Text Mining for Pest and Disease Identification on Rice Farming with Interactive Text Messaging

Edio da Costa[1], Handayani Tjandrasa[2], Supeno Djanali[3]
[1,2,3]Department of Informatics, Sepuluh Nopember Institute of Technology, Surabaya, Indonesia
[1]Department of Computer Science, Dili Institute of Technology, Dili, Timor Leste

Article Info

Article history:

Received Jan 13, 2018
Revised Apr 29, 2018
Accepted Mar 8, 2018

Keyword:

Interactive text messaging
Knowledge-based system
Rice pests and diseases
Text mining
Timor Leste

ABSTRACT

To overcome pests and diseases of rice farming, farmers always rely on information and knowledge from agricultural experts for decision making. The problem is that experts are not always available when the farmers need and the cost is quite high. Pests and diseases elimination is hard to be done individually since the farmers are lack of knowledge about the pest types that attack the rice fields. The objective of this study is to build a knowledge-based system that can identify pests and diseases interactively based on the information that has been told by the farmers using SMS communication services. The system can provide a convenience way to the farmers in delivering pests and disease problem information using a natural language. The text mining method performs tokenizing, filtering and porter stemming that used to extract important information sent by a SMS service. The method of Jaccard Similarity Coefficient (JSC) was used to calculate similarities of each pest and disease based on symptoms that are sent by the farmers through SMS. The corpus database usedin this study consists of 28.526 root words, 1.309 stop wordsand 180 words list. Pest and disease database reference in this study was obtained from the Ministry of Agriculture and Fisher (MAF) Timor-Leste. The result of the experiment shows that the system is able to identify the symptoms based on the keywords identified with the accuracy of 81%. The result of pest and disease identification has the accuracy of 86%.

Corresponding Author:

Edio da Costa,
Department of Informatics,
Sepuluh Nopember Institute of Technology,
Surabaya, Indonesia.
Email: ediocosta73@gmail.com

1. INTRODUCTION

Agriculture is one of the main means of livelihood for the people of the Timor Leste. Data shows that agricultural productivity in Timor Leste is very low compared with other country in Southeast Asia and 45% of farmers surveyed in Timor- Leste suffer from food shortages because of crop failure every year [1]. Nearly 76% of consumption needs to be imported because of the low productivity level. The consumption of rice in Timor-Leste is 135,000 tons per year, while farmers produce only 65,000 tons, so that every year the government has to import rice from Vietnam [2]. Farmers lose an estimated average of 37% of their rice crop to pests and diseases every year. Farmers' yields are likely to experience changes in the last few years. Eradication of pests and diseases is difficult to be implemented to the maximum since most farmers do not understand the type of pests that attack rice crops. All kinds of symptoms areconsidered similar, so farmers use only one type of pesticide to eliminate a variety of symptoms [3]. To overcome the problems of pests and diseases, some of the web-based agricultural extension systems [4], and web-based smartphone applications [5] have been developed, but only a few groups of farmers understand the technology of its complexity.

Pests and diseases identification have been developed bythe International Rice Research Institute [6]. The study has built a pest and disease consulting application called "Rice Doctor" with consulting services for web-based and android-based mobile phone. Selection of symptoms in this study was done by marking a check list of symptoms to be delivered. The problems arise when symptoms are numerous, so that farmers have to read one by one the existing symptoms until finally found the symptom referred. While, the study in the [7] has built the SMS consultation services to solve the rice farming problem such as the pests and diseases identification. The same study using SMS was conducted by [8], the identification process used some parameters and must followed the complex procedure, so the farmers did not easily deliver information about the pests and diseases problems. The growing use of SMS service as a communication medium becomes one of the platforms to solve interactive problems.

It is estimated that over half of mobile phone users globally will have smartphones in 2018 [9]. The mobile phone has established themselves as the most invading communication media in the developing countries. The SMS service can be of great use to the farmers in the rural community where they lack advanced internet services and computing technology. The use of SMS service can be utilized by the farmers to communicate with experts available remotely with much ease in a cost-effective manner and with prompt response [10]. The use of mobile phone in Timor-Leste continues to increase each year, the data show that in 2015 mobile phone usage increased by 90% [11]. Surveys conducted by the United Nations Integrated Mission in Timor-Leste (UNMIT) showed that 58% of farmers choose to use the SMS service to send and receive messages [12]. Farmers choose to use SMS because internet packet data service is still very expensive. The mobile penetration rate has been increasing rapidly during the last few years but the network infrastructure is still not available to all the districts in Timor Leste so that farmers have limited access to the internet by using smartphones.

Therefore, the objective of this study is to build a knowledge-based system that can identify pests and diseases interactively based on the information that has been told by the farmers using SMS communication services. The system can be used to link farmers and a knowledge-based system which acts as an expert. Our approach is based on developing an innovative and interactive ICT to enable an agricultural knowledge-base system with the help of experts. Farmers easily send SMS using a natural language, i.e. the text messages associated with the problems of pests and diseases of rice. Then the proposed system processes the messages and replies the user's request interactively.

The paper consists of five sections: Following this first section, the second section describes the proposed method for pests and diseases data extraction. Natural language processing using text mining includes tokenizing, filtering to select keywords and the Porter stemming to get the root words. The third section describes the experimental results for symptoms identification, pests and diseases identification using Jaccard Similarity Coefficient (JSC). Finally, this paper is concluded in section four.

1.1. Text mining

Text mining is a process of discovery of new information or terms that were not revealed previously. Text mining has been applied in several fields such as health [13], telecommunications and marketing [14]. While, in Indonesian language text mining has also been implemented in some cases, such as sentiment analysis on social media [15], health services [16] and security [17]. Figure 1 represents the steps in the process of text miningwich consists of: Tokenizing, Filtering, Stemming, Tagging and Analysis [18]. However, this study only used three processes, namely: Tokenizing, Filtering, and Stemming. Tokenizing is the stage of segmenting string input into words. Filtering is an important information collection stage from the process of tokenizing. This process eliminates the non-functional charactersthat consists of [19]:

a. Remove the subsequent characters if they are followed by a space,
b. Remove the symbols,
c. Eliminate the following pairs of brackets,
d. Eliminate the single and double quotation symbol,
e. Eliminate the slash.

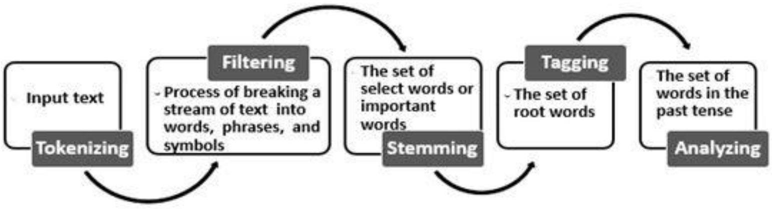

Input text
Tokenizing
Filtering
- Process of breaking a stream of text into words, phrases, and symbols
The set of select words or important words
Stemming
Tagging
- The set of root words
The set of words in the past tense
Analyzing

Figure 1. The stages of text mining

Stemming is the stage to get the root words based on the filtering results. The stemming process in Indonesian language text is more complicated than English because there are variations of affixes that must be removed to get the root word [20]. The structure of Indonesian morphology has a higher level of complexity than the English language. Porter stemming is one method of stemming for Indonesian language text that requires a shorter time compared to the other stemming algorithm.

In Indonesian language text, the suffix, prefix, and confix (combination of the prefix and suffix) must be removed to get the root word. There are four groups of rules to conduct Porter stemming method for the Indonesian language [20]. Example, removing prefix "*meng"* in the word "*men*guning→*meng*-uning" the prefix "*meng"* is removed and plus "k" before "u" become "*kuning"* (Table 1). If a word begins with *mem*-pattern (vowel/no vowels) by removing the prefix "*mem"* from the word *"mem*busuk→*mem*-busuk", it becomes "*busuk"* (Table 1). The morphology of Indonesian language words can comprise both inflectional and derivational structures. Inflectional is the simplest structure which is expressed by suffixes. Removing the inflection particle "pun" from the word "akarpun" it becomes "akar" (Table 2).

Table 1. Rule of First Order Derivational Prefix

Prefix	Word	Root Word	Meaning
meng-	Menguning	Kuning	Yellow
meny-	Menyerang	Serang	Attack
mem-	Membusuk	Busuk	Rotten
pem-	Pembungaan	Bunga	Flower

Table 2. Rule of Inflection Particle

Suffix	Word	Root Word	Meaning
-Pun	Akarpun	Akar	Root
-lah	Batanglah	Batang	Stem
-kah	Apakah	Apa	what

Table 3. Rule of Inflection Possessive Pronouns

Suffix	Word	Root Word	Meaning
-ku	Bungaku	Bunga	Flower
-mu	Padimu	Padi	Rice
-nya	Daunnya	Daun	Leaf

Table 4. Rule of Derivational Suffix

Suffix	Word	Root Word	Meaning
-kan	Menyebabkan	Sebab	Cause
-an	Makanan	Makan	Food
-i	Mendapati	Dapat	Can

Suffix removing consists of 2 categories: possessive pronouns and derivational suffix (Table 3 and Table 4). Removing each category should only be done once. Stemming algorithm combined with stop word corpus can provide the high keywords identification accuracy. The research by [21] shows the accuracy of 90% with the extraction keywords using corpus which has been preprocessed by removing the stop words and by the stemming process, compared with the accuracy of 82% which is preprocessed without removing the stop words and using stemming process.

1.2. Short message service (SMS)

SMSbecomes one of the services most used by the user because it is cheap and easy to use. SMS service has been implemented in some public services, such as sexual health education [22]. The SMS messages were used for health information sharing purpose like communication between patient and health clinics, sexual health education and so on. While, in the agriculture, a research for corn farmers had been implemented [23]. The research focussed on the how to use SMS services to help farmers to be able to identify the best date to start planting, best date to harvest and optimal water availability as well as the projected crop yield. Another work by [24] evaluated mobile-phone based consulting service to the farmers. The study has built a mobile-phone based agronomic information service called "Avaaj Otalo". Many farmers used this servicefor getting expert advice regarding cotton farming. Some countries, such as India, Tanzania and Kenya have also implemented a knowledge-based SMS service to provide easiness for the farmers to consult with experts [25].

We chose to use SMS service because:

a. Low cost and SMS services are available in all smartphones,
b. Easy to used by farmers because of it's simplicity,
c. GSM mobile service is available widely than other services like GPRS [10].

The Attention Command (ATC) used to send and receive SMS are described in Table 5 [26]. Connect GSM or CDMA modem to the computer and then using ATC for sending text messages.

Mobile phone or GSM/CDMA modem will respond by giving Protocol Data Unit (PDU) of the desired SMS, which included the number of the sender, sending time, and the content of the SMS sent. The task of PDU is to encode SMS-Center data, so that the messages sent by the system will be received by the user.

Table 5. Send and Receiving SMS AT commands

Sending SMS		Receiving SMS	
Attention commands	Usage	Attention commands	Usage
AT+CMGS	Send SMS	AT+CNMI	Identify new message
AT+CMSS	Send SMS from storage	AT+CMGL	List all the message
AT+CMGW	Write SMS to storage	AT+CMGR	Read the message
AT+CMGD	Delete SMS	AT+CNMI	Identify new message

1.3. Jaccard similarity coefficient (JSC)

JSC has been applied in various fields and is one ofthe similarity indexes most widely used on binary data (0 and 1) [27]. The formula for calculating the similarity between objects A and B are as follows:

$$J(A,B)=\frac{A\bigcap B}{A\bigcup B}$$

Where $J(A,B)$ is similarity between A to B, $|A\bigcap B|$ (A Intersection B) is the set containing just those elements common to both A and B. And $|A\bigcup B|$ (A Union B) is is the set containing everything in either A or B or both.

2. RESEARCH METHOD

Pests and diseases data in this study were obtained from interviews with the experts of the pest and disease research department, Ministry of Agriculture and Fisher (MAF) Timor-Leste. Figure 2 describes all process of the system to connect the farmers with knowledge-based systems to identify pests and diseases of rice plants. This process consists of knowledge-basedcreating, text mining SMS testing, symptoms identification, pests and diseases identification and interactivity.

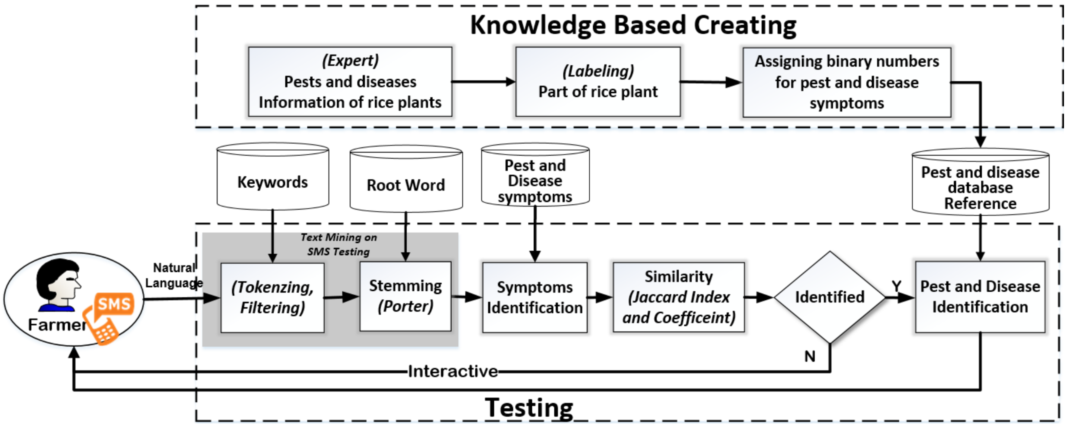

Knowledge Based Creating
(Expert) Pests and diseases Information of rice plants
(Labeling) Part of rice plant
Assigning binary numbers for pest and disease symptoms
Keywords
Root Word
Pest and Disease symptoms
Pest and disease database Reference
Farmer
SMS
Natural Language
Text Mining on SMS Testing
(Tokenzing, Filtering)
Stemming (Porter)
Symptoms Identification
Similarity (Jaccard Index and Coefficeint)
Identified
Y
N
Pest and Disease Identification
Interactive
Testing

Figure 2. System architecture for pests and diseases identification

2.1. Knowledge-based creating

The process of knowledge-based creating consists of three steps: data collecting of pests and diseases from experts, labeling based on rice plant components from the growth process until harvest andassigning binary numbers for pests and diseases symptoms, then stored in the database reference. The results of labeling based on the part of rice farming problems of the growth process until harvesting, which consist of 14 parts is shown in Table 6.

Table 6. Labeling based on The Part of Rice Farming

Morphology	Meaning	Labeling
Daun	Leaf	a
Batang	Stem	b
Akar	Root	c
Biji	Seed	d
Bibit	Seeding	e
Malai	Tassel	f
Pucuk	Shoot	g
Pembungaan	Flowering	h
Anakan	Tillering	i
Pelepah	Midrib	k
Kecambah	Sprout	l
Tangkai	Stalk	m
Pembijian	Seeding	n
Tunas	Bud	o

The next process was labelingbased on the symptoms of each morphology that were attacked by pests and diseases. Part ofthe classification results are shown in Table 7.

Table 7. Labeling Symptoms based on Morphology in Part of Rice Farming

Sub Morphology	Meaning	Labeling
Daun	Leaf	a
daun busuk	rotten leaf	a_1
daun kering	dry leaf	a_2
daun gulung	curling leaf	a_3
…	…	…
daun putih	white leaf	a_{89}
Batang	Stem	b
batang pendek	terse stem	b_1
batang kering	droughty stem	b_2
batang kuning	yellow stem	b_3
…	…	…
batang hitam	black stem	b_{16}
Akar	Root	c
akar busuk	rotten root	c_1
akar hitam	black root	c_2
akar coklat	brown root	c_3
…	…	…
akar kasar	coarse root	c_9
…	…	…
Tunas	Shoot	p

The last step of the knowledge-based creating process is to build a rule-based system in the form of columns using the binary values 0 and 1, part of them is shown in Table 8. The values of 1 indicate that the symptoms occur for the pests and diseases, whereas the values of 0 indicate that the symptoms do not occur. In this case, it is assumed that all the pests and diseases have the same weighting factors. That is, there are no symptoms that have the higher value than the other symptoms.

Table 8. Rule based Pests and Diseases

Pests and diseases	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	…	p_3
Hama putih	0	0	1	0	0	0	0	0	…	0
Hama putih	0	0	1	0	0	0	0	0	…	0
Hama putih	0	0	1	0	0	0	0	1	…	0
Garis coklat daun	1	0	0	0	1	1	0	1	…	0
Garis coklat daun	1	0	0	0	1	0	0	1	…	0
Garis coklat daun	1	0	0	0	1	1	0	0	…	0
Bercak coklat sempit	0	0	0	0	1	1	1	1	…	0
Bercak coklat sempit	0	0	0	0	1	0	1	1	…	0
Bercak coklat sempit	0	0	0	0	1	1	1	1	…	0

Eachof pest and disease consists of three variants obtained from three experts. There is some similarity of symptoms owned by the three variants. For an example, there are similarity of the pest "Hama putih” (leaffolder) symptoms, i.e. a_3=“daun gulung” (curling leaf) which is owned by the three variants. But there are also symptoms that are not shared by the two other variants, such as symptoms of a_8=“daun bergaris” (striped leaf) which is only found in the last variant.

2.2. Text mining on sms testing

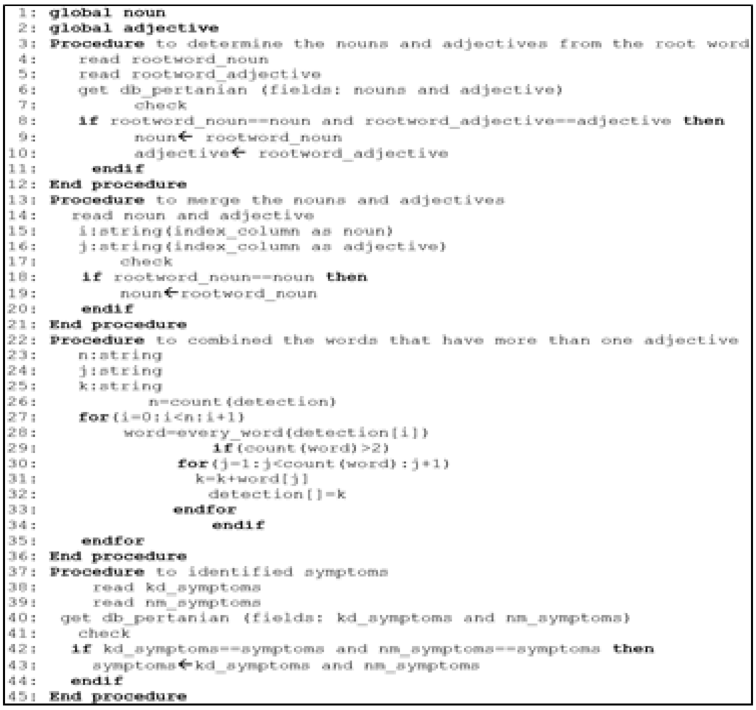
SMS extraction was used to collect important information based on SMS content delivered by farmers. To extract unstructure SMS, the authors adopt the Tokenizing, Filtering, and Stemming algorithm as proposed in [19], [28]. The purpose of those three steps was to improve the effectiveness of the system. Keywords obtained from tokenizing and filtering were matched with the list of keywords in the database of symptoms of rice pests and diseases. If there was a match then the system proceeded to the process of stemming. Conversely, when there was no match, the system sent a message of failure and directed farmers to resubmit problems related to pests and diseases of rice. The final stage of the SMS extraction process was stemming in order to obtain root word and keywords from the previous process. Corpus Indonesian language in this study consisted of 28, 526 root words and 1.309 stop words [29]. The word list of pests and diseases consisted of 179 words obtained from the extraction the symptoms of pests and diseases of rice.

2.3. Symptoms identification

The result obtained from the root word would be matched with symptoms database to identify pests and diseases. The level of difficulty in this process is how to translate the natural language that has been told by the farmers using SMS communication services. In general, the language of communication used by farmers is unstructured, like nouns, adjectives, and complements. Thus it requires a knowledge-based system that is able to interpret the content of the submitted SMS. The steps author suggested for identifying symptoms of pests and diseases based on natural language are:

a. Send SMS using natural language
b. Removing punctuation and words that are not important
c. Determine the nouns and adjective from the root word
d. Merge the nouns and adjectives
e. Combined the words that have more than one adjective
f. Matching symptoms based combination word that has been generated
g. Pest and disease symptoms identification

Referring to suggested steps from the previous process, then we proposed algorithm for identifying the symptoms of pests and diseases are shown in Figure 3.



```
global noun
global adjective
Procedure to determine the nouns and adjectives from the root word
   read rootword_noun
   read rootword_adjective
   get db_pertanian (fields: nouns and adjective)
       check
   if rootword_noun==noun and rootword_adjective==adjective then
       noun← rootword_noun
       adjective← rootword_adjective
    endif
End procedure
Procedure to merge the nouns and adjectives
  read noun and adjective
   i:string(index_column as noun)
   j:string(index_column as adjective)
      check
   if rootword_noun==noun then
      noun←rootword_noun
   endif
End procedure
Procedure to combined the words that have more than one adjective
   n:string
   j:string
   k:string
        n=count(detection)
   for(i=0;i<n;i+1)
      word=every_word(detection[i])
            if(count(word)>2)
          for(j=1;j<count(word);j+1)
            k=k+word[j]
             detection[]=k
          endfor
             endif
   endfor
End procedure
Procedure to identified symptoms
    read kd_symptoms
    read nm_symptoms
 get db_pertanian (fields: kd_symptoms and nm_symptoms)
   check
  if kd_symptoms==symptoms and nm_symptoms==symptoms then
    symptoms←kd_symptoms and nm_symptoms
  endif
End procedure
```

Figure 3. Symptoms identification algorithm

Here is one case of pests and diseases problems that farmers sent using natural language: "*tadi pagi saya ke sawah, tanaman padi saya kok daunnya kecoklatan dan bercak serta daun mengambang seperti ketupat. Bagaimana cara mengatasinya?*".

That case subsequently solved using the suggested steps and algorithms. The first step, was tokenizing and filtering to remove a punctuation and unimportant words to get the important words. After that, stemming was done using Porter algorithm to obtain the root word, the result are shown in Table 9. The next process was marking words consist of noun and adjective. The noun refers to 14 parts of rice main problem from growth process until harvesting (Tabel 6). The adjective is a type of symptoms which existed in every symptom of pest and disease, such as *coklat* (brown), *bercak* (spotting), and *mengambang* (floating). So, the result of marking the word is shown in Tabel 10.

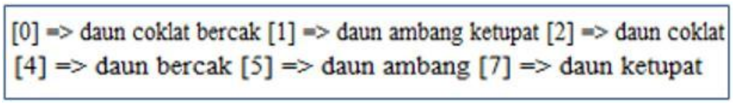
Table 9. The Result of Word List and of Root Word

Word list	Root word	Meaning
daunnya	daun	Leaf
kecoklatan	coklat	Brown
bercak	bercak	Spotting
daun	daun	Leaf
mengambang	ambang	Floating
ketupat	ketupat	Rhomb

Table 10. The Result of Marking Word

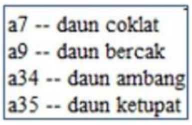
Word	Meaning	Marking
daun	leaf	noun
coklat	brown	adjective
bercak	spotting	adjective
daun	leaf	noun
ambang	floating	adjective
ketupat	rhomb	adjective

The next process was combining a noun with an adjective and setting back compound words that have more than one adjective. During the marking process, adding array depended on object keyword. When the compound words consisted of more than two words, looping was done then added new array. Example "*daun coklat bercak"* (spotted brownish leaf) more than two words. To match it with a database, the number of array looping done was twice "*daun coklat*" (brownish leaf) and "*daun bercak*" (spotted leaf), so the result is as follows:



```
[0] => daun coklat bercak [1] => daun ambang ketupat [2] => daun coklat
[4] => daun bercak [5] => daun ambang [7] => daun ketupat
```

The final stage of symptoms identification process was to match with pests and diseases database reference. So that the result of symptoms identification is as follows:



```
a7 -- daun coklat
a9 -- daun bercak
a34 -- daun ambang
a35 -- daun ketupat
```

Next, the symptoms were matched with the database reference to identify pests and diseases.

2.4. Pest and disease identification

There were four (4) symptoms obtained from SMS extraction, a_7="daun coklat" (brown leaf), a_9="daun bercak" (spotted leaf), a_{34}="daun ambang" (float leaf) and a_{35}="daun ketupat" (rhomb leaf). Those four identification symptoms were matched to symptoms from two diseases namely "Bercak daun coklat" (Brown leaves spot) (P_1) and Blast (P_2), so the illustration result is shown in Figure 4.

There are two symptoms that have a match with the "Bercak daun coklat" (Brown leaf spot) disease (P_1) that is a_7="daun coklat" (brown leaf) and a_{35}="daun ketupat" (rhomb leaf). So that the value of the Intersection=2, while of Union=11, then JSC obtained is 2/11=0.18. While "Blast" (Blast) pests (P_2) has 4 matches, namely a_7="daun coklat" (brown leaf), a_9="daun bercak" (spotted leaf), a_{34}="daun ambang" (float leaf) and a_{35}="daun ketupat" (rhomb leaf), therefore the value of Intersection=4, while value of Union=4, then JSC was 4/4=1.00. Both of the illustration concluded that the highest similarities value was Blast (P_2).

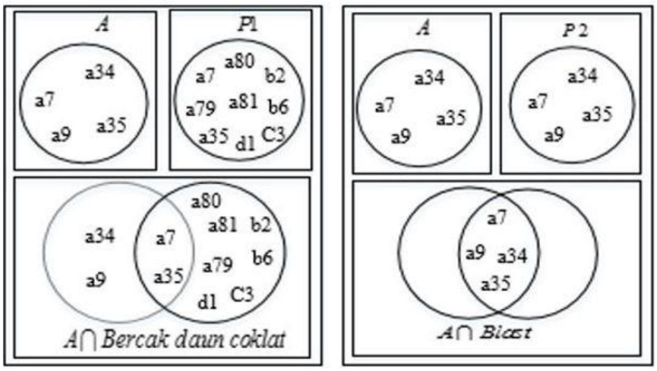


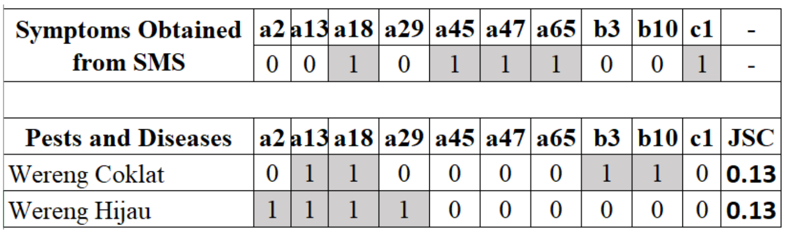
Figure 4. Pests and diseases identification with JSC

2.5. Interactive process

The same result of JSC causes ambiguity system to identify pests and diseases because there are similar symptoms between of one disease with another. To solve this problem, we propose the steps for the interactive process:

a. Show the pests and diseases that have the same JSC
b. Forming of the dynamic questions
c. Ask the main symptoms.
d. Send the symptoms, if the answer "yes" shows the result ofidentification, *else*s end the next symptoms.

There were five (5) symptoms obtained from SMS extraction (Figure 5), and one symptom that has a match with both of the pests and diseases.The result shows the pests and diseases identification that have the same value of JSC. To facilitate the identification process the system will send the main symptoms between both of the pests and diseases, e.g., a_2="daun kering" (dried leaf), or b_3="batang kuning" (yellow stem).



Symptoms Obtained from SMS	a2	a13	a18	a29	a45	a47	a65	b3	b10	c1	-
	0	0	1	0	1	1	1	0	0	1	-
Pests and Diseases	**a2**	**a13**	**a18**	**a29**	**a45**	**a47**	**a65**	**b3**	**b10**	**c1**	**JSC**
Wereng Coklat	0	1	1	0	0	0	0	1	1	0	**0.13**
Wereng Hijau	1	1	1	1	0	0	0	0	0	0	**0.13**

Figure 5. The same value of JSC

The next process is a chain construction for interactive questions. A chain of the questions is dynamically arranged based on the symptoms in the SMS sent by the farmers, so the illustration result is shown in Figure 6.

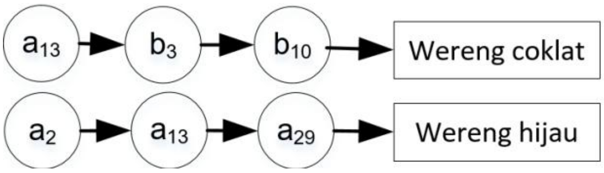


Figure 6. Illustration of interactive questions

The six symptoms in the illustration will be sent interactively, andthe farmers only answer yes or no. If the symptoms b3 = "batang kuning" and b10 = "batang kerdil" that are sent interactively, and answered "yes" then the conclusion is "Wereng Coklat", otherwise go to the symptoms a2 = "daun kering" and a_{29}= "daun busuk".

2.6. Analysis of testing result

The test was conducted to identify possible pests and diseases based on the information by the farmers. The test was done twice with the following formulations:

a. The test was conducted to measure the accuracy of expected symptoms from a user input. Symptoms identification is performed to measure the accuracy of symptoms according to the user input. Accuracy is calculated using the formula:

$$Symptoms\ Accuracy = \frac{Number\ of\ symptoms\ identified}{Number\ of\ symptoms\ on\ SMS} x100\%$$

b. The accuracy of the pests and diseasesidentification based on the symptoms is calculated using the formula as follows:

$$Accuracy = \frac{The\ number\ of\ true\ test\ result}{The\ number\ of\ testing\ data} x100\%$$

3. RESULTS AND ANALYSIS

3.1. The result of symptom identification

The experiment of symptom identification was done for 100 sentences. The experiment was performed using 40 sentences collected from the experts and 60 sentences collected from the farmers in the survey by Ministry of Agriculture and Fisher (MAF) Timor Leste. The experiment was done to test the accuracy of the system to identify symptoms based on the actual number of the symptoms. The experiment was conducted on all of 179 symptoms of pests and diseases and achieve the accuracy of 81%. Table 11 shows some examples and results of symptoms identification that has been generated by the systemreaching an accuracy of 100%.

Table 11. Examples of Input Sentence and Results of Symptoms Identified

Example of data input	Number of actual symptoms	Symptoms identified (%)
Tadi pagi saya ke sawah, tanaman padi saya kok daunnya kecoklatan dan bercak serta daun mengambang seperti ketupat. Bagaimana cara mengatasinya?	4	100%
daun tanaman kering dan mati. Tanaman ada yang menjadi kerdil, bagian pucuk berwarna kuning hingga kuning kecoklatan.	5	100%
Bagaimana mengatasi tanaman padi yang daun bercak cenderung lebih sempit, batang lebih pendek dan berwarna gelap. Berukuran berwarna coklat gelap.terima kasih.	5	80%
Apa yang menyebabkan daun padi menggulung dan kuning kecoklatan?	3	100%
Daun tanaman seperti bercak-coklat. bagian daun ada yang menjadi kerdil dan pucuk berwarna kuning hingga kuning kecoklatan.	4	75%
Selamat pagi, kami bingung karena tanaman padi kami daunnya bintik bintik dan lubang kecil pada daunnya, bagaimana mengatasinya?	2	100%
Selamat sore, tanaman padi kami malainya menjadi coklat, daun berbintik-bintik dan akar membusuk.	3	66.67%

Symptoms misidentification was due tothe symptoms that have more than one suffix, for example, “berbintik-bintik” (spotted), “bercak-coklat” (brown spot), “kekuning-kuningan” (yellowish). These words might be considered as root words during the stemming process, so it was not identified as symptoms. Misidentification of the second symptoms is a synonym problem, for example, the words “buah berserakan” (scattered fruit) while the intention of the user is “biji berserakan” (scattered seeds). Table 12 shows the result of the identification of symptoms obtained from the input sentence produced by the process of text mining.

For 100 trials, there were 81 trials identify the symptoms with the success rate of 100%, 9 trials reached the success rate of 80%, 75% for 6 trials, and 66.67% for 4 trials.

Table 12. The Result of Symptoms Identification

Amount of data	Number of actual symptoms	The number of errors symptoms identified
81	2, 3, 4,5,6,10	-
9	5	1
6	4	1
4	5	2

3.2. The result of pests and diseases identification

The database reference for pests and diseases identification consists of 60 data and 179 symptoms of pests and diseases. The selected data are the dominant pests and diseases in Timor-Leste. The data were obtained from Ministry of Agriculture Timor-Leste in the pest and disease research department and interviews with the experts. The data are used as a reference database to identify pests and diseases of rice plants from farmers by SMS.

The testing data of pests and diseases were obtained from the identification of the symptoms in the previous process from the text messages. The number of text messages was 100. Pests and diseases identification using 2 scenarios: the first scenario (A) used 40 data and the second scenario (B) used 60 data. The objectives of these two scenarios were to analyze the possible effects of the amount of data on the system's performance to identify pests and diseases based on JSC recommendation. JSC value is influenced by the number of the symptoms in the SMSsent by the farmers. The higher value of JSC recommendation means the higher possibility of rice plants affected by pests and diseases. The accuracy both of the scenarios are 86%.

The JSC recommendations consist of three outputs. The results of recommendation were sorting based on JSC value. The first recommendation shows the higher JSCscore and so on. But there were some pests and diseases have the same JSC recommendation value because has similar symptoms. The JSC recommendation in scenario A (Table 13) is obtained a value range of 0.08-0.75 and achieved the accuracy of 87.5%. The result of the experiment shows that there were 12 trials which had the same JSC with the same of pests and diseases, 23 trials had the different value of JSC same with the different pests and diseases, 4 trials had the same value of JSC with different pest and diseases.

Table 13. Results Identification Pests and Diseases (Scenario A)

Count of Symptoms	Pest and Diseases Identification	JSC Recommendation
5	Penggerek batang padi	0.75
	Penggerek batang padi	0.67
	Penggerek Batang Padi	0.43
4	Walang sangit	0.43
	Walang sangit	0.36
	Walang sangit	0.36
6	Wereng coklat	0.43
	Wereng hijau	0.40
	Wereng coklat	0.30
3	Fusarium	0.50
	Fusarium	0.30
	Bercak daun coklat	0.08

Table 13 shows some of the results of the recommendation of pests and diseases based on the three higher values ofJSC. The recommendation result on the JSC for diseases Penggerek batang padi =0.75, Penggerek batang padi =0.67 and Penggerek batang padi =0.43. So the result of identification is a disease “Penggerek batang padi”. The next results showed that the Fusarium appeared two times with the highest similarity value are 0.50 and 0.30, so it can be concluded that the likely outcome is a Fusarium disease identification with a value of JSC0.50. The three recommendation of JSC with 3 output the same pests and diseases can easy the system for decision making because one of the three outputs is concluded as a result of the identification.

The result of testing JSC recommendation in scenario B (Table 14) obtained a value range of 0.07-0.67 and achievedthe accuracy of 85%. From the results of 60 trials, there were 51 trials which had three outputs with the highest value in the first recommendation, 8 trials had the same output recommendations, whereas 2 trials were not identified. The results also show that there are some output results that have same value JSC recommendation, i.e. pests called “Wereng coklat”=0.17, “Wereng hijau”=0.17 and “Ulat tentara”=0.17. The problem causes ambiguity in decision-making. Then, the system will send interactive questions to the farmers toobtain additional symptoms to make a decision.

Table 14. Results Identification Pests and Diseases (Scenario B)

Count of Symptoms	Pest and Diseases Identification	JSC Recommendation
3	Wereng coklat	0.17
	Wereng hijau	0.17
	Ulat tentara	0.17
4	Blast	0.40
	Blast	0.27
	Penyakit kresek	0.27
3	Penyakit kresek	0.50
	Wereng hijau	0.40
	Wereng coklat	0.25
5	Blast	0.11
	Fusarium	0.11
	Fusarium	0.07
4	Wereng coklat	0.13
	Wereng hijau	0.13
	Hama putih	0.11

4. CONCLUSION

By utilizing SMS services, the farmers easily send and receive SMS interactively with the system. The system used SMS service technology to send and receive messages, and it does not require expensive devices, low cost effective, and easy to use by rural farmers. The application of text mining and similarity method as machine learning can automatically identify pests and diseases delivered by the farmers. The system succeeded to identify symptoms keyword of pests and diseases, and achievedthe accuracy of 81%.The system successfully identified pests and diseases with three outputs recommendation based on the expert knowledge. Pest and diseases identification achieved a value of 0.08 to 0.75 of JSC recommendations for scenario A and 0.07 to 0.67 for scenario B. The both accuracies of the scenarios are 87.5% and 85%.

In this study, there are still some problems in identifying symptoms with more than one suffix in the Indonesian language. In the process keyword identification, the system does not check errors typing and abbreviations by the farmer, therefore it is necessary to combine several other methods to solve the problems. The future studies are expected to include each symptom weights to make the system become more intelligent.

ACKNOWLEDGEMENTS

Edio da Costa thanks to Mr. Americo Alves, Head of the Department of Plant Protection and Agriculture. Mr. Justinho dos Santos Silva, the National Director of Food and Security, Ministry of Agriculture and Fisher (MAF) Timor-Leste, for providing the data.

REFERENCES

[1] Costa E, Tjandrasa H, Djanali S, “A conceptual Information of Technology Framework to Support Rice Farming in Timor Leste”, *International Conference on Information & Communication Technology and Systems (ICTS), 2015;* pp. 209-213.

[2] Horta M, “Produsaun Rai Laran Seidauk Sufsiente, Timor Leste Nafatin Importa Sasan Husi Rai Liur”, 2014. [Online]. Available: http://jornal.suara-timor-lorosae.com/produsaun-rai-laran-seidauk-sufsiente-tl-nafatin-importa-sasan-husi-rai-liur/. [Accessed: 02-Dec-2016].

[3] Amaral J, Brito A, “Integrated Pest Management Options for Sustainable Rice Production in Timor Teste”, *Tamil Nadu Rice Research Institute*, 2016.

[4] Istiadi, Sulistiarini EB, “Representing Knowledge Base Into Database for WAP and Web-based Expert System”, *International Conference on Information Systems for Business Competitiveness*, 2013, pp. 81-85.

[5] Navulur S, Sastry ASCS, Prasad MNG, “Agricultural Management through Wireless Sensors and Internet of Things”, *International Journal of Electrical and Computer Engineering (IJECE)*, 2017, vol. 7, no. 6, pp. 3492-3499.

[6] IRRI. How to manage pests and diseases. 2014. [Online]. Available: http://www.knowledgebank.irri.org/step-by-step-production/growth/pests-and-diseases. [Accessed: 05-Jan-2018].

[7] Nguyen CN, Thai-NgheN. *An Agricultural Extension Support System on Mobile Communication Networks.* International Conference on Advanced Technologies for Communications. 2015; 534-539.

[8] Istiadi, Sulistiarini EB, Putra GD, “Enhancing Online Expert System Consultation Service With Short Message

Service Interface”, *International Conference on Information Technolgy Computer and Electrical Engineering*, 2014 Nov, pp. 266-271.

[9] 2 Billion Consumers Worldwide to Get Smart (phones) by 2016, 2014. [Online]. Available: https:// www.emarketer.com/ Article/2-Billion-Consumers-Worldwide-Smartphones-by-2016/1011694. [Accessed: 05-Jan-2018].

[10] Devkota B, Adhikari B, Shrestha D, “Integrating Romanized Nepali Spellchecker with SMS Based Decision Support System for Nepalese Farmers”, *International Conference on Software, Knowledge, Information Management and Applications*, 2015.

[11] “Word Bank Data,” 2015. [Online]. Available: https://data.worldbank.org/indicator/IT.MLT.MAIN.P2?end=2016&locations=TL&start=2003&view=chart,[Accessed: 05-Jan-2018].

[12] E. Soares D. Doradi, “Timor-Leste Communication and Media Survey, United Nations Integr”, Mission Timor-Leste, 2011.

[13] Jonnagaddala J, Liaw ST, Ray P, Kumar M, Chang NW, Dai HJ, “Coronary Artery Disease Risk Assessment from Unstructured Electronic Health Records Using Text Mining”, *Journal of Biomedical Informatics*, 2015, vol. 58, pp. 203-210.

[14] Hashimi H, Hafez A, Mathkour H, “Selection Criteria for Text Mining Approaches”, *Computer in Human Behavior,* 2015, vol. 51, pp. 729-733.

[15] Naradhipa AR, Purwarianti A, “Sentiment Classification for Indonesian Message in Social Media”, *International Conference on Electrical Engineering and Informatics*, 2012, vol. 11, pp. 1-5.

[16] Suwarningsih W, Purwarianti A, Supriana I, “Extraction of Predicate-Argument Structure From Sentence Based on PICO Frames”, *International Conference on Automation and Information Technology*, 2015, pp. 91-95.

[17] Margono H, Yi X, Raikundalia GK, “Mining Indonesian Cyber Bullying Patterns in Social Networks”, *Proceedings of the Thirty-Seventh Australasian Computer Science Conference*, 2014, pp. 115-124.

[18] Mooney RJ, “Machine Learning Text Categorization”, University of Texas Austin, 2006.

[19] Bano S, Rao KR, “Partial Context Similarity of Gene/Proteins in Leukemia UsingContext Rank Based Hierarchical Clustering Algorithm”, *International Journal of Electrical and Computer Engineering (IJECE),* 2015, vol. 5, no. 3, pp. 483-490.

[20] Tala FZ, “A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia”, Thesis, Institute for Logic, Language and Computation, Universiteit Van Amsterdam The Netherlands, 2003.

[21] Veritawati I, WasitoI I, Basaruddin T, “Text Preprocessing using Annotated Suffix Tree with Matching Keyphrase”, *International Journal of Electrical and Computer Engineering (IJECE),* 2015, vol. 5, no. 3, pp. 409-420.

[22] Lim MS, Hocking JS, Hellard ME, Aitken CK, “SMS STI: A Review of the Uses of Mobile Phone Text Messaging”, in *Sexual Health International Journal of STD & AIDS*, 2008, vol. 199, no. 5, pp. 287-290.

[23] Trogo R, Bagtasa G, Tongson E, Balderama O, ”SMS-based Smarter Agriculture Decision Support System for Yellow Corn Farmers in Isabela”, *Canada International Humanitarian Technology Conference*, 2015.

[24] Cole SA, Fernando AN, “The Value of Advice: Evidence from Mobile Phone-Based Agricultural Extension”, *International Conference of the African Association of Agricultural Economists*, 2013.

[25] Brugger F, “Mobile Applications in Agriculture”, *Syngenta Foundation*, Switzerland, 2011.

[26] Cockings C, “Application Note 010 GSM AT Command Set Cambridge Technology Centre Melbourn”, 2001.

[27] Choi S, Cha S, Tappert CC, “A Survey of Binary Similarity and Distance Measures”, *Journal of Systematics Cybernetics and Informatics*, 2010, vol. 8, no. 1, pp. 43-48.

[28] Abuzir Y, Sabbah T, “First Token Algorithm for Searching Compound Terms Using Thesaurus Database”, *Journal of Computer Science*, 2012, vol. 8, no. 1, pp. 61-67.

[29] Setiawan E, Kamus Besar Bahasa Indonesia (KBBI) [Online]. Available: http://kbbi.web.id/ [Accessed:25-April-2017].

BIOGRAPHIES OF AUTHORS

Edio da Costa has received his BSc. from Dili Institute of Technology, Timor-Leste in 2010 and M.Cs. from Satya Wacana Christian University, Indonesia. Currently, he is a Doctoral Candidate at Sepuluh Nopember Institute of Technology, Surabaya, Indonesia. His research interests are in Text Mining and Artificial Intelligence. He has worked for Dili Institute of Technology, Timor Leste since 2010.

Handayani Tjandrasa is a professor at the Department of Informatics, Sepuluh Nopember Institute of Technology, Surabaya, Indonesia. She received her M.Sc. and Ph.D. from University of Wisconsin-Madison, USA. Her research interests are in the areas of computational intelligence, digital image processing, and biomedical engineering.

Supeno Djanali is a Professor of Network Architecture and Design in Department of Informatics, Sepuluh Nopember Institute of Technology, Surabaya, Indonesia. He earned his master and Ph.D. degrees from University of Wisconsin-Madison, USA. His research areas are primarily network security and mobile computing.