Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 269−276
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2041

Question Entailment on Developing Indonesian Covid-19 Question Answering System

Muhammad Zaky Aonillah*, Hasmawati, Ade Romadhony

Fakultas Informatika, Program Studi Informatika, Telkom University, Bandung, Indonesia
Email: 1,*mhmdzaky@student.telkomuniversity.ac.id, 2hasmawati@telkomuniveresity.ac.id,
3aderomadhony@telkomuniversity.ac.id
Submitted: **22/07/2022**; Accepted: **23/08/2022**; Published: **30/08/2022**

Abstract− Despite the severe impact of COVID-19 on humans has already decreased, people still need to be aware of the recent disease information. A continually updated Frequently Asked Questions (FAQ) system could help the public get valid and relevant information. To maintain a FAQ system manually needs much effort, hence an approach to develop the system automatically is needed. Question Answering System (QAS) is a system that can accept input in question sentences and produces an answer quickly, concisely, and relevantly, and could be used to provide COVID-19 information to the public. One method on developing a QAS is Recognizing Question Entailment (RQE). RQE is a form of relationship based on a cause-and-effect relationship between two pieces of text called text (T) and hypothesis (H). We present a study on developing Covid-19 QAS in Bahasa Indonesia using RQE. The datasets are collected from reputable sources and consist of 725 pairs of questions and answers. The experimental results show that the best performance results were obtained using the Logistic Regression model in training set 1, which contains 54.2% of positive question pairs and 45.8% of negative question pairs with an f-measure value of 83.65%. These results indicate that the RQE method can identify the entailment between new questions and questions in the dataset well.

Keywords: Covid-19; Indonesia; Question Answering System; Question Entailment; Supervised Learning

1. INTRODUCTION

Since December 2019, the whole world has been shocked by the emergence of the Covid-19 outbreak, and Indonesia is no exception. The World Health Organization (WHO) reports an increase in the number of cases of coronavirus(2019-nCov)-infected pneumonia (NCIP) or Covid-19 that were first identified in Wuhan City, Hubei Province, China, with the first 4 cases reported in December 29, 2019 [1]. On March 2, 2020, the Ministry of Health of the Republic of Indonesia reported two confirmed cases of Covid-19, which increased to 1,285 cases on March 29, 2020. The emergence of the Covid-19 outbreak is a major challenge for the Indonesian people always to receive the latest and most relevant information regarding the development of COVID-19 in Indonesia [2].

Efforts to collect information about the development of Covid-19 are closely related to the search engine. This can also be seen on several websites that have created a particular page for general questions or commonly called frequently asked questions (FAQ). However, people may find it challenging to find answers on search engines or the FAQ page because the results are still in the form of documents and cannot provide answer information to users directly. Therefore, the existence of an information system known as the Question Answering System (QAS) that can accept user input in the form of question sentences and produce an answer quickly, concisely, and relevantly is essential.

The Question Answering System (QAS) is designed to be able to meet human information needs quickly and precisely, such as situations when talking to virtual assistants, interacting with search engines, and performing queries on databases [3] . Question answering system has two main paradigms: Information Retrieval-based (IR) and Knowledge-based. IR-based relies on several documents from the web or collections of scientific papers to find the part of the answer that is relevant to the question. Meanwhile, Knowledge-based will build a semantic representation of the query, such as a logical representation. Knowledge-based also has two main paradigms, including graph-based and Semantic Parsing [4].

Recognizing Question Entailment (RQE) is one method that can be implemented in the Question Answering System. According to Ben Abacha et al. , RQE can allow the system to identify the correct answer more accurately than using keywords or pattern-based methods [5]. Question Entailment is a form of relationship based on a cause-and-effect relationship between two pieces of text called text (T) and hypothesis (H) which in this study, the (T and H) refers to a pair of questions (Q1 and Q2). Based on this definition, the main focus of Question Entailment is to determine whether Q1 entails Q2 if every answer to Q2 is also a correct answer to Q1 exactly or partially [6]. Here is an example from datasets:

1. Q1 (Community Question):
   Hello doctor, I have had a fever for 3 days, is this a symptom of COVID-19?
2. Q2 (Dataset):
   What are the symptoms of covid-19?
3. Q1 → Q2

FAQ FINDER is one of the question-and-answer systems that implement searches using similar questions and answers from existing data [7][8] . FAQ FINDER will retrieve existing answers from the Frequently Asked

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 269−276
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2041

Question (FAQ) data set. This research is based on semantic knowledge sourced from WORDNET, a semantic collection of English words that aims to improve performance in matching questions and answers. Researchers used several methods, including the SMART Information Retrieval System as a feature selection method, TF-IDF as a feature extraction method to assess the overall similarity of user questions with pairs of questions and answers, then recall and rejection as an evaluation method. At the experimental stage of the test results, there is a conclusion that the semantic score has a lower recall value, which is 55%, compared to the statistical score with a recall value of 58%. However, the semantic score has a good rejection performance. This shows that the application of semantic knowledge can be explicitly used to increase the rejection of unanswered questions.

Similar studies that apply the similarities between these questions include the research conducted by Jeon et al. [9][10] with the title Finding Similar Questions in Large Questions and Answers. This research is based on semantic similarity, which estimates the probability of a translation-based model. Researchers used four similarity measurement methods to calculate the distance between answers: Cosine Similarity, Negative KL Divergence between Language models, output score of query-likelihood (LM-HRANK), and Okapi Model. In the experimental stage, the researchers compared the translation model performance with three baseline retrieval models, including the vector space model with cosine similarity, Okapi BM25, and the quey-likelihood (LM) language model. Using precision and recall as evaluation methods, the translation model outperforms other baseline models such as cosine similarity, LM, and Okapi. The researchers used MAP (Mean Average Precision) as an advanced evaluation method. The result is that the translation model performs the best among other baseline models. The performance gap of the translation model is statistically significant with a 95% confidence level.

This paper uses a method derived from research conducted by Ben Abacha by Ben Abacha and Demner-Fushman [11] with the title A Question-Entailment Approach to Question Answering. There are two main focuses of this research. The first focus of this research is to compare the evaluation results of 2 methods, namely logistic regression and deep learning models for Recognizing Question Entailment (RQE), using various datasets, including textual inference, question similarity, and entailment from open-domain and clinical-domain. Furthermore, the second focus of this research is to combine the Information Retrieval (IR) model with the best RQE method to select the required questions and rank the answers taken. Researchers tested methods on several datasets from SNLI, multi-NLI, Quora, and Clinical-QE. At the experimental stage, the test results concluded that the deep learning model using GloVe embeddings gave the best results on three datasets (SNLI, multi-NLI, and Quora). Logistic regression gives the best accuracy on the Clinical-RQE dataset with a value of 98.60%. When tested on test data, logistic regression trained in Clinical-QE gave the best performance with an accuracy of 73.18%.

2. RESEARCH METHODOLOGY

2.1 Proposed System Overview

In this study, we built a question and answer system with the Covid-19 domain based on the similarity between the input of new questions and questions in the dataset. Several steps must be done before creating a model for question entailments, such as question preprocessing, semantic textual similarity, feature extraction, and data splitting. The model was tested using several supervised learning algorithms, and the results of each tested algorithm will be evaluated. The description of the system to be built can be seen in Figure 1.

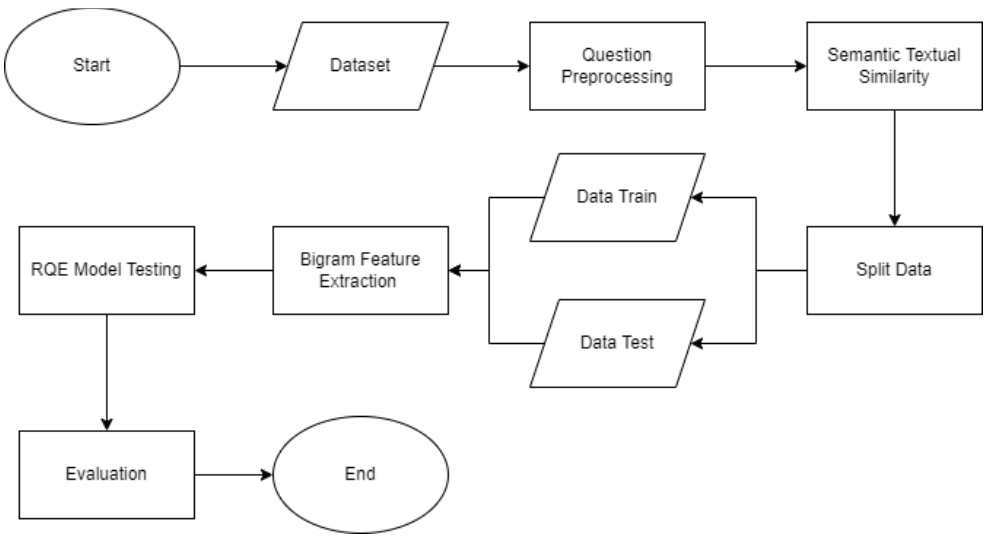

Figure 1. System Development Stages

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 269−276
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2041

2.2 Dataset

The dataset used in this study is a collection of pairs of questions and answers originating from credible Twitter accounts about Covid-19, such as doctors, the Covid-19 task force, and the Ministry of Health. Answers in our dataset are collected from several credible websites in answering medical questions such as www.who.int, www.alodokter.com, www.covid19.go.id , and other official health websites. This study referenced a dataset of questions collected from March 2020 to October 2021, with 725 pairs of questions and answers collected. Then each question was grouped manually into 17 different question categories, such as questions about vaccines, transmission, symptoms, and others [12]. The number of questions in each category can be seen in Table 1.

Table 1. Number of Categories in the Dataset

Question Category	Number of Question
Vaksin (Vaccine)	126
Transmisi (Transmission)	94
Gejala (symptom)	93
Pengobatan (Treatment)	66
Tes (Test)	55
Terkena Covid (Infected Covid)	52
Original (Original)	46
Perbandingan (Comparison)	36
Spekulasi (Speculation)	32
Tanggapan Masyarakat (Community Response)	25
Pencegahan (Prevention)	25
Efek Sosial (Social Effect)	20
Biaya (Cost)	15
Lokasi (Location)	13
Pemakaman (Burial)	9
Bansos (Social Assistance)	9
Respon Individual (Individual Response)	8

2.3 Preprocessing

Preprocessing is carried out on questions on the dataset, which aims to eliminate words that are not needed in building the system. Figure 2 shows the steps that need to be done at the preprocessing stage.

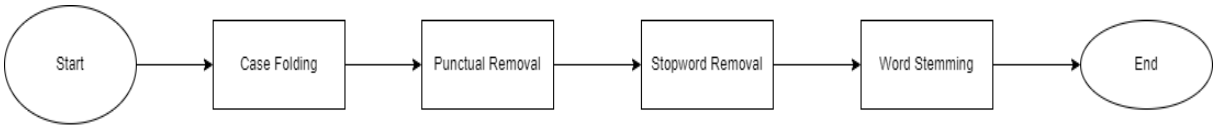

Figure 2. Question Preprocessing

In Figure 2, it can be seen that the first stage in preprocessing is case folding which is used to change every word in the sentence into lowercase letters. Then punctual removal is used to remove punctuation marks such as commas, periods, question marks, and so on in a text or sentence. The next stage in preprocessing is stopword removal, which removes common words that do not significantly affect sentences or texts. The last stage in preprocessing is word-stemming which functions to change a word with affixes into its base word. Examples of input and output in preprocessing can be seen in Table 2.

Table 2. Input and Output Example of Preprocessing

Preprocessing	Input	Output
Case folding	*Dokter, apakah covid dapat menular dari hewan yang sakit ??* (Doctor, can covid be transmitted from sick animals ??)	*dokter, apakah covid dapat menular dari hewan yang sakit ??* (doctor, can covid be transmitted from sick animals ??)
Punctual Removal	*dokter, apakah covid dapat menular dari hewan yang sakit ??* (doctor, can covid be transmitted from sick animals??)	*dokter apakah covid dapat menular dari hewan yang sakit* (doctor can covid be transmitted from sick animals)
Stopword Removal	*dokter apakah covid dapat menular dari hewan yang sakit* (doctor can covid be transmitted from sick animals)	*dokter apakah covid dapat menular dari hewan sakit* (doctor covid infects sick animal)
Word Stemming	*dokter apakah covid dapat menular dari*	*dokter apakah covid dapat tular dari*

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 269−276
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2041

Preprocessing	Input	Output
	hewan sakit (doctor covid infects sick animal)	*hewan sakit* (doctor covid infectious sick animal)

2.4 Semantic Textual Similarity

After the data is processed, the next step is the semantic textual similarity stage, which aims to obtain relevant features between the query and the tested data. In this study, the cosine equation is used to measure the similarity between two documents so that the calculation results are used as features. The calculated cosine equation will calculate the cosine value between two vectored documents [13], where the two documents in this research are the original question text and the topic question text. The following is the formula for calculating the cosine similarity value.

$$CosineSimilarity(q_t, d_t) = \frac{\overrightarrow{q_t} \cdot \overrightarrow{d_t}}{|\overrightarrow{q_t}| \times |\overrightarrow{d_t}|} \quad (1)$$

Description:
$CosineSimilarity(q_t, d_t)$ = the cosine value between the question document and the new question
$\overrightarrow{q_t}$ = m-dimensional vector of new question
$\overrightarrow{d_t}$ = m-dimensional vector of the question document
An example of applying cosine similarity to get the similarity value between the query and the dataset can be seen in Table 3.

Table 3 Example of Cosine Simarity Score Between Original Question and Question Topic

Question	Question Topic	Cosine Similarity
Dok, biaya untuk tes covid-19 berapa ya? (Doc, how much does the Covid-19 test cost?)	*berapa biaya tes Covid-19?* (how much does a Covid-19 test cost?)	0.967
Gejala utama covid apa saja dok? (doc, what are the main symptoms of covid?)	*Apa gejala dari Covid-19?* (What are the symptoms of Covid-19?)	0.852
Apakah asap rokok dapat menularkan covid-19? (Can cigarette smoke transmit COVID-19?)	*Covid-19 menular melalui apa?* (What is Covid-19 transmitted through?)	0.655

2.5 Splitting Data Train

The data train contains several pairs of questions which are distinguished between pairs of positive questions and pairs of negative questions. A pair of positive questions will use questions in the original form (original question) and a short form (short question). Meanwhile, to get pairs of negative questions, original questions will be used with questions in different categories (cross-category questions). In this study, following the dataset split setting on [6], we set 5 different dataset split as follows:

a. Training set 1: consists of 54.2% positive pairs and 45.8% negative pairs.
b. Training set 2: consists of 26% positive pairs and 74% negative pairs.
c. Training set 3: consists of 33.4% positive pairs and 66.6% negative pairs.
d. Training set 4: consists of 25% positive pairs and 75% negative pairs.
e. Training set 5: consists of 50% positive pairs and 50% negative pairs.

The following is an example of positive and negative question pair:

a. <pair id= "1", value= "**positive**">
   1. <t> Doc, I want to ask, if covid can be transmitted through fluids, it means that there is a possibility of transmission through mosquitoes, right? </t>
   2. <h> Can COVID-19 be transmitted through mosquitoes? </h>
b. <pair id= "2", value= "**negative**">
   1. <t> How long does the COVID-19 virus survive in the air? </t>
   2. <h> *How long can a COVID-19 patient be declared cured?* </h>

2.6 Feature Extraction (TF-IDF Bigram)

Term Frequency Inverse Document Frequency (TF-IDF) is an algorithm to get the weight value of all the words in a document. This study combines the TF-IDF feature extraction with N-grams to produce the maximum weight value. The N-gram used in this study is a bigram that will calculate the probability of a current word with the previous word [14]. The following is the formula for calculating the TF-IDF value.

$$TFIDF_{jk} = term_{jk} \times \log(D/df_k) \quad (2)$$

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 269−276
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2041

Description:
$TFIDF_{jk}$ = Word weighting result ($term_k$) against a document (d_j)
$term_{jk}$ = The number of occurrences of the word ($term_k$) in a document (d_j)
D = Total documents on the dataset
df_k = Amount containing $term_k$

2.7 RQE Model Testing

Several supervised learning algorithms are tested to detect whether the input questions entail questions in the training set. The supervised learning algorithms used in this research include SVM, Logistic Regression, Naïve Bayes, and J48. All of these algorithms are trained on five different training sets so that the output of each model can be evaluated.

SVM aims to find the best hyper lane in N-dimensional space by maximizing the margin between the training pattern and the decision boundary [15][16]. Hyperlane is a function that can be used to separate classes, and the outermost data closest to the hyper lane is called a support vector. Logistic Regression works by calculating the probability ratio to assign a value as negative (0) or positive (1) based on the relationship between the independent input variable (feature) and the dependent variable (target) [17]. Naive Bayes comes from a probabilistic approach to Bayes' theorem, which predicts future opportunities based on previous experience. The Naïve Bayes approach assumes that the strong (nave) ideas on the evaluated features are independent [17]. Then the J48 algorithm works by building a decision tree based on labeled input data, where the concept of a decision tree is to process data into a hierarchical tree structure [18]. J48 is the development of the ID3 algorithm (interactive Dichotomize 3) with good classification accuracy [19]. The stages for testing the RQE model can be seen in Figure 3.

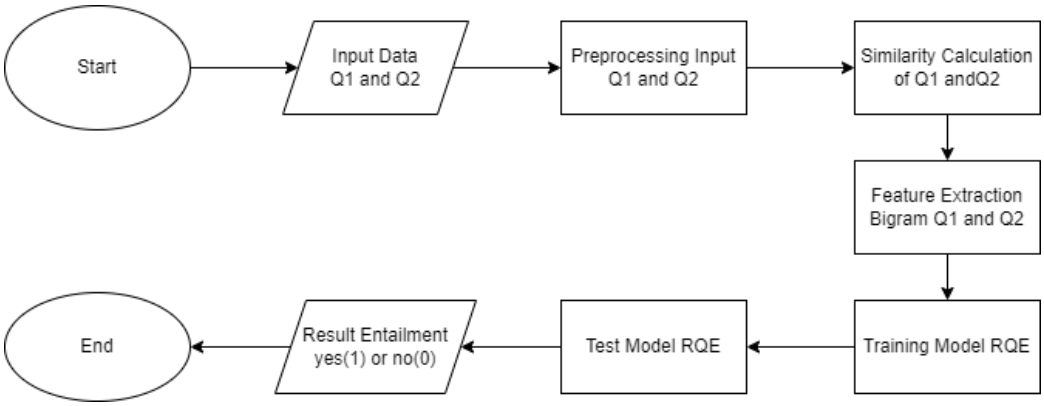

Figure 3. Process of Testing RQE Model

The model testing phase begins with receiving two questions which Q1 and Q2 represent. Q1 is a new question to be tested, and Q2 is a question contained in the dataset. The two questions will first go through the preprocessing stage and then calculate the similarity value between Q1 and Q2. Before training the RQE model, the two questions must be feature extracted to get the weight value of each word and produce a vector that can be processed at the training stage. Furthermore, the data will be trained with four supervised learning algorithms to detect whether Q1 entails Q2. The model will produce an output of 1 or 0, where 1 indicates both questions entail and 0 does not. Table 4 shows an example of the testing result of RQE model.

Table 4. Example RQE Testing Result

Q1 (New Question)	Q2 (Dataset Question)	Entailment
Doc, if there is a mother who is pregnant, will it affect the fetus?	Can pregnant women infected with COVID-19 transmit COVID-19 to the fetus?	Yes

3. RESULTS AND DISCUSSION

The evaluation phase of the research was carried out on a dataset of questions about Covid-19 in Indonesian. The test scenario of this research was carried out on each different supervised learning model for RQE, namely SVM, logistic regression, nave Bayes, and J48. Model testing is also carried out on five training sets so that each model will produce several evaluation metrics, including precision, recall, and f-measure. This test aims to determine which RQE model is the best in classifying whether a new question entails or not against the existing questions in the dataset.

In the first test scenario, the entire training set was tested with the whole test model: SVM, Logistic Regression, Naïve Bayes, and J48. The experimental results in the first test scenario can be seen in Table 4.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 269−276
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2041

Table 5. F-Measure Value in All Training Sets

Training Set	SVM	Logistic Regression	Naïve Bayes	J48
Set 1	81.90	83.65	56.60	80.16
Set 2	76.74	72.50	11.38	76.47
Set 3	71.30	69.81	17.93	66.66
Set 4	76.54	64.78	6.61	57.39
Set 5	74.73	77.83	21.90	79.62

Based on the test results in Table 4, the training set 1 produces the highest f-measure value compared to other training sets. Furthermore, all RQE models will be tested on the best training set. Namely, the training set 1, to see each model's evaluation values of precision, recall, and f-measure. The experimental results in training set 1 can be seen in Table 5.

Table 6. Testing Scenario Result on Training Set 1

Algorithm	Precision	Recall	F-Measure
SVM	72.88	**93.47**	81.90
Logistic Regression	**96.66**	73.72	**83.65**
Naïve Bayes	63.82	50.84	56.60
J48	79.83	80.50	80.16

In the second test scenario, the best classification model was taken from the first test and then tested on a different training set to compare the resulting precision, recall, and f-measure values. Based on the first test, the best results were obtained using a logistic regression model with an F-Measure value of 83.65%. In this test, it is also possible to determine the effect of comparing pairs of positive questions and pairs of negative questions in each training set. The experimental results in the second test scenario can be seen in Table 6.

Table 7. Results of Logistic Regression Classifier in Five Training Sets

Training Data	Precision	Recall	F-Measure
Set 1	96.66	73.72	**83.65**
Set 2	96.66	57.99	72.50
Set 3	92.50	56.06	69.81
Set 4	100	47.91	64.78
Set 5	94.73	66.05	77.83

Based on the test results for the two test scenarios, it can be seen that each supervised learning model tested for RQE produces varying evaluation values. This test shows that training set 1 provides the best F-Measure value in the overall RQE model. The test results in the first scenario can be seen in Figure 4. Based on the test results for the two test scenarios, it can be seen that each supervised learning model tested for RQE produces varying evaluation values. This test shows that training set 1 provides the best F-Measure value in the overall RQE model. The test results in the first scenario can be seen in Figure 4.

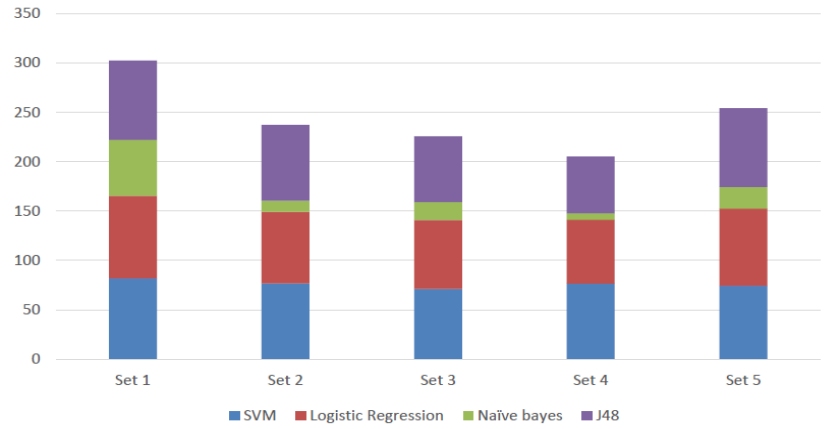

Figure 4. Analysis of Model Testing Results for All Training Sets

In the previous test scenario, it can be seen that the use of a training set 1 and the use of the logistic regression algorithm is the RQE model that produces the best evaluation value. These results were obtained because logistic regression can provide a good prediction of the entailment of new questions compared to questions in the dataset. Furthermore, the logistic regression model was tested against five training sets that had been built previously to obtain other evaluation values, namely precision, and recall. The test results in the second scenario can be seen in Figure 5.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 269−276
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2041

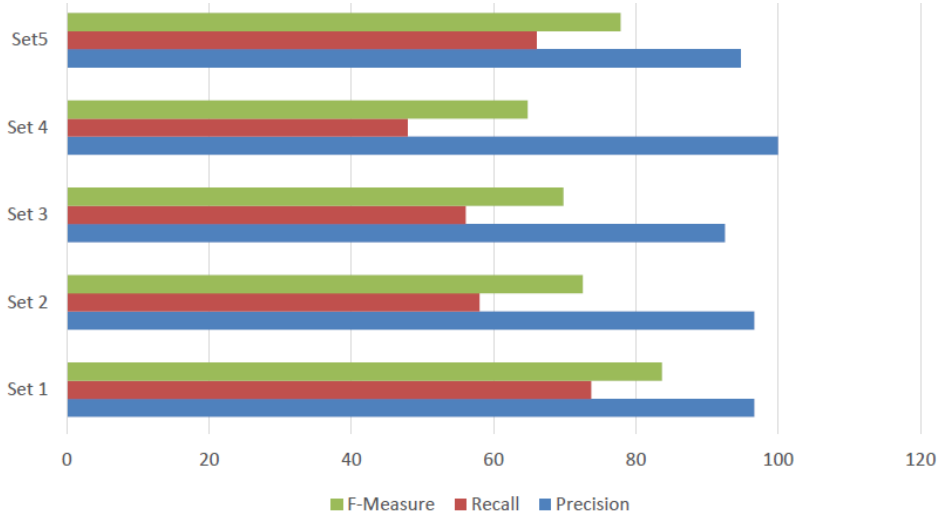

Figure 5. Logistic Regression Model Testing Results on the Overall Training Set

Based on all the model testing scenarios and training set variations that have been carried out previously, the next evaluation is to perform an error analysis on data that is incorrectly classified as entailment or not between the new question and the topic of the question in the dataset. This evaluation was carried out on the RQE model, and the best variation of the training set in the previous test, namely Logistic Regression in training set 1. The test data used in training set 1 amounted to 218 of 725 data in the entire dataset. In the experiment, 34 data were misclassified from 218 test data. Examples of misclassified data can be seen in Table 7.

Table 8. Results of Error Analysis on Test Data

No	Question	Question Topic	Label	Prediction
1	*temen2 nanya tes2 covid 19 bayar gak si* (guys 1 want to ask, do you pay for the covid 19 test?)	*Apakah tes covid 19 gratis* (is the covid-19 test free?)	1	0
2	*dok kalo badan ngrasa lemes trsa ky masuk angin tp ga dingin kepala berat gejala covid 19 dok* (Doc, if your body feels weak, it feels like you have a cold but your head is not cold, is it a symptom of covid 19?)	*apakah badan lemas dan kepala berat termasuk gejala covid 19* (Is the body weak and the head heavy, including the symptoms of covid 19?)	0	1
3	*tes covid bayar ga dok tes nya mana aja* (do you pay for the covid test, doc? where is the test?)	*dimana tempat tes covid 19* (where is the place to test covid-19?)	1	0
4	*dok kena covid 19 fungsi paru paru turun efektivitas nya paruparu orang normal* (Doc, got covid 19, lung function has decreased the effectiveness of normal people's lungs)	*apakah fungsi paru paru akan menurun efektivitasnya jika sudah pernah terkena covid 19* (Will lung function decrease in effectiveness if you have been exposed to COVID-19?)	0	1
5	*dok dokter covid 19 bs tular lwat sample darah ato tdk* (doc, covid 19 can be transmitted through blood samples or not?)	*apakah covid 19 dapat menular melalui sampel darah* (Can covid 19 be transmitted through blood samples?)	1	0

In Table 7, it can be seen that the labels on the dataset with the predictions made by the model produce different results. This result can occur because people's questions in Indonesian do not always use standard language. For example, there are still many abbreviated words (question pair number 5: *"lwat", "bs", "tdk",* etc) that cannot be detected by stopword removal or word-stemming at the preprocessing stage, so this causes the question sentence to not contains words that point to the topic of the question.

4. CONCLUSION

Based on the results of testing and analysis that have been carried out using several test scenarios for the question entailment method, it can be concluded that varying the number and types of positive and negative pairs in the corpus training greatly affects the evaluation results obtained. This result provides a good start for building

ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 269−276
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2041

a Question Entailment model to find questions in the dataset similar to new questions. Referring to the test scenarios carried out on all RQE models and the entire training set, it was found that training set 1 (54.2% positive question pairs and 45.8% negative question pairs) gave the best f-measure value compared to the other four training sets. Then based on the test results, it was also found that the logistic regression model gave better evaluation results than other RQE models, with an f-measure value of 83.65 in training set 1. This could be because logistic regression can calculate the probability ratio of a sample with a binary approach, such as classifying whether new questions entail or not to existing questions in the dataset. In this study, we built a Question Answering System (QAS) on the Indonesian Covid-19 Question Answering System using Recognizing Question Entailment (RQE) model with several supervised learning algorithms. Suggestions that can be made in further research are to expand the training dataset including expand the testing dataset with more diverse sources such as questions asked by experts or health workers and FAQs collected manually from several official websites on health. Then the preprocessing stage needs to be developed, such as normalizing words in Indonesian because there are still words that are not standard and use abbreviated words that affect the results of the model prediction. And the use of various algorithm models can provide different performances to determine the effect of the model on RQE.

REFERENCES

[1] R. Tosepu *et al.*, "Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia," *Sci. Total Environ.*, vol. 725, 2020, doi: 10.1016/j.scitotenv.2020.138436.

[2] Keputusan Menteri Kesehatan Republik Indonesia, "Keputusan Menteri Kesehatan Republik Indonesia Nomor HK.01.07/MenKes/413/2020 Tentang Pedoman Pencegahan dan Pengendalian Corona Virus Disease 2019 (Covid-19)," *MenKes/413/2020*, vol. 2019, p. 207, 2020.

[3] B. McCann, N. S. Keskar, C. Xiong, and R. Socher, "The Natural Language Decathlon: Multitask Learning as Question Answering," 2018, [Online]. Available: http://arxiv.org/abs/1806.08730.

[4] D. Jurafsky and J. H. Martin, *Speech and language processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition Third Edition draft*. 2021.

[5] A. Ben Abacha, C. Shivade, and D. Demner-Fushman, "Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering," *BioNLP 2019 - SIGBioMed Work. Biomed. Nat. Lang. Process. Proc. 18th BioNLP Work. Shar. Task*, pp. 370–379, 2019, doi: 10.18653/v1/w19-5039.

[6] A. Ben Abacha and D. F. Dina, "Recognizing Question Entailment for Medical Question Answering," *AMIA ... Annu. Symp. proceedings. AMIA Symp.*, vol. 2016, pp. 310–318, 2016.

[7] R. D. Burke, K. J. Hammond, V. Kulyukin, S. L. Lytinen, N. Tomuro, and S. Schoenberg, "Question answering from frequently asked question files: Experiences with the FAQ FINDER system," *AI Mag.*, vol. 18, no. 2, pp. 57–66, 1997.

[8] M. De Bruyn, E. Lotfi, J. Buhmann, and W. Daelemans, "MFAQ: a Multilingual FAQ Dataset," pp. 1–13, 2021, doi: 10.18653/v1/2021.mrqa-1.1.

[9] J. Jeon, W. B. Croft, and J. H. Lee, "Finding similar questions in large question and answer archives," *Int. Conf. Inf. Knowl. Manag. Proc.*, pp. 84–90, 2005, doi: 10.1145/1099554.1099572.

[10] S. Bahri, S. Sumpeno, and S. M. S. Nugroho, "An information retrieval approach to finding similar questions in question-answering of Indonesian government e-procurement services using TF∗IDF and LSI model," *Proc. 2018 10th Int. Conf. Inf. Technol. Electr. Eng. Smart Technol. Better Soc. ICITEE 2018*, pp. 626–631, 2018, doi: 10.1109/ICITEED.2018.8534856.

[11] A. Ben Abacha and D. Demner-Fushman, "A question-entailment approach to question answering," *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–23, 2019, doi: 10.1186/s12859-019-3119-4.

[12] J. Wei, C. Huang, S. Vosoughi, and J. Wei, "What Are People Asking About COVID-19? A Question Classification Dataset," 2020, [Online]. Available: http://arxiv.org/abs/2005.12522.

[13] M. Benard Magara, S. O. Ojo, and T. Zuva, "A comparative analysis of text similarity measures and algorithms in research paper recommender systems," *2018 Conf. Inf. Commun. Technol. Soc. ICTAS 2018 - Proc.*, pp. 1–5, 2018, doi: 10.1109/ICTAS.2018.8368766.

[14] A. Alatawi, W. Xu, and J. Yan, "The Expansion of Source Code Abbreviations Using a Language Model," *Proc. - Int. Comput. Softw. Appl. Conf.*, vol. 2, pp. 370–375, 2018, doi: 10.1109/COMPSAC.2018.10260.

[15] X. Deng, Y. Li, J. Weng, and J. Zhang, "Feature selection for text classification: A review," *Multimed. Tools Appl.*, vol. 78, no. 3, pp. 3797–3816, 2019, doi: 10.1007/s11042-018-6083-5.

[16] D. A. Pisner and D. M. Schnyer, *Support vector machine*. Elsevier Inc., 2019.

[17] H. H. Rashidi, N. K. Tran, E. V. Betts, L. P. Howell, and R. Green, "Artificial Intelligence and Machine Learning in Pathology: The Present Landscape of Supervised Methods," *Acad. Pathol.*, vol. 6, 2019, doi: 10.1177/2374289519873088.

[18] S. Diwandari and N. A. Setiawan, "Perbandingan Algoritme J48 Dan Nbtree Untuk Klasifikasi Diagnosa Penyakit Pada Soybean," *Semin. Nas. Teknol. Inf. dan Komun.*, vol. 2015, no. Sentika, pp. 205–212, 2015.

[19] H. Hong *et al.*, "Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China)," *Catena*, vol. 163, no. January, pp. 399–413, 2018, doi: 10.1016/j.catena.2018.01.005.