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

Recommender System using Content Based Filtering for News Portal in Indonesia

Sri Hesti Mahanani, Valentinus, Dennis, Tuga Mauritsius

Abstract*: Nowadays there is much news on the internet. It makes the reader become information overload. The reader does not know the most important news for them. The digital era, especially in Indonesia, generated data in Bahasa very fast that referred to as big data. Data mining by process big data can collect the data insight that the reader already read. This paper proposes a new model to proceed with Bahasa news and use the TF-IDF method to collect the feature of the article. Cosine similarity from the news article used to rank the new unknown articles to recommend articles based on their preference. we can filtering the stream of information and highlight the most likely article they will read but based on their preference that we already collect implicitly from the article that they read it, it's a scroll depth of the article they read.Then we can serve the news more personalized from what they love to read.

***Keywords* : Data minning, tf-idf, cosine similarity, article recommender, big data, bahasa**

I. INTRODUCTION

Mass media industry had a lot of change in twenty years, the biggest challenge in media industry is the number of article. As there were many news, information phenomena overload happened, in which the readers are difficult to get the most important news for them [15]. Customers do not know what they really want to know in a large of news displayed.

Nowadays, the information technology is very important in the industry, especially in industry Mass media. Information technology is used as a tool to increase the effectivity and efficiency in the industry. To increase the power of business competition, with the present the competitor business.

Going to digital is a new strategy of any kind of industry included mass media to increase revenue. Data is being generated very fast that is refered as a big data [1]. database will be important to make the data to be more useful for the company to increase the profit. Data mining method can be used to process the data in grouping each customer and study customer's behaviour in every news that seen.

Manuscript received on February 10, 2020.
Revised Manuscript received on February 20, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author
Sri Hesti Mahanani, Master's, Department Information System Management, Bina Nusantara University, Jakarta
Valentinus, Master's, Department Information System Management, Bina Nusantara University, Jakarta
Dennis, Master's, Department Information System Management, Bina Nusantara University, Jakarta
Tuga Mauritsius, Master's, Department Information System Management, Bina Nusantara University, Jakarta

Data mining itself refers to search the hidden information from a part of a data that was difficult to do it manually [10]. In addressing this phenomenon article recomendation will guide customer to new unknown experienced articles that may be relevant to the customers current task [8].

In this paper, we introduce a simple model to examine the relevance between the recommended article and the preferences of users as an article recommender. Article recommender using data mining can be a tool to help the user to overcome the overloaded information, by filtering the stream of information that enter based on the preferensi of user to make personalized news content [12]. Article recommender will increase the number of customers and benefit sellers, but whether they benefit customers by providing relevant article.

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 [3]. 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 [3].

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 [3].

Processing data stages according to [7] explain the definition about Tokenizing is the stage of separating sentences into words, deleting special characters and punctuation.

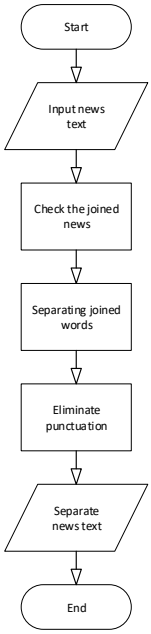

Fig. 1.Tokenizing Processes Phase [7]

According to [6] 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.

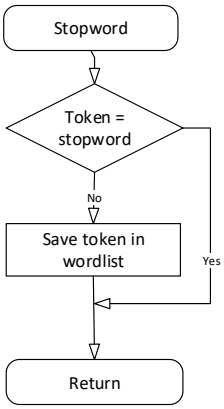

Fig. 2.Phase Stopword Processes [6]

The definition of Stemming according to [5], 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 [5] 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- I: 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 :

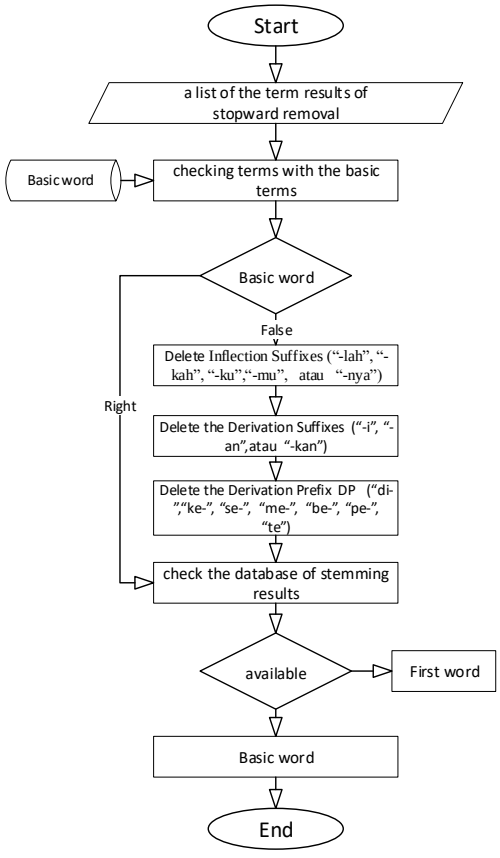

Fig. 3.Phase of Stemming Processes[5]

Retrieval Number: F7231038620/2020©BEIESP
DOI:10.35940/ijrte.F7231.038620
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

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 [4]. 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 [2].

D. Consine Similarity

Vector space model is a model used to measure similarity between a document and a query. In this model, query and document considers as vectors in a room dimensions, where n is a total from all term contained in leksicon. Leksicon is a list of all terms that are in the index. One way to overcome this in the vector space model is by extending the vector. The expansion process can be done on query vectors, document vectors or on both vectors.

The algorithm vector space model used a formula to look up cosinus between the angle from the two vector (WD) of each document weight (WK). The formula used in the vector space model is as follows [14] :

$$Cosinus \rightarrow sim(d_j, q) = \frac{d_j \cdot q}{|d_j| \cdot |q|} = \frac{\sum_i W_{ij} \cdot W_q}{\sqrt{\sum_i W_{ij}^2} \cdot \sqrt{\sum_i W_q^2}}$$

Fig. 4.The formula of Vector Space Model [14]

Note :
W : word weight i in document j
Wq : query weight

Calculation of the cosine value of the angle between the two vectors is known as the method Cosine Similarity. The cosine angle value between two vectors determines the similarity of two objects compared where the smallest value is 0 and the largest value is 1 [13].

III. METHODOLOGY

The objective of this paper is to implement the mass media article recommendation system using a text mining method that can provide article recommendations to each user based on historical articles that are often read by that user. 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 recommendations for articles that are appropriate for each user.

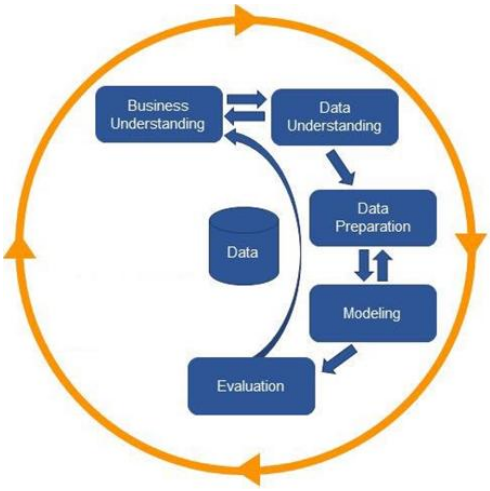

Fig. 5. The Methodology of Text Mining [11]

- Business Understanding

The first step we look at the problem of the mass media industry in this case there is a problem of information overload in Indonesia people. We also search from the literature review from other business like e-commerce, supermarket and any other industry other than mass media industry.

- Data Understanding.

Dataset was taken from kompas.id company. Data taken from mongodb which store costumer's detail profile and list of articles read by costumers.Atrributes needed in this paper are article title, article category, article tag, article summary, article scroll dept, and customer id. One customer profile detail and historical article read used to build model. List of new article which does not read before used to test model built.

- Data Preparation

We prepare the data for building the model. We combine the data from article title, category, tag, summary to become one data, And we used stemming to stem the word from the article and we remove the stopword from the article using sastrawi library in python. To ensure we can collect the feature from the article. We used UUID to indentified the user that access. And we used title to indentified the article.

- Modeling

We build model using TF-IDF of the article that already combine all of the attributes and the data already stemmed and remove the stopword. We use TF-IDF to count the word that the reader like to read. We cobine the TF-IDF with the scroll depth of the user to build the user profile of the article. Than we can use cosine similarity to look at the similarity of the article that we want to recommend to the user.

- Evaluation

We evaluate the article in qualitative method to know what the people like to read based on their historical preference. And we try to suggest using cosine similarity from the deck of news article to know what they like to read.

IV. RESULT

	UUID	Kategori	Tag	Summary	scroll depth	judul
0	00252af7-89ba-4fb8-8d32-a73468a2aceb	utama cerpen-hiburan sastra	kompasminggu	la yang diingat sebagai yang lebih tua melunc...	70	Monolog si Kembar Setelah Menyebar
1	00252af7-89ba-4fb8-8d32-a73468a2aceb	utama buku humaniora	djoko-pekik menyusu-celeng sindhunata gramedia...	Orang-orangan sawah itu menggendong seekor cel...	60	Menjadi Celeng Bersama Sindhunata
2	00252af7-89ba-4fb8-8d32-a73468a2aceb	utama buku humaniora	djoko-pekik menyusu-celeng sindhunata gramedia...	Orang-orangan sawah itu menggendong seekor cel...	50	Menjadi Celeng Bersama Sindhunata
3	00252af7-89ba-4fb8-8d32-a73468a2aceb	utama ekonomi	natal konsumsi transportasi tahun-baru liputan...	Data-data arus pergerakan warga selama empat h...	40	Kebutuhan Berlibur yang Tetap Tinggi
4	00252af7-89ba-4fb8-8d32-a73468a2aceb	utama buku humaniora	djoko-pekik menyusu-celeng sindhunata gramedia...	Orang-orangan sawah itu menggendong seekor cel...	50	Menjadi Celeng Bersama Sindhunata

Fig. 6. Result of Tokenizing

From the data set we get 5 column:

1) UUID
   UUID is the user profile of the user of the website.
2) Category
   Category is the category of the article of the data in the website. The category consist of politic, utama, cerpen, book, ekonomi, humaniora, opinion, research, nusantara, sport, photography and many more.
3) Tag
   Tag used to classify the article so we can get the meaning of the data from the article. Tag used to mark the article
4) Summary
   Summary is the summary of the article it's consist of the three hundred words that written by an editorial department.
5) Scroll depth
   scroll depth is the data from the userbehaviour that scrolled into the bottom of the article that they read. The number if from 0 until 100. But we change it by divide it by 20 to make the data more precision.
6) Title
   The title of the article that can be used for primary key of the article. Title is used to give the summarize of the article.

	UUID	scroll depth	judul	CombineAll
0	00252af7-89ba-4fb8-8d32-a73468a2aceb	70	Monolog si Kembar Setelah Menyebar	utama cerpen-hiburan sastra kompasminggu la ya...
1	00252af7-89ba-4fb8-8d32-a73468a2aceb	60	Menjadi Celeng Bersama Sindhunata	utama buku humaniora djoko-pekik menyusu-celen...
2	00252af7-89ba-4fb8-8d32-a73468a2aceb	50	Menjadi Celeng Bersama Sindhunata	utama buku humaniora djoko-pekik menyusu-celen...
3	00252af7-89ba-4fb8-8d32-a73468a2aceb	40	Kebutuhan Berlibur yang Tetap Tinggi	utama ekonomi natal konsumsi transportasi tahu...
4	00252af7-89ba-4fb8-8d32-a73468a2aceb	50	Menjadi Celeng Bersama Sindhunata	utama buku humaniora djoko-pekik menyusu-celen...

Fig. 7. Result of Stemming

We prepare the data by combining all of the category, tag, summary and title and become one column that called combine all. And we delete the category, tag and summary column. So we can focus to produce the TF-IDF from the combine all column.

Modelling TF-IDF

Fig. 8. Modelling TF-IDF [7]

After that we used TF-IDF to make a bag of word that become our feature in the image above. And from the feature above we get TF-IDF to become like this.

	token	relevance
0	celeng	0.175187
1	iklim	0.157327
2	ekonomi	0.155611
3	orang	0.124592
4	ubah	0.107497
5	laku	0.104909
6	ubah iklim	0.103275
7	ajar	0.096893
8	mogok	0.096893
9	mogok sekolah	0.096893
10	sekolah	0.096893
11	jadi	0.094409
12	digital	0.089754
13	kontribusi	0.088325
14	gendong	0.087593
15	orang sawah	0.087593
16	sawah	0.087593
17	sindhunata	0.087593
18	tingkat	0.080798
19	global	0.079424

Fig. 9. Result of TD-IDF from combine all column

We used the scroll to build the user profile of tf-idf that become like this:

Cosine Similarity	Article title
0.11107693	Dua Raja Bertemu di Keraton Yogyakarta
0.0854122	Di Balik Perjumpaan Dua Raja di Keraton "Jogja"
0.08243335	Sembunyikan 70 Kilogram Sabu di Ban Pengedar Ditembak Mati
0.07650437	Ada 291 306 Juta Rekening Simpanan di Bank
0.07334038	Menilik Piknik Sang Raja di Borobudur...
0.06749902	Papua dan Imajinasi Antarbudaya
0.06090463	Pengkritik Proses Seleksi Dilaporkan Diduga Upaya Melemahkan Pengawalan Publik
0.0599402	Penyelundupan Benih Lobster Terus Berulang
0.05895653	Bali United Kunci Juara Paruh Musim Liga 1
0.05654763	Ratusan Warga Kalideres Terpaksa Hidup di Pengungsian
0.05650609	Pindah Ibu Kota ASN Jangan Terlalu Khawatir
0.05595156	Polisi Tetapkan Satu Tersangka Ujaran Kebencian ke Mahasiswa Papua
0.05471232	Tidak Ada Kejahatan yang Sempurna...
0.0532928	Di Jakarta Pasokan dan Harga Beras Stabil
0.05172768	Jubir KPK Dilaporkan ke Polda Metro Jaya
0.05161894	Pekerja Lanjut Usia
0.05084161	Tertinggal dari Singapura dan Malaysia Indonesia Harus Genjot Inovasi
0.05051882	Jalan Prestasi Anak Berkebutuhan Khusus
0.04932486	Pengalaman Konsumen Jadi Kunci Kesuksesan Bisnis
0.04879074	Beijing Tak Puas pada G-7
0.04833414	Menyiasati Harga Rumah yang Makin Tinggi
0.04826928	Bahasa Inggris di Era Disrupsi untuk Pacu Berpikir Kritis
0.0481604	Asa Petani Tumbuh di Ogan Komering Ilir
0.04773546	Terkait Kebakaran Lahan Hukum Harus Tajam bagi Semua Kalangan

Fig. 10. Result of Testing Article

	token	relevance
0	ekonomi	0.186719
1	iklim	0.166128
2	celeng	0.144512
3	digital	0.138829
4	ubah	0.115344
5	orang	0.114003
6	ubah iklim	0.112513
7	laku	0.106237
8	ajar	0.097690
9	mogok	0.097690
10	mogok sekolah	0.097690

Fig. 11. Result of TF-IDF combine with scroll depth

The TF-IDF combine it with the scroll depth and from the data we know that the people like to read economic article, celeng, digital and many more. We test of the article using 200 different article that the never read it before. And we used cosine similarity of the article

From the testing article we can suggest to recommend the article that they must read is “dua raja bertemu di keratin jogja” this article contain the humanity and economic perspective and also the two king from Malaysia and Jogjakarta who discuss about the economy from two different kingdom and the culture.

V. CONCLUSION

Using data that we collect from the big data we mine the data and become more feature using tf-idf. We collect the feature from tf-idf and we build the recommendation system for the article to help the customer to a new unknown experienced article that relevant to the their using cosine similarity of the new article that they never read it before. So we can filtering the stream of information and highlight the most likely article they will read but based on their preference that we already collect implicitly from the article that they read it, it’s a scroll depth of the article they read. Then we can serve the news more personalized from what they love to read.

REFERENCES

1. Anam Sardar, J. F. (2017). Recommender System for Journal Articles using Opinion Mining and Semantics. (IJACSA) International Journal of Advanced Computer Science and Applications.
2. Chen, K. Z. (2016). Turning from TF-IDF to TF-IGM for term weighting in text classification. Expert Systems With Applications, 245–260.
3. Feldman, R. (2007). The text mining handbook: Advanced approaches in analyzing unstructured data. England: Cambridge.
4. 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.
5. 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.
6. Hapsari, R. K. (2015). STEMMING ARTIKEL BERBAHASA INDONESIA DENGAN Pendekatan Confix-Stripping. Prosiding Seminar Nasional Manajemen Teknologi XXII, 1-8.
7. 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.
8. Mehrbakhsh Nilashi, K. B. (2013). Collaborative Filtering Recommender Systems . Research Journal of Applied Sciences, Engineering and Technology.
9. Purnomo, J. A. (2010). Analisis perbandingan beberapa metode pembobotan kata terhadap performansi kategorisasi teks.
10. Zuha, F. &. (2016). Analysis of Data Mining Techniques and its Applications. International Journal of Computer Applications, 140(3), 6–14.
11. Chapman, et al., (2000). CRISP-DM 1.0: Step-by-step data mining guide. Computer Science.
12. Karimi, Jannach, Jugovac, (2018). News recommender systems – Survey and roads ahead. Information Processing & Management.
13. Firdaus, et al,. (2014). Retweet prediction considering user's difference as an author and retweeter. IEEE
14. 14. Sidorov Grigori, et al,. (2014). Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model. Computacion y sistemas.

15. Tingting Jiang, Fang Liu, Yu Chi, (2015). Online information encountering: modeling the process and influencing factors. Emerald Group Publishing Limited

AUTHORS PROFILE

Sri Hesti Mahanani, Information System Management Department, BINUS Graduate Program – Master of Information System Management, Bina Nusantara University, Jl. Raya Kb. Jeruk No. 27, Jakarta 11480
Sri.mahanani@binus.ac.id

Valentinus, Information System Management Department, BINUS Graduate Program – Master of Information System Management, Bina Nusantara University, Jl. Raya Kb. Jeruk No. 27, Jakarta 11480
Valentinus@binus.ac.id

Dennis, Information System Management Department, BINUS Graduate Program – Master of Information System Management, Bina Nusantara University, Jl. Raya Kb. Jeruk No. 27, Jakarta 11480
Dennis@binus.ac.id

Tuga Mauritsius, Information System Management Department, BINUS Graduate Program – Master of Information System Management, Bina Nusantara University, Jl. Raya Kb. Jeruk No. 27, Jakarta 11480
tmauritsus@binus.edu