International Journal of Engineering & Technology, 7 (4.36) (2018) 1382-1385

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

N-Gram Accuracy Analysis in the Method of Chatbot Response

Dhebys Suryani Hormansyah[1]*, Eka Larasati Amalia[2], Luqman Affandi[3], Dimas Wahyu Wibowo[4], Indinabilah Aulia[5]

s[1]State Polytechnic of Malang
[2]State Polytecnic of Malang
[3]State Polytecnic of Malang
[4]State Polytecnic of Malang
[5]State Polytecnic of Malang
**Corresponding author E-mail: dhebys.suryani@polinema.ac.id*

Abstract

Chatbot is a computer program designed to simulate interactive conversations or communication to users. In this study, chatbot was created as a customer service that functions as a public health service in Malang. This application is expected to facilitate the public to find the desired information. The method for user input in this application used N-Gram. N-gram consists of unigram, bigram and trigram. Testing of this application is carried out on 3 N-gram methods, so that the results of the tests have been done obtain the results for unigram 0.436, bigram 0.28, and trigram 0.26. From these results it can be seen that trigrams are faster in answering questions.

Keywords: Chatbot,TF-IDF,Cosine Similarity, N-gram, Bot Line

1. Introduction

Information needed by the community will continue to increase. Time efficiency and delivery greatly support the accuracy of information. The development of information technology will facilitate the society in getting the information needed. As information media develops, more cities develop in the future. Public Service is a medium provided by the government to be able to provide up-to-date information to the public. With the availability of public services, the community will be facilitated in finding information desired and expected by the community to be able to utilize the service to the maximum extent possible.There is a website to provide public service information in the city of Malang, namely www.malangkota.go.id. Website visitors will get information about health services on the website, and can see the information provided. But in the delivery of information it feels less interactive, where visitors are required to carefully sort out the available information data where they need. Submission of information on the website only has a table and there is no customer service. Customer service will be very useful on a public service website supported by the existing technology trends.

There is an application system that can be used as a substitute for customer service in the form of a chatbot application system, where the application will be placed on the website. Based on the existing system, it will be developed again and also increase the response results of the system that previously could not display results in accordance with the details desired by the user community. The development of information systems follows the needs of the user community, namely ChatBot Line.

Therefore, to overcome this problem a system can be created that can be used as a substitute for customer service in the form of a chatbot application system. We often know that customer service is required to standby 24 hours non-stop. Chatbot itself is a computer program designed to simulate an interactive conversation or communication to a user (human) through text, sound and visual forms. By using the TF-IDF method as a question and answer response and the N-gram method as a method that processes the input information from the user / community of users, which will be applied in the execution of the Chatbot Line application on public health services in Malang City**.**

2. Basic Theory

a. Chatbot

Chaybot is one of Natural Languange Processing (NLP) branch. NLP learns communication between humans and computers through natural language [1]. Chatbot allows humans to communicate with machines using everyday language. The form of communication that occurs is through conversation using written media. Conversations with chatbots can be in the form of regular chat or chat on certain themes involving other disciplines. Conversations that occur between computers and humans is a form of response from programs that have been declared in the program database on the computer [2]. The ability of computers to store large amounts of data without forgetting even one of the information stored in combination with practicality in asking information sources directly compared to finding information themselves and the learning abilities that it has causes chatbot is a reliable customer service

Example conversation with chatbot:
User: what's your name?
Bot: my name is Bot

Chatbot will answer according to what is available in the dataset, where the dataset is built by the author with the help of method calculations so that the system can analyze the similarity between questions must be answered with what kind of response.

b. Line

Line is an application used for sending messages (messenger / chat) for free on a smartphone device. However, Line applications can actually be referred to as social networking applications because there is a timeline feature as a place to share status, voice messages, videos, photos, contacts and location information. With the Line application we can also make voice calls and video calls in real time and for free. Line is provided on all smartphone devices and on all mobile operating systems: Android, iPhone / iOS, Nokia / Windows Phone, Blackberry and also PC (computers that have Mac OS or Windows system) [4]. The line on the chatbot system this time will act as an intermediary media or user interface that will relate directly to the user / community user. Where the chatbot is made will be in the form of bot chat in a group-room or room-chat itself.

c. Messaging API Line

Line messaging API allows data to be sent through the system application server bot with the Line platform. When the system bot sends a system bot message, a webhook will be triggered and the Line platform will send a request to the URL of the webhook system bot. The system bot server will then send a request to the Line platform to respond to the user. Requests are sent via HTTPS in JSON format. Here is the flow of the Messaging Fire Line system on a bot system: [4]

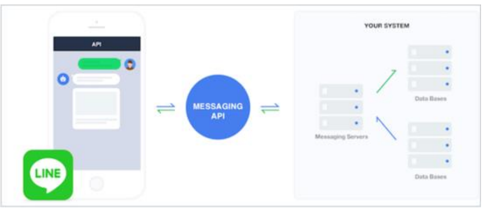

Figure 1 Flow of Messaging API Line

d. Natural Languange Processing

Natural Language Processing (NLP) is a branch of AI that focuses on processing natural language. Natural language is a language that is commonly used by humans in communicating with each other. The language received by the computer needs to be processed and understood in advance so that the intent of the user can be understood properly by the computer.

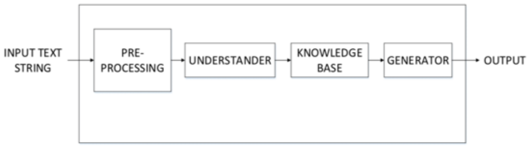

Figure 2 Flowchart NLP

There are various applied applications from NLP. Among them are Chatbot (an application that allows users to communicate with a computer), Stemming or Lemmatization (cutting words in a particular language into the basic form of recognition of the function of each word in a sentence), Summarization (summary of reading), Translation Tools (translating language) and other applications that allow computers to be able to understand language instructions inputted by users [1].

TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) method is a method to calculate weight of the relationship of a word (term) to a document.

1. Formula to calculate TF (2.1)

$$\boldsymbol{tf} = \boldsymbol{tf_{ij}} \ \textbf{(2.1)}$$

With tf is the term frequency, and $tfij$ is the number of occurrences of term ti in the document dj, Term frequency (tf) is calculated by counting the number of occurrences of the term ti in the document dj.

2. Formula to calculate idf (2.2)

$$\boldsymbol{idf_i} = \boldsymbol{log}\ \mathrm{N}/\boldsymbol{df}\textbf{(2.2)}$$

$idfi$ is inverse document frequency, N is the number of documents retrieved by the system, and $idfi$ is the number of documents in the collection where term ti appears in it.

3. Formula to calculate TF IDF (2.3)

$$\boldsymbol{W_{ij}} = \boldsymbol{tf_i} \times \boldsymbol{log}\ \textbf{(D/}\boldsymbol{df}\textbf{) (2.3)}$$

With Wij is the document weight, N is the number of documents retrieved by the system, $tfij$ is the number of occurrences of the term ti in the documentdj, and dfi is the number of documents in the collection where term ti appears in it. Document weight (Wij) is calculated for obtaining a weight resulting from the multiplication or a combination of term frequency ($tfij$) and Inverse Document Frequency (dfi) [8].

e. N-Gram

Extraction will be based on the N-gram division algorithm. Here N means the value of the word to be considered as a whole to relate its metadata. For example, for the phrase "the cow jumps over the moon". If N = 2 (known as bigrams), then n-grams will be:

- the duck
- duck walks
- walks over
- over the
- the river

If X = number of words in sentence (K), the number of ngrams for sentence K will be:

$$\textbf{N grams}_{\mathbf{K}} = \mathbf{X - (N\text{-}1)}$$

f. Cosine Similarity

Calculation of cosine similarity using the equation

$$\mathrm{Sim}\ (\mathrm{q,d_j}) = \frac{q.dj}{|q|\times|dj|} = \frac{\sum_{i=1}^{t} W_{iq} \times W_{ij}}{\sqrt{\quad}\ \sqrt{\quad}}$$

Similarity between query and document or Sim (q, dj) is directly proportional to the number of query weights (q) multiplied by the weight of the document (dj) and inversely proportional to the root of the sum of squares q (| q |) multiplied by the root of the square of the document (| dj |) . Similarity calculation results in the weight of the document that is close to 1 or produces a document weight that is greater than the value generated from the calculation of the inner product [5].

3. Implementation

System functional testing is needed in testing the performance of the system that has been built. This test is done by running every feature in the application and seeing the results as expected

- Question-Answering

Dataset :

1. Alamat rumah dokter Julie Kun Widjajano
2. Layanan Pembayaran rs Panti Nirmala
3. Alamat rumah dokter Saiful Burhan

Query Question:

alamat rumah dokter Saiful Burhan

(*Unigrams*)

- alamat [1]
- rumah [2]
- dokter [3]
- Saiful [4]
- Burhan [5]

(*Bigrams*)

- alamat rumah [1]
- rumah dokter [2]
- dokter Saiful [3]
- Saiful Burhan [4]

(*Trigrams*)

- alamat rumah dokter [1]
- dokter Saiful Burhan [2]

In the example above there is no space, if there is a space then the space will use the character "_" in front or at the end of the word. The use of N-gram (N) with Unigram (N = 1), Bigram (N = 2), and Trigram (N = 3), the number of sentences in X, the results of the formula in Ngrams as the repetition word limit, produce the following formula

$$Ngrams_k = X - (N - 1) \tag{3.1}$$

(*Unigrams*)

- X = 5; N=1
- Ngrams = 5

(*Bigrams*)

- X = 5; N=2
- Ngrams = 4

(*Trigrams*)

- X = 5; N=3
- Ngrams = 3

In this system n-gram is used as a compression medium or narrows the data space only the relevant ones are taken from the database that will be processed by the system. The above case example from the third calculation of N-Grams from the dataset finds the same relevant data, as follows:

Relevant data:

1. Alamat rumah dokter Julie Kun Widjajano
2. Alamat rumah dokter Saiful Burhan

After the weight (W) of each document is known, the sorting / sorting process is carried out where the greater the W value, the greater the level of similarity of the document to the keyword, and vice versa. Examples of implementations of Tf-Idf after the N-Gram process are as follows:

Relevant Data:

1. Alamat rumah dokter Julie Kun Widjajano
2. Alamat rumah dokter Saiful Burhan

The following is the TF-IDF calculation table:

term	kk	d1	d2	d3	df	n/df	log(n/df)
alamat	0	1	1	0	2	1	0
praktik	0	1	0	0	1	2	0,30103
rumah	0	1	1	0	2	1	0

term	kk	d1	d2	d3	df	n/df	log(n/df)
alamat	0	1	1	0	2	1	0
praktik	0	1	0	0	1	2	0,30103
rumah	0	1	1	0	2	1	0
dokter	1	1	1	1	4	0,5	-0,30103
julie	0	1	0	0	1	2	0,30103
kun	0	1	0	0	1	2	0,30103
widjajano	0	1	0	0	1	2	0,30103
saiful	1	0	1	0	2	1	0
burhan	1	0	1	0	2	1	0

W		
kk	d1	d2
0	1	0
0	2	0
0	1	0
-0,30103	0,5	-0,30103
0	2	0
0	2	0
0	2	0
0	0	0
0	0	0
-0,30103	12,5	-0,30103

Weight for (W) d1 = 12.5

Weight for (W) d2 = -0,3010

Then it is calculated using the Cosine Similarity method which has the following formula:

$$Sim(q, d_j) = \frac{q, d_j}{|q| \times |d_j|} = \frac{\sum_{i=1}^{t} W_{iq} \times W_{ij}}{\sqrt{\sum_{i=1}^{t}(W_{iq})^2} \times \sqrt{\sum_{i=1}^{t}(W_{ij})^2}}$$

Similarity between query and document or Sim (q, dj) is directly proportional to the number of query weights (q) multiplied by the weight of the document (dj) and inversely proportional to the root of the sum of squares q (| q |) multiplied by the root of the square of the document (| dj |) . So that the table is obtained as follows:

Similarity (a)	
d1	d2
0	0
0	0
0	0
-0,150515	0,09061906
0	0
0	0
0	0
0	0
0	0
-0,150515	0,09061906

The above calculates the resemblance of a vector with the result of the weight (w) of the keyword multiplied by the weight (w) of each document. The yellow column is the number of each document.

Vector length (b)		
kk	d1	d2
0	1	0
0	4	0
0	1	0
0,09061906	0,25	0,09061906
0	4	0
0	4	0
0	4	0
0	0	0
0	0	0
0,30103	4,5	0,30103

The above calculates the vector length of the keyword and each document, where the result of the weight (w) of the keyword and each document is squared, then summed down, and the result of the sum is squared each to calculate the cosine value of the document (in the colored column yellow).

Cosine Similarity	
d1	-0,1111111
d2	1

The above calculates the cosine value of each document, which in the vector similarity table, which is divided by the results of the length of the query vector query table is multiplied by the result of the squared number (each document) in the vector length. From the calculation results in Table Cosine Similarity (3) it can be seen that the document d2 has the highest level of similarity to the keyword.

3.1. Testing of N-Gram Response

Testing of the ngram which will later be used for the system. The results of testing on this system are shown in the following table.

number	Data
1	Dimana alamat praktik dokter gatot waluyo
2	List dokter spesialis bedah plastic
3	Berapa nomor telepon persada hospital
4	Layanan pembayaran apa yang ada di rs panti waluya
5	Berapa jumlah tempat tidur di rsi malang

The data in the table above will be used in testing the system response to questions from users. Testing is done on N-Gram (Unigram, Bigrams, Trigrams) so that the results of the system response are the fastest in answering questions from users. The test results are shown in the following table.

Number	N-Gram (second)		
	Unigram	Bigrams	Trigrams
1	0,99	0,72	0,68
2	0,69	0,61	0,58
3	0,04	0,002	0,01
4	0,09	0,004	0,038
5	0,37	0,0116	0,0104
Rata^2	0,436	0,28	0,26

From the results of testing the system response to the questions from the user obtained by the number of seconds. The smaller the nominal seconds on the results of the system response to questions from users, then it can be said that the system has responded quickly and accordingly. On the average in the test results above it can be concluded that by using Trigrams the speed of the system in answering is faster than Unigram and Bigrams

4. Conclusion

From the results of research and design that has been made, it can be concluded that Question-Answering in the form of Chatbot using N-Gram, TF-IDF and Cosine Similarity can communicate and convey information. This system uses the N-Gram method, based on tests that have been carried out from several test data using unigram, bigrams, and trigrams get faster execution time using trigrams

Acknowledgement

The author would like to thank to the Malang State Polytechnic who have supported this research project.

References

[1] Bayu Setiaji. 2016. “Chatbot Using A Knowledge in Database Human-to-Machine Conversation Modeling”. International Conference on Intelligent Systems, Modelling and Simulation. https://doi.org/10.1109/ISMS.2016.53

[2] Nirmala Shinde, 2018. “Chatbot using TensorFlow for small Businesses”. Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018).

[3] Bhavika R. Ranoliya. 2017. “Chatbot for University Related FAQs”. IEEE

[4] Varvara Logacheva. 2018. “A Dataset of Topic-Oriented Human-to-Chatbot Dialogues”. Bayan Abu Shawar. 2011. “A Chatbot as a Natural Web Interface to Arabic Web QA”. International Journal of Emerging Technologies in Learning. Vol. 6, No. 1. http://dx.doi.org/10.3991/ijet.v6i1.1502

[5] Aniket Dole. 2015. “Intelligent Chat Bot for Banking System”. International Journal of Emerging Technologies in Learning. Volume 4, Issue 5(2).

[6] Shunichi Ishihara. 2014. “A Comparative Study of Likelihood Ratio Based Forensic Text Comparison Procedures”. Fifth Cybercrime and Trustworthy Computing Conference. https://doi.org/10.1109/CTC.2014.9

[7] Ranjeet Kumar. 2014. “A Trigram Word Selection Methodology to Detect Textual Similarity with Comparative Analysis of Similar Techniques”. Fourth International Conference on Communication Systems and Network Technologies. https://doi.org/10.1109/CSNT.2014.82

[8] Sixing Wu. 2018. “A Fully Character-level Encoder-Decoder Model for Neural Responding Conversation”. IEEE International Conference on Computer Software & Applications.