Literation Hearing Impairment (I-Chat Bot): Natural Language Processing (NLP) and Naïve Bayes Method

Merry Anggraeni1 Mohammad Syafrullah2 Hillman Akhyar Damanik3
1Dept. Master of Computer Science, Faculty of Information Technology Budi Luhur University, Jakarta, Indonesia
2Faculty of Information Technology Budi Luhur University, Jakarta, Indonesia
3Dept. Master of Computer Science, Faculty of Information Technology Budi Luhur University, Jakarta, Indonesia

Abstract. A part from advances in Artificial Intelligence and Natural Language Processing, designed CHATBOT will try to understand user requests accurately, so that they do not give the wrong answer or no response. The difficulty of getting specific information about hearing impairments that can be asked properly is asking someone who understands that it still cannot be found on search engines on the internet which is a favorite tool for most people in obtaining information. By using Machine Learning and with the help of Artificial Intelligence, a question and answer (conversation) interaction mechanism is created to gain literacy knowledge that supports the educational process. Therefore, the CHATBOT application was developed in this study as a media for retrieving information about hearing loss. Using the NLP method and the Naïve Bayes Algorithm for classifications used to get input classes to I-Chat Bot, as well as to test hypotheses using the Technology Acceptance Model developed (extended). The result is an I-Chat Bot with artificial intelligence that understands user input and provides an appropriate response and produces a preferred and easy system model to be used in the search for information about required hearing impairments. This result paper also gets the value of the test accuracy with Precision 98.6%, Recall 88.75% and Accuracy 88.75%.

1. Introduction
A part from advances in Artificial Intelligence and Natural Language Processing, designed CHATBOT will try to understand user requests accurately, so that they do not give the wrong answer or no response. An effective solution to overcome this problem is to involve human ability in CHATBOT operations to understand user requests [1]. In general, chat bots are designed using 3 approaches, namely pattern matching, through algorithms and neural networks [2]. In this study, researchers provide insight into how to design the I-Chat Bot system that is different from previous studies that use a pattern matching approach that usually uses AIML (Artificial Intelligent Markup Language), in this study researchers used an approach through algorithms, namely Algorithms Naïve Bayes will be used at the text classification stage with the NLTK (Natural Language Tool Kit) module on Natural Language Processing (NLP) from Python, so CHATBOT can be trained, making CHATBOT able to handle user requests in terms of information about hearing loss accurately.

2. Literature Review
A. Natural Language Processing (NLP)
In this study CHATBOT (an application that allows users to communicate with a computer), Stemming or Lemmatization (cutting words in a particular language into the basic form of recognition of the function of each word in a sentence), Summarization (summary of reading), Translation Tools (translating language) and other applications that allow computers to be able to understand language instructions entered by users [3]. PUSTEJOVSKY and Stubbs explain that there are several main areas of research in the NLP field, including [4]: Question Answering Systems (QAS), Summarization, Machine Translation, Speech Recognition, Document classification, Identify the types of entities extracted, Using Sentiment Analysis to identify sentiments from a series of texts, Information Retrieval

B. Naïve Bayes Method

The advantage of using this method is that it only requires a small amount of training data to determine the parameter estimates needed in the classification process [5]. Because it is assumed to be an independent variable, only variants of a variable in a class are needed to determine the classification, not the whole of the covariance matrix [6].

Naive Bayes process steps
1. Calculate the number of classes / labels
2. Calculating the Number of Cases Per Class
3. Multiply All Class Variables
4. Compare Results Per Class

Taking into account the class variables and dependent feature vectors, Bayes Theorem states the following relationship:

\[ P (A |B) = \frac{P(B |A)P (A)}{P(B)} \]  

Where are A and B an event and P (B)?

- Basically, trying to find the probability of event A, given the event B is correct. Event B is also referred to as evidence.
- P (A) is priori A (previously possible, example the probability of an event before the evidence is seen). Evidence is the attribute value of an unknown example (here, it is event B).
- P (A | B) is a posteriori probability B, which is the probability of an event after the evidence is seen.

Now, with regard to our data collection, we can apply Bayes theorem in the following ways:

\[ P (Y | X) = \frac{P(X | Y)P (Y)}{P(X)} \]  

Where;

- Y is a class variable
- X is a dependent feature vector (size n) where:

\[ X= (x_1, x_2, x_3, ..., x_n) \]  

So basically, P (X | y) here means, possibly "y" given that the condition "X". The naive assumption of Bayes theorem, which is the independence between its features. First the evidence is divided into independent parts. If there are two independent A and B events, then:

\[ P(A,B) = P(A)P(B) \]  

Therefore, the results achieved are which can be expressed as:

\[ P (y|x_1, ..., x_n) = \frac{\prod_{i=1}^{n} P (x_i | y)}{P (x_1)P (x_2)...P (x_n)} \]  

Now, because the denominator remains constant for the given input, the term can be removed:

\[ P (y|x_1, ..., x_n) \propto P (y) \prod_{i=1}^{n} P (x_i | y) \]
Now, it is necessary to make classifying models. For this, we find the probability of the input set given for all possible values of the class \( y \) variable and take the output with the maximum probability. This can be stated mathematically as:

\[
y = \arg \max_y P(y) \prod_{i=1}^{n} P(x_i | y)
\]

So, finally, just count \( P(y) \) and \( P(x_i | y) \). Please note that \( P(y) \) is also called class probability and \( P(x_i | y) \) is called conditional probability. The naive Bayes classifiers differ primarily by assumptions made regarding the distribution of \( P(x_i | y) \).

3. Research Method

In general, the system that will be chosen to be designed in solving research problems will use Architecture with a Retrieval Based Model, because chat can describe the same message with different expressions. CHATBOT can have a separate response module and response [7] [8], as shown in Figure 1 below:

![Figure 1. Architecture adopted for system design](image)

System Design: I-Chat Bot: Interactive Question and Answer Using Natural Language Processing and Bayesian Algorithms for Literary Literacy.

![Figure 2. System Design: I-Chat Bot](image)
In this research set data there are 5 Classifications of Classes / Categories in the corpus database to be used as training data written in YAML files with an extension (.yml). When the application starts, the training data in the corpus is placed in the SQLite database [9] [10]:

When there are input questions, the steps taken by the Naive Bayes Algorithm in the Logic Adapter are as follows:

| Doc | Class/Category |
|-----|----------------|
| Data Training | what hearing impairment | Greetings |
|      | what is hearing impairment ? | Greetings |
|      | Want to ask about Greetings hearing impairment | Greetings |

a) In the corpus has a set of reviews (documents) and classifications.

b) Input to the adapter logic: "What does the hearing impairment mean?" to process the response starting from determining what class the question is.

| Doc | Class/Category |
|-----|----------------|
| Data Training | what hearing impairment | Greetings |
|      | what is hearing impairment ? | Greetings |
|      | Want to ask about Greetings hearing impairment | Greetings |
| Data Test | what does that mean hearing impairment ? | ? |

c) Represent each document with word vectors, one attribute per word position in the document

| Doc | Greetings hearing what what ask ask about purpose Class/Category |
|-----|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
|     | what hearing impairment | what is hearing impairment | Want to ask about Greetings hearing impairment | what hearing impairment | what is hearing impairment | Want to ask about Greetings hearing impairment | what hearing impairment | what is hearing impairment | Want to ask about Greetings hearing impairment |

| Doc | Greetings hearing what what ask ask about purpose Class/Category |
|-----|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
|     | Greetings | what hearing impairment | what is hearing impairment | Want to ask about Greetings hearing impairment | what hearing impairment | what is hearing impairment | Want to ask about Greetings hearing impairment |
| Data Test | what does that mean hearing impairment ? | ? |

d) Calculate Prior Probability P(Greetings) and P(Hearing)

\[
P(\text{Greetings}) = \frac{N_g}{N} = \frac{2}{3} = 0.67 \\
P(\text{Hearing}) = \frac{N_h}{N} = \frac{1}{3} = 0.33
\]

e) Calculate Likelihood, where \( n \) = number of words in a Class of Disruption or Greetings. \( n_k \) = how many times the word \( k \) occurs in this case

\[
V = \text{vocabulary size} \\
P(\text{wk} | \text{Greetings}) = \frac{n_k + 1}{n + |V|} \\
P(\text{wk} | \text{Hearing}) = \frac{n_k + 1}{n + |V|}
\]

Result Likelihood:

\[
P(\text{Greetings} | \text{Greetings}) \approx (2+1)/(8+8) = 0.1875 \\
P(\text{Hearing} | \text{Greetings}) \approx (2+1)/(8+8) = 0.1875 \\
P(\text{it} | \text{Greetings n}) \approx (2+1)/(8+8) = 0.1875 \\
P(\text{purpose} | \text{G Greetings}) \approx (0+1)/(8+8) = 0.0625 \\
P(\text{what} | \text{Greetings}) \approx (2+1)/(8+8) = 0.1875
\]
\[ P(\text{Greetings} | \text{Greeting}) = \frac{1+1}{8+8} = 0.153846154 \]
\[ P(\text{Hearing} | \text{Greeting}) = \frac{1+1}{8+8} = 0.153846154 \]
\[ P(\text{it} | \text{Greeting}) = \frac{0+1}{8+8} = 0.076923077 \]
\[ P(\text{about} | \text{Greeting}) = \frac{0+1}{8+8} = 0.076923077 \]
\[ P(\text{what} | \text{Greeting}) = \frac{0+1}{8+8} = 0.076923077 \]

\( a) \) Select class: \( y = \text{argmax}_y P(y) \prod_{i=1}^n P(x_i | y) \)
\[ P(\text{Greeting} | \text{doc4}) = 5.17559E-05 \]
\[ P(\text{Greeting} | \text{doc4}) = 3.55514E-06 \]

Then the results of the class / category obtained for the question: "What does the hearing impairment mean?" Goes into the class = Interference

4. Result and Discussion

4.1 Testing

4.1.1 Accuracy Testing with Confusion Matrix

The accuracy of the answers is done to find out how accurate the response of the answers given by the application to the questions of the user. Of the 5 Database Categories or classes and 344 statements / statements, 80 random questions were tested between classes / categories. Based on the rule of thumb of information retrieval, the minimum number of searches for data in a test is 50 [7]. The relevance of the document is determined by one judge as a measurement point by following the standard measurement of unraked retrieval sets. Measurement uses a binary value, which is a value of "0" to indicate an irrelevant document and a value of "1" to indicate the relevant document. The following is a calculation of Recall and Precision according to the relevance test results denoted in Table 1.

| Prediction Value | True Value | False Value |
|------------------|------------|-------------|
| TRUE             | 71         | 0           |
| FALSE            | 1          | 9           |

From the test results obtained several possibilities are:

1. True Positive (TP) is the answer that the system produces correctly.
2. False Positive (FP) is the answer that is generated wrong or the system does not produce an answer
3. True Negative (TN) that is the question submitted does not comply with the provisions and the system does not produce an answer
4. False Negative (FN) is the question submitted does not comply with the provisions but the system produces an answer.

The following formula is used to calculate precision and recall along with accuracy in the system:

Recall, True Positive Rate (Sensitivity) \( \frac{\text{TP}}{\text{TN}+\text{FP}} \)
Recall, True Negative Rate (Specificity) \( \frac{\text{TN}}{\text{N}} \)

Precision, Positive Predictive Value = \( \frac{\text{TP}}{\text{TP}+\text{FP}} \)
Precision, Negative Predictive Value = \( \frac{\text{TN}}{\text{TN}+\text{FN}} \)
This can indicate that accuracy is a function of sensitivity and specify Accuracy = Sensitivity \( \frac{P}{P+N} \)
+ Specificity \( \frac{N}{P+N} = \frac{TP+TN}{P+N} = \frac{TP}{\text{Jumlah Populasi(Classification as)}} \)

Overall Comparison Test Results are in Table 2.

Table 2. Accuracy Test Results with Confusion Matrix

| Precision | recall   | Accuracy |
|-----------|----------|----------|
| 98,61111% | 88,75%   | 88,75%   |

4.1.2 Reliability Testing

Table 3. Instrument Reliability Test Results

|                | Odd Amount | Even Number |
|----------------|------------|-------------|
| Odd Amount     | 1          |             |
| Even Number    | 0.912083551| 1           |

The results of this correlation coefficient are then included in the Spearman Brown formula: 
\( r = \frac{(2r)(1+r)}{1+r} \) \( r = 0.954020603 \). Based on the results of this value test has been reliable, because it is greater than 0.600. So the instrument used is reliable, so the instrument can be used for measurement in the context of data collection.

4.1.3 Hypothesis Testing

Hypothesis testing is done by using Path Analysis which is one of the statistical analysis techniques used in quantitative research, in this study to find out how the relationships and influences occur: 1. Between exogenous variables (perceived ease of use (PEOU) and perception of usefulness (PU)) towards endogenous variables (attitudes towards the use of I-Chat Bot (ATU)) 2. Between exogenous variables (perception of ease of use (PEOU) and perception of usefulness (PU) and Attitudes towards use (ATU)) to endogenous variables (actual I-Chat Bot (AU) usage) which is carried out with MS analysis tools Excel.

A. The relationship and influence that occurs between the exogenous variables of perception of ease of use (PEOU) / (X1) and perceived usefulness (PU) / (X2) of endogenous variables on the use of I-Chat Bot (ATU) / (Y) are presented in the figure 3. the following:

![Figure 3. Multiple Regression Model Pathway Analysis of X1 and X2 against Y](image)

This output summary table reports the strength of the relationship between the model (independent variables) and the dependent variable.
Table 4. SUMMARY OUTPUT PEOU and PU towards ATU

| Regression Statistics |
|-----------------------|
| Multiple R           | 0.753922439 |
| R Square             | 0.568399045 |
| Adjusted R Square    | 0.536428603 |
| Standard Error       | 0.786529107 |
| Observations         | 30          |

Table 5. ANOVA Simultaneous Effect Test of PEOU and PU on ATU

| df | SS     | MS      | F          | Significance F |
|----|--------|---------|------------|----------------|
| Regression | 2 | 21,99704303 | 10,99852151 | 17.77889276    | 0.000012        |
| Residual    | 27 | 16,70295697 | 0.618628036 |                |                |
| Total       | 29 | 38,7      |            |                |                |

Based on table 5 shows the F value of 17.77889276 which is known as F count in hypothesis testing compared with F table value in this study obtained F table with 30 correspondents of 3.340385558. Because F count> F table, it can be stated that simultaneously (together) PEOU and PU have a significant effect on ATU. Besides that, it can also be compared between the real level and the p-value (in Excel terms is Significance F). If the real level> of the p-value is the same conclusion as above. In this case the real level is set at 5%. Because p-value (Significance F) = 0.000012, it can be concluded that PEOU and PU jointly have a significant effect on ATU.

Table 6. Coefficient Regression of T PEOU and PU Tests on ATU

| Coefficients | Standard Error | t Stat | P-value |
|--------------|----------------|--------|---------|
| Intercept    | 2.3954         | 1.1414 | 2.0986  | 0.0453  |
| Amount PU    | 0.1916         | 0.1006 | 1.9049  | 0.0675  |
| Amount PEOU  | 0.1496         | 0.0786 | 1.9047  | 0.0675  |

Based on the T Calculate (t Stat) and T Table comparisons. Calculated T value can be seen in column t stat. A variable is said to have a significant effect if the T value is calculated> T table. The T value of the table can be known briefly using the formula on MS. Excel (without having to manually open T Table). The probability is 0.05 and deg freedom is 28 (the amount of data - the number of independent variables - 2) is obtained T Table = 1.7011. So it can be concluded that Perceived ease of use and perceive usefulness partially affect the attitude toward using.

Figure 4. Residual Output Predicted Data Points

Implications
Relationships and influences that occur between perceptions of ease of use (PEOU) / (X1) and perceived usefulness (PU) / (X2) of exogenous variables on the use of I-Chat Bot (ATU) / (Y), the results can be concluded the formulation of the hypothesis as follows:

1. Testing about Perceived ease of use has a partial effect on attitude toward using.
   \[ \text{Decision Making Basics:} \]
   \[ T \text{ value} > T \text{ Table (1.7011)} \]
   \[ \text{then Ha is accepted and Ho is rejected} \]
   \[ T \text{ value is calculated} < T \text{ table (1.7011)} \]
   \[ \text{then Ho is accepted and Ha is rejected} \]
   \[ \text{Decision:} \]
   \[ T \text{ value} = 1.9047 > 1.7011 \]
   \[ \text{then Ha is accepted and Ho is rejected} \]
   \[ \text{Conclusion:} \]
   Perceived variable ease of use has a partial effect on the attitude toward using the I-CHATBOT system

2. Testing about Perceived usefulness has a partial effect on attitude toward using.
   \[ \text{Decision Making Basics:} \]
   \[ T \text{ value} > T \text{ Table (1.7011)} \]
   \[ \text{then Ha is accepted and Ho is rejected} \]
   \[ T \text{ value is calculated} < T \text{ table (1.7011)} \]
   \[ \text{then Ho is accepted and Ha is rejected} \]
   \[ \text{Decision:} \]
   \[ T \text{ value} = 1.9049 > 1.7011 \]
   \[ \text{then Ha is accepted and Ho is rejected} \]
   \[ \text{Conclusion:} \]
   Perceived usefulness variables have a partial effect on the attitude toward using the I-CHATBOT system

3. Tests on Perceived usefulness and Perceived ease of use have a simultaneous effect on attitude toward using.
   \[ \text{Basic Decision Making:} \]
   \[ \text{Significance value} < 0.005 \text{ then Ha is accepted and Ho is rejected} \]
   \[ \text{Significance value} > 0.005 \text{ then Ho is accepted and Ha is rejected} \]
   \[ \text{Decision:} \]
   \[ \text{Significance value} = 0.000012 < 0.005 \text{ then Ha is accepted and Ho is rejected} \]
   \[ \text{Conclusion:} \]
   Perceived usefulness and Perceived variables ease of use have a significant effect simultaneously on the attitude toward using the I-CHATBOT system

B. Relationships and influences that occur between the exogenous variables of perception of ease of use (PEOU) / (X1) and perceptions of usefulness (PU) / (X2) and attitudes toward use (ATU) / (Y) towards the endogenous Actual Use (AU) / (Z) in the I-Chat Bot system is presented in figure 5. the following:

![Figure 5. Multiple Regression Model Pathway Analysis X1 and X2 and Y to Z](image)

Implications
Relationships and influences that occur between perceptions of ease of use (PEOU) / (X1) and perceived usefulness (PU) / (X2) of exogenous variables on the use of I-Chat Bot (ATU) / (Y), the results can be concluded the formulation of the hypothesis as follows:

1. Testing regarding Perceived ease of use has a partial effect on actual using.

Basic Decision Making:

Calculated T value> T Table (1.7011) then Ha is accepted and Ho is rejected

T value is calculated < T table (1.7011) then Ho is accepted and Ha is rejected

Decision:

Calculated T value = 1.3684 < 1.7011 then Ho is accepted and Ha is rejected

Conclusion:

Perceived variable ease of use does not have a significant partial effect on the attitude toward using the I-CHATBOT system

2. Tests on Perceived usefulness have a partial effect on actual using Hypothesis 2:

Basic Decision Making:

Calculated T value> T Table (1.7011) then Ha is accepted and Ho is rejected

T value is calculated < T table (1.7011) then Ho is accepted and Ha is rejected

Decision:

Calculated T value = 1.3935 < 1.7011 then Ho is accepted and Ha is rejected

Conclusion:

Perceived usefulness variables do not have a significant partial effect on the actual use of the I-CHATBOT system

3. Testing regarding attitude toward using effect partially on actual using

Basic Decision Making:

Calculated T value> T Table (1.7011) then Ha is accepted and Ho is rejected

T value is calculated < T table (1.7011) then Ho is accepted and Ha is rejected

Decision:

Calculated T value = 1.5407 < 1.7011 then Ho is accepted and Ha is rejected

Conclusion:

Variable attitude toward using does not have a significant partial effect on actual using on the use of the I-CHATBOT system

4. Testing about Perceived usefulness and Perceived ease of use and attitude toward using the effect simultaneously on actual using.

Basic Decision Making:

Significance value < 0.005 then Ha is accepted and Ho is rejected

Significance value > 0.005 then Ho is accepted and Ha is rejected

Decision:

Significance value = 0.00001 < 0.005 then Ha is accepted and Ho is rejected

Conclusion:

Perceived usefulness and Perceived variable ease of use and attitude toward using having a significant simultaneous effect on actual using on the use of the I-CHATBOT system.
5. Conclusion
The conclusion obtained from this study is that an intelligent interactive question-and-answer system developed using the concept of Naïve Bayes and Natural Language Processing (NLP) has been able to answer user questions regarding hearing impairments in an interactive way using the chat agent used. This system also uses text to text and the development of learning models to improve the user interface. The approach used is also able to categorize the database itself in an efficient way. Based on the tests performed, the I-CHATBOT application produces an I-Chat Bot with artificial intelligence that understands user input and provides an appropriate response and produces a preferred and easy system model to be used in finding information about required hearing impairments (Test Results TAM), this study also obtained the value of the accuracy test with 98.6% Precision, 88.75% Recall and 88.75% Accuracy.

References
[1] Louister, Hansen, "Chatbots Application for Diagnosing Toxoplasma Gondii Using Aiml (Artificial Intelligence Markup Language". Thesis, Jakarta: Gunadarma University [2] Bayu Setiaji, Ferry Wahyu Wibowo, “Chatbot Using A Knowledge in Database Human-to-Machine Conversation Modeling”, 2016 7th International Conference on Intelligent Systems, Modelling and Simulation, 2166-0670/16 © 2016 IEEE.
[3] Bapat, R. (2017), Helping Chatbots To Better Understand User Requests Efficiently Using Human Computation.
[4] Pustejovsky, A Stubbs, “ Natural Language Annotation for Machine Learning: A guide to corpus-building for applications”, "O'Reilly Media, Inc.". Copyright ©2012.
[5] Nitin Hardeniya, Jacob Perkins, Natural Language Processing: Python and NLTK, Packt Publishing, Copyright ©2016
[6] Olson & Delen. Advanced Data Mining Techniques. USA: Springer-Verlag Berlin Heidelberg. 2008
[7] William E Lewis, Software Testing and Continuous Quality Improvement, 2nd Edition 2009, (2009:134).
[8] Daniel Jurafsky & James H. Martin, “Speech and Language Processing” Copyright ©2017.
[9] Min Wu, Xiaoyu Zheng, Michelle Duan, Ting Liu and Tomek Strzalkowski, “Question Answering By Pattern Matching, Web-Proofing, Semantic Form Proofing”, ILS Institute, Computer Science Department SUNY Albany
[10] Nitin Hardeniya, Jacob Perkins, Natural Language Processing: Python and NLTK, Packt Publishing, Copyright ©2016