the name *disambiguation* system with the calculation of *precision, recall, and accuracy*. As well as grouping using clustering algorithms from the similarity of characteristics or similarity between name *entities*.

Related Studies

This section contains the results of theoretical studies conducted by the author. The following theories are the theories used and related to this thesis, such as *Named Entity Disambiguation, Mining Evidence, Context Similarity, Cosine Similarity, Document Label Association, Word Label Entity Association, Word2Vec, Confusion Matrix, Clustering* (Farnham & Rowland, 1968).

Name Entity Disambiguation

Disambiguation of a named entity is (Bunescu & Pas, n.d.; Farnham & Rowland, 1968) the process of identifying names that refer to the same *entity* in a context, in order to eliminate the ambiguity of two *entities* of the same name.

Datasets

The data set used was taken from a collection of 101 *Sahih Bukhari's hadiths*(Dia et al., n.d.). Hadith data is manually labeled to be extracted as a dataset. Other data used is taken from *Wikipedia* on a *website* (text input) as much as 50 data.

Mining Evidence

Mining Evidence (Farnham & Rowland, 1968) is a method used as proof of an *entity's name*, whether the same or not based on existing data. *Mining Evidence* functions to increase the value of data performance in the process of eliminating the obligation of named *entities*.

Context Similarity

Context Similarity(Cucerzan, 2007; Farnham & Rowland, 1968) is the process of calculating the value of *similarity* as a comparison of entities in a text. Where the named *entity* is built a text that is related to the word representation. The approach used in this process is the *label word association*. The approach has a relationship between an entity with the word representation or a relationship formation between an *entity* with the word representation. The results of the *entity label word association* in the form of word *vector* or representation of the content of the *hadith*.

Cosine Similarity

Cosine Similarity (Bunescu & Pas, n.d.; Farnham & Rowland, 1968; Guntara, 2019) is a method for calculating the *similarity* between two *vectors* or documents in a *vector space*. The calculation is done by calculating the cosine angle of each pair of word *vectors*. If a vector document with a *vector* query is the same then the document can be seen increasingly in accordance with the *query*. The equation for calculating cosine similarity between two *vectors* is as in equation 1.

$$Sim = \cos(\theta) = \frac{A.B}{\| A \| . \| B \|} \text{...............(1)}$$

Description :
A = document
B = query

Two-word *vectors* in an *entity* that has *similarities* are likely to have high similar values. Similarity-based on the *cosine similarity* method gives a better recommendation if the value of *sim* (a, b) is between 0.0 and 1.0. Value 1.0 indicates that the two *entities* are the same.

Word2Vec

Word2Vec(Gupitasari, 2019) is the process of predicting target words based on their context in order to get the value of similarity *Word2Vec* using *cosine similarity*. *Word2Vec* can process words from very large *datasets* in a relatively short time with better *accuracy*.

Document Label Association

Document Label Association (Farnham & Rowland, 1968) is the process of *labeling* an *entity* that represents the document to get the value of *similarity*. Label provisions can be seen from the topic of the document. *Document-label associations* help express *entity labels* for *labeled documents*.

Label Entity Association

Label Entity Word Association (Farnham & Rowland, 1968) is a process of labeling name *entities* to calculate the estimated value of each word that has been labeled in order to get a *similarity* value. The higher the value, the word has lost its *ambiguity*.

Confusion Matrix

Confusion matrix (Chairulloh et al., n.d.; Ginting et al., n.d.) is one method that can be used to measure the performance of a system with the classification method or referred to as a confusion matrix. The confusion matrix is a representation of the results of the classification process in the form of actual and predicted values. In performance measurement using the confusion matrix, there are 4 terms as a representation of the results of the classification process. The four terms are *True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN)*(Chairulloh et al., n.d.). Table 1 shows the terms from the table *confusion matrix*.

Tabel 1. Confusion Matrix

	Actual	
Prediction	Positive	Negative
Positive	TN	FP
Negative	FN	TP

Sumber : (Ginting et al., n.d.)

Score *True Negative (TN)* is the amount of negative data that was detected correctly, whereas *False Positive (FP)* is negative data but detected as positive data, *True Positive (TP)* is positive data that was detected correctly and *False Negative (FN)* is the opposite of True Positive, so the data is positive, but detected as negative data.

By utilizing the terms of the method confusion matrix is able to get value *precision, recall, and accuracy*. In equations 2,3, and 4 are the Calculations confusion matrix.

Precision

Precision (Chairulloh et al., n.d.; Hoffart et al., 2011) is used to measure the *accuracy* of the system in determining the relevant documents in the search for documents received, in other words the result of *precision* is the value of the ability of the system in making decisions. The formula below is a calculation *precision*.

$Precision = \frac{TP}{TP+FP}$..............(2)

Description :
TP = True Positive
FP = False Positive

Recall

Recall (Chairulloh et al., n.d.; Cucerzan, 2007; Hoffart et al., 2011) is used to measure the *accuracy* of the system in determining relevant documents in search of all relevant documents, in other words the result of recall is the value of the accuracy of the system to get the actual relevant documents. The formula below is a calculation *recall*.

$Recall = \frac{TP}{TP+FN}$..............(3)

Description :
TP = True Positive
FN = False Negative

Accuracy

Accuracy (Farnham & Rowland, 1968) is used to get the *accuracy* of decision making by the system to determine the relevance of documents. As for how to calculate *accuracy*.

$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$..............(4)

Description :
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

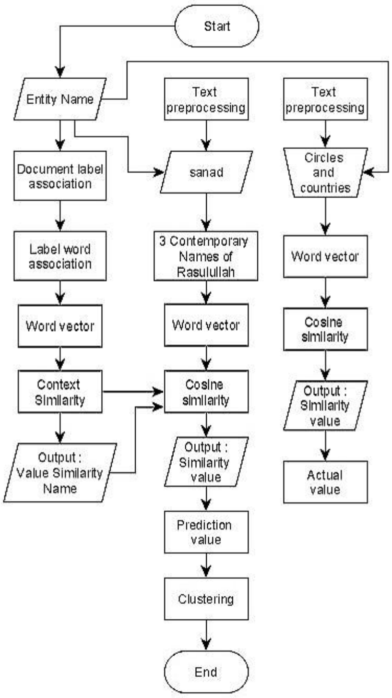
Clustering

Clustering (Gupitasari, 2019) is a method used to group data to produce a data representation that represents a pattern formed by the relationships that exist between the data. *Clustering* is an *unsupervised machine learning* method, where the collection of words in other words that have similar characteristics based on the similarity function for calculating the distance of words. As for several methods of clustering, namely *k-Means.*

K-Means (Gupitasari, 2019)is a method of partition clustering and aims to minimize data objects with a number of k *centroids*.

RESEARCH METHOD

The system built aims to eliminate the *ambiguity* of the same name *entity* using the *Mining Evidence* method and *Cosine Similarity*(Farnham & Rowland, 1968) in order to get the value of *similarity*. Figure 1 is the stages carried out in eliminating the ambiguation of a *named entity*:



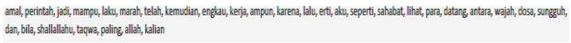
Picture 1. System Overview

The picture above has 3 processes. The first process is the name extraction for each entity obtained from the *Bukhari Hadith*. Extraction is done manually using the *document label association* and the *word association label*. The extraction produces narrators' data. The second process, is the result of extraction in the form of narrator chain data from the first process. The narrator's chain was re-processed to get a friend's name who was a *contemporary* of the Prophet Muhammad. The process of getting it by sorting each named entity that has a chain of narrators with the suffix Rasulullahu'alaihi wassalam. Then, take 3 last names before the Messenger of Allah. The process can be seen in 4.

The third process, is the result of extraction in the form of data between the state based on name *entities*. The results of all three processes, each one is processed using the *cosine similarity* and word *vector* methods to get the similarity value. The *similarity* value is in order to get the actual and predicted values using the confusion matrix. As the results of the hadith data are grouped based on the same name *entity* using the *clustering* and a *scatter plot algorithm*. While the *Wikipedia* data uses a clustering algorithm with *k-means* and *scatter plot* approaches to visualize the grouping.

Context Similarity

This process uses the *word association label*(Farnham & Rowland, 1968) approaches obtained from the results of the *text preprocessing*. Another approach used is to consider the topic of the text of a hadith number.



amal, perintah, jadi, mampu, laku, marah, telah, kemudian, engkau, kerja, ampun, karena, lalu, erti, aku, seperti, sahabat, lihat, para, datang, antara, wajah, dosa, sungguh, dan, bila, shallallahu, taqwa, paling, allah, kalian

Sumber : (Dia et al., n.d.)

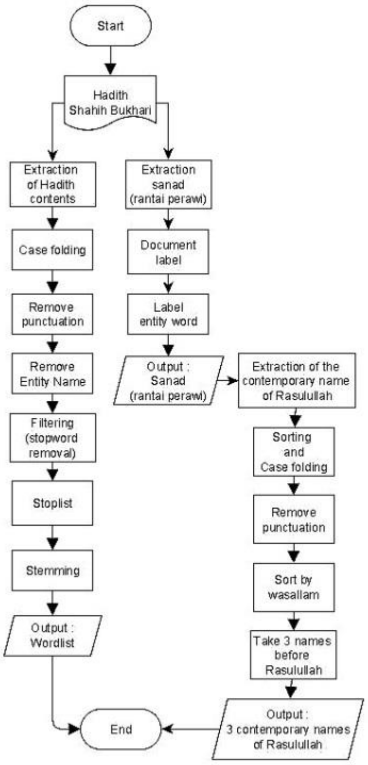
Picture 2. Hadith Data Extraction Result

Picture 2 is an example of the *word association label* extraction results. These results are in the form of unique words that represent *ambiguous* name *entities*, then proceed by using *Cosine Similarity* to get the *context similarity* value with a range of values between 0 and 1.

Text Preprocessing

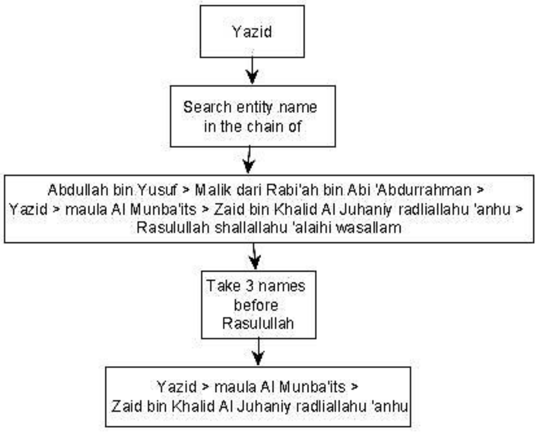
Text preprocessing (Farnham & Rowland, 1968; Guntara, 2019) is the stage of selecting the data to be processed. In Figure are the stages of *text preprocessing* which are divided into three processes.

The first process is the extraction of the name or *word association label* obtained from the Hadith Sahih Bukhari. The process is carried out for each named entity, in order to extract the contents of the hadith in accordance with the number of the hadith. Then a step is made that changes all the letters in the document to lowercase and eliminates punctuation, or referred to as a *folding case*. Next, process *filtering* or *stoplist* by removing words that are less important to connect an entity with the same word representation so as to produce output in the form of unique words from the text of the hadith. The process is called *stemming*.


Start
Hadith Shahih Bukhari
Extraction of Hadith contents
Extraction sanad (rantai perawi)
Case folding
Document label
Remove punctuation
Label entity word
Remove Entity Name
Output : Sanad (rantai perawi)
Extraction of the contemporary name of Rasulullah
Filtering (stopword removal)
Sorting and Case folding
Stoplist
Remove punctuation
Stemming
Sort by wasallam
Output : Wordlist
Take 3 names before Rasulullah
End
Output : 3 contemporary names of Rasulullah

Picture 3. Text Preprocessing Process

The second process is the extraction of the chain of narrators, wherein this process the data is labeled manually, to separate the entity names, hadith numbers, and chain of narrators from the hadith text. From this output, to obtain names that are already unambiguous are taken based on the same-named *entity*. The third process, namely getting a *contemporary* name with the Messenger of Allah, which is obtained from the chain of narrators. This is done based on the science of hadith. Figure 3 is an example explanation of the process of getting 3 names that were *contemporary* with the Messenger of Allah.


Yazid
Search entity name in the chain of
Abdullah bin Yusuf > Malik dari Rabi'ah bin Abi 'Abdurrahman > Yazid > maula Al Munba'its > Zaid bin Khalid Al Juhaniy radliallahu 'anhu > Rasulullah shallallahu 'alaihi wasallam
Take 3 names before Rasulullah
Yazid > maula Al Munba'its > Zaid bin Khalid Al Juhaniy radliallahu 'anhu

Picture 4. The Process of Obtaining The Name of Rasulullah

Tabel 2. Hadith Data Testing Result

Hadith Data Testing Result								
	Hadith Number	Content	Topic	Circles	Country	Contemporary	Prediction	Actual
Abdul Wahid	2326	0,6	0	1	1	1	Y	Y
Abdul Wahid	2316	0,4	0	1	1	1	Y	Y
Abdul Wahid	122	0,5	0	1	1	1	Y	Y
Ibnu Umar	59	0,4	1	1	1	1	Y	Y
Ibnu Umar	80	0,4	0	1	1	1	Y	Y
Ibnu Syihab	5278	0,4	0	1	0,9	1	Y	Y
Ibnu Syihab	2283	0,4	0	1	1	1	Y	Y
Ibnu Syihab	23	0,4	0	1	1	1	Y	Y
Maula	2251	0,5	1	1	1	1	Y	Y
Maula	2249	0,4	1	1	0,9	1	Y	Y
Maula	2256	0,3	1	1	1	1	Y	Y
Nafi’	2338	0,5	1	0.7	1	1	Y	Y
Nafi’	2360	0,4	0	0.75	0	0	N	N
Nafi’	2311	0,3	0	0.4	0	0	N	N
Qutaibah	2409	0,2	0	1	1	1	Y	Y
Qutaibah	2319	0,2	0	0	0	0	N	N
Qutaibah	59	0,3	0	1	1	1	Y	Y
Yazid	2251	0,4	1	1	1	1	Y	Y
Yazid	2249	0,8	1	1	0,88	1	Y	Y
Yazid	2256	0,4	1	1	1	1	Y	Y

From the picture 4, to find out the same name entity friends who are *contemporaries* with the Messenger of Allah or not can be seen by looking at the chain of narrators of the name *entities* ending in the Prophet. Then do the checking by sorting the last 3 names before Rasulullah. Sorting based on the word wasallam. From the results of the last 3 names obtained, compared each *entity* to get the *similarity* value.

Contemporaries

Contemporaries (Chairulloh et al., n.d.)are at the same time or at the same time. In this process an examination of the two *entities* with the same name from the Prophet's companions, circles, and countries of origin obtained from the hadith encyclopedias. In the same name *entity*, friends, circles, and countries are examined by *cosine similarity* to get an output value of 1 if friends, circles, and countries are the same or output 0 if friends, circles, and countries are different.

RESULTS AND DISCUSSION

In this section explains the analysis of test results and system test results.

Test Result

The testing process was carried out using test data totaling 101 data from *the text preprocessing* data of the Hadith Sahih Bukhari(Dia et al., n.d.). From 101 data where each name consists of several of the same names. For example there is the name Na'f in the hadith number 2338, the number hadith 2360, and the number hadith 2311. The name entities are tested with each other to eliminate the obligation of the name.

The following attached table 2 is some of the results of testing that has been done for *Hadith* data. There is a name, hadith number, context of the content, context of the topic, circles, country, contemporary, prediction, and actual. This is an example of some of the results of a *hadith test data* of 100 test data

Other testing processes use *Wikipedia* data as much as 50 data obtained from the *Wikipedia* Market (*text input*). Tests on *wiki* data are not much different from hadith data, both of which aim to eliminate the obligation of a *named entity*. From 50 data where each name consists of several names that have the same prefix. The name cannot be said to be the same even though the prefix name looks the same. So, the writer explores further from the text. For example, there is the name of *Michael Jordan* with *Michael B Jordan* (Farnham & Rowland, 1968). The name *entities* are tested with each other to eliminate name ambiguation with the *cosine similarity* and word *vector* methods to get the value of *similarity* or *similarity* between *entities*.f the hadith contained in the appendix.

Table 3 on the below for *Wikipedia* data consist of number, title, comparison(Comp) between entity names, data of an entity, context name(CN), context of the content(CC), actual(A), and prediction(P). This is an example of some of the results of a *Wikipedia test* data of 50 test data.

Tabel 3. Wikipedia Data Testing Result

Wikipedia Data Testing Result						
	Compare	Data	Context Name	Context of Content	Actual	Prediction
Bryan Cranston	Bryan Spicer	American actor, film	0,5	0,30	N	Y
Bryan Spicer	Bryan Adams	American film, singer	0,5	0,32	N	Y
Bryan Adams	Bryan Cranston	Singer, Actor	0,5	0,69	Y	Y
David R Ellis	David Lynch	Film director, film maker	0,4	0,6	Y	Y
David Lynch	David Fincher	Film director, film maker	0,5	0,7	Y	Y
David Beckham	David Villa	Professional footballer	0,5	0,7	Y	Y
James Gandolfini	James Dean	American actor	0,5	0,6	Y	Y
James Dean	James Woods	American actor	0,5	0,7	Y	Y
John Cusack	John Stamos	American actor	0,5	0,5	Y	Y
John Stamos	John Wayne	American actor	0,5	0,6	Y	Y
John Wayne	John Cusack	American actor	0,5	0,6	Y	Y
Kevin Spacey	Kevin Costner	American actor	0,5	0,6	Y	Y
Kevin Martin	Keith Urban	Singer	0,40	0,49	Y	Y
Peter Coyote	Peter Berg	Director, Actor	0,5	0,665	Y	Y
Robert Aldrich	Robert Duvall	Actor, Director	0,5	0,54	Y	Y
Ribert Duvall	Robert Z. Leonard	American actor	0,40	0,58	Y	Y
Michael Jackson	Michael Bolton	American singer	0,5	0,6	Y	Y
Michael Jordan	Michael B.Jordan	Basketball player	0,8	0,6	Y	Y
Tom Cruise	Tom Hanks	American actor	0,5	0,7	Y	Y
William Hanna	Will Graham	Animator, Character	0	0,55	Y	N

Based on the overall results of the test data, then use the *confusion matrix* to get the value of *precision, recall, and accuracy*. Table 4 results of *Hadith* data. Table 5 results of *Wikipedia* data. The tables are a combination of *predicted* values and *actual* values.

Tabel 4. Result Confusion Matrix Hadith Data

	Actual	
Prediction	Positive	Negative
Positive	16	7
Negative	9	70

The table above shows the results of the matrix of the hadith data. *Truth Positive (TP)* hadith data of 70 names that were predicted to be the same person. *Truth Negative (TN)* Hadith data of 16 names that were predicted to be true were different people. *False Positive (FP)* Hadith data as many as 7 names that are predicted to be wrong are different people in the actual also prediction. *False Negative (FN)* hadith data as many as 9 people who are the same but predicted as different people.

Tabel 5. Result Confusion Matrix Wikipedia Data

	Actual	
Prediction	Positive	Negative
Positive	0	4
Negative	7	40

Table 5 shows the matrix results from *Wikipedia* data. *Truth Positive (TP)* wiki data of 40 names that were predicted to be the same person. *Truth Negative (TN)* Wiki data of 0 names that were predicted to be true were different people. *False Positive (FP)* Wiki data of 4 names that are predicted to be wrong are different people in the prediction as well as actual. *False Negative (FN)* wiki data of 7 same people but predicted as different people.

Analysis of Test Result

The results of this test indicate the existence of the same *name entity* values varying with the given *threshold* of more than 0.4 to 1. The *threshold* applies to the results of the hadith and *Wikipedia* data. The results of this test are obtained from the actual and predicted values to get *precision*, *recall*, and *accuracy*. Attached to table 6 is the result of a systematic evaluation of *Hadith* data and *Wikipedia* data consist of *TN, FP, FN, TP, Precision, Recall, and Accuracy.* (Farnham & Rowland, 1968; Hoffart et al., 2011; Nguyen & Cao, 2008).

Tabel 6. Evaluation Result

	TN	FP	FN	TP	Prec	Rec	Acc
Hadith	16	7	9	70	0.9	0.843	0.843
Wiki	0	4	7	40	0.9	0.85	0.8

Judging from the test, that the evaluation results using confusion matrix, where the value of *precision*, *recall*, and *accuracy* approaching 100% is high because the combined results of the predicted value, the actual value and can also be seen in *Truth Positive (TP),* *False Positive (FP), False Negative (FN)* and *True Negative (TN).*

The results of the confusion matrix show that the values of *precision, recall*, and *accurately* predicted by the system are the same person. These results apply to the data of the Hadith. That is because two entities have the same name if people and countries are the same, it can be ascertained that these two names are contemporaneous with the Messenger of Allah. Therefore, the average actual value can reach more or less 80%.

In *Wikipedia* data, the values of *precision*, *recall*, and *accuracy* are the same person. Because the two name *entities* being compared are thought to have fields and text that are similar, the result of *similarity* is also high.

Factors that influence the predictive value of the Hadith data, namely the value of *similarity* and *threshold*. The resulting *similarity* value is very small because the word list that accompanies each named entity is unique. The *threshold* used for prediction values is more than 0.4 because it is based on experiments of several values, with a *threshold*
0.4 obtained the best prediction results.

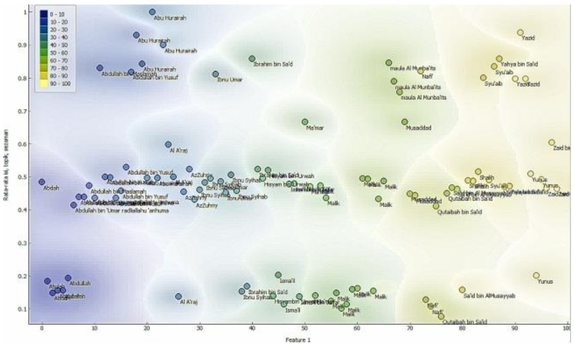
Another factor that influences the predictive value of *Wikipedia* data is the *threshold*. Because the name *entities* in *Wikipedia* only have the same prefix name. Thus, the similarity value for the name is very small.
Thus, additional professional data and text from each *named entity* are used to obtain high *similarity* and *accuracy* values. In this data, the *threshold* provisions used are the same as the date of the Hadith.

From the above test, this system produces an average value of *precision* of 0.9, an average value of *recall* of 0.8, and an average *accuracy* value of 0.8. The value obtained shows that the same *name entity* is the same person and also this system is good enough because the value obtained is high. As a result of the system, the hadith data and wiki data are grouped according to the class each entity uses.

In calculating performance on *Hadith* data and *Wikipedia* data. The calculations used are *precision*, *recall*, *accuracy* and the confusion matrix is the actual value and predictive value. Predictive value is obtained from the combined results of the context value similarity, topic, and narrator contemporaries with the Prophet. The pre-determined and actual values have two outputs, namely the output "Y" for the overall value more than the *threshold* or "N" for the overall value less than the *threshold*. The *threshold* is a level of *similarity* of more than 0.4 obtained from the results of several value experiments. If the overall value is greater than the *threshold*, then the two *entities* with the same name are the same person. If the overall value is less than the *threshold*, then the two entities with the same name are different people. Then for the actual value obtained from the same circles and countries for each *entity name*. If, circles and countries are the same, then it can be ensured that the *entity* is the same as the Prophet.

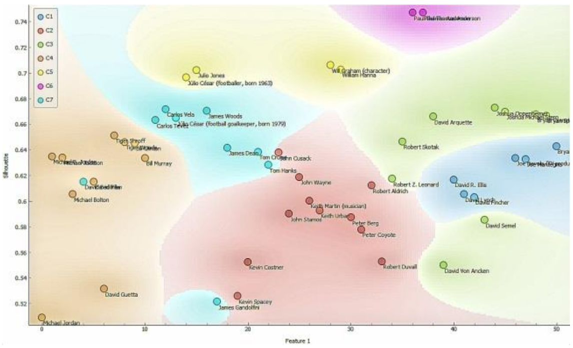
Clustering

The *clustering* process is carried out to classify a *Hadith* data and *Wikipedia* data in order to produce a data representation that is formed due to existing relations between the data. The data used is obtained from the results of the *similarity* that has been done. So that objects that have the same *similarity* are grouped in the same cluster and objects that have different *similarities* are grouped into another *cluster*. Picture 5 is the result of clustering, visualization of the hadith data.



Picture 5. Clustering of Hadith Data

Data it is formed where the collection of words with other words that have similar characteristics based on the *similarity* function for calculating the distance of these words.



Picture 6. Clustering of Wikipedia Data

Picture 6 is the result of the *Wikipedia* data *cluster*. Data grouping is taken based on the same *similarity* value without calculating other values such as actual and predicted values. *Clustering* grouping is assisted with a *k-means* approach which aims to minimize data objects with a number of k *centroids* using *scatter plots* as visualization. So that the grouping can be seen based on the color and similarity results for each *entity* with the same name.

CONCLUSION

The conclusion that can be drawn from the test results is the performance of the system to eliminate the *ambiguation* of a *named entity* from the *Sahih Bukhari Hadith* and *text input* from *Wikipedia Page* using the *cosine similarity* method to obtain a confusion matrix value from the *similarity* results(Farnham & Rowland, 1968). From these results it shows that the system has obtained values of *precision, recall*, *accuracy* with a *threshold* of more than 0.4. Thus, the average *accuracy* obtained is 0.8431372549 for hadith data and 0.8 for *Wikipedia* data. The final data can then be grouped by *Name entity* using a *clustering algorithm*. This shows that the system performance is good. In this test, the data presented are only 101 hadith from the many hadith of Bukhari, and 50 *Wikipedia* texts from many other *Wikipedia*.

Suggestions for further development is the need for the addition of hadith and *Wikipedia* data to further improve the performance of this system. Because the *accuracy* value can vary depending on a lot of data used.

REFERENCE

Bunescu, R., & Pas, M. (n.d.). *Using Encyclopedic Knowledge for Named Entity Disambiguation*.

Chairulloh, M. R., Bijaksana, M. A., & Wahyudi, B. A. (n.d.). *Analisis Name Matching untuk Nama Arab Menggunakan Metode N-gram dan Jaccard Similarity Pendahuluan Studi Terkait Hadis Pedoman Transliterasi Aksara Arab ke Latin*. 1–7.

Cucerzan, S. (2007). Large-scale named entity disambiguation based on Wikipedia data. *EMNLP-CoNLL 2007 - Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, *June*, 708–716.

Dia, L., Maka, M., & Tidak, S. (n.d.). *No Title*.

Farnham, J. E., & Rowland, R. E. (1968). The retention of 133Ba in beagles. ANL-7615. *ANL [Reports]. U. S. Atomic Energy Commission*, 32–38.

Ginting, M. F., Bijaksana, M. A., Wahyudi, B. A., & Telkom, U. (n.d.). *Analisis Pencocokan Nama Arab Dengan Terjemahan Nama Indonesia Menggunakan Metode Jaro Winkler*.

Guntara, F. F. (2019). *Pembangunan Daftar Kata Terkait pada Kosa Kata Al-Qur ' an Berdasarkan Kesamaan Distribusiaonal Proposal Tugas Akhir Program Studi Sarjana Informatika Fakultas Informatika Universitas Telkom Bandung*.

Gupitasari, L. (2019). *Pembangunan Synonym Set untuk Tesaurus Al-Quran dengan Pendekatan Kamus Monolingual dan WordNet Proposal Tugas Akhir Program Studi Sarjana Informatika Fakultas Informatika Universitas Telkom Bandung*.

Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., & Weikum, G. (2011). Robust disambiguation of named entities in text. *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 782–792.

Nguyen, H. T., & Cao, T. H. (2008). Named entity disambiguation on an ontology enriched by Wikipedia. *RIVF 2008 - 2008 IEEE International Conference on Research, Innovation and Vision for the Future in Computing and Communication Technologies*, *00*(c), 247–254. https://doi.org/10.1109/RIVF.2008.4586363

AUTHOR PROFILE

Adelya Astari, the year 2018 graduated from the University of Telkom Informatics Management Diploma Program in Bandung. Currently, he is continuing his bachelor's degree in Informatics Engineering from Telkom University.