Multinomial Naïve Bayes and Rapid Automatic Keywords Extraction for Taharah (Purify) Law Chatbot

Rizkhita Habib Muhtar[1], Yana Aditia Gerhana[2], Dian Sa'adillah Maylawati[3], Cepy Slamet[4], Cecep Nurul Alam[5], Wahyudin Darmalaksana[6], Muhammad Ali Ramdhani[7]
{ kikirizkhita@gmail.com[1], yanagerhana@uinsgd.ac.id[2], diansm@uinsgd.ac.id[3], cepy_lucky@uinsgd.ac.id[4], yudi_darma@uinsgd.ac.id[5], m_ali_ramdhani@uinsgd.ac.id[6]}

Department of Informatics, UIN Sunan Gunung Djati Bandung, Indonesia[1,2,3,4,5,7]
Department of Information and Communication Technology, Asia e University, Malaysia[2,4,5]
Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka
Department of Ilmu Hadist, UIN Sunan Gunung Djati Bandung, Indonesia[3]

Abstract. The aim of this study is to utilize the Natural Language Processing (NLP) technology, one of them is in the form of a chatbot. Chatbot has the ability to answer the questions as a conversational search engine. The methods that used on chatbot's machine are Multinomial Naïve Bayes (MNB) with TF-IDF vectorization to classify the intent, and Rapid Automatic Keywords Extraction (RAKE) to classify the entity. The methods are implemented for thaharah (purify) law as one of Muslim's daily life that cannot be separated from Islamic law. It is important for Muslims to know the thaharah law. The experiments of the methods against chatbot have used a total of 132 data trains and 44 data tests. Results represented by the Confusion Matrix showed the implementation of methods has the overall accuracy 97% with an average precision 90% and recall 97%, which means MNB and RAKE can give the answer well.

Keywords: Chatbot, Multinomial Naïve Bayes, Natural Language Processing, Rapid Automatic Keywords Extraction, *Thaharah*, Text Mining

1 Introduction

Today, in the Industrial 4.0 era, Artificial Intelligent (AI) is utilized in many sectors and case studies for making human activities easier and more efficient. AI is a technology that make computer more intelligent [1]–[3], not only for computing, but also for predicting, detecting, recognizing, analyzing, and doing activities likes humans do. So that AI capable and reliable to solve many case in health [4]–[6], economic and bussiness that common called with Bussiness Intelligent [7], [8], games [9]–[11], education [10], [12], [13], robotics [14], and common AI technology is combined with internet that popular with Internet of Things (IoT) terms [15], [16]. Moreover, for language processing, there is a specific technique in AI that called Natural Language Processing (NLP). NLP is a process for analyze and discover the insight knowledge of data that containing the language, such as text data and speech data [17]–[19]. NLP is a field of AI that studies communication between computers and humans through natural language [20], where one form of NLP application is chatbot.

The ability of chatbot to answer question (question answering) with natural language can be used as a conversational search engine, that is, search engines with conversational language. From the Figure 1, in generally, to analyze the questions from a user, the chatbot identifies the intent and entity of questions first. Intent on chatbot is the purpose that the user wants to know. Whereas, entity represents the specific context of the intent. For example, when a user enters the question "Where is the building A?", The intent or what is meant by the user is "location" while the entity or specific context is "location of building A". The conversational ability of search engines on chatbots should also allow back-and-forth communication or back and forth communication between users and bots. Back-and-forth communication in question is the return question given by the chatbot to get more specific questions from the user. However, from the results of a literature study conducted in several previous studies, the chatbot designed has not been able to back-and-forth communication.

Fig. 1 Work flow of chatbot in quetion answering system

In common and many research NLP with text data is combine with Text Mining technique and also Machine Learning (ML) inside. ML that is also an AI approach that is widely used to imitate human behavior in solving problems or doing automation [21], [22]. However, ML can process the various data type, either structured data, semi-structured data, or unstructured data. Text is unstructured data that usually process with specific approach that called Text Mining, where can implement the ML technique in the mining phase. Chatbot with text data type has been proven better as question answering system using ML with similarity approach than without ML, moreover with the accuracy of the answer around 90% [23]. Many ML algorithm that can be used either for unsupervised learning (clusterisation), supervised learning (classification) or semi-supervised learning (clustering combined with classification). However, for this research classification is suitable method. Classification is the method in ML that is used to classify data objects as one of the categories (classes) that have been previously defined [20], [22]. Classification is used by machines to sort objects based on certain characteristics, as humans try to distinguish objects from one another.

From many algorithm for classification, Naive Bayes is the basic and commonly used. The Naive Bayes algorithm uses probability and statistical methods found by scientists named Thomas Bayes in the 18th century. Naive Bayes algorithm works more optimally than other algorithms [24]. In addition, the Naive Bayes algorithm achieved better accuracy results than the Support Vector Machine algorithm [25]. There are several classification models from Naive Bayes, including Multinomial Naive Bayes (MNB) and Multivariate Bernouli Naive Bayes (MBNB) . In the previous study, MNB algorithm for classification can handle large amounts of vocabulary quite well as so that the MNB algorithm can be used on the chatbot engine to identify the intent of the question entered [26]. Whereas to identify entities, chatbot machines can use other ML approaches, including the Keyword extraction method. Based on previous research,

the Rapid Automatic Keyword Extraction (RAKE) works faster than other keyword extractions, such as TextRank [27].

Several related research about chatbot, among others: (1) Designing a chatbot Application Information on Bandung Tourism Objects with a Natural Language Processing Approach [28]. In this study the development of a chatbot application for information on tourism objects in the city of Bandung was carried out. The aim is to make it easier for visitors to find the information needed, namely information about the address of tourist attractions; (2) Designing a chatbot Student Information Center Using Artifical Intelligence Markup Language (AIML) as a Web-Based Virtual Assistant [29]. This study aims to build a chatbot that has the goal of being a Virtual Assistant that provides information to students through data stored on the system that contains information about informatics engineering courses and the addition of new knowledge if stored data is not found; (3) Analysis and Design of Chatbot Reminder Interaction with User-Centered Design [30]; (4) An Arabic Question Classification Method Based on New Taxonomy and Continuous Distributed Representation of Words [31]. The system under study provides precise answers to questions that are formulated in natural languages for Arabic; (5) A Hindi Question Answering (QA) System using Machine Learning Approach [23]. This study discusses the implementation of the Hindi QA system which was developed using a Machine Learning approach.

Chatbot technology using NLP and ML approach that explained above will be implemented for *thaharah* (purify) as one of Islamic law. Indonesia is a country with the majority of the people are Muslim. Muslims in their daily lives cannot be separated from Islamic law, one of which is in the context of *thaharah.* In language, the word *thaharah* comes from the word "*thahara-yathuru-thaharatan*" which means to purify [1]. The law of *thaharah* is obligatory, and not valid if it is not in accordance with the Shari'a of the Qur'an and the Sunnah. There are many media that provide information about *thaharah* such as formal books, tutorial videos, articles and so on. But nowadays, technologies that support computers are acting smarter, including AI. So that NLP as one of AI technology is utilized to provide information to Muslims about *thaharah* using chatbot. Therefore, from the explanation and previous reseach above, this research aims to build chatbot for *thaharah* using MNB and RAKE algorithm as ML algorithm and part of NLP and AI technology.

2 Methodology

2.1 Waterfall Software Development Life Cycle

The method used in this research (that described in the Figure 2) has main activities based on Waterfall Software Development Life Cycle (SDLC). Waterfall SDLC is used not only for rigid or critical system, but also for simple system that all of requirements have been clearly defined [32], [33]. Activities of Waterfall SDLC among others: Requirement elicitation, Analysis, Design, Implementation, Testing, Deployment, and Maintenance. In this research all of Waterfall SDLC is conducted and adapted to the need to build chatbots for *thaharah*. Such as in the Requirement elicitation phase is done collecting data and knowledge with scraping mobile application and studying the documents that related with Islamic Law for *thaharah*. Then, in Analysis and Design phase is analyzed, elaborated, and modeled the functional requirement of the system using structured software modeling. Structured software modeling that used is Data Flow Diagram (DFD). For implementation phase is used Python language

programming for implementing NLP process, MNB, and RAKE. For the Testing phase has two scenarios, which is blackbox testing to evaluate the functionality of the chatbot, and evaluation texting for the performance of chatbot (MNB and RAKE) with calculate the Precision, Recall, and Accuracy value. Then, the result of testing evaluation is represented using Confusion Matrix. Last, for Deployment, the system can be accessed online, and Maintenance phase is flexible if there is new question and knowledge that is inputed into chatbot.

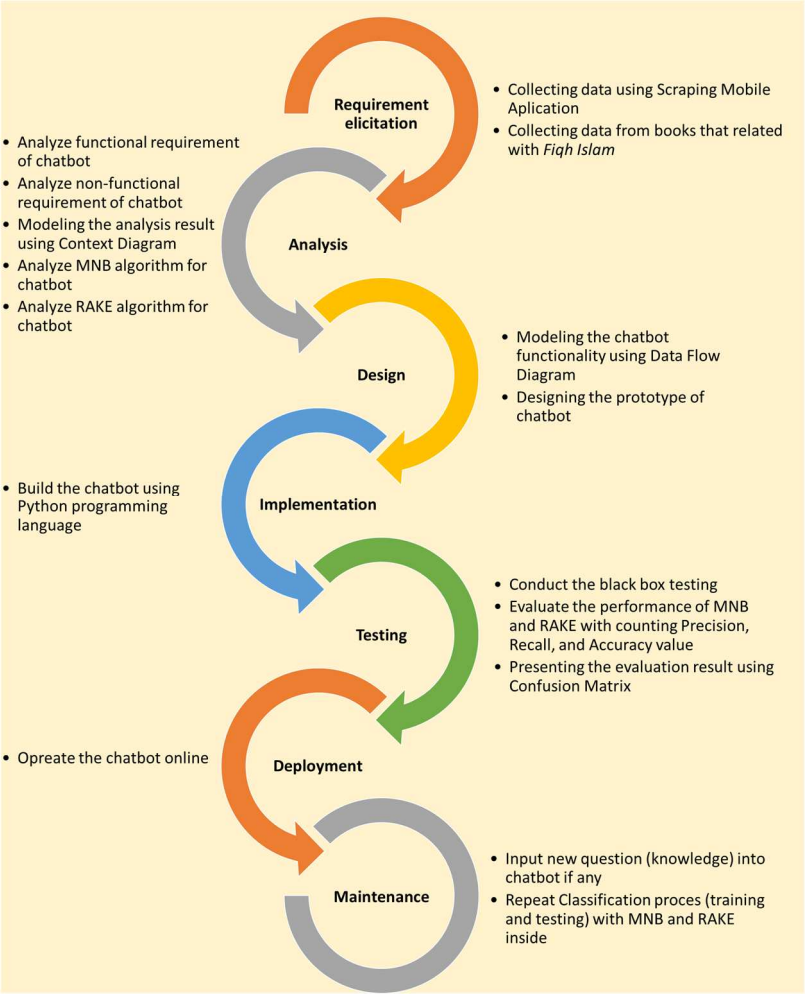

Fig. 2 Work flow of chatbot in quetion answering system

Scraping Mobile Aplication for collecting data about *thaharah* from "Kitab Fikih Imam Syafi'i" application that developed by Ely's Studio. While, collecting data from books that related with *Fiqh* Islam especially *thaharah*, among others: (1) "125 Masalah Thaharah", that discuss the common problems of *thaharah* by an expert that often faced by his interpreters in this book, so that the majority of data is taken from this book [34]; (2) "Fiqih Taharah" is one of special book because the author is a scholar who has been recognized for the author's

expertise in the Islamic world, both *fiqh*, *aqeedah* and others [35]; and (3) "Ilmu Fiqih (*Safinatunnaja*) Berikut Penjelasannya" that contain translation the book of *Safinatunnaja*, and this book is widely studied everywhere, both in Islamic boarding schools, *majlis ta'lim* (islamic studies), and the mosques [36].

In the Implementation phase, to build the chatbot there are several text pre-processing that conducted before running MNB and RAKE algorithm. Text data is unstructured data that must be transformed into structured data, so that the text pre-processing is an importan phase to get good result [37], [38]. Text pre-processing means to prepare initial text data that is still various to be used as regular data that can be subjected to or applied by several existing text mining methods [39]. Those text pre-processing process among others: (1) Tokenizing, tokenization process is useful for breaking every sentence from all knowledge documents into terms by using tab delimiters and space characters [40], [41]; (2) Case Folding, which is a step that changes all the letters in the document into letters of uniform size, usually lowercase letters [42]; (3) Filtering and Stopwords Removal, which is removing all of the stopwords (unimportant words or words that do not store meaning) that frequently appear in document and will affect the quality of result [43]; then (4) Stemming process, which is stemming process is useful to change a word into its basis so that will increase the quality of result [44].

2.2 Multinomial Naive Bayes (MNB)

MNB assumes independence between the appearance of words in a document, regardless of the word order and context of information in sentences or documents in general. In addition, this method takes into account the number of occurrences of words in the document [45]. The multinomial model takes into account the frequency of each word that appears in the document. For example there are documents d and sets of classes c. To calculate the class from the document d, it can be calculated using the formula (1) [46]. While, prior probability from class c is calculated using formula (2). Then, in statistics, smoothing additives, also called laplacian smoothing, or lidstone smoothing, are techniques used to smooth categorical data is used to find the probability of the n^{th} word is determined using the laplacian smoothing technique (formula (3)).

$$\text{P(c|term dokumen d)} = \text{P(c)} \times \text{P(t1 |c)} \times \text{P(t2 |c)} \times \ldots \times \text{P(tn |c)} \quad (1)$$

Where,
P(c) = prior probability from class c
tn = the n^{th} of word in documen d
P(c|term dokumen d) = probability of document will be classified in class c
P(tn |c) = Probability the n^{th} word that found as class c

$$P(c) = \frac{Nc}{N} \quad (2)$$

Where,
Nc = total class c from all of documents
N = total documents

$$P(tn \mid c) = count(tn\ x\ c) + 1\ /\ count(c) + |V| \quad (3)$$

Where,
$count(tn\ x\ c)$ = total of term tn which is found in all of training data and classified as class c

$count(c)$ = total of terms in all of training data that classified as class c
V = total of all terms in training data

2.3 Rapid Automatic Keyword Extraction (RAKE)

Keyword extraction is quite important in text mining applications. Generally, there are two approaches used in the keyword extraction algorithm model, which is supervised learning that requires training data and unsupervised learning that does not require training data. The unsupervised learning approach is divided into 4 categories, namely graph-based ranking, topic-based, simultaneous learning and language modeling [47]. RAKE algorithm is one of methods that use unsupervised learning approaches. In the keyword extraction process, RAKE uses stoplist to get a list of candidate keywords from a document. Then the candidate-keywords score is calculated using graph-based ranking. First, RAKE will be counted word frequency (freq(w)), word degree (deg(w), and *keyword* score with formula (4).

$$Keyword\ Score\ =\ deg(w)\ /\ freq(w) \quad (4)$$

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

The TF-IDF method is a method for calculating the weight of each word that is most commonly used in information retrieval. This method is also known to be efficient, easy and has accurate results [48]. The Term Frequency-Inverse Document Frequency (TF-IDF) method is a method of giving the weight of a term to a document. This TF-IDF is a statistical measure used to evaluate how important a word is in a document or in a group of words. For a single document each sentence is considered a document. The frequency of occurrence of words in the given document shows how important the word is in the document [49], [50]. At Term Frequency (TF), there are several types of formulas that can be used, namely [51]: binary *tf*; raw *tf*, logaritpic *tf*; and normalization *tf*. In this research uses mormalization *tf* with formula (5). Whereas, IDF can be counted using ormula (6). Thus the general formula for TF-IDF is a combination of the raw *tf* calculation formula with the *idf* formula, by multiplying the *tf* value by the *idf* value such as formula (7) and (8) [52]. Based on the IDF formula, regardless of the value of tfi_j, if $D = df_j$, the result will be 0 (zero) for the calculation of IDF. For that it can add a value of 1 on the idf side, so the calculation of the weight becomes as formula (9) [52].

$$tf\ =\ 0.5 + 0.5\ x\ (tf\ /\ \max tf) \quad (5)$$

$$idfj\ =\ log\ (\ D\ /\ dfj\ ) \quad (6)$$

Where,
D = total documents in the database
df_j = total documents that containing term t_j

$$wij\ =\ tfij\ x\ idfj \quad (7)$$

$$wij = tfij\ x \log(\frac{D}{dfj}) \quad (8)$$

Where,
wij = weight of term t_j to d_i
tfj = appearance total of term t_j in document d_i

$$wij = tfij\ x \log(\frac{D}{dfj}) + 1 \quad (9)$$

2.5 Confusion Matrix

Confusion Matrix is a method commonly used for calculating accuracy. In testing the accuracy of the search results will be evaluated the value of recall, precision, accuracy, and error rate [50], [53], [54]. Where precision evaluates the system's ability to find the most relevant ranks, and is defined as the percentage of documents that are interpreted and are truly relevant to the query. Recall evaluates the system's ability to find all relevant items from a document collection and is defined as the percentage of documents relevant to the query. Accuracy is a comparison of cases that are correctly identified by the total number of cases and an error rate is a case that is identified incorrectly by the number of cases.

3 Results and Discussions

3.1 Requirement Elicitation, Analysis, and Design of Chatbot

The result of requirement elicitation process is the collection of questions, and each question labels or class and answer that related with the question. Total of question is 176 questions with 6 label or class that described in Table 1 (the label is presented in Indonesian Language and the term in the Islamic law of purification). An an explanation in Section 2.1, the data collection is collected from scraping mobile application and books that related with *tharah* law.

Table. 1 Total Data Collection for Thaharah Chatbot

Label	Total Question and Answer Data
Wudhu (Ablution)	31
Tayammum	24
Darah Wanita (Women Blood)	28
Air dan Najis (Water and Unclean)	61
Mandi (Bath)	17
Istinja	15

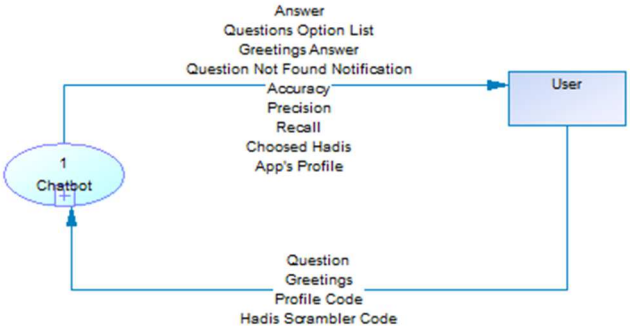

Fig. 3 Wcontext Diagram of Chatbot

Figure 3 presents Context Diagram that described the functionality of chatbot generally. Based on the analysis result, the functional requirements of chatbot among others: (1) chatbot can read data from database; (2) chatbot can recieve input from users; (3) chatbot can show the information and instruction for use; (4) chatbot provide citation of *hadith* that related with *thaharah* randomly; (5) chatbot provides optional question if user question is not available in database but has the same label or class; (6) chatbot can answer the question from user if the question available in database or exacly same with the classification result; (7) chatbot can give error message if can not find the answer of user question; (8) chatbot can show the evaluation of accuracy, precission, and recall value of MNB and RAKE algorithm that presented using Confusion Matrix. While, the non-functional requiremnets of chatbot among others: the chatbot is a web-based application that build using Flash and the time limit for chatbot to answer the question is 15 second. Then, for the analysis of MNB and RAKE that implemented in the chatbot is described in Figure 4 and will be explained more with the example in Section 3.1.

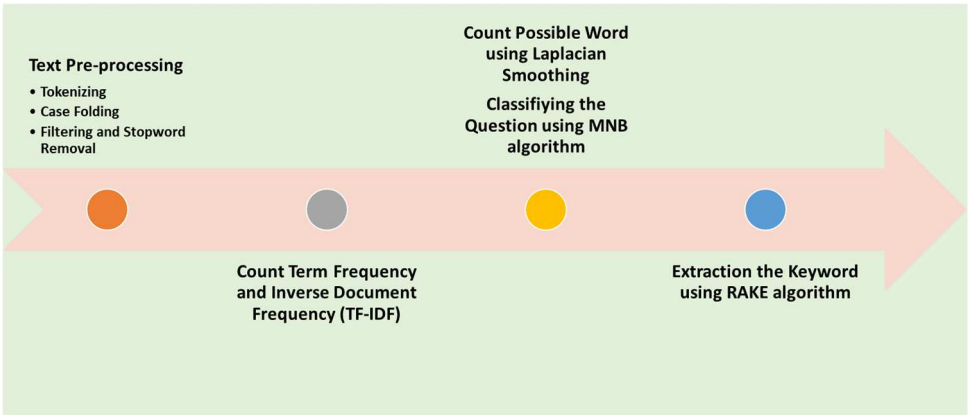

Fig. 4 Process of Implementation of NLP, MNB algorithm, and RAKE algorithm in the chatbot

3.2 Implementation of Multinomial Naive Bayes and Rapid Automatic Keyword Extraction

Based on Figure 4, the implementation of MNB and RAKE for *thaharah* chatbot is begin from conduct text pre-processing (among others: tokenizing, case folding, and stopword removal, in this research is not conducted the stemming process in accordance with the chatbot need), count TF-IDF, count the possible word using Lampacian Smooting, classify the question using MNB algorithm, and the las extract the keyword to get related answer based on the class or label prediction using RAKE algorithm. The illustration of those process is conducted with the example dataset that available in Table 2 (present in Indonesian language). While, Table 3 shows the result of text preprocessing, begin from tokenizing that separating the word, case folding that make words in uniform size, stopword removal and punctuation.

Table. 2 Example of Question Dataset

Label	Question
Wudhu	*Apa yang dapat membatalkan wudhu?* (What can cancel ablution?)
Wudhu	*Bagaimana whudu jika terkena tinta?* (How is the ablution affected by ink?)
Wudhu	*Bolehkah wudhu telanjang?* (Can ablution be naked?)
Tayammum	*Apa saja rukun tayammum?* (What are the pillars of *tayammum*?)
Tayammum	*Apa sebab disyariatkannya tayammum?* (What is the reason for *tayammum*?)

The result of text pre-processing is a structured representation of text data that can be processed computionally in the next process [55]. Therefore, the resul of text preprocessing from Table 3 is calculated the value of the TF-IDF using formula in the Section 2.4. Table 4 shows the calculation result of TF-IDF, while Table 5 shows the comparison of vector value of each words before and after TF-IDF calculation. The vector value of each words of features will be calculated in MNB algorithm begin Laplacian smooting using formula (3). And the result of MNB algorithm is available in Table 6.

Table. 3 Example of Text Pre-processing Result

Document ID	Words or Features
Doc 1	<membatalkan, wudhu>
Doc 2	<wudhu, terkena, tinta>
Doc 3	<wudhu, telanjang>
Doc 4	<rukun, tayammum>
Doc 5	<disyariatkannya, tayammum>

Table. 4 TF-IDF Calculation Result from the Example of Table 3

Words	TF Value using Normalzation					IDF Value +1	TF x IDF Value				
	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Log (N/DF)+1	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5
disyariatkannya	0	0	0	0	1/2	1.6999	0	0	0	0	0.85
membatalkan	1/2	0	0	0	0	1.699	0.85	0	0	0	0
rukun	0	0	0	1/2	0	1.699	0	0	0	0.85	0
tayammum	0	0	0	1/2	1/2	1.349	0	0	0	0.67	0.67
telanjang	0	0	1/2	0	0	1.699	0	0	0.85	0	0
terkena	0	1/3	0	0	0	1.699	0	0.57	0	0	0
Tinta	0	1/3	0	0	0	1.699	0	0.57	0	0	0
wudhu	1/2	1/3	1/2	0	0	1.233	0.62	0.41	0.62	0	0

Table. 5 Comparison of Vector Value Before and After TF-IDF Value from the Example of Table 3

Words	Before TF-IDF Calculation					After TF-IFD Calculation				
	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5
disyariatkannya	0	0	0	0	1	0	0	0	0	0.85
membatalkan	1	0	0	0	0	0.85	0	0	0	0
rukun	0	0	0	1	0	0	0	0	0.85	0
tayammum	0	0	0	1	1	0	0	0	0.67	0.67
telanjang	0	0	1	0	0	0	0	0.85	0	0
terkena	0	1	0	0	0	0	0.57	0	0	0
tinta	0	1	0	0	0	0	0.57	0	0	0
wudhu	1	1	1	0	0	0.62	0.41	0.62	0	0

Laplacian smoothing isused for counting the possible word from the new data againts with the data that available in database. For example, if given a new question "*Debu apa yang bisa digunakan tayammum?* (What dust can be used for tayammum?)", then it will be determined including what label the question is. Based on the example data from Table 3 the possible word is [*disyariatkannya, membatalkan, rukun, tayammum, telanjang, terkena, tinta, wudhu*] and the total weight of the TF-IDF possible words is 8 The probability that sentence including ablution is 0.0003072, while the probability of the question itself is 0.0006869. With the same process of text pre-processing, the question "*Debu apa yang bisa digunakan tayammum?*" is continued with Laplacian smoothing calculation, with the result of text preprocessing and Laplacan smoothing available in Table 6. Then, all of the probability of words is multiplied using formula (1). Therefore, based on the calculation of MNB algorithm, the question "*Debu apa yang bisa digunakan tayammum?*" is classified as "*Tayammum*".

Table. 6 The Result of Laplacian Smoothing and MN Algorithm of "*Debu apa yang bisa digunakan tayammum*"

Word	P(Word | *Wudhu*)	P(Word | *Tayammum*)
debu	(0+1)/(4,473+(1x8)) = 0,080	(0+1)/(3,046+(1x8)) = 0,090
digunakan	(0+1)/(3,008+(1x8)) = 0,080	(0+1)/(3,046+(1x8)) = 0,090
tayammum	(0+1)/(3,008+(1x8)) = 0,080	(1.348+1)/(3,046+(1x8))= 0.212
P(*debu, digunakan, tayammum* | *Wudhu*) = P(*debu* | *Wudhu*) x P(*digunakan*|*Wudhu*) x P(*Tayammum* | *Wudhu*) x P(*Wudhu*) = 0.0003072		
P(*debu, digunakan, tayammum* | *Tayammum*) = P(*debu* | *Tayammum*) x P(*digunakan*| *Tayammum*) x P(*Tayammum* | *Tayammum*) x P(*Tayammum*) = 0.0006869		

RAKE algorithm is used as another technique to extract keywords from document automatically, in this case keyword from new question that inputed in chatbot. If the question is unclassiified, then RAKE algorithm will be extracted the related keywords and show the optional questions that similar with new question. For example, if there is new question "*Bagaimana hukumnya buang air besar di sebuah toilet yang menghadap kiblat?* (How is the law to defecate in a toilet facing the Qibla?)", then do the text pre-processing, the each word will be represented with co-occurance graph keywods. Actually, co-occurrence graph is the same as TF-IDF matrix with one extra count of each word that appears in a phrase. Next, with formula (4) every keywords will be counted. The highest keyword score means that the keyword related or similar, and chatbot will show the optional question that related with it. In this example, the highest keyword scrore is "*hukumnya buang air*" with keyword score 9.0 from deg(w) of each word is 3, and freq(w) of each word is 1.

3.3 Implementation of *Thaharah* Chatbot using MNB and RAKE Algorithm

Implementation of *thaharah* chatbot is conducted using Ubuntu 16.04 LTS 32-bit Operating System with Phython 3.5.2 programming language. Figure 5 show the design of chatbot user interface and its implementation.

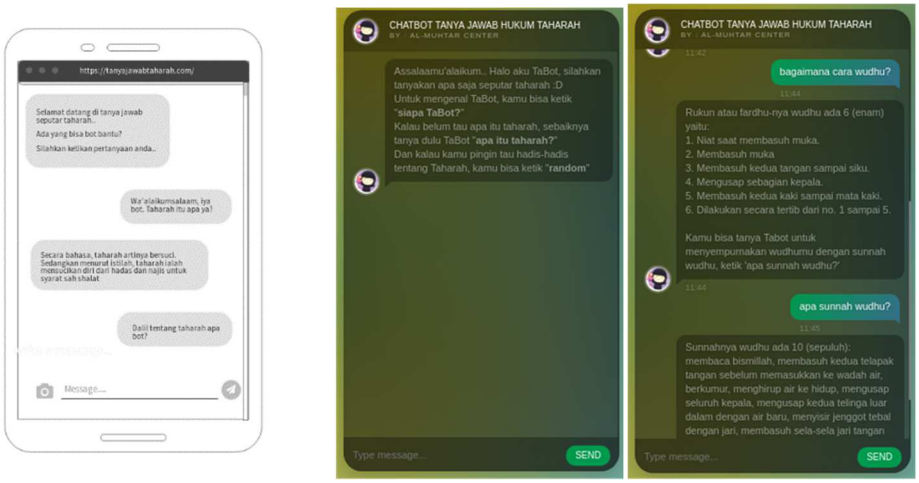

Fig. 5 User interface Design of Chatbot and Its Implementation

3.4 Testing and Evaluation Result

As an explanation in Section 2.1., this research using blackbox testing to make sure all of the functionality of chatbot run well and confusin matrix to evaluate the performance of MNB and RAKE algorithm. Based on the blackbox testing result, all of functions of chatbot had run in accordance with an expected result. It means in quality, *thaharah* chatbot had met the quality factor of functionality and correctness [32], [56]. Whereas, the testing of Multinomial Naïve Bayes classification method is done using cross validation technique, by partitioning existing data into 2 groups into training data and testing data from total 176 data. Testing with the number of training data as many as 132 questions, data testing as many as 44 questions and labels as many as 6 produce confusion matrix as presented using Confusion Matrix in Table 7.

Table. 7 Confusion Matrix of Prediction Result

Label	Prediction Result					
	Air dan Najis	*Darah Wanita*	*Istinja*	*Mandi*	*Tayammum*	*Wudhu*
Air dan Najis	**16**	0	0	0	0	0
Darah Wanita	0	**7**	0	0	0	0
Istinja	0	0	**3**	0	0	0
Mandi	0	0	0	**4**	0	0
Tayammum	2	0	0	0	**4**	0
Wudhu	2	0	0	0	0	**6**

Table. 8 Precison, Recall, and Accuracy Value

Evaluation	Label						Average
	Air dan Najis	*Darah Wanita*	*Istinja*	*Mandi*	*Tayammum*	*Wudhu*	
Pecision	80%	100%	100%	100%	100%	100%	**97%**
Recall	100%	100%	100%	100%	67%	75%	**90%**

Table. 9 Evaluation Result of RAKE Algorithm

Question	Extracted Keywords	Keyword Score
Apa itu air mutlak? (Is that absolute water?)	*'air mutlak'*	4.0
Bagaimana cara memandikan jenazah? (How do you bathe the body?)	*'memandikan jenazah'*	4.0
Apakah ada pembagian hukum air selain dari kesuciannya? (Is there a legal distribution of water aside from its purity?)	*'pembagian hukum air'*	9.0
	'kesuciannya'	1.0
Bolehkah istijmar menggunakan kotoran binatang? (Can the *istijmar* use animal manure?)	*'istijmar'*	1.0
	'kotoran binatang'	4.0
Apa saja indikator kenajisan dalam air? (What are the indicators of impurity in water?)	*'indikator kenajisan'*	4.0
	'air'	1.0

In the confusion matrix in Table 7, the number of data testing from each real label is the number of numbers in each row. While the amount of data predicted in each label, is the number of numbers in each column. Numbers with bold numbers are questions that are predicted with labels that are not suitable so that they are not predicted correctly. While bold numbers are just

the number of questions whose labels are predicted correctly. For example, on the line "t*ayammum*" and the column "*tayammum*" shows the number 4, meaning there are 4 questions about *tayammum* are predicted as *tayammum*, so the questions are predicted correctly. Whereas in the "*tayammum*" line and the "*air dan najis*" column shows the number 2, meaning that 2 questions about *tayammum* are predicted as *air dan najis*, so the questions are not predicted correctly. Then, using recall, precision, and accuracy formula [50], [53], [54], it was gotten the accuracy values around 90%, with 97% of precision value and 90% of recall value (available in Table 8). While, for RAKE algorithm, there are 5 example question that used as an input data. From those testing showed that RAKE algorithm had been run well in extracting the keyword based on the calculation of keyword score (described in Table 9).

In testing the classification method, a training process was carried out with a total of 6 labels and a total amount of 176 data. The results of the tests showed that the average label produced Precision by 97%, Recall at 90% and Overall Accuracy by 90%. The resulting accuracy value is certainly influenced by various things, including data complexity, the amount of training data on each label and also influenced by how the model studies the data provided. And in the machine learning model, the more data the better the ability of the model to predict the label or class of each given case.

5 Conclusions

The implementation of the MNB algorithm in the chatbot's intent classifier has been successfully carried out by managing existing data in the form of a list of questions along with a label on each question. List questions are used to train the model in predicting the label of each question entered by the user when using the chatbot. On user questions, text preprocessing is done. Then, the question was changed to the TF-IDF vector and entered into the MNB calculation process using the lapling smoothing technique to predict the label. After the label is predicted, the label is entered into the chatbot's entity recognizer. RAKE algorithm in the chatbot's entity recognizer has also been successfully performed. RAKE is used to extract keywords from user questions. The extraction process begins with taking words based on the stopword position as keywords, then the keyword with the highest score is selected. The keyword is checked in the list of entities labeled the same as the predicted label. If the keyword is in exactly one entity then the entity is a question that is used by the user. If the keywords are in more than one entity, they are returned to the user as a question option. If the entity matches the user's question, the chatbot will send the answer. If it doesn't match, the chatbot will display an unregistered question notification. Based on the evaluatin using confusion matrix, MNB that combine with RAKE have accuracy quite good that reach percentage around 90%.

In further works, an additional amount of data can be made to improve model knowledge. To improve the quality of the extraction of keywords, you can use the Part-Of-Speech Tagging or POS Tagging method to retrieve keywords based on the type of words such as nouns or nouns and verbs or verbs. In addition to POS Tagging, other methods can be used such as the TextRank method from Google. And to improve the quality of the accuracy of the classification model, other techniques can be used such as K-Fold validation. Then, for a variety of further research, the Deep Learning approach can also be used.

Acknowledgement

Our thanks for the Research and Publishing Center of UIN Sunan Gunung Djati Bandung which has provided support and funded the publication of this research.

References

[1] M. NEGNEVITSKY, *Artificial Intelligence: A Guide to Intelligent Systems*. 2017.

[2] S. Russel and P. Norvig, *Artificial intelligence—a modern approach 3rd Edition*. 2012.

[3] T. Sutojo, E. Mulyanto, and V. Suhartono, *Kecerdasan Buatan*. Semarang: Penerbit Andi, 2010.

[4] P. Hamet and J. Tremblay, "Artificial intelligence in medicine," *Metabolism.*, 2017.

[5] Y. A. Gerhana, W. B. Zulfikar, A. H. Ramdani, and M. A. Ramdhani, "Implementation of Nearest Neighbor using HSV to Identify Skin Disease," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 288, no. 1, p. 012153, 2018.

[6] K. R. Pradeep and N. C. Naveen, "Lung Cancer Survivability Prediction based on Performance Using Classification Techniques of Support Vector Machines, C4. 5 and Naive Bayes Algorithms for Healthcare Analytics," *Procedia Comput. Sci.*, vol. 132, pp. 412–420, 2018.

[7] T. P. Michalak and M. Wooldridge, "AI and Economics," *IEEE Intell. Syst.*, 2017.

[8] H. Lu, Y. Li, M. Chen, H. Kim, and S. Serikawa, "Brain Intelligence: Go beyond Artificial Intelligence," *Mob. Networks Appl.*, 2018.

[9] H. K. W. Silvia Rostianingsih, Gregorius Satia Budhi, "GAME SIMULASI FINITE STATE MACHINE UNTUK PERTANIAN DAN PETERNAKAN," *Sci. reposity J.*, 2017.

[10] I. Yunanto, A. A., Herumurti, D., & Kuswardayan, "Kecerdasan Buatan Pada Game Edukasi Untuk Pembelajaran Bahasa Inggris Berbasis Pendekatan Heuristik Similaritas," *J. Sist. dan Inform.*, 2017.

[11] M. Abdi, D. Herumurti, and I. Kuswardayan, "Analisis Perbandingan Kecerdasan Buatan pada Computer Player dalam Mengambil Keputusan pada Game Battle RPG," *JUTI J. Ilm. Teknol. Inf.*, 2017.

[12] M. J. Timms, "Letting Artificial Intelligence in Education out of the Box: Educational Cobots and Smart Classrooms," *Int. J. Artif. Intell. Educ.*, 2016.

[13] M. M. Islam and A. S. M. L. Hoque, "Automated Essay Scoring Using Generalized Latent Semantic Analysis," *J. Comput.*, vol. 7, no. 3, 2012.

[14] L. Hu, Y. Miao, G. Wu, M. M. Hassan, and I. Humar, "iRobot-Factory: An intelligent robot factory based on cognitive manufacturing and edge computing," *Futur. Gener. Comput. Syst.*, 2019.

[15] B. Yong *et al.*, "IoT-based intelligent fitness system," *J. Parallel Distrib. Comput.*, 2018.

[16] S. Ionita, "Autonomous vehicles: From paradigms to technology," in *IOP Conference Series: Materials Science and Engineering*, 2017.

[17] A. Kulkarni and A. Shivananda, "Deep Learning for NLP," in *Natural Language Processing Recipes*, 2019.

[18] F. Chaubard, G. Genthial, R. Socher, M. Fang, R. Mundra, and K. Clark, "Natural Language Processing with Deep Learning," *CS224n*, 2017. .

[19] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: An introduction," *J. Am. Med. Inform. Assoc.*, 2011.

[20] Kusumadewi, *Artificial Intelegence (Teknik dan Aplikasinya)*. Yogyakarta: Graha Ilmu, 2003.

[21] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, Fourth Edi. United States of America: Morgan Kaufman, 2012.

[22] H. Jiawei, M. Kamber, J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2006.

[23] G. Nanda, M. Dua, and K. Singla, "A Hindi Question Answering System using Machine Learning approach," in *2016 International Conference on Computational Techniques in Information and Communication Technologies, ICCTICT 2016 - Proceedings*, 2016.

[24] D. Xhemali, C. J. Hinde, and R. G. Stone, "IJCSI IJCSI Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages," *IJCSI Int. J. Comput. Sci. Issues*, 2009.

[25] S. N and M. Govindarajan, "Mining Movie Reviews using Machine Learning Techniques," *Int. J. Comput. Appl.*, vol. 144, no. 5, 2016.

[26] M. Andre and N. Kamal, "A Comparison of Event Models for Naive Bayes Text Classificatio," *AAAI-98 Work. Learn. text Categ.*, 1998.
[27] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic Keyword Extraction from Individual Documents," in *Text Mining: Applications and Theory*, 2010.
[28] E. N. S. C. P and I. Afrianto, "Rancang Bangun Aplikasi Chatbot Informasi Objek Wisata Kota Bandung Dengan Pendekatan Natural Language Processing," *J. Ilm. Komput. dan Inform.*, 2015.
[29] Maskur Maskur, "Perancangan Chatbot Pusat Informasi Mahasiswa Menggunakan AIML Sebagai Virtual Assistant Berbasis Web ," *Kinetik*, 2016.
[30] A. A. Akhsan and F. Faizah, "Analisis dan Perancangan Interaksi Chatbot Reminder dengan User-Centered Design," *J. Sist. Inf.*, 2017.
[31] A. Hamza, N. En-Nahnahi, K. A. Zidani, and S. El Alaoui Ouatik, "An arabic question classification method based on new taxonomy and continuous distributed representation of words," *J. King Saud Univ. - Comput. Inf. Sci.*, 2019.
[32] R. S. Pressman, *Software Engineering: A Practitioner's Approach*, 7th ed. New York: McGraw-Hill, 2011.
[33] I. Sommerville, *Software Engineering*. 2010.
[34] A. S, *125 Masalah Taharah*. Solo: Tiga Serangkai, 2008.
[35] Al-Qardhawi, *Fikih Thaharah*. Jakarta: Pustaka Al-Kautsar, 2004.
[36] Al-Hadhrami, *Ilmu Fiqih (Safinatunnajah) Berikut Penjelasannya*. Bandung: Sinar Baru Algesindo, 2011.
[37] H. Mahgoub, D. Rösner, N. Ismail, and F. Torkey, "A Text Mining Technique Using Association Rules Extraction," *Int. J. Comput. Intell.*, vol. 4, no. 1, pp. 21–28, 2008.
[38] D. S. Maylawati, H. Aulawi, and M. A. Ramdhani, "Flexibility of Indonesian text pre-processing library," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 13, no. 1, pp. 420–426, 2019.
[39] A. T. J. H, "Preprocessing Text untuk Meminimalisir Kata yang Tidak Berarti dalam Proses Text Mining," *Inform. UPGRIS*, 2015.
[40] M. Robani and A. Widodo, "Algoritma K-Means Clustering Untuk Pengelompokan Ayat Al Quran Pada Terjemahan Bahasa Indonesia," *J. Sist. Inf. BISNIS*, 2017.
[41] D. Setiawati, I. Taufik, J. Jumadi, and W. B. Zulfikar, "Klasifikasi Terjemahan Ayat Al-Quran Tentang Ilmu Sains Menggunakan Algoritma Decision Tree Berbasis Mobile," *J. Online Inform.*, 2016.
[42] Jumadi, D. S. Maylawati, B. Subaeki, and T. Ridwan, "Opinion mining on Twitter microblogging using Support Vector Machine: Public opinion about State Islamic University of Bandung," in *Proceedings of 2016 4th International Conference on Cyber and IT Service Management, CITSM 2016*, 2016.
[43] S. M. Weiss, N. Indurkhya, T. Zhang, and F. J. Damerau, "Information Retrieval and Text Mining," *Springer Berlin Heidelb.*, no. Fundamentals of Predictive Text Mining, pp. 75–90, 2010.
[44] R. Setiawan, A. Kurniawan, W. Budiharto, I. H. Kartowisastro, and H. Prabowo, "Flexible Affix Classification for Stemming Indonesian Language," in *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 2016.
[45] I. Destuardi and S. Sumpeno, "Klasifikasi Emosi Untuk Teks Bahasa Indonesia Menggunakan Metode Naive Bayes," *Semin. Nas. Pascasarj. Inst. Teknol. Sepuluh Nop.*, 2009.
[46] A. Rahman, "Online News Classification Using Multinomial Naive Bayes," *ITSMART*, 2017.
[47] K. S. Hasan and V. Ng, "Automatic Keyphrase Extraction: A Survey of the State of the Art," 2015.
[48] A. A. Maarif, "Penerapan Algoritma TF-IDF untuk Pencarian Karya Ilmiah," *Dok. Karya Ilm. | Tugas Akhir | Progr. Stud. Tek. Inform. - S1 | Fak. Ilmu Komput. | Univ. Dian Nuswantoro Semarang*, 2015.
[49] Melita, "Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Syarah Umdatil Ahkam)," *J. Tek. Inform.*, vol. 11, no. 2, 2018.
[50] V. Amrizal, "PENERAPAN METODE TERM FREQUENCY INVERSE DOCUMENT FREQUENCY (TF-IDF) DAN COSINE SIMILARITY PADA SISTEM TEMU KEMBALI INFORMASI UNTUK MENGETAHUI SYARAH HADITS BERBASIS WEB (STUDI KASUS: HADITS SHAHIH BUKHARI-MUSLIM)," *J. Tek. Inform.*, 2019.

[51] R. Mandala, “Bahan Kuliah Sistem Temu Balik Informasi,” Bandung, 2004.
[52] S. Robertson, “On event spaces and probabilistic models in information retrieval,” *Inf. Retr. Boston.*, 2005.
[53] K. M. Ting, “Confusion Matrix,” in *Encyclopedia of Machine Learning and Data Mining*, 2017.
[54] A. K. Santra and C. J. Christy, “Genetic Algorithm and Confusion Matrix for Document Clustering,” *IJCSI Int. J. Comput. Sci. Issues,* 2012.
[55] S. Vijayarani, J. Ilamathi, and M. Nithya, “Preprocessing Techniques for Text Mining - An Overview,” *Int. J. Comput. Sci. Commun. Networks*, vol. 5, no. 1, pp. 7–16, 2015.
[56] J. A. McCall, P. K. Richards, and G. F. Walters, “Factors in Software Quality - Volume 1 - Concept and Definitions of Software Quality,” *Def. Tech. Inf. Cent.*, 1977.