AUTOMATIC QUESTION ANSWERING SYSTEM USING RELEVANCE VECTOR MACHINE CASE STUDY: FRONT OFFICE

Gingga Ismu Muttaqin Hadiko[1], Ken Kinanti Purnamasari[2]

[1,2]Universitas Komputer Indonesia
Jl. Dipatiukur No. 112-116, Bandung 40132
E-mail : gingga_ismu_m@email.unikom.ac.id[1], ken.kinanti@email.unikom.ac.id[2]

ABSTRACT

Question Answering System is a system on a computer that can answer questions using natural language that humans normally use automatically. The answers obtained from the system contain information that comes from a database source. A question and answer system, is expected to help people to get more information about certain things without having to ask others. The main purpose of an automatic answer system is to find the right answer using natural language that is determined by humans. Previous studies using the SVM (Support Vector Machine) method resulted in a classification accuracy of 46% and an answer accuracy of 39%. This study uses the RVM (Relevance Vector Machine) method to classify questions into 7 classes. Before entering the main stages, the stages in preprocessing consist of case folding, cleansing (symbol removal), tokenizing, stopword removal, and stemming. The results showed that the classification accuracy was 71.21% and the answer accuracy was 66.67%.

Keywords : Question Answering System, Relevance Vector Machine, One VS One, Cosine Similarity.

1. INTRODUCTION

An information can be obtained by asking. But everyone has limitations in terms of information and the time they have to answer. So, an automatic answer system is built with the main goal is to find the right answer using natural language determined by humans [5]. The answers obtained from the system contain information that comes from a database source [4].

Previous research, the question and answer system in the UNIKOM front office case has been carried out using the SVM method. The study resulted in an answer accuracy of 39%. One of the factors causing the low accuracy of answers lies in the question classification stage. At this stage SVM classifies questions with an accuracy of only 46% [2].

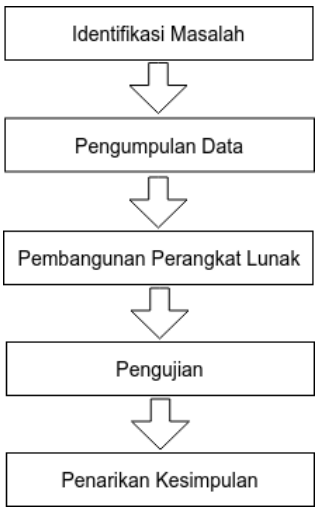
Therefore this research will try to use a different classification method. Based on research conducted by Rafi and Shaikh [3], the RVM method has better accuracy than SVM with the results of the f-measure percentage for RVM is 92.33%, while for SVM is 90.29%. Based on the advantages of RVM compared to SVM, this study will analyze the accuracy of the classification of questions using RVM. So it is hoped that this method can produce better classification accuracy and answers.

2. RESEARCH CONTENTS

This section explains the research conducted, covering research methods, system architecture, preprocessing.

2.1 Research Method

The method that will be used in this study is the experimental method. This method was chosen because hypothesis testing will be carried out using a series of experiments. The description of the stages in this study is as follows.

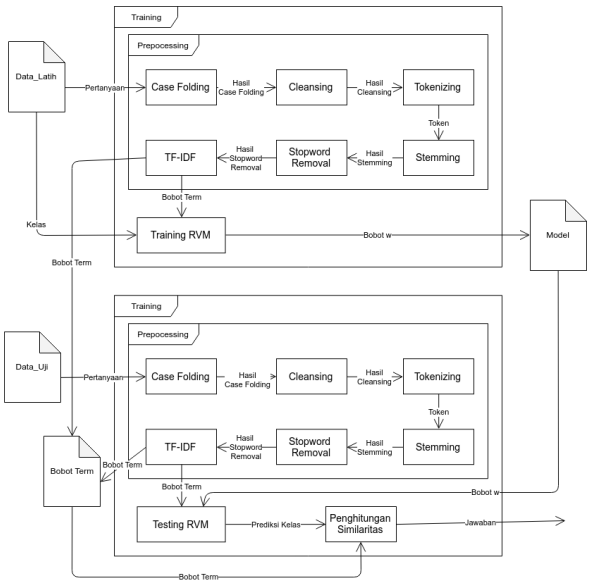


Gambar 1. Tahapan Penelitian

2.2 System Architecture

This system consists of two main parts, namely training and testing. Data input in the form of question text. Then the data is processed in the preprocessing stage to get the weight value. At the preprocessing stage consists of case folding, cleansing, tokenizing, stopword removal, and weighting. In the training section, the term weight values will be mapped to each class using the

RVM method. The classes are grouped by type of question. Then the results of the training in the form of w weights are stored in a file. While in the testing section, the term weight value will be predicted by the class based on the model that has been made. The class prediction will be used as a determinant of which class of data will be used at the cosine similarity stage. The following is a general description of the system being built.



2.3 Preprocessing

Preprocessing is the process of processing raw data before it is processed at the main stage. The stages in preprocessing in this study consisted of case folding, cleansing (symbol removal), tokenizing, stopword removal, and stemming.

2.3.1 Case Folding

Case folding is the process of uniforming letters in the form of capital letters (uppercase) or lowercase letters (lowecase) [6]. The following is an example of a folding case.

Tabel 1. Example of Case Folding

Before	ada berapakah fakultas di unikom?
After	ada berapakah fakultas di unikom

1.1.1 *Cleansing*

Cleansing is the process of removing punctuation [6]. The following is an example of cleansing.

Tabel 2. Example of Cleansing

Before	ada berapakah fakultas di unikom?
After	ada berapakah fakultas di unikom

2.3.2 Tokenizing

Tokenizing is the process of breaking a sentence into several words [6]. The following is an example of tokenizing.

Tabel 3. Example of Tokenizing

Before	ada berapakah fakultas di unikom
After	ada
	berapakah
	fakultas
	di
	unikom

2.3.3 Stemming

Stemming is the process of turning a word into its basic word [6]. The following example stemming.

Tabel 4. Example of Stemming

Before	**After**
ada	ada
berapakah	berapa
fakultas	fakultas
di	di
unikom	unikom

2.3.4 Stopword Removal

Stopword Removal is the process of removing words that are considered not important. The following example semoval stopword [6].

Tabel 5. Example of Stopword Removal

Before	**After**
ada	
berapa	berapa
fakultas	fakultas
di	
unikom	unikom

2.3.5 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 [7]. The first step in calculating weights with TF-IDF is to count the number of words in a question, term frequency (TF). Here is an example of TF on the word "kapan".

Term	tf																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
kapan	1	1	0	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0

Gambar 2. Example of Term Frequency

After that, count the number of questions that contain the word "kapan", document frequency (DF). Here's an example of DF.

Term	tf																				df
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
kapan	1	1	0	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	6

Gambar 3. Example of Document Frequency

Then calculate the value of Inverse Document Frequency (IDF) using the value of DF. The following is an example of calculating the IDF value for the word "kapan".

$$\begin{aligned} IDF &= \log\left(\frac{D}{df}\right) \\ &= \log\left(\frac{21}{6}\right) \\ &= 0.5441 \end{aligned}$$

Where:

D = Total document

df = Number of documents containing a word

In the last step, calculate the TF-IDF weight value by multiplying the TF and IDF values [6]. Following is an example of TF-IDF calculation for the word "kapan".

$$\begin{aligned} w_1 &= tf * IDF \\ &= 1 * 0.5441 \\ &= 0.5441 \end{aligned}$$

Where:

tf = Term frequency or word frequency

w = Weight of a document of a word

2.4 Relevance Vector Machine

Relevance Vector Machine (RVM) was introduced by Michael E. Tipping in 2000. RVM is a machine learning method adapted from the Bayesian Framework. Besides RVM has similarities with Support Vector Machine (SVM) in terms of function models. Like SVM, RVM was developed for binary analysis [8].

RVM is included in supervised learning. As with other supervised learning methods, RVM also requires training data that consists of a vector set mapped to a target, where the target is a value for regression and class label for classification. The purpose of supervised learning is to train a model using a number of training data, so that it is expected that an input x can be predicted as accurately as possible in the grade or grade. Predictions using the RVM method can be calculated using equation [9]:

$$y(x;w) = \sum_{i=1}^{M} w_i \varphi_i(x) + w_0 = w^T \varphi(x) + w_i$$

where w is a weight vector, φi (x) is the kernel's function of data x, and wo is biased.

The kernel function used is the RBF (Radial Base Function) kernel. This function is calculated using the equation:

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$$

where x and x 'are input data.

The classification in this case is that the prediction model takes the form of a linear combination of the base / kernel function which is changed by the sigmoid logistic function [9].

$$\begin{aligned} y(x, w) &= \sigma(x; w) \\ &= \sigma(w^T \varphi(x)) \end{aligned}$$

where σ (.) is a sigmoid logistic function defined by the equation.

$$\sigma(-y) = \frac{1}{1 + \exp(-y)}$$

Based on the definition of the Bernoulli distribution, likekihood is defined as follows:

$$p(t|w) = \prod_{t=1}^{N} \sigma\{y(x_n; w)\}^{t_n} \{y(x_n; w)\}^{1-t_n}$$

for the target $t_n \in \{0,1\}$.

The likelihood equation is equipped with a priority to the parameter (weight) in the form

$$p(w|a) = \prod_{t=1}^{N} \frac{\sqrt{a_i}}{2\pi} \exp\left(-\frac{a_i w_i^2}{2}\right)$$

where α = (α1, α2, ... αn) T is a hyperparameter that is introduced to control the strength and priors of the associated weight parameters, N is a lot of sentences, and w is a weight parameter.

For certain α values, the posterior distribution of weights against data can be calculated using the Bayes rule, with the equation:

$$p(w|t, a) = \frac{p(t|w)(p|a)}{p(t|a)}$$

For certain α values, the posterior distribution of weights against data can be calculated using the Bayes rule, with the equation:

$$p(w|t,a)=\frac{p(t|w)(p|a)}{p(t|a)}$$

where p (t | w) is likelihood, p (w | α) is prior, and p (t | α) is evidence.

Model weight parameters cannot be obtained by analytical means, so Laplacian approximation is used. Since p (w | t, α) is linearly proportional p (t | w) x p (w | α), it is possible to find the maximum of the equation

$$\ln p(w|t,\alpha)=\ln\{p(t|w)p(w|\alpha)\}+\ln p(t|\alpha)$$
$$=\sum_n^N\{t_n\ln y_n+(1-t_n)\ln(1-y_n)\}+\frac{1}{2}w^TAw$$

2.5 Accuracy

RVM performance can be known by calculating accuracy. Accuracy is the number of correct predictions divided by the amount of data. This study calculates 2 pieces of accuracy, namely class prediction accuracy and answer prediction accuracy. The following formula for calculating accuracy [10].

$$Akurasi=\frac{Jumlah\ Prediksi\ Benar}{Jumlah\ Data}*100\ \%$$

2.6 Hasil Pengujian

Testing is done by using 66 test data in the form of questions. The following are examples of some test data used.

Tabel 6. Example of Test Data

No	Questions	Class
1	Kapan pendaftaran dibuka?"	3
2	berapa biaya semester di unikom?	6
3	Kapan Jadwal PMB UNIKOM?	3
4	Persyaratannya apa saja ?	2
5	Berapa biaya pendaftarannya?	6
...	…	...
62	berapa biaya autodebet?	7
63	Tanggal Autodebet?	1
64	kapan autodebet yang pertama?	1
65	apa saja persyaratan untuk mendaftar	2
66	tanggal berapa mulai test untuk gelombang 2	3

After going through the classification stage, the prediction results obtained for each class as follows.

Tabel 7. Result of Classification

No	Real Class	Prediction Class	Result
1	3	3	Benar
2	6	6	Benar
3	3	2	Benar
4	2	2	Benar
5	6	6	Benar
	...	...	
62	7	6	Salah
63	1	5	Benar
64	1	1	Benar
65	2	2	Benar
66	3	3	Benar

Based on the tests conducted, the system performs 47 class predictions correctly. Next, calculate the classification accuracy using the following formula.

$$Akurasi=\frac{Jumlah\ Prediksi\ Kelas\ Benar}{Jumlah\ Data}*100\ \%$$
$$=\frac{47}{66}*100\ \%$$
$$=71{,}21\%$$

After getting the classification results, then determine the answer based on class predictions. So we get the following prediction answers.

Tabel 8. Result of Prediction Answer

No	Test Question	Closest Question	Answer
1	Kapan pendaftaran dibuka?	Kapan di buka pendaftaran pmb	Pendaftaran Tahun akademik 2017/2018 dibuka mulai tanggal 1 Maret 2017
2	berapa biaya semester di unikom?	berapa biaya semester di unikom?	Untuk Angkatan 2017 sebesar 6 juta
3	Kapan Jadwal PMB UNIKOM?	Ada berapa jurusan di Unikom?	ada 26 Jurusan
...	...	...	...
64	kapan autodebet yang pertama?	batas akhir autodebet pertama kapan?	Batas waktu autodebet terakhir adalah tanggal 15 Agustus 2016
65	apa saja persyaratan untuk mendaftar	persyaratan untuk pendaftaran apa saja?	Syarat Pendaftaran : 1. lulusan smu/sederajat 2. membayar biaya pendaftaran sebesar Rp. 350.000. 3. mengisi formulir pendaftaran di situs pmb unikom (pmb.unikom.ac.id)
66	tanggal berapa mulai test untuk	Kapan mulai daftar?	Pendaftaran dimulai tanggal 1 maret 2016-19 juli 2016

	gelombang 2		

The test results showed 44 relevant answers from 66 data tested. So the answer accuracy can be calculated as follows.

$$Akurasi = \frac{Jumlah\ Prediksi\ Jawaban\ Relevan}{Jumlah\ Data} * 100\,\%$$
$$= \frac{44}{66} * 100\,\%$$
$$= 66{,}67\%$$

This research resulted in a classification accuracy of 71.21% and an answer accuracy of 66.67%.

2.7 Analysis of Test Results

System accuracy is influenced by various factors. The following analysis of the accuracy of the system has been obtained.

1. Misclassification of questions on the system occurs because there are questions in the test data that are not clear which class belongs.

2. Another thing that causes misclassification is the low weight on keywords for a class.

3. The main factor in predicting an irrelevant answer is an error in the class classification stage.

4. Prediction of the answer is determined by the answer data contained in the training data. This causes the answer to be irrelevant if there is no answer in the training data that matches the question.

3. CLOSING

3.1 Conclusions

Based on the results of research that has been done, it is known that the application of RVM in the question and answer system in the case of the front office gets a classification accuracy of 71.21% and an answer accuracy of 66.67%.

3.2 Suggestions

Based on the analysis of test results, the accuracy of the system built can still be improved. Therefore, it is expected that further research can be further refined and developed. The following are suggestions for further research.

1. Add more training data with more varied questions.

2. Some words that have the same meaning (synonyms) are counted as one word, so that they have greater weight.

REFERENCES

[1] D. Zhang and W. S. Lee, ‘Question Classification using Support Vector Machines’, p. 7.

[2] T. E. Hutapea, ‘Penerapan Metode SVM Untuk Sistem Tanya Jawab Pada Kasus Front-Office’, p. 6.

[3] M. Rafi and M. S. Shaikh, ‘A comparison of SVM and RVM for Document Classification’, Procedia Computer Science, p. 6, 2013.

[4] E. J. Wantroba and R. A. Romero, ‘An interactive question-answer system with dialogue for a receptionist avatar’, in 2015 12th Latin American Robotics Symposium and 2015 3rd Brazilian Symposium on Robotics (LARS-SBR), 2015, pp. 360–365.

[5] L. Hirschman and R. Gaizauskas, ‘Natural language question answering: the view from here’, Natural Language Engineering, vol. 7, no. 04, Dec. 2001.

[6] A. R. Sentiaji and A. M. Bachtiar, ‘Analisis Sentimen Terhadap Acara Televisi Berdasarkan Opini Publik’, Bandung: Universitas Komputer Indonesia, 2014.

[7] A. A. Maarif, ‘Penerapan Algoritma TF-IDF Untuk Pencarian Karya Ilmiah’, Jurnal. Jurusan Teknik Informatika. Fakultas Ilmu Komputer. Universitas Dian Nuswantoro Semarang, 2015.

[8] G. M. Foody, ‘RVM-based multi-class classification of remotely sensed data’, International Journal of Remote Sensing, vol. 29, no. 6, pp. 1817–1823, Mar. 2008.

[9] M. E. Tipping, ‘Sparse Bayesian learning and the relevance vector machine’, Journal of machine learning research, vol. 1, no. Jun, pp. 211–244, 2001.

[10] B. Santosa and A. Umam, Data Mining dan Big Data Analytics: Teori dan Implementasi Menggunakan Python & Apache Spark. Yogyakarta: Penebar Media Pustaka, 2018.