2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)
April 09-11. 2021, ISTTS Surabaya, Indonesia

Predicting Frequently Asked Questions (FAQs) on the COVID-19 Chatbot using the DIET Classifier

1st Wistiani Astuti
Faculty of Engineering
Universitas Negeri Malang
Malang, Indonesia
wistiani.astuti@umi.ac.id

2nd Desy Pratiwi Ika Putri
Faculty of Engineering
Universitas Negeri Malang
Malang, Indonesia
Desypratiwi407@gmail.com

3rd Aji Prasetya Wibawa
Electrical Engineering
State University of Malang
Malang, Indonesia
aji.prasetya.ft@um.ac.id

4th Yulita Salim
Faculty of Computer Science
Universitas Muslim Indonesia
Makassar, Indonesia
yulita.salim@umi.ac.id

5th Purnawansyah
Faculty of Computer Science
Universitas Muslim Indonesia
Makassar, Indonesia
purnawansyah@umi.ac.id

6th Anusua Ghosh
ASCEE Australia
Adelaide, Australia
anusua@ascee.org

2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT) | 978-1-6654-0514-0/20/$31.00 ©2021 IEEE | DOI: 10.1109/EIConCIT50028.2021.9431913

***Abstract*—A popular dialogue system in the field of natural language processing (NLP) is the chatbot. Chatbots aim to create conversations between humans and machines. COVID-19 is a member of the Coronavidae (CoV) family of the Corona viirinae family which causes the respiratory system to become severe in humans. This paper predicts chatbot answers to questions about COVID-19 with the RASA framework and uses the DIET Classifier pipeline for 300 training data. The test results with the DIET Classifier model on *rasa.core.test* and *rasa.nlu.test* provided confidence values of F1-Score, precision, and accuracy for the correct answer to the question about COVID-19, namely 1.0 with a percentage of around 85%.**

Keywords—chatbot, NLP, nlu, DIET classifier

I. INTRODUCTION

In 2020, on March 11, the world has been experiencing problems in the health sector. The World Health Organization (WHO) sets the viral disease corona 2019 (COVID-19) as its global pandemic [1]. COVID-19 is one of the members of Coronavidae (CoV) of the family Coronaviirinae which is part of the order Nidovirales [2]. COVID-19 is the virus that causes the function of respiratory severe or acute accounted since 11 February 2020 [3]. The respiratory issue caused by viruses COVID-19 spread throughout the world and is very contagious, where the disease is having a blast on a large scale. In the beginning, COVID-19 was traced in Hubei Province China in the main city of Wuhan in January as the cause of the specific of the blast outbreak of the virus [2]. The COVID-19 pandemic has been disrupting all human life activities in the world and has a huge impact in almost all aspects.

The development of Natural Language Processing (NLP) which is currently widely used is the chatbot. Chatbots aim to create conversations between humans and machines. The chatbot program can provide user satisfaction with its ability to respond to a variety of different questions quickly. Chatbots can store information without forgetting any information stored and combined with practicality in finding and providing information [4]. Among several types of chatbot, the main type is called virtual assistant to support the needs of users in various domains and sectors [5]. NLU is a challenging subtopic in NLP, which can extract and understand specific conversations. NLU uses an algorithm to reduce natural language to a structured ontology that allows entity extraction [6].

The NLU RASA is an open-source NLP library for intent classification and entity extraction in chatbots which can help in building custom NLP for chatbots. In RASA, there are two components, namely RASA NLU and RASA Core [7]. Several studies related to chatbots such as researchers [8] who conducted literature studies and analyses on the open source chatbot framework, RASA and concluded that RASA-based chatbots have many capabilities compared to other open-source. Researcher [9] who uses the WhatsApp bot as a data provider for COVID statistics using PHP, Flask, and MySQL and produce a chatbot application via WhatsApp and provide information and statistical data on COVID-19 in Indonesia.

Dual Intent and Entity Transformer (DIET) with Open Source RASA 1.8.0 can use trained embeddings from the BERT Language model in the NLU Rasa pipeline. DIET is a multi-task transformer architecture for simultaneous classification and identification of entities. A DIET is made of many components which allows it to have the flexibility to exchange different components. One of its main features is the ability to combine different word embeddings such as BERT and GloVe or trained words from the model language and combine them with sparse words and features the n-gram character level in plug-and-play mode. Many trained language models are very heavy in the sense that they require large powerful computation and long inference times so that despite their strong performance they are not designed for AI conversational applications, whereas DIET is different. DIET is a modular architecture which allows software developers to have more flexibility in their experiments, matches well-trained language models in terms of accuracy, and outperforms current SOTA and trains 6X faster [10]. This study implements a chatbot with the RASA framework to predict intent and entity of each word related to COVID-19 using the DIET Classifier model to provide information related to COVID-19. There are 11 intents used in predicting chatbot answers for Covid-19 such as greet, deny, goodbye, corona_intro, corona_spread, high_risk, warm_weaher, corona_food_spread, bot_challenge, mood_great, and mood_unhappy.

II. LITERATURE REVIEW

A. COVID-19

COVID-19 is a member of Coronavidae (CoV) from the Coronavirinae family which is part of the order Nidovirales

[2]. About 40 subfamily varieties of single-stranded RNA viruses that nest in animals in bats and other wild birds can evolve to infect humans, mammals, and birds. The COVID-19 can mutate and infect various species of cell types. The COVID-19 virus continues to emerge and develop, which causes humans and animals to be exposed to the virus [11]–[13]. Seven common types of COVID-19 in humans are as: 229 E (alpha coronavirus), HKU1 (beta coronavirus), MERS CoV (beta coronavirus), NL63 (alpha coronavirus), OC43 (beta coronavirus), SARS CoV (beta coronavirus), and COVID-19 (SARS CoV-2) (beta coronavirus). In Fig. 1 shows the coronavirus or COVID-19 with the projections of the trimeric glycoprotein Spike from the surface, and image B depicts the 3-D structure of the S-Protein in the coronavirus.

B. RASA NLU

RASA Natural Language Understanding (NLU) is used to understand chatbot language and AI systems that focus on the classification of meaning and extraction of entities. The "spacy_sklearn" pipeline consists of different components using several NLP libraries such as spaCy, scikit-learn, and sklearn-crfsuite. With these components, RASA NLU can analyse messages [5]. Table I represents some commands contained in RASA NLU [14].

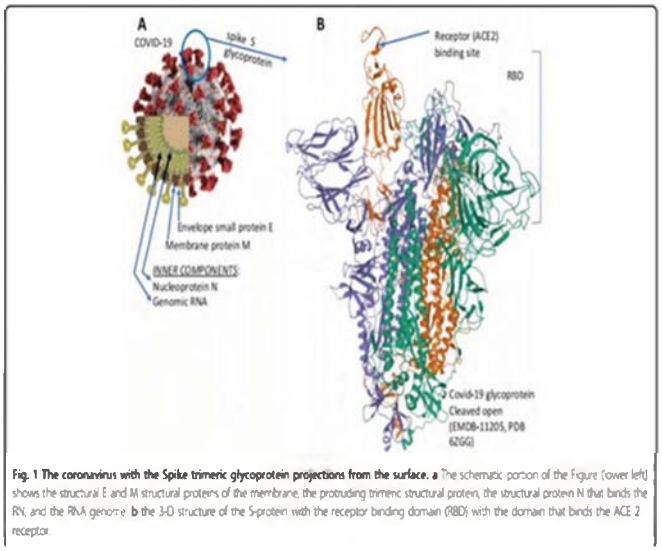

Fig. 1 The coronavirus with the Spike trimeric glycoprotein projections from the surface. a The schematic portion of the Figure (lower left) shows the structural E and M structural proteins of the membrane, the protruding trimeric structural protein, the structural protein N that binds the RN, and the RNA genome **b** the 3-D structure of the S-protein with the receptor binding domain (RBD) with the domain that binds the ACE 2 receptor.

Fig. 1. COVID-19 (Source:[2])

TABLE I. COMMANDS CONTAINED IN RASA NLU

Command	Function
rasa init	Create new projects such as training data, configuration files, and actions
rasa train	Train models using data from the NLU and store trained models in ./models
rasa interactive	Create new training data to start an interactive learning session
rasa shell	Loads your trained model
rasa run	Starts a server with your trained model
rasa run actions	Starts an action server using the Rasa SDK
rasa visualize	Generates a visual representation of your stories.
rasa test	Tests a trained Rasa model on any files starting with test
rasa data split nlu	Performs a 80/20/ split of your NLU training data
rasa data convert	Converts training data between different formats
rasa data validate	Checks the domain, NLU and conversation data for inconsistencies.
rasa export	Export conversation from a tracker store to an event broker
rasa x	Launches Rasa X in local mode
rasa -h	Shows all available commands

Interpreting the user's speech will generate a new task which is often called natural language comprehension [15]–[17] then NLU extracts the information from the user's sentence. Information from one or more consecutive words is defined as named entities (for example, date, and location). Consider the following sentence: "I want to book a ticket to London tomorrow", and the sentence meant by the user is referring to "ordering a ticket". The system needs to know the destination of the user and when the user wants to use the chatbot. The entity identifier named NER can be used to find information. Named entity classifiers can be trained to classify London as a destination. Most of the systems allow entities to be defined by instances and regular expressions and such sample tickets can be used for keyword matching with a simple system [18].

RASA NLU pipeline: RASA NLU has an entity search. The interaction between the bot and the user through the RASA or RASA X framework where the users search for information, and messages will be sent to the bot and the RASA NLU abstract entity from the message, then the bot will get the message intent from the user and respond correctly according to the RASA NLU translator machine. The data is organized into sections such as synonyms and regex features. In Fig. 2, an example of the pipeline principle on the RASA NLU is illustrated, and Fig. 3 shows the architecture of the RASA NLU.

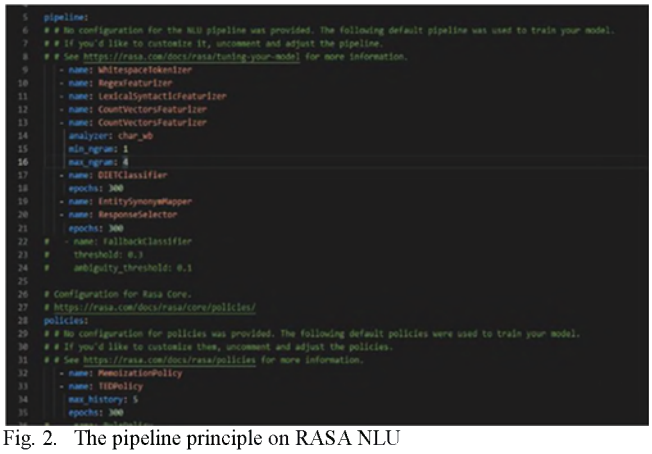

Fig. 2. The pipeline principle on RASA NLU

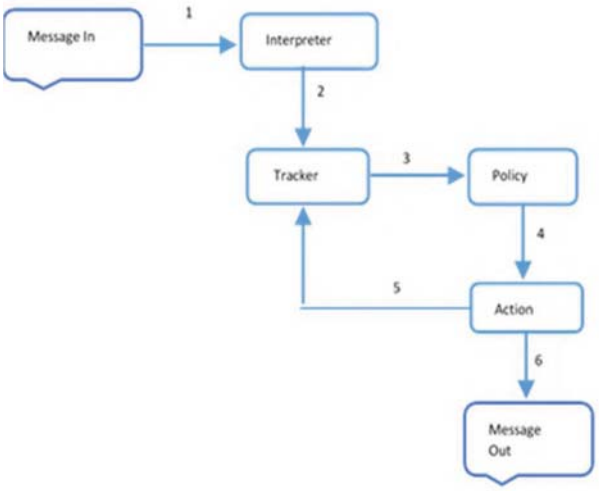

Fig. 3. RASA NLU architecture Source:[19]

C. DIET Classifier

Dual Intent and Entity Transformer (DIET) is a new multi-task architecture for entity classification and recognition which has a main feature to able to incorporate trained word embeddings from the model language and combine them with sparse words and features n-gram character level in plug-and-play mode. A DIET without trained embeddings will improve complex NLU data sets. In addition, adding a pre-trained embedding of words and sentences from the Language model further improves the overall accuracy of all tasks [10]. The representation of the DIET architecture in Fig 4 consists of several important parts. The phrase "play ping pong" has the intention of playing a game and the entity name is "ping pong" with the weight of the feed-forward layer distributed to the token.

III. Result and Discussion

The initial stages conducted in the process of making the COVID-19 chatbot with the RASA framework, namely:

- activation of RASA,
- Create a Chatbot folder to store all files and insert them into the folder
- install RASA
- After the installation process is complete, the init RASA process is carried out,
- Create Intent and Entity on nlu.md and stories.md to enter all words that will be used later by the Bot and user,
- Create a response for each utter on domain.yml
- Configure the DIET Classifier in config.yml
- Conduct training data
- Test the chatbot

For light configuration, Count Vectors Featurizer is used in config.yml which creates a bag-of-word representation for each incoming message or at word and character level. In the conversion configuration, the ConveRT components work in the same way where they have their tokenizers and feature zero. The pipeline created and the training data collected will be easy to get predictions of intent and entity. To assess the accuracy of the COVID-19 chatbot with the NLU RASA using the DIET Classifier model, 200 epochs are used. The intents made in nlu.md are shown in Fig. 5.

Testing of the COVID-19 Chatbot on the RASA NLU has been performed using the RASA test and RASA shell NLU commands. RASA test will show the results of the DIET Classifier model on rasa.core.test and rasa.nlu.test by evaluating 7 stories and the results in END-TO-END level namely Correct, F1-Score, Precision, Accuracy, In-data fraction, Confusion Matrix without normalization on *rasa.core.test* and the test results are in Fig. 6. In addition, Fig.7 depicts the DIET Classifier test results with *rasa.nlu.test* while Fig. 8 represents the COVID-19 chatbot interface.

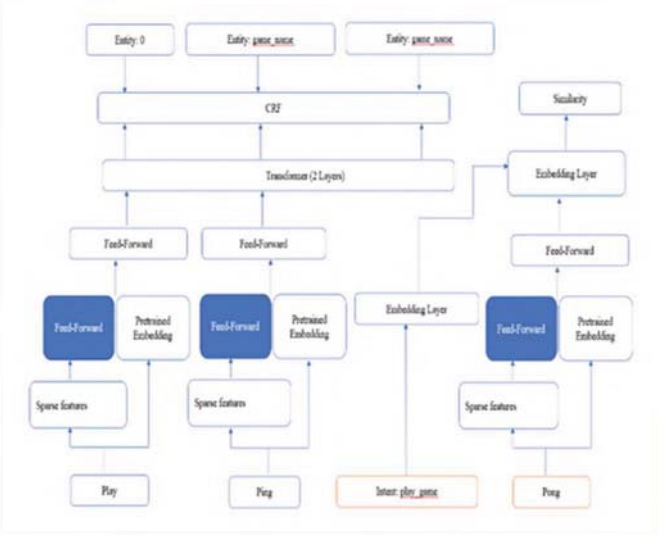

Fig. 4. DIET architecture (Source:[10])

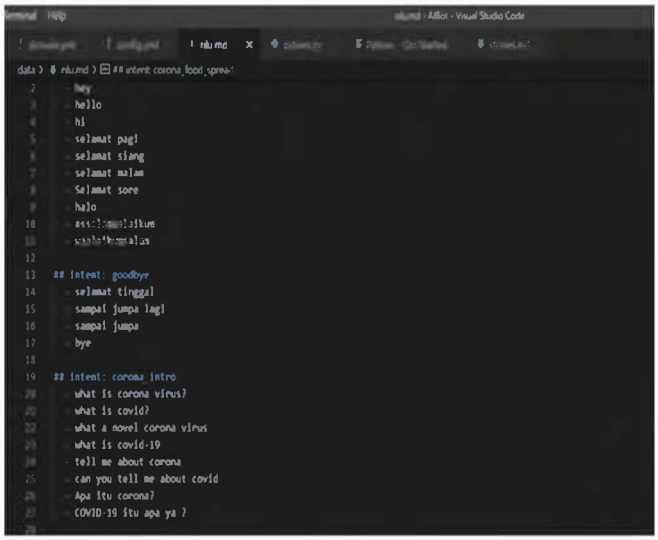

Fig. 5. Intents of COVID19 chatbot

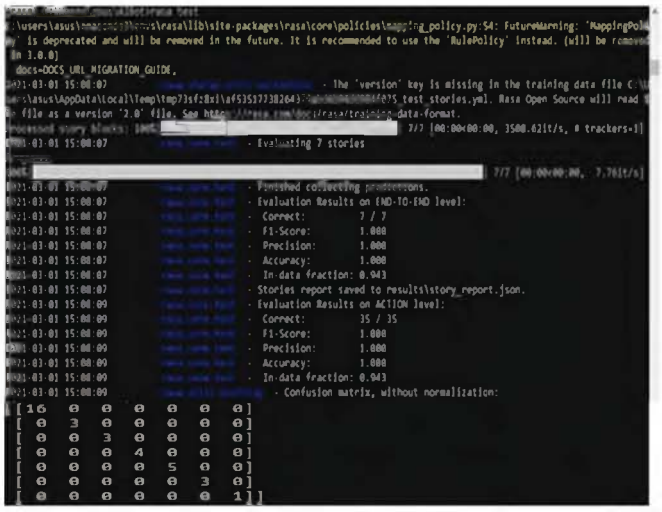

Fig. 6. Testing of the DIET Classifier with rasa.core.test

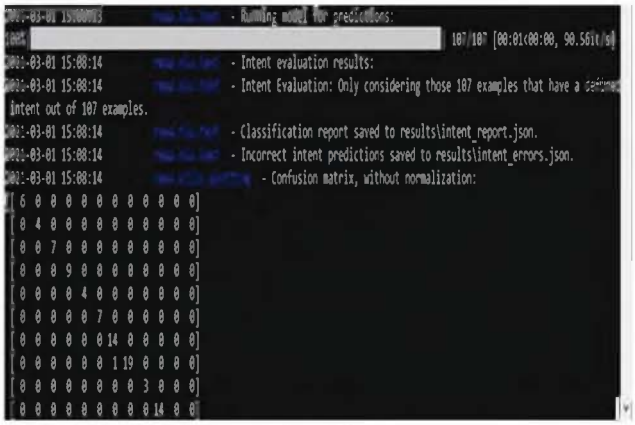

Fig. 7. Testing of the DIET Classifier with rasa.nlu.test

The prediction of the COVID-19 chatbot with the DIET Classifier model had a different confidence value from each word or sentence in the chatbot. Fig. 9 shows the confidence value in the word of "corona" with a value of 0.9779. Moreover, it was also predicted for other classifications, but the confidence value is relatively smaller such as the "high_risk" classification with a value of 0.0002 and "corona_food_spread" which was only 0.0046.

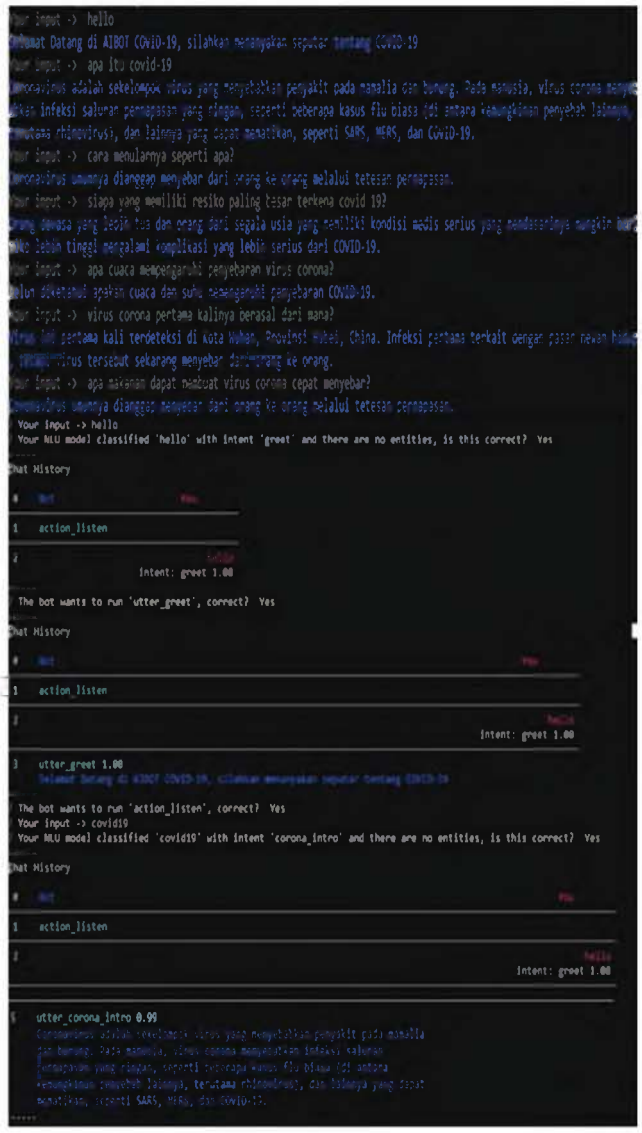

Fig. 8. COVID-19 Chatbot

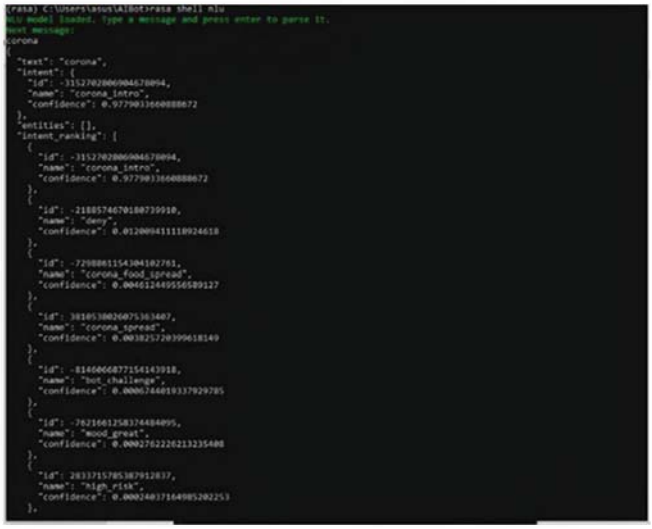

Fig. 9. Conclusion Confidence value prediction using DIET Classifier for the word "corona"

CONCLUSION

The DIET Classifier model in the RASA framework predicts answers on the chatbots related to COVID-19 information which uses around 300 epochs without normalizing data. DIET Classifier model can predict each word or sentence from the chatbot with different confidence values for all of the 11 intents used. The confidence value of the Intent which is the entity of the question produced an answer with an adequately high value of around 0.979 and for the intent value that is not an entity of the question it will have a moderate or small value of around 0.0002. The effectiveness of the DIET model pre-training with the use of embeddings always provides the best result among different data sets. Furthermore, DIET can work with large-scale embeddings as well although without trained embeddings, it can still obtain a competitive performance. Testing on the covid-19 chatbot with the RASA framework was carried out on rasa.core.test and rasa.nlu.test. The results of the taste.core.test for F1-Score, Precision, Accuracy have a value of 1.0 for the correct answer in accordance with the intent and entity that has been used. The results for rasa.nlu.test are also provided. In the future development of the COVID-19 chatbot, other models such as BERT, GloVe are recommended as a comparison of the prediction results obtained for the DIET Classifier model.

REFERENCES

[1] S. Koven, "They Call Us and We Go," *N. Engl. J. Med.*, vol. 382, no. 21, pp. 1978–1979, 2020.

[2] S. Platto, T. Xue, and E. Carafoli, "COVID19: An Announced Pandemic," *Cell Death and Disease*. Springer US, pp. 5690–5694, 2020, doi: 10.1038/s41419-020-02995-9.

[3] B. K. Romanov, "Coronavirus Disease COVID-2019," *Saf. Risk Pharmacother.*, vol. 8, no. 1, pp. 3–8, 2020, doi: 10.30895/2312-7821-2020-8-1-3-8.

[4] E. Ricciardelli and D. Biswas, "Self-improving Chatbots Based on Reinforcement Learning Self-improving Chatbots Based on Reinforcement Learning," in *In 4th Multidisciplinary Conference on Reinforcement Learning and Decision Making*, 2019, no. May, pp. 1–4.

[5] A. Jiao, "An Intelligent Chatbot System Based on Entity Extraction Using RASA NLU and Neural Network," *J. Phys. Conf. Ser.*, vol. 1487, no. 1, p. 012014, 2020, doi: 10.1088/1742-6596/1487/1/012014.

[6] P. Lauren and P. Watta, "A Conversational User Interface for Stock Analysis," in *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, 2019, pp. 5298–5305, doi: 10.1109/BigData-47090.2019.9005635.

[7] S. Raj, *Building Chatbots with Python*. 2019.

[8] R. K. Sharma and M. Joshi, "An Analytical Study and Review of open Source Chatbot framework," *Int. J. Eng. Res.*, vol. 9, no. 06, pp. 1011–1014, 2020.

[9] R. Parlika, S. I. Pradika, A. M. Hakim, and K. R. NM, "Bot Whatsapp Sebagai Pemberi Data Statistik," *J. Inform. dan Sist. Inf.*, vol. 1, no. 2, pp. 282–293, 2020.

[10] T. Bunk, D. Varshneya, V. Vlasov, and A. Nichol, "DIET: Lightweight Language Understanding for Dialogue Systems," *arXiv*. 2020.

[11] M. M. Lai and D. Cavanagh, "The Molecular Biology of Coronaviruses," *Adv. Virus Res.*, vol. 48, pp. 1–100, 1997, doi: 10.1016/s0065-3527(08)60286-9.

[12] J. Ziebuhr, "The Coronavirus Replicase," *Curr. Top. Microbiol. Immunol.*, vol. 287, pp. 57–94, 2005, doi: 10.1007/3-540-26765-4_3.

[13] N. Kuljić-Kapulica and A. Budisin, "Coronaviruses," *Srp. Arh. Celok. Lek.*, vol. 120, no. 7–8, pp. 215–218, 1992, doi: 10.4161/rna.8.2.15013.

[14] A. Singh, K. Ramasubramanian, and S. Shivam, *Introduction to Microsoft Bot, RASA, and Google Dialogflow*. 2019.

[15] D. Braun, A. H. Mendez, F. Matthes, and M. Langen, "Evaluating Natural Language Understanding Services for Conversational Question Answering Systems," in *SIGDIAL 2017 - 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Proceedings of the Conference*, 2017, no. August, pp. 174–185, doi: 10.18653/v1/w17-5522.

[16] A. Jaech, L. Heck, and M. Ostendorf, "Domain Adaptation of Recurrent Neural Networks for Natural Language Understanding," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 08-12-Sept, pp. 690–694, 2016, doi: 10.21437/Interspeech.2016-1598.

[17] X. Yang, Y. N. Chen, D. Hakkani-Tür, P. Crook, J. Li, X., Gao, and L. Deng, "End-To-End Joint Learning of Natural Language Understanding and Dialogue Manager," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2017, pp. 5690–5694, doi: 10.1109/ICASSP.2017.-7953246.

[18] R. Huijzer, "Automatically Responding to Customers." p. 58, 2019.

[19] T. Bocklisch, J. Faulkner, N. Pawlowski, and A. Nichol, "Rasa: Open-Source Language Understanding and Dialogue Management," *arXiv*, pp. 1–9, 2017.