ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse86942020.pdf
https://doi.org/10.30534/ijatcse/2020/86942020

Chatbot Smart Assistant Using N-Gram and Bi-Gram Algorithm

Taqwa Hariguna[1], Yusup Efendi[2]
[1]Department of Information System, Faculty of Computer Science, Universitas Amikom Purwokerto, Purwokerto, Indonesia, taqwa@amikompurwokerto.ac.id
[2]Department of Informatics, Faculty of Computer Science, Universitas Amikom Purwokerto, Purwokerto, Indonesia, efendiyusup520@gmail.com

ABSTRACT
Information is one of the things that cannot be released from human life. The researchers design and build a chatbot application as a machine learning-based information service media by implementing the jaccard similarity algorithm and the N-Gram algorithm feature, Bi gram, to get an appropriate response based on user questions compared to corpus. As a result of this research project, it is expected that the process of finding information gets clear information accuracy and time effectiveness of the search for users in searching for information.

Keywords: Chatbot, N-Gram, Bi-Gram, Jaccard

1. INTRODUCTION

In today's digital era, human dependence on technology is increasingly apparent, one of which is information technology. Information is one of the things that cannot be released from human life.
The development of information technology today is utilizing artificial intelligence (AI). AI is a process in which a computer or machine can do or respond to an incident around it to maximize the results to be achieved with a mindset like humans.
Machine learning is a part of (AI). Machine learning is clustering and classification. Clustering is an activity that aims to group data based on the proximity of the features it has, while classification aims to separate data into certain classes [7][8][9].
One application of machine learning is Chatbot. Chatbot or robot chat is a conversation machine that is specifically designed to respond to an input made by the user (human) by using natural language so that there is a conversation interaction worthy of a conversation between two individuals. Chatbot itself is a computer program that is designed to have a conversation using natural language or language used by humans based on a topic that is in the chatbot knowledge model. This means that chatbot must be able to recognize every word entered by the user[10][11].
One of the detection of words or strings is N-gram. N-gram is a chunk of n characters in a certain string or n numbers in a certain sentence [12][13]. One of the benefits of N-gram in identifying words and strings is its resistant nature in recognizing errors in writing made by humans. So that errors in the string only affect some N-grams. In fact human language always has a word whose frequency is higher than other words, it also forms the basis of the N-gram method [14][15]. As a result of the frequency of letters appearing can be varied, for example the letter "a" appears highest in the Indonesian text, while in English the vowel "e" is the letter with the highest number of words. The difference in the appearance of letters or words indicates that the N-gram of each word is unique so that it can be made a profile of each language.
In the N-gram is divided into several features but of all the features that have N-gram, Bi-gram is the most significant feature in checking a word or string.
Although the features of the N-gram, especially the bigram, are very good in correcting words and strings, Bi-Gram cannot distinguish the closeness of a string. This is because each user in communication differs from one another even though in conveying the information has the same purpose. For example by comparing two sentences: the cat is very hungry. The black cat is hungry. Of the two sentences have the same resemblance of a hungry cat but different in conveying it for that use jaccard similarity to find out the closeness of a sentence which is later compared with a list of sentences owned by Chattbot, so as to produce an appropriate response based on the input asked by the user against chatbots.

2. METHODS

2.1 Chatbot design models

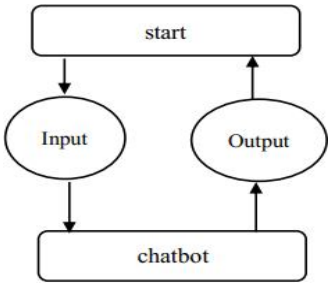

Figure 1: Chatbot Model

Figure 1 is the stage how users interact with chatbots

.2.2 Proses Response Chatbot

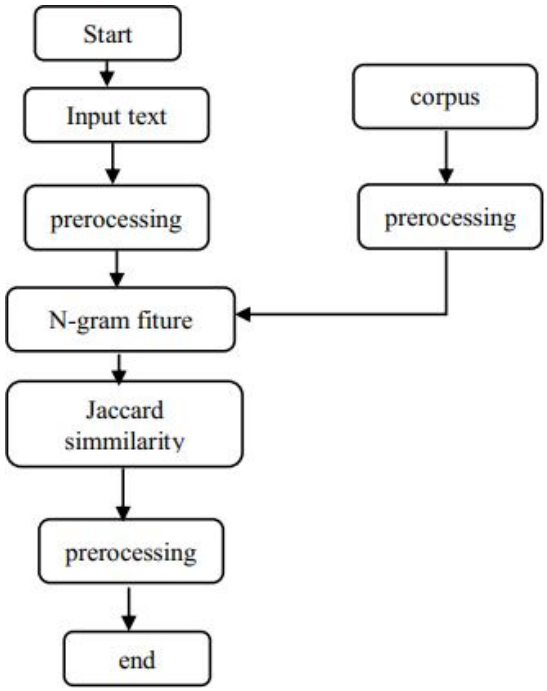

Figure 2: Jaccard Index Chatbot Process Flowchart on N-Gram Features.

In Figure 2 is the workflow of chatbot using the N-gram and jaccard methods, which are started from the text input process that is done by the user, then text processing will be carried out and then a character division of n is done, in this research only using n = 2 or in call it the Bi-gram, then the text will be indexed using the jaccard method, it is also done from the knowledge base to measure the closeness of the questions the user inputs with questions in the knowledge base.

2.3 Processing Jacard Similarity in N-Gram Fitur Bi-Gram

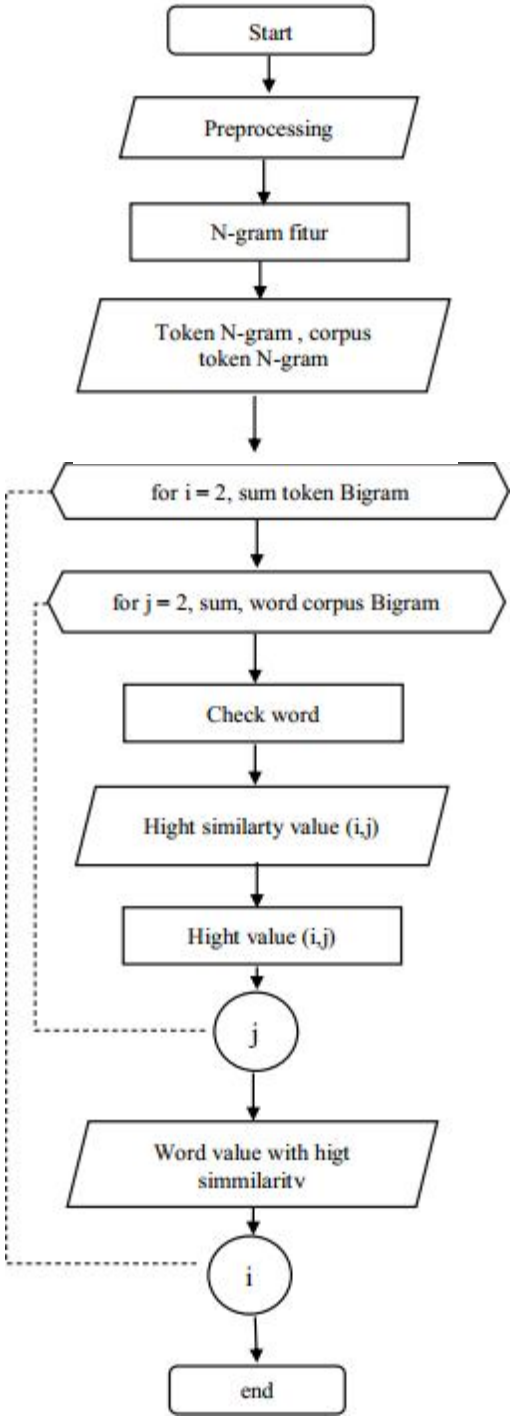

Figure 3: Preprocessing

In Figure 3, the word correction stage starts with the tokenization of words into N-gram words, from each set of tokens the level of similarity is calculated with words or sentences in the knowledge base or corpus chatbot, by implementing jaccard similarity to look for the highest similarity value of a question that is in question inputted by

the user with a list of intelligence in the chatbot, by comparing questions in the chatbot intelligence data with questions input by the user, to find the best answer expected by the user.

3. DISCUSSION AND COMMENTS

3.1 Implementation of Methods and Systems

A. Bi-Gram

After performing the text-processing stage the character truncation is performed with n = 2, the following is a result of bi-gram characters

Table 1: Truncation of Bi-Gram Characters

No	world	Bi-gram
1	when	_w,wh,he,en,n_
2	Amikom	_a,am,mi,ik.ko,om,m
3	Purwokerto	_p,pu,ur,rw,wo,ok,ke,er,rt,to,o_
4	university	_u,un,ni,iv.ve,er,rs,si,it,ty, y_
5	student	_s,st,tu,ud,de,en,nt,t_
6	registration	_r,re,eg,gi,is,st,tr,ra,at,ti,io,on,n_

The word contained in the knowledge base or corpus also undergoes a Bi-gram process.

B. Jaccard index

Jaccard and N-gram implementation in this chatbot is in the process of searching strings or sentences based on keywords that already exist in the chatbot database to be able to provide answers based on input made by the user. Then from the input will be matched based on keywords and categories of user input[16][17][18]. The following is an implementation of the Jaccard index on N-gram text.

d1 = when is the registration student of university of amikom purwokerto.

d2 = student registration of university of amikom purwokerto

d3 = registration in amikom?

At this stage manual calculations are performed with the Jaccard index on and N-gram text. The following is the specific value of Bi-gram with n = 2.

d1 = when is the registration student of university of amikom purwokerto?

d2 = student registration of university of amikom purwokerto

d3 = registration in amikom?

J = (d1/d2) = |A n B / A U B | = 4 / 5 = 0.8

d1 = {when is amikom, amikom purwokerto, purwokerto university, university student, student registration}

The number of members of the set n on one is 6.

d2 = {amikom purwokerto, purwokerto university, university student, student registration}. The number of members of set n on d2 is 4.

Then d1 n d2 amounts to N = 4 because there are the same number of words as much as 2, and d1 U d2 is the sum of all members of the set d1 and d2 = 6.

J = (d2/d3) = |A n B / A U B | = 0/ 6 = 0

d2 = {amikom purwokertp, purwokerto university, university student, student registration}. The number of members of the set N on d2 is 4.

d3 = {registration in, in amikom}

The number of members of the set N on d3 is 0.

From the two calculations above, it is concluded that the similarity of words that have a similarity is depressed, namely between d1 and d2, with a score of 0.8

Figure 4: Display of the ASA Chatbot

In Figure 4 is the initial display chatbot the jaccard test and Bigram response are shown in Figure 5 below:

Figure 5: Chatbot Response

Chatbot is also designed to give a default answer if the chatbot cannot write an appropriate answer from the question that the user input[19][20]. It can be seen in Figure 6 below:

Figure 6: Default Chatbot Response

In Figure 6 is the default answer from chatbot if the chat bot does not find an answer that matches the user input question by giving a message "sorry I don't understand, please contact us via whatsapp at 085845 for more info, sis".

3.2 System Testing

Table 2: Testing Table for Chatbot Black Box Testing

No	Component Testing	Input	Output	Status
1	Chatbot message	displays the chatbot page	system is capable of displaying chatbot pages	accepted
2	Input message	input message in the text file	system is capable of responding input	accepted
3	response	question response	system is able to respond well to questions	accepted
4	Response	the accuracy of chatbot answers based on user input	chatbot is able to provide answers that match the questions asked by users	accepted
5	Response	Default answer	chatbot is able to provide default answers when answers are not in the knowledge data	accepted
6	algorithm performance	algorithm performance biagram and jaccord similarity	chatbot or system running	accepted

Table 3: Response Chatbot Testing

No	Input	Keyword search	System response
1	When amikom registration is opened	Amikom registration opened	Amikom Purwokerto student registration starts from 2 December 2019 - 12 September 2020 which is divided into several waves as follows: 1. Period of achievements opened from December 2 2019 - 2 March 2020 2. Special Period are opened from 1 January 2020 - 31 March 2020 3. Period 1 is opened from 1 April 2020 - 30 May 2020 4. Period 2 is opened from 1 June 2020 - 31 July 2020 5. Period 3 is opened from 1 August 6. 2020 - 12 September 2020
2	What are the benefits of registering at Amikom Purwokerto if done at the beginning?	The benefits of registering at Amikom Purwokerto	Without a written test (the opportunity is accepted 0at Amikom University Purwokerto is greater, more affordable fees, direct interview test
3	Amikom registration requirements	Registration requirements	Registration requirements: 1. Photocopy of KTP/KK 2. Photocopy of diploma/transcript

From the results of the black box testing scenario the system can run well starting with menus and displays in accordance with what has been made and testing that has been carried out in accordance with their respective functions, following the test results from the Chatbot Smart Assistant system and for testing the response of Chatbot with Chatbot testing is done by inputting 30 questions.

Testing is done by running the Chatbot application, and then entering a message in the text file provided, then press the send button to get a response or answer[21][22].

Based on the 30 questions entered by the user, 25 questions were answered accordingly, based on these results, to find out the level of success of the Bi-gram algorithm and jaccard similarity is to use the following calculation:

Accuracy = The number of successful responses is correct x 100% Number of questions tested

Accuracy = 30/25 X 100%

Accuracy = 83,3%

So the percentage of success of Chatbot by using the Bi-gram algorithm and Jaccard Index is 83.3%.

Based on the results of the calculation above the success rate of the Bi-gram algorithm and jaccard similarity is 83.3%, this is good enough although it cannot provide answers to questions asked by the user with perfect accuracy, but the Bi-gram and jaccard similarity algorithms can recognize words based on the similarity of words entered by the user with intelligence data from Chatbot.

4. CONCLUSION

Based on the analysis and testing of the chatbot application it can be concluded that the chatbot smart assistant application can respond properly according to the questions submitted with a success rate of 0. This shows that the chatbot using Bigram features and jaccard similarity works well because the presentation of success is above 80%. Expand twining data to produce even more intelligent Chatbot and maximum results.

ACKNOWLEDGEMENT

The authors would like to thank the financial support from Department of Information System, Faculty of Computer Science, Universitas Amikom Purwokerto.

REFERENCES

1. Campagna, G., R. Ramesh, S. Xu, M. Fischer, and M.S. Lam, **"Almond: The Architecture of an Open, Crowdsourced, Privacy-Preserving, Programmable Virtual Assistant",** International World Wide Web Conferences Steering Committee, 2017. https://doi.org/10.1145/3038912.3052562
2. Benedictus, R. R., Wowor, H., & Sambul, A. (2017). **Rancang Bangun Chatbot Helpdesk untuk Sistem Informasi Terpadu Universitas Sam Ratulangi**. E-Journal Teknik Informatika.
3. Alpaydın, E. (2010). **Introduction to Machine Learning**. London: The MIT Press.
4. Utama, P. K. (2018). **Bot Chat: Customer Relation dengan Teknologi Artificial Intelligence**. JURNAL ILMIAH ILMU AGAMA DAN ILMU SOSIAL BUDAYA, 81-87
5. Lisangan, E.A., 2015, **Implementasi n-Gram Technique dalam Deteksi Plagiarisme pada Tugas Mahasiswa**. Jurnal Tematika, Vol. 1 No. 2, ISSN: 2303-387824-30.
6. Hamzah, A. (2010). **Deteksi bahasa untuk dokumen teks berbahasa Indonesia.** Dalam prosiding Dukungan ICT dalam bidang industry dan manajemen ESDM. Halaman A-5 – A-13
7. Padr´o, M.,dan Padr´o, L. (2004). **Comparing methods for language identification**. Dalam prosiding Procesamiento del Lenguaje Natural. Halaman 155–162.
8. Indriani A dkk (2018). **Implementasi Jaccrad index dan Ngram pada rekayasa Aplikasi koreksi kata Bahasa Indonesia**. SEBATIK 1410-3737 Hal 95-101
9. Nugraheny D (2015). **Metode Nilai jarak guna kesamaan atau kemiripan ciri suatu citra (kasus deteksi awan colonimbus menggunakan komponen principal componet analisis**) jurnal amgkasa vol VII 21-20
10. Chawla, R., & Anuradha, J. **COUNSELLOR CHATBOT. Computer Science**, 5, 126-136.
11. Kumar, L., & Bhatia, P. K. (2013). **Text mining: concepts, process and applications**. Journal of Global Research in Computer Science, 4 (3), 36-39.
12. K. Jwala, G. S. (2019). **Developing a Chatbot using Machine Learning**. International Journal of Recent Technology and Engineering (IJRTE), vol. 8 89-92.
13. Niwattanakul, S., Singthongchai, J., Naenudorn, E., & Wanapu, S. (2013, March). **Using of Jaccard coefficient for keywords similarity**. In Proceedings of the international multiconference of engineers and computer scientists (Vol. 1, No. 6, pp. 380-384).
14. Mumu, J., & Tanujaya, B. (2018). **Desain pembelajaran materi operasi pada himpunan menggunakan permainan "Lemon Nipis"**. Journal of Honai Math, 1(1), 14-23.
15. Hariguna, T., Adiandari, A. M., & Ruangkanjanases, A. (2020). Assessing customer intention use of mobile money application and the antecedent of perceived value, economic trust and service trust. International Journal of Web Information Systems. http://doi.org/10.1108/IJWIS-12-2019-0055
16. Hariguna, T., & Ruangkanjanases, A. (2020). Elucidating E-satisfaction and Sustainable Intention to Reuse Mobile Food Application Service , Integrating Customer Experiences , Online Tracking , and Online Review, XXIX, 122–138. http://doi.org/10.24205/03276716.2020.704
17. Hariguna, T., Rahardja, U., & Ruangkanjanases, A. (2020). The impact of citizen perceived value on their intention to use e-government service: An empirical study. Electronic Government, an International Journal, 16(1), 1. http://doi.org/10.1504/eg.2020.10028551
18. Hariguna, T., Maulana, W., & Nurwanti, A. (2019). Sentiment Analysis of Product Reviews as A Customer Recommendation Using the Naive Bayes Classifier

Algorithm. International Journal of Informatics and Information Systems, 2(2), 48–55.
19. Imron, M., Hasanah, U., & Humaidi, B. (2020). Analysis of Data Mining Using K-Means Clustering Algorithm for Product Grouping. International Journal of Informatics and Information Systems, 3(1), 12–22.
20. Hariguna, T., & Rachmawati, V. (2019). Community Opinion Sentiment Analysis on Social Media Using Naive Bayes Algorithm Methods. International Journal of Informatics and Information Systems, 2(1), 33–38.
21. Santiko, I., & Subarkah, P. (2019). Comparison of Cart and Naive Bayesian Algorithm Performance to Diagnose Diabetes Mellitus. International Journal of Informatics and Information Systems, 2(1), 9–16.
22. Hariguna, T., Hung, C., & Sukmana, H. T. (2019). The antecedent of citizen intention use of e-government service. TELKOMNIKA (Telecommunication Computing Electronics and Control), 17(1), 202–209. http://doi.org/10.12928/TELKOMNIKA.v17i1.11588