International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

Recommender System for Topic Articles Based on Forum Trending using Multilayer Perceptron

Sri Hesti Mahanani, Tuga Mauritsius

Abstract*: SehatQ is a portal and application that helps manage personal and family health. One of SehatQ's services is providing information and directories in the form of articles. To improve relations with web visitors, SehatQ also provides services in the form of discussion forums. The forum actually contains a variety of topics and changes very quickly over time, so to identify a topic from a collection of forums is very difficult and time-consuming if done manually by humans. But unfortunately the SehatQ editorial team has limited time and human resources in sorting out information sourced from the SehatQ forum to draw conclusions as a topic in the article. This research will offer a solution in analyzing Topic modeling using text mining with the Multilayer perceptron algorithm to provide trending information on the topics most frequently discussed at the forum at a certain time.

Keywords* : Data minning, tf-idf, multilayer perceptron, dv-ngram, n-gram, topik modeling, big data, bahasa

I. INTRODUCTION

SehatQ is a portal and application that helps manage personal and family health. One of the features found in SehatQ is that there are health articles and discussion forums provided on the SehatQ platform. Every day SehatQ's editorial team produces various health articles that are ready to be published. The article will appear on the website www.sehatq.com. with the article can increase the number of website visitors. But there is no doubt that as a medical expert SehatQ has a team to review every content that will be published on SehatQ to ensure that everything that comes out of the portal is the most valid and best health content that can be found in Indonesia.

Manuscript received on February 10, 2020.
Revised Manuscript received on February 20, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author
Sri Hesti Mahanani, Department Information System Management, Nusantara University, Jakarta, Indonesia 11480
Tuga Mauritsius, Department Information System Management, Nusantara University, Jakarta, Indonesia 11480

In addition to articles, a feature found on the SehatQ website is a discussion forum. Discussion forums are a feature provided by SehatQ to accommodate questions from website visitors. In the forum visitors can interact with both the doctor and with other visitors. Existing discussion forums are quite active, as evidenced from 1 January 2019 to 25 January 2020 recorded 10707 records of SehatQ forums submitted by visitors. The forum actually contains a variety of topics and changes very quickly over time, so to identify a topic from a collection of forums is very difficult and time-consuming if done manually by humans. In fact, the collection of posts in the forum is a source of data that has the potential to provide information on what is happening in the healthy portal Q. But unfortunately the SehatQ editorial team has limited time and human resources in sorting out information sourced from the SehatQ forum to draw conclusions as a topic in the article.

In the forum, both questions and answers from readers contain information that often contains many symbols and non-standard word elements. This makes it difficult for the editorial team to interpret manually thousands of texts in the existing forum. Text mining is a method for analyzing large amounts of text data. Text mining techniques are needed to find an interesting pattern in finding trends based on text in the healthy forum Q.

Based on the above problems, this research will offer a solution in analyzing Topic modeling using text mining with the Multilayer perceptron algorithm to provide trending information on the topics most frequently discussed at the forum at a certain time. The results of the trending topic can be used as a reference by the editorial team in determining the topic of the next article. [12] The company can maintain direct relationships and connections between the forum and the article, using the data.

II. LITERATURE REVIEW

A. Text Mining

Text mining is the process of analysis data in the form of text and the source of data is obtained from documents [2]. The concept of text mining usually use in the classification of textual documents where the documents will be classified according to the topic of the document. With text mining an article can be known the types of categories through the words contained in the article. The contents of the article are analyzed and matched in the keyword of database that has been predetermined. So, text mining can help to group a word in the document with a short time. The stages of analyze text mining are collecting data and extracting the features that will be used [2].

According to [17], based on data processing with the Data or text Mining technique, where it gets several algorithms that are used to produce a comparison between actual predictions with actual conditions it can be designed Decision Support System or system recommendations.

B. Processing Data

Data Processing intend to get a dataset that can be processed quickly and produce the suitable conclusions. One of the process of data processing is feature selection. There are several stages in the selection of features, including : Tokenizing, is the stage of cutting the input string to separate sentences into words. And the Stopword, the process of removing words that are not important in the text is carried out. This process need a dictionary of words that can store bag of words to be removed. The last one is Stemming, this stage that carries out the process of turning derivative words into basic words [2].

Processing data stages according to [6] explain the definition about Tokenizing is the stage of separating sentences into words, deleting special characters and punctuation.

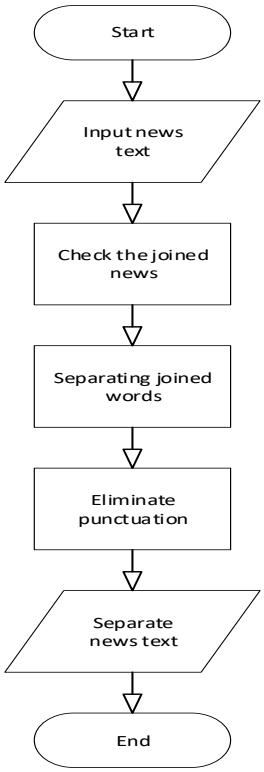

Fig. 1.Tokenizing Processes Phase [6]

According to [5] define the meaning of stop word. At this stage it is a continuation of the tokenizing process by taking important words with the stop list algorithm (eliminate words that are not important) and word list (save important words). The stages of stopword process can be summarize :

1) Compare the results of tokenizing with stopword data
2) Check the token data with the stopword data
3) If there is the same data it will be deleted
4) If the data are not the same then the data will be displayed because it is an important word.

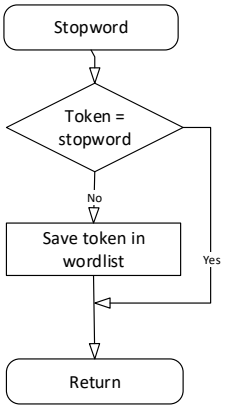

Fig. 2.Phase Stopword Processes [5]

The definition of Stemming according to [4], this stage is the process of eliminating derivative words that still have prefixes into basic words. In the stemming process we use the Nazief and Adriani algorithm [4] with the following step :

1) Words that have not been stemmed are searched from in the dictionary. When the word is found immediately, it means the word is a basic word. The word will be returned and the algorithm terminated.
2) Inflection Suffixes ("-lah", "-kah", "-ku","-mu", or "-nya") will be remove. When the word contains particle ("-lah", "-kah", "-tah", or "-pun") this step will be repeated again to remove a Possesive Pronouns ("-ku", "-mu", atau "- nya"), if any.
3) Delete the Derivation Suffixes ("-i", "-an",atau "-kan"). If the word is found in the dictionary, the algorithm stops. In case the word not found then continue to step 3a. If "-an", was removed and the last letter of the word is "-k", then "-k" will remove too. When the word is found in the dictionary, the algorithm stops. In case the word not found then continue to step 3b. The deleted suffix ("-i", "-an",atau "-kan") returned, continue to step 4.
4) Delete the Derivation Prefix ("di-","ke-", "se-", "me-", "be-", "pe-", "te") with the maximum iteration is three times.

   a. The step 4 will stop in case :

   • Forbidden combinations of prefixes and suffix occur as in Table 1 below.

Table- 1: Combination of suffix prefix that is not allowed

Prefix	Suffix that is not permitted
be-	-i
di-	-an
ke-	-i, -kan
me-	-an
se-	-i, -kan

• Three prefixes have been removed

b. The prefix type is determined through the following steps :

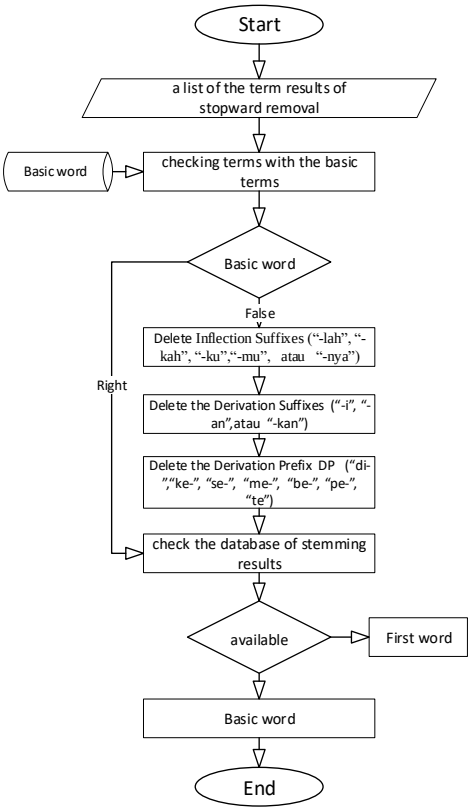

Fig. 3.Phase of Stemming Processes[4]

1) When the prefix : "di-", "ke-",atau "se-", then the prefix type in a row is "di-", "ke-", or "se-".
2) When the prefix "te-", "me-","be-", atau "pe-", then additional process is needed to determine the type of prefix.
3) Find the word that has been omitted in the dictionary. When the word not found, then step 4 is repeated again. In case the word found, the whole process stops.
4) After there are no more affixes left, then the algorithm is stopped and the base word is searched in the dictionary, when the base word is found it means the algorithm is successful but when the base word is not found in the dictionary, then recoding is performed.
5) All of the steps have been taken but the root word is not found in the dictionary as well so this algorithm returns the original word before stemming.

C. Term Frequency Inverse Document Frequency (TF-IDF)

Data that has gone through the preprocessing stage must be numeric. To convert the data into numeric we use the TF-IDF weighting method. Term Frequency Invers Document Frequency (TF-IDF) method is used to define the text (term) related with the document from weighting each word. TF-IDF method combine two concept which is frequency a word in a document and inverse document frequency in a word [3]. In calculating TF-IDF using weighting method, first calculated TF value with the weight of each word is 1. While IDF value formulated in equal $IDF(word) = log\frac{td}{df}$. IDF(word is the IDF value of each) the word to be searched for, td is total of document available, df number of words in all documents.

Term Frequency Invers Document Frequency (TF-IDF) method is a method used to determined how far the word (term) related to a document with every weight of each word. In text preprocessing, term weighting is the most important stages. This stage is done with the aim to give a value or weight to the terms contained in a document.

The weight given to a term depends on the method used to weight it. In text mining, there are several types of weighting methods which include TF, TF-IDF and WIDF. The output is compared to the performance of text categorization. There are parameters used as benchmarks for comparing performance text categorization, which are precision, recall and f-measure. To test the weighting result, we can used tools of data classification namely Weka, with Naïve Bayes and Naïve Bayes Updateable as that methode of classified. Based on test result, found that the WIDF weighting method has better performance than the other weighting methods (TF dan TF-IDF). Generally, WIDF outperform other methods in some of the tests conducted.

The frequency with a term appears in a document and normalizes it throughout the entire document, make this method better than the others [9]. Two new term weighting schemes is SQRT_TF-IGM and TF-IGM generated from the moment of reserve gravity are proposed to improve the weight behaviour TF-IGM [1].

D. Artificial Neural Network

Multilayer perceptron (MLP), also known as feedforward neural network. The term "ANN" hereinafter refers to MLP and more complex architecture[9]. Looking at the illustration in Figure 4, the multilayer perceptron literally has several layers. In general there are three layers: input, hidden, and output layer. The input layer accepts the input (without carrying out any operation), then the input value (without passing to the activation function) is given to hidden units. (In hidden units, the input is processed and a calculation of the activation function results for each neuron, then the results are given to the next layer. Output from the input layer will be received as input for the hidden layer. Likewise, the hidden layer will send the results to the output layer[10].

[11]This activity is called feed forward. The same applies to artificial neural networks with more than three layers. Neuron parameters can be optimized using the gradient-based optimization method. MLP is a combination of many non-linear functions. This combination of many non-linear functions is more powerful than a single perceptron. As shown in Figure 4, each neuron is connected to all the neurons in the next layer. This configuration is referred to as fully connected. MLP generally uses a fully connected configuration.

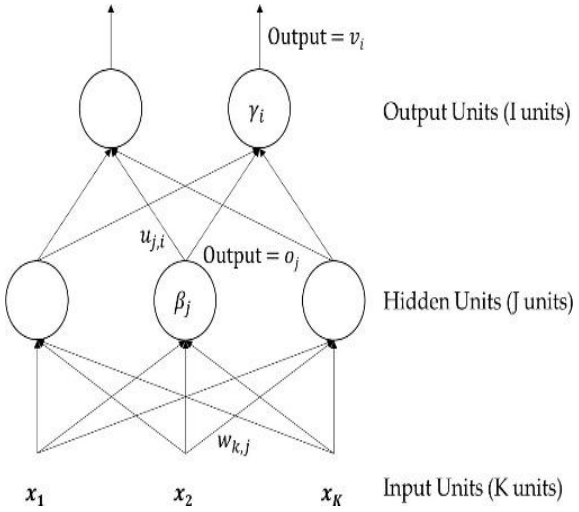

Fig. 4. Multilayer Perceptron 2 [11]

To practice MLP, the algorithm that is generally used is back propagation. The meaning of the word backpropagation is difficult to translate into Indonesian. We update the parameters (synapse weights) gradually (from the output to the input layer, because it is called backpropagation) based on error / loss (output compared to desired output). The point is to correct the synapse weight from the output layer to the layer life, then the error is propagated to the previous layer. That is, changes in synapse weight in a layer are affected by changes in synapse weight in the subsequent layers. Backpropagation is a gradient-based optimization method applied to ANN.

(2) Hidden to Output

$$v_i = \sigma\left(\sum_{j=1}^{J} o_j u_{j,i} + \gamma_i\right)$$

(3) Output to Hidden

$$\delta_i = (y_i - v_i)v_i(1 - v_i)$$
$$\Delta u_{j,i} = -\eta(t)\delta_i o_j$$
$$\Delta\gamma_i = -\eta(t)\delta_i$$

(1) Input to Hidden Layer

$$o_j = \sigma\left(\sum_{k=1}^{K} x_k w_{k,j} + \beta_j\right)$$

(4) Hidden to Input

$$\varphi_j = \sum_{i=1}^{I} \delta_i u_{j,i} o_j(1 - o_j)$$
$$\Delta w_{k,j} = -\eta(t)\varphi_j x_k$$
$$\Delta\beta_j = -\eta(t)\varphi_j$$

Fig.5. MLP using backpropagation [11]

III. METHODOLOGY

The objective of this paper is analyzing Topic modeling using text mining with the Multilayer perceptron algorithm to provide trending information on the topics most frequently discussed at the forum at a certain time. System using a text mining method that can provide topic modeling based on trending topic from forum discussion. The application of the text mining method consists of several stages, including: Stages of Data Processing (Tokenizing, Stopword and Stemming), the application of the TF IDF method for weighting each word and applying the cosine similarity model to measure the similarity between a document and a query. The method of text mining is a solution to provide topic modeling that are appropriate for each forum discussion.

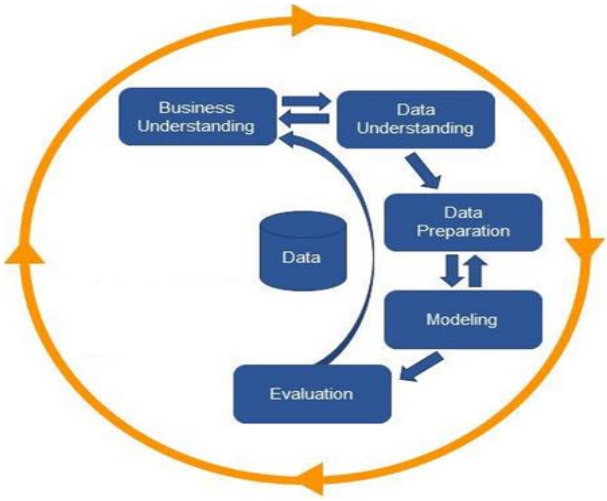

Fig.6. The Methodology of Text Mining [8]

- **Business Understanding**

In the first stage we look at problems with companies, especially editors who have difficulty filtering out information contained in sehatQ forum discussion to draw conclusions into a topic that will be discussed in the article because both questions and answers from readers contain information that often contains many symbols and elements of the word no raw. This makes it difficult for the editorial team to interpret manually thousands of texts in the existing forum.

- **Data Understanding.**

At this stage discussing the data used for the data mining process. A data set taken from the healthy SQL database in the form of forum data, article data, and tag data will be used in this study. Data posted by visitors in the form of questions or answers will be stored in the forum data. While article data is data created by a team of editors to explain a particular topic in a series of articles. Then for the tag data itself is a set of topics or keywords that have been determined by the editorial team that serves as a marker on an article. If there are articles that discuss the same problem then the same tag can be given.

- **Data Preparation**

We prepare the data for building the model. We combine the data from article, forum and list tag summary to become one data, And we used stemming to stem the word from the article, forum and list tag and we remove the stopword from the article using sastrawi library in python. To ensure we can collect the feature from the article. After we process the token from that with dv-ngram to get more valid phrases.

- **Modeling**

We build model using MLP to learning modeling for generate topic. We use IDF to count the frequent topic that appears.

- Evaluation

We evaluate the topic in accuracy and loss method to find out the quality of the topics produced.

- Deployment

We implement a model into the system to search for trending topics and recommend topics based on trending forum topics generated.

IV. RESULT

4. 1 Data Preparation

The following is a summary of the process in the preparation phase in this study:

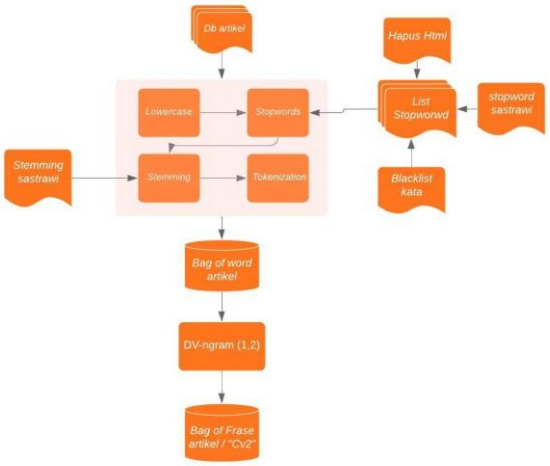

Fig.7. Diagram flow Data Preparation

From Figure 4.11 we can see the order to clean up existing data. Starting from the initial data selection then after the initial data collected will be converted to lowercase, then proceed to eliminate the data that is in the stopword list where the stopword data list is the deletion of data containing html code, literary stopword data list, and word blacklist from the internal team. After the deletion phase, the stemming phase will be continued, ie the initial erasure and the unfolding of words that are not needed. This is so that the tokenizing process can be cleaner. The process will produce a token or commonly called bag of words. In this study it is not enough to just one token, but researchers are trying to form tokens using dv-2-gram to get a more tangible collection of phrases.

4.1.1 Initial Dataset

Raw data in the .sql format that is processed in the MySQL database management system (DBMS) is converted to the .xls format to make it easier to process. The data taken is forum data, article data and tags which are a collection of topics in general contained in healthy articles Q. The following pictures are for article and forum data:

	A	B	C	D	E
1	articles_id	articles_title	articles_content	articles_published_date	
2	4820	Operasi Caesar	<h2>Apa itu Op	28/01/2020 10.12.57	
3	4818	Efek Samping IUD, Per	<p><span style="font-w	28/01/2020 09.26.17	
4	4816	Pola Makan Jadi Kunci	<p><span style="font-w	27/01/2020 19.30.00	
5	4814	Bayi Baru Lahir Sering	<p><span style="font-w	28/01/2020 09.00.00	
6	4813	Penyebab Novel Coron	<p>Wabah <a href="http	28/01/2020 08.30.00	
7	4811	Manfaat Sinar X dalam	<p>Namanya terbaca se	27/01/2020 18.15.00	
8	4810	Bye-bye Bad Hair Day!	<p>Rambut yang menge	27/01/2020 18.00.00	
9	4809	Mengenal Likopen, Si F	<p>Buah dan sayuran b	27/01/2020 18.36.00	
10	4808	Cek Tagihan BPJS den	<p><span style="font-w	27/01/2020 16.40.00	
11	4807	Perawatan Wajah untu	<p>Masa-masa kehamil	27/01/2020 19.00.00	
12	4806	Apa Saja Bagian-Bagia	<p>Pencernaan tidak ha	27/01/2020 16.16.24	
13	4805	Beragam Cara Menghil	<p><span style="font-w	27/01/2020 15.45.00	

Fig.8. Sample data article

	A	B	C	D
1	askdoctors_id	askdoctors_title	askdoctors_question	created_at
2	10737	Cyclo progynova	Dok say konsumsi cyclo progynova ta	28/01/2020 11.51.17
3	10736	Testing	Testing	28/01/2020 09.26.02
4	10735	kaku pada lutut	pagi dok, saya mau tanya jika setelah	28/01/2020 08.12.28
5	10734	tanya obat penyakit kus	selamat pagi dok, saya mau tanya kal	28/01/2020 07.58.29
6	10733	Perut gendut apa hamil?	Dok saya tidak berhubungan itim tpi ha	28/01/2020 00.18.30
7	10732	Ukuran janin	Dok usia kandungan saya 21minggu ja	27/01/2020 22.43.59
8	10731	Berat janin	Dok usia kandungan saya 21 minggu E	27/01/2020 22.40.12
9	10730	Kenapa Selangkangan S	Dok Saya mau Tanya, bberapa tahun t	27/01/2020 21.19.38
10	10729	Apa saya harus konsulta	dok, sy smg merasa sedih & tertekan	27/01/2020 20.48.38
11	10728	Apa penyebab sakit ping	Dok mau tanya sakit pinggang sebelah	27/01/2020 15.52.39
12	10727	Berhubungan Intim Kem	Dokter saya mau tanya, saya selamar	27/01/2020 15.34.54
13	10726	Apa penyebab gatal dan	Selamat siang dok, Dok, saya ada ke	27/01/2020 14.53.56
14	10725	Bagaimana posisi saat t	Salam kenal dokter, saya mau tanya.	27/01/2020 14.53.18
15	10724	batu ginjal, radiologi / im	Dear dokter, mau tanya tentang hasil r	27/01/2020 14.20.54
16	10723	Haid tidak teratur seben	hai dokter. saya mens sering sekali cu	27/01/2020 13.00.24
17	10722	Melakukan hubungan se	Dok..Apakah melakukan hubungan sel	27/01/2020 09.49.35
18	10721	Apa penyebab sakit pay	Payudara sebelah kanan terasa sakit s	27/01/2020 06.51.03

Fig. 9. Sample data Forum

	A	B
1	articletags	articletags_name
2	1	bpjs kesehatan
3	2	asuransi swasta
4	3	penyakit paru-paru
5	4	penyakit
6	5	operasi
7	6	batu ginjal
8	7	jus jeruk
9	8	makeup untuk kulit berminyak
10	9	operasi caesar
11	10	kulit berminyak
12	11	makeup
13	12	kulit sehat
14	13	diabetes
15	14	komplikasi diabetes
16	15	perawatan kulit
17	16	klinik tumbuh kembang anak

Fig. 10. Sample data tag

4.1.2 Data Cleansing

articles_id	title&content	published	title&content_stem	tags	tagsInArticle
4820	operasi caesar apa itu	28/01/2020 10.12.57	operasi caesar apa operasi	['operasi caesar', 'lahir', 'siap lahir', 'caesar']	['anestesi' 'batuk' 'bayi' 'caesar' 'cedera' ...]
4818	efek samping iud adalah	28/01/2020 09.26.17	efek samping iud timbang pilih ...	['sakit wanita', 'sehat wanita', 'sakit rahim', 'kontrasepsi']	['cuci tangan' 'darah' 'demam' 'haid' 'hamil' 'ibu susu' ...]
4816	pola makan jadi kunci ...	27/01/2020 19.30.00	pola makan jadi kunci 8 cara bakar lemak ...	['protein', 'turun berat badan', 'serat', 'benjol lemak', 'perut', 'pola hidup sehat']	['alkohol' 'diabetes' 'hormon' 'hormon kortisol' ...]

Fig. 11. Data Result from text preprocessing article

Fig. is a sample table of data processed by preprocessing text on article and tag data combined into one. The articles_id column is the article id, the title & content column contains the confusion between the article title and the article content, publish is the date of the published article, title & content_stem is a combination of title and content data obtained from the stemming process, tags are a general topic in articles that have been determined by the editor , and taginarticle is a collection of tags or topics contained in the title & content_stem column that has passed the dv-2 gram stage. So clearly seen tokon contained not only one katu but there are consisting of 2 words or bigram.

The process of cleaning the data is also carried out on the askdoctor or forum data set. For forum data sets, a file named create_database_forum.xlsx will be formed. The data will be saved as the table below:

created_at	title&question&answer	title&question&answer_stem
28/01/2020 10.12.57	cyclo progynova dok konsumsi cyclo progynova tapi saya juga mengkonsumsi obat ….	cyclo progynova dok konsumsi cyclo progynova konsumsi obat …..
27/01/2020 10.12.57	kaku pada lutut pagi dok saya mau tanya jika setelah jongkok agak lama ….	kaku lutut pagi dok tanya jongkok lama …
26/01/2020 10.12.57	ukuran janin dok usia kandungan saya 21minggu janin saya mempunyai ukuran …	ukur janin dok usia kandung 21minggu janin punya ukur …

Fig.12. Data Result from text preprocessing Forum

Fig is a sample table of data processed by preprocessing text in forum data. The created_at column is the date the forum was created by visitors, the title & question & answer column contains a combination of article titles, questions and answers in one forum, the title & question & answer is a combination of title, question and answer data obtained from the stemming process.

4. 2 Methods for Topic Modeling

The following is a flow chart on the modeling topic in this study:

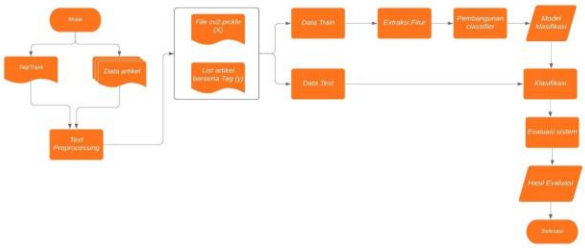

Fig.13. Flow diagram Model MLP

The picture above is an illustration of the process of forming a modeling topic using the MLP model. Starting from the preprocessing text on article and tag data. First for preprocessing articles generate cv2 files in pickle format, this cv2 file will be used as input for the model to be created. Then preprocessing on the combined article and tag data will produce a list of article tags. This article tag list will be used as the value of y in the formation of the model. Then x and y data are divided for training data and testing data with a percentage of 67% for training data and 33% for testing data, the data will be taken randomly by the system.

The data selected as the data train will be forwarded to the feature extraction process, ie the model will create a formula in the data so that it can be used for the cluster builder process. Clusters here are a number of tags contained in the healthyQ database tag list. After that, a modeling topic will be formed into a cluster. For the data selected as test data, the cluster will be determined directly. After that it can be evaluated to determine the cluster of the data train results and test data by measuring the loss and accuracy in each row.

In this study the structure of the model used is where there are three layers: input layer, hidden layer, and output layer.

Table- 2: Structure model MLP

Model MLP	Keterangan
Input layer	Sesuai dengan jumlah dari hasil ekstraksi fitur berdasarkan hasil olah dv gram
Hidden layer	Satu hidden layer dengan jumlah 50 neuron
Output layer	Jumlah 1717 neuron
Epoch/iterasi	50
Fungsi aktivasi Hidden layer	relu
Fungsi aktivasi output	sigmoid
Dense/Hidden layer	100
Optimizer	rmsprop
loss	binary_crossentropy
metrics	accuracy

```
setArticle=set(cv2.get_feature_names())

cv2.dtype=np.int8
inputData=cv2.transform(create_database[
"title&content_stem"])

inputData=inputData.toarray()

create_database2=create_database.copy()
create_database2["vector"]=create_databa
se2.tags.map(lambda x:
get_vector_from_array(x))
dataArticle=pd.DataFrame(create_database
2["vector"].to_list(),columns=TagsRemove
Stopword.Tags,index=create_database2.art
icles_id)
numDataArticle=dataArticle.sum()
idf_article=np.log(dataArticle.shape[0]/
numDataArticle)
X=inputData
y=dataArticle.to_numpy()
# fix random seed for reproducibility
seed=7
np.random.seed(seed)
# split into 67% for train and 33% for test
X_train, X_test, y_train, y_test =
train_test_split(X,
y, test_size=0.33,
random_state=seed)
```

```
# create model
model = Sequential()
model.add(Dense(100,
input_dim=X_train.shape[1],use_bias=Fals
e, activation='relu'))
model.add(Dense(y_train.shape[1],
input_dim=50, use_bias=False))
model.add(Activation('sigmoid'))
# Compile model
model.compile(loss='binary_crossentropy'
, optimizer='rmsprop',
metrics=['accuracy'])
# Fit the model
history=model.fit(X_train, y_train,
validation_data=(X_test,y_test),
epochs=50, batch_size=64)
```

Code.1. Source codemodel MLP

4.3 4.4 Evaluation of Topic Models

This evaluation is done by calculating the value of loss and accuracy on the model. To evaluate the model the writer divides between train data and test data with a ratio of 67% for train and 33% for test. Each data was chosen randomly to get the train and test data. The following results from the topic modeling process:

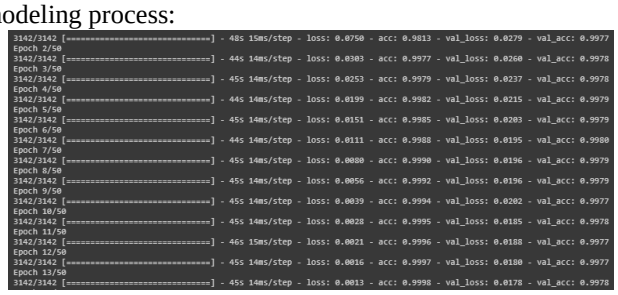

Fig.14. Running model

After the model is run it produces a process like the picture above. The figure provides information that from the iteration process as much as 50 times the system produces the greatest loss value is 0.750 and the accuracy value of 0.9813 each scale 1. If visualized in graphical form is as follows:

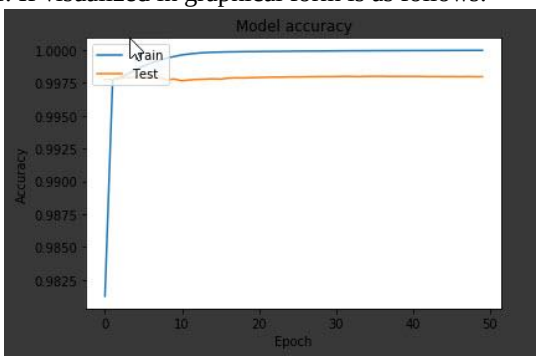

Fig.15. Comparison of accuracy

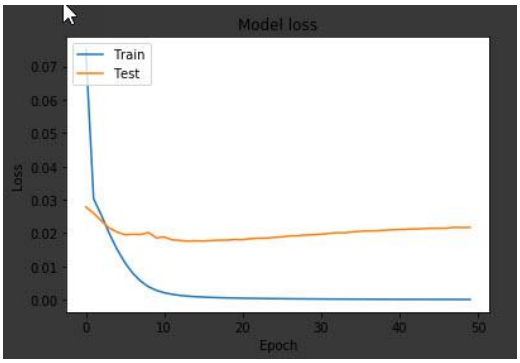

Fig.16. Comparison of Loss

From the results above it can be seen that the more experiments the value of aquaculture is increasing even stable. As for the loss, the more trials the loss value decreases and tends to be stable.

The loss function used in this study is binary_crossentropy. Cross-entropy will calculate a score that summarizes the average difference between the actual and predicted probability distributions for prediction class 1. The score is minimized and a good cross-entropy value is 0.

4.5 Deployment

4.5.1 Designing the topic trending system in the forum

The purpose of making this model is to provide the top 10 topics that are trending on the sehatQ Forum. The following is an overview of the process of the trending system topic created:

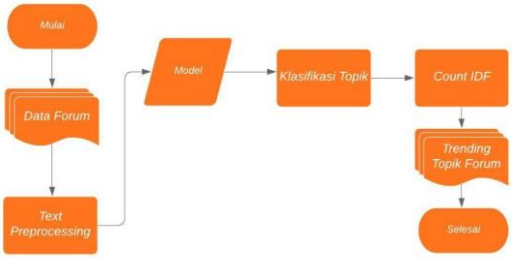

Fig.17. Alur sistem Trending topik forum

- After the data has been through a text forum and dv-ngram preprocessing of data stored in the file noHp.xlsx and will then use the process modeling topics will be formed clusters of topics each forum. Relationships Forum tags are stored in the file database_forum_tags.xlsx. Use the following code:

```
%%time
tagsForumAllPro=[]
tagsForum=[]
create_database_forum2=create_database_f
orum[~create_database_forum["noHP"]]
create_database_forum2=create_database_f
orum2.drop("tagsInArticle",axis=1)
create_database_forum2=create_database_f
orum2.dropna()
#
create_database_forum2["tagsForumAllPro"
]=""
create_database_forum2["tagsForum"]=""
for n in
tqdm(create_database_forum2.index):
# if n>3:
#   break
X=cv2.transform([create_database_forum2.
loc[n,"title&question&answer_stem"]]).to
array()
A=model.predict_proba(np.array([X[0]]))
B=sorted(A[0,idx_tags],reverse=True)[5]
if sum(A[0,idx_tags])>0:
C=[setTags[i] for i in
np.where(A[0,idx_tags]>B)[0] if
setTags[i] not in
remove_tags_from_trending]
# C=[setTags[i] for i in np.where(A>B)[1]]
# print(C)
else:
C=[]
# print(C)
```

```
#
create_database_forum2.loc[n,"tagsForumA
llPro"]=A[0]
#
create_database_forum2.loc[n,"tagsForum"
]=C
tagsForumAllPro+=[A[0]]
tagsForum+=[C]
```

Code.2. Code trending topic forum

• After the process of forming a tag on the forum already berlajan writer can determine the value of TF-IDF on each phrase in represntasikan in the form of Vector Space Model. TF-IDF consists of Term Frequency (TF) and Inverse Document Frequency (IDF). Where TF-IDF utilize the data tag on the forum that have been collected from the previous process. Where the formula used in the TF is as follows:

TF (t) = (Number of times term t Appears in a document) / (Total number of terms in the document).

And the formula used in the IDF are as follows:

IDF (t) = log_e (Total number of documents / Number of documents with the term t in it).

So the formula for determining the TF-IDF is:

TF-IDF = TF (t) x IDF (t)

In this case researchers will use only the IDF value because it can provide information about the intensity level of the emergence of a phrase. The lower the frequency level idf signifies the emergence of a phrase.

Here's the code that is used to determine the idf of each tag forum:

```
%%time
databaseForum2=databaseForum.copy()
databaseForum2["vector"]=databaseForum2.
setTags.map(lambda x:
get_vector_from_array(x))
dataForum=pd.DataFrame(databaseForum2["v
ector"].to_list(),columns=TagsRemoveStop
word.Tags,index=databaseForum2.index)
numDataForum=dataForum.sum()
idf_forum=np.log(dataForum.shape[0]/numD
ataForum)
TagsRemoveStopword["idf_forum"]=idf_foru
m
```

Code.3. Code idf tag code determination forum

The results of this system is in the form of a top 10 list of topics that often appear on the forum sehatQ following picture:

```
[ ]  Tags_all_idf[Tags_all_idf.Tags_stem.isin(C)].sort_values(by='idf_forum')
```

	Tags_ori	Tags_stem	idf_forum	idf_article
508	hidup sehat	hidup sehat	1.83505	2.5236
759	kesehatan gigi	sehat gigi	2.73497	4.32605
444	gigi berlubang	gigi lubang	3.4127	4.98745
1402	sakit gigi	sakit gigi	3.63244	4.76431
766	kesehatan mulut	sehat mulut	3.92579	5.56282
450	gigi sensitif	gigi sensitif	4.01487	5.96828
1606	tips kesehatan	tips sehat	4.64193	4.95668
5	abses gigi	abses gigi	5.51276	6.66143
449	gigi putih	gigi putih	6.69838	6.66143
683	karang gigi	karang gigi	7.01683	7.35458

Fig.18. The results of the forum system Trending topics

4.5.1Designing the system on the topic of the article

The second form of implementation of the model that has been made is in the form of the system on the topic of the article. The system described the plot as shown below:

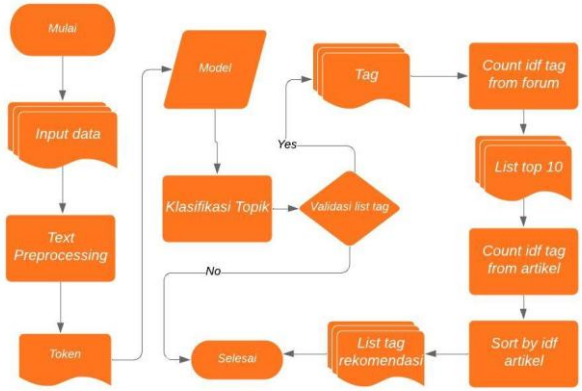

Fig.19. The process of system recommendations article topics

Starting from the user input data and then do the text preprocessing then will form a token. The token will be determined through the model to enter into any tag. Then going through the process of validation of the list of existing tags. If the tag is generated are included in the set list tag will count its idf value based on data from the tag forum, take the top 10 data only. Then from 10 data tags will look for its idf value of the data of the article. Results ahirnya be sorted based on the value of idf The top 10 take its course to be issued as a recommendation. The following code is used to determine the outcome of recommendations:

Fig.20. Output system on the topic of the article

From the figure above shows that the phrase or words input by the user is maintaining dental health. Recommendations generated topic is 'abses gigi', 'gigi lubang', 'gigi putih', 'gigi sensitif', 'hidup sehat', 'karang gigi', 'sehat gigi', 'sehat mulut', 'sakit gigi', 'tips sehat'.

V. CONCLUSION

Telah dilakukan eksperimen Topic modeling dengan metode MLP. Eksperimen yang dilakukan dengan model sequential dengan hidden layer menggunakan 100 neuron, fungsi aktivasi relu dan sigmoid, optimizer menggunakan rmsprop dan fungsi loss menggunakan binary crossentropy dapat mengahasilkan suatu model untuk menentukan suatu topik dengan baik. Telah dilakukan evalusi dari model topik yang telah dibuat dengan mengukur akurasi dan nilai loss. Dengan menggunakan data train 67% dan data test 33% di ambil secara acak dapat menghasilkan nilai akurasi 0.9813 dan nilai loss 0.0750 dengan skala terbesar adalah 1. Model yang dihasilkan dapat diimplementasi untuk sistem penentuan trending topik dengan menambahkan proses penghitungan Idf untuk menghasilkan output yang sesuai yaitu mencari nilai seringnya tingkat kemunculan sebuah tag.

REFERENCES

1. Chen, K. Z. (2016). Turning from TF-IDF to TF-IGM for term weighting in text classification. Expert Systems With Applications, 245–260.
2. Feldman, R. (2007). The text mining handbook: Advanced approaches in analyzing unstructured data. England: Cambridge.
3. Fitrianah, D. &. (2016). Audit Sistem Informasi/Teknologi Informasi Dengan Kerangka Kerja Cobit Untuk Evaluasi Manajemen Teknologi Informasi Di Universitas Xyz. Jurnal Sistem Informasi, 4(1), 37.
4. Guerreiro, J. &. (2018). Journal of Hospitality and Tourism Management How to predict explicit recommendations in online reviews using text mining and sentiment analysis. Journal of Hospitality and Tourism Management, 1-4.
5. Hapsari, R. K. (2015). STEMMING ARTIKEL BERBAHASA INDONESIA DENGAN Pendekatan Confix-Stripping. Prosiding Seminar Nasional Manajemen Teknologi XXII, 1-8.
6. Herwijayanti, B. R. (2017). Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDF dan Cosine Similarity. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 2, no. 1, 306-312.
7. Purnomo, J. A. (2010). Analisis perbandingan beberapa metode pembobotan kata terhadap performansi kategorisasi teks.
8. Chapman, et al., (2000). CRISP-DM 1.0: Step-by-step data mining guide. Computer Science.
9. Grace, Tika. (2019). Klasifikasi Topik Berita Berbahasa Indonesia menggunakan Multilayer Perceptron. e-Proceeding of Engineering : Vol.6
10. Akhmad, Rohim, et al. (2019).Convolution Neural Network (CNN) Untuk Pengklasifikasian Citra Makanan Tradisional. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer
11. Joshua, Foer. (2011). Moonwalking with Einstein: The Art and Science of Remembering Everything. Vol. Feedforward Neural Network. Allen Lane an imprint of penguin books
12. Mauritsius Tuga, et al, (2019). Bank Marketing Data Mining using CRISP-DM Approach.International Journal of Advanced Trends in Computer Science and Engineering
13. Franky.R, Tuga.M. (2019).Condition-base maintenance using data mining techniques on internet of things generated data.Journal of Theoretical and Applied Information Technology

AUTHORS PROFILE

Sri Hesti Mahanani Information System Management Department, BINUS Graduated Program – Master Information System Management Bina Nusantara University, Jakarta, Indonesia 11480
Sri.mahanani@binus.ac.id

Tuga Mauritsius Information System Management Department, BINUS Graduated Program – Master Information System Management Bina Nusantara University, Jakarta, Indonesia 11480
tmauritsus@binus.edu