Journal of Multimedia Trend and Technology - JMTT
Vol. 2, No. 1, April 2023, ISSN 2964-1330
https://journal.educollabs.org/index.php/jmtt/

Implementation of Smart Communication BOT Model Using Machine Learning with the N-Gram Method

Irfan Santiko[1], Muhammad Arifin[2]

[1] Informatic Departement, Universitas Amikom Purwokerto
[2] Information System Department, Universitas Muria Kudus
Jawa Tengah, Indonesia
Email: [1] irfan.santiko@amikompurwokerto.ac.id, [2] arifin.m@umk.ac.id

ARTICLE INFO

History :

Submit on 30 December 2022
Review on 8 January 2023
Accepted on 15 February 2023

Keyword :

Machine Learning,
BOT,
N-Gram,
Communication

ABSTRACT

The tight world of higher education in providing educational services, makes a lot of effort made by the service department at each institution. These efforts are often encountered in certain moments such as before the acceptance of new students. But how is time other than that moment used? Certainly this is related to how to make a strategy so that the market share of these educational institutions is wider. Service and handling is a priority for a company engaged in the world of services. With the ability to develop information technology, it will be able to provide support for service methods and handling to customers. The point is to make service and handling more optimal compared to the usual method. Cannot be separated from the rules, policies, processes and procedures that are owned by every service company. Systematically the technology that will be used also follows the outline of the rules, policies and procedures previously mentioned. This research will be conducted at higher education institutions where an institution is engaged in educational services. This data processing uses machine learning with the N-Gram method. This system will run on the front office or customer service side. The purpose of the results of this research is to be more useful and to assist customer service in higher education institutions in providing services to prospective students or student guardians.

Corresponding Author:

Irfan Santiko
Informatic, Universitas Amikom Purwokerto Jawa Tengah
Email : irfan.santiko@amikompurwokerto.ac.id

INTRODUCTION

In today's digital era, human dependence on technology is increasingly evident, one of which is information technology. Information is one thing that cannot be separated from human life [1]. The development of information technology today is by utilizing artificial intelligence. AI is a process where a computer or machine can do or respond to events around it to maximize the results to be achieved with a human-like mindset. Machine learning is a part of. Machine learning is clustering and classification [2]. Clustering is an activity that aims to group data based on the proximity of its features, while classification aims to separate data into certain classes [3].

The world of information is getting wider and growing fast. The public's need for information is also very high [4]. This is evident from the results of a survey sourced from the Republika Online website, which states that 74 million Indonesians or 28% of the entire Indonesian population consider themselves "Netizen". Many users means consuming information in cyberspace (internet) at least 6-8 hours every day. The remaining 72% consume information via mobile & smart phone [5].

The tight world of higher education in providing educational services, makes a lot of effort made by the service department at each institution [6]. These efforts are often encountered in certain moments such as before the acceptance of new students/students [7]. But how is time other than that moment used? Certainly this is related to how to make a strategy so that the market share of these educational institutions is wider [8].

The institutions that are the object of research in this research paper are no exception, namely universities [9]. Universities also do the same thing, namely how to make a strategy in expanding market share or in this case the target market is prospective students [10]. In its own character, Higher Education tends to apply points from the B2C and B2B characters, meaning that the institution seeks to carry out automation, collaboration and electronic communication focused on the aspects of customer relations, service and handling. On the other hand, Universities want their customers to not only be a source of company income, but also hope that customers are also long-term assets that need to be managed and maintained using a Customer Relationship Management strategy [11].

There are many such or similar applications that are used for certain needs, such as ElBot, Jaberwacky, Pandora, Mitsuku and currently what is becoming popular is OpenAI Chat GPT. This is the basis of the ideas obtained to take advantage of these artificial intelligence features [6][12] [8].

One application of machine learning is chatbots. Chatbot or communication robot is a conversation machine that is specifically designed to respond to input by the user (human) by using natural language processing so that a conversational interaction occurs like a conversation between two individuals [10]. Chatbot itself is a computer program designed to carry out conversations using natural language or language used by humans based on a topic that is in the Chatbot knowledge model. This means that the chatbot must be able to recognize every word inputted by the user [13].

One of the word or string detection is N-gram. N-grams are pieces of n characters in a certain string or pieces of n in a certain sentence. One of the advantages of the N-gram in identifying words and strings is that it is resistant to recognizing errors in human writing [14]. So that the error in the string only results in a portion of the N-grams [15]. The fact that human language always has words with a higher frequency of occurrence than other words is also the basis for using the N-gram method [16]. As a result, the frequency of occurrence of letters can also vary, for example the letter "a" appears highest in Indonesian text, while in English, the vowel "e" is the letter with the highest frequency of occurrence. The difference in the occurrence of letters or words indicates that the N-gram of each word is unique so that it can be used as a profile for each language [13][7]. In N-gram it is divided into several features but of all the features that N-gram has, Bi-gram is the most significant feature in examining a word or string.

Even though the features possessed by N-grams, especially bigrams, are very good at correcting words and strings, Bi-Grams cannot distinguish the proximity of a string, this is because each user communicates differently from one another even though conveying the information has the same purpose. For example by comparing two sentences [17]. the cat is very hungry. The black cat is hungry. the two sentences have the same resemblance, that is, the cat is hungry, but it is different in conveying it [17]. For this reason, jaccard similarity is used to find out the proximity of a sentence which will later be compared with the list of sentences owned by the chatbot, so that it can produce an appropriate response based on the input asked by the user against chatbots [18].

METHOD

In the method in this paper, the author arranges it into several model frameworks, then arranges the concept into a series of code written using PHP syntax. In the first modeling the authors arrange the framework in the index diagram as follows:

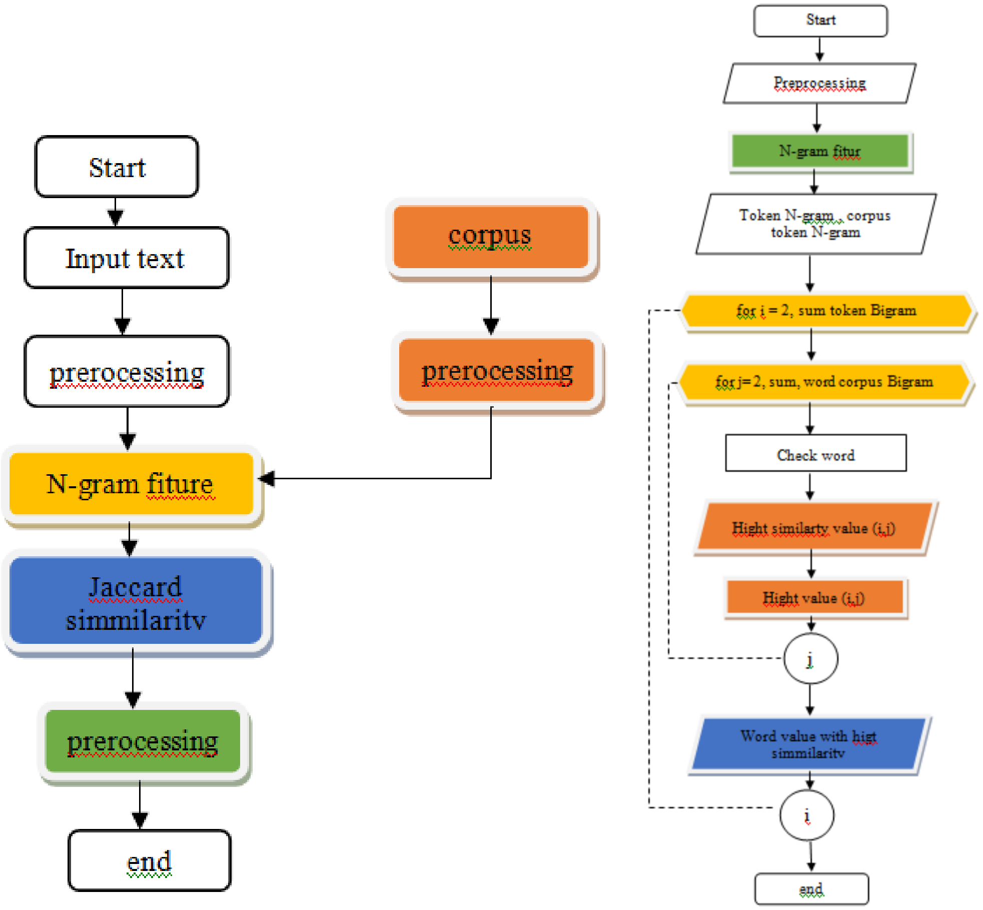


Figure 1, Model design flow and program plan

In figure 1 is the workflow of the chatbot using the N-gram and jaccard methods, which starts from the input text process carried out by the user, then text processing will then be carried out by dividing the characters in a number of n, in this research only using n =

2 or the call Bi-gram, then the text will be indexed using the jaccrd method, this is also done from a known base to measure the closeness of the questions inputted by the user to questions in the known base.

In Figure 1, the word correction stage begins by tokenizing words into N-gram words, from each token the level of similarity is calculated with words or sentences in the kenoledge base or corpus chatbot, by implementing Laccard similarity to find the highest similarity value of a question which is input by the user with the intelligence list in the chatbot, by comparing the questions in the chatbot intelligence data with the questions entered by the user, to find the best answer expected by the user.

RESULT & DISCUSSION

Creating a Chatbot by implementing the feature N-gram algorithm, namely Bi-gram and Jaccard similarity, is because each user has a different language in making conversations but has the same aims and objectives. For this reason, the implementation of the Bi-gram and Jaccard similarity algorithms is very suitable for solving problems and identifying text entered by users. In the first discussion, the author carries out a stage called preprocessing, which is where each data is collected to be processed using a learning machine. After carrying out the text-processing stage, the characters are cut with n = 2, the following is the result table for bigram character cuts.

Table 1, Bi-Gram Notation

No	world	Bi-gram
1	*when*	_w,wh,he,en,n_
2	*college*	_c,co, col, ll,le, leg, ge_
3	*info*	_i,in,nf,fo,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 implementation of Jaccard and N-gram in this chatbot is in the process of searching for 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 this input a match will be made based on keywords and categories from user input. The following is an implementation of the Jaccard index on N-gram text.

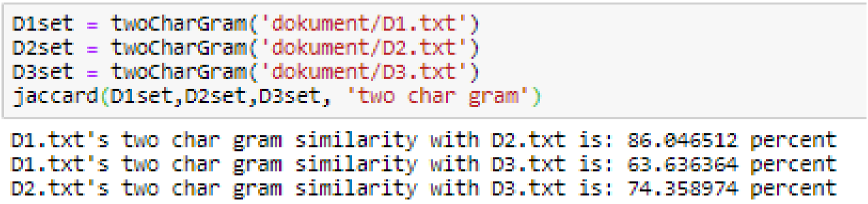
d1 = when is a campus or university student.
d2 = campus or university student registration.
d3 = registration in university?

At this stage, manual calculations are carried out with the Jaccard index on and N-gram text. The following is the special case for Bi-grams with the number n = 2.

d1 = when is a campus or university student.
d2 = campus or university student registration.
d3 = registration in university?

J = (d1/d2) = |A n B / A U B | = 4 / 5 = 0.8
d1 = { when is a campus, campus university, university, university student, student registration }
The number of members of the set n in one is 6
d2 = { campus university, university, university student, student registration}
the number of members of the set n in d2 is 4
then d1 n d2 totals N = 4 because there are 2 of the same words, and d1 U d2 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 = { campus university, university, university student, student registration }
The number of members of the set N in d2 is 4
d3 = {registration in, in university}
the number of members of the set N on d3 is 0.



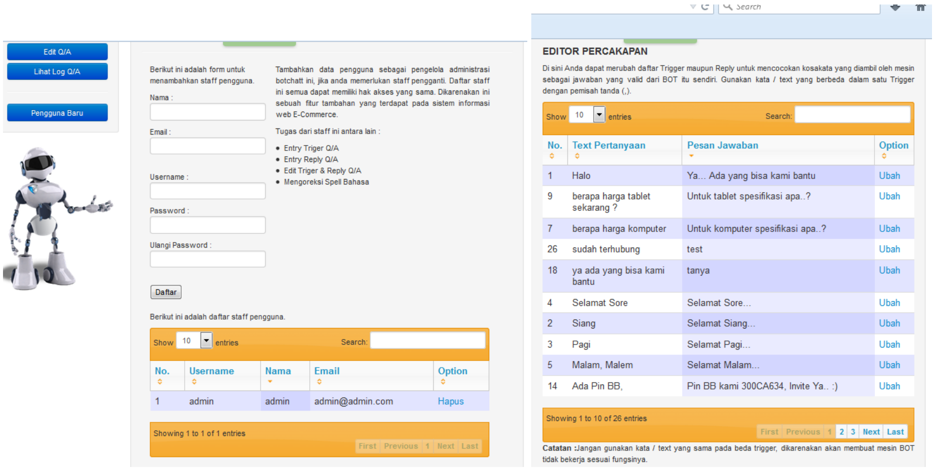
```
D1set = twoCharGram('dokument/D1.txt')
D2set = twoCharGram('dokument/D2.txt')
D3set = twoCharGram('dokument/D3.txt')
jaccard(D1set,D2set,D3set, 'two char gram')
```

```
D1.txt's two char gram similarity with D2.txt is: 86.046512 percent
D1.txt's two char gram similarity with D3.txt is: 63.636364 percent
D2.txt's two char gram similarity with D3.txt is: 74.358974 percent
```

Figure 2, Calculations based on Jaccard Similarity

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

The next stage is to compile it into a programming language. The author translates into a website-based program, namely PHP and in storing data using MySQL Server. The following display results that can be seen. The next stage is to compile it into a programming language. The author translates into a website-based program, namely PHP and in storing data using MySQL Server. The following display results that can be seen. In this case the author uses "Bahasa" because this platform is tested at universities in Indonesia.



Berikut ini adalah form untuk menambahkan staff pengguna.

Nama :

Email :

Username :

Password :

Ulangi Password :

Tambahkan data pengguna sebagai pengelola administrasi botchatt ini, jika anda memerlukan staff pengganti. Daftar staff ini semua dapat memiliki hak akses yang sama. Dikarenakan ini sebuah fitur tambahan yang terdapat pada sistem informasi web E-Commerce.

Tugas dari staff ini antara lain :

- Entry Triger Q/A
- Entry Reply Q/A
- Edit Triger & Reply Q/A
- Mengoreksi Spell Bahasa

Berikut ini adalah daftar staff pengguna.

Show 10 entries Search:

No.	Username	Nama	Email	Option
1	admin	admin	admin@admin.com	Hapus

Showing 1 to 1 of 1 entries

First Previous 1 Next Last

EDITOR PERCAKAPAN

Di sini Anda dapat merubah daftar Trigger maupun Reply untuk mencocokan kosakata yang diambil oleh mesin sebagai jawaban yang valid dari BOT itu sendiri. Gunakan kata / text yang berbeda dalam satu Trigger dengan pemisah tanda (,).

Show 10 entries Search:

No.	Text Pertanyaan	Pesan Jawaban	Option
1	Halo	Ya... Ada yang bisa kami bantu	Ubah
9	berapa harga tablet sekarang ?	Untuk tablet spesifikasi apa..?	Ubah
7	berapa harga komputer	Untuk komputer spesifikasi apa..?	Ubah
26	sudah terhubung	test	Ubah
18	ya ada yang bisa kami bantu	tanya	Ubah
4	Selamat Sore	Selamat Sore...	Ubah
2	Siang	Selamat Siang...	Ubah
3	Pagi	Selamat Pagi...	Ubah
5	Malam, Malem	Selamat Malam...	Ubah
14	Ada Pin BB,	Pin BB kami 300CA634, Invite Ya.. :)	Ubah

Showing 1 to 10 of 26 entries

First Previous 1 2 3 Next Last

Catatan :Jangan gunakan kata / text yang sama pada beda trigger, dikarenakan akan membuat mesin BOT tidak bekerja sesuai fungsinya.

Figure 3, Platform Model BOT Communication

From the results of the black box testing scenario, the system can run properly starting with the menus and displays that are in accordance with those that have been made and the tests that have been carried out in accordance with their respective functions. The following are the test results from the Chatbot Smart Assistant system and for testing the response from 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 pressing the send button to get a response or answer. Based on the 30 questions entered by the user, 25 questions were answered accordingly. Based on these results, to determine the success rate of the Bi-gram and Jaccard similarity algorithms, we use the following calculations:

$$Accuracy =$$
$$Accuracy = 30/25 \; X \; 100\%$$
$$Accuracy = 83.3\%$$

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

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

CONCLUTIONS

Based on the results of the research that has been done, several conclusions can be obtained, including, 1) It has been successful in designing and building chatbots by being able to answer or respond to questions properly. This is done by conducting tests using black box testing, and it can be concluded that chatbots can run as expected. 2) Regarding the Bi-gram and Jaccard simlarity algorithms used, it works as expected. 3) The Chatbot success rate is 83.3% of the number of questions tested. 4) The level of service in the STMIK AMIKOM Purwokerto environment becomes more controlled, efficient, and has a fast response, and 5) Can be an added value for the performance of the marketing department, especially in serving prospective students.

Acknowledgement

In writing this paper the author would like to express his gratitude to all partners who have collaborated to help design this platform. As many users know, AI currently dominates every application based on information technology. This platform is only part of a function adopted for certain needs, especially in Higher Education. With the presence of platforms such as Chat GPT issued by Open AI, this is the biggest challenge in Artificial Intelligence technology.

REFERENCE

[1] M. Casillo, F. Clarizia, G. D'Aniello, M. De Santo, M. Lombardi, and D. Santaniello, "CHAT-Bot: A cultural heritage aware teller-bot for supporting touristic experiences," *Pattern Recognit. Lett.*, vol. 131, pp. 234–243, 2020, doi: https://doi.org/10.1016/j.patrec.2020.01.003.

[2] M. Poongodi, V. Vijayakumar, L. Ramanathan, X.-Z. Gao, V. Bhardwaj, and T. Agarwal, "Chat-bot-based natural language interface for blogs and information networks," *Int. J. Web Based Communities*, vol. 15, no. 2, pp. 178–195, 2019, doi: 10.1504/IJWBC.2019.101048.

[3] A. Prisco *et al.*, "A Facebook chat bot as recommendation system for programming problems," in *2019 IEEE Frontiers in Education Conference (FIE)*, 2019, pp. 1–5. doi: 10.1109/FIE43999.2019.9028655.

[4] A. Hassanzadeh, F. Kanaani, and S. Elahi, "A model for measuring e-learning systems success in universities," *Expert Syst. Appl.*, vol. 39, no. 12, pp. 10959–10966, 2012, doi: 10.1016/j.eswa.2012.03.028.

[5] I. Santiko, "Desain Dan Implementasi Semantik Pada Fitur Pencarian Di Aplikasi Perpustakaan Stmik Amikom Purwokerto," *Semin. Nas. Teknol. Informasi, Bisnis, dan Desain*, pp. 265–271, 2017.

[6] T. Holtgraves and T.-L. Han, "A procedure for studying online conversational processing using a chat bot," *Behav. Res. Methods*, vol. 39, no. 1, pp. 156–163, 2007, doi: 10.3758/BF03192855.

[7] I. Roll and R. Wylie, "Evolution and Revolution in Artificial Intelligence in Education," *Int. J. Artif. Intell. Educ.*, vol. 26, no. 2, pp. 582–599, 2016, doi: 10.1007/s40593-016-0110-3.

[8] *et al.*, "Investigating the acceptance of applying chat-bot (Artificial intelligence) technology among higher education students in Egypt," *Int. J. High. Educ. Manag.*, vol. 08, no. 02, pp. 1–13, 2022, doi: 10.24052/ijhem/v08n02/art-1.

[9] N. Amin, S. Singh, and A. Walavalkar, "College Enquiry Chatbot System," *Indian J. Comput. Sci.*, vol. 7, no. 2, p. 16, 2022, doi: 10.17010/ijcs/2022/v7/i2/169681.

[10] P. Thosani, M. Sinkar, J. Vaghasiya, and R. Shankarmani, "A Self Learning Chat-Bot From User Interactions and Preferences," in *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2020, pp. 224–229. doi: 10.1109/ICICCS48265.2020.9120912.

[11] P. Punithavathi and S. Geetha, "Disruptive smart mobile pedagogies for engineering education," *Procedia Comput. Sci.*, vol. 172, no. 2019, pp. 784–790, 2020, doi: 10.1016/j.procs.2020.05.112.

[12] S. S and H. Wang, "Naive Bayes and Entropy based Analysis and Classification of Humans and Chat Bots," *J. ISMAC*, vol. 3, no. 1, pp. 40–49, 2021, doi: 10.36548/jismac.2021.1.004.

[13] L. Chen, P. Chen, and Z. Lin, "Artificial Intelligence in Education: A Review," *IEEE Access*, vol. 8, pp. 75264–75278, 2020, doi: 10.1109/ACCESS.2020.2988510.

[14] M. Vaziri, L. Mandel, A. Shinnar, J. Siméon, and M. Hirzel, "Generating Chat Bots from Web API Specifications," in *Proceedings of the 2017 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*, 2017, pp. 44–57. doi: 10.1145/3133850.3133864.

[15] D. Krisbiantoro, S. F. Rohim, and I. Santiko, "Perbandingan Algoritma N-gram dan Algoritma Knuth Morris Pratt untuk Mengukur Tingkat Akurasi Plagiarisme pada Dokumen Abstrak Skripsi Berbasis Website," *JITU J. Inform. Technol. Commun.*, vol. 5, no. 1, pp. 30–39, 2021, doi: 10.36596/jitu.v5i1.390.

[16] D. Yang *et al.*, "DevOps in practice for education management information system at ECNU," *Procedia Comput. Sci.*, vol. 176, pp. 1382–1391, 2020, doi:

10.1016/j.procs.2020.09.148.

[17] A. Abu-Al-Aish and S. Love, “Factors influencing students’ acceptance of m-learning: An investigation in higher education,” *Int. Rev. Res. Open Distance Learn.*, vol. 14, no. 5, pp. 82–107, 2013, doi: 10.19173/irrodl.v14i5.1631.

[18] Y. A. Adenle, E. H. W. Chan, Y. Sun, and C. K. Chau, “Exploring the coverage of environmental-dimension indicators in existing campus sustainability appraisal tools,” *Environ. Sustain. Indic.*, vol. 8, no. June, p. 100057, 2020, doi: 10.1016/j.indic.2020.100057.