JSE Journal of Science and Engineering, Vol. 1, No. 1 (page: 23-29)
e-ISSN: 2723-3944 September 2023
DOI: https://doi.org/10.30650/jse.v1i1.3680

Latent Dirichlet Allocation Utilization as a Text Mining Method to Elaborate Learning Effectiveness

Netri Alia Rahmi[1]*, Rudiman[2], Rafik Septiana[3] , Christeigen Theodore Suhalim[4]
[1] Teknologi Sains Data, Fakultas Teknologi Maju dan Multidisiplin, Universitas Airlangga
[2] Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Muhammadiyah Kalimantan Timur
[3,4] Teknologi Sains Data, Fakultas Teknologi Maju dan Multidisiplin, Universitas Airlangga
* Corresponding Email: netri.alia.rahmi-2021@ftmm.unair.ac.id

Abstract – Learning method is a way to explain the lesson materials to students so that the learning process can occur in students as an effort to achieve the goals. Learning methods can be said to be a success if students are active, both physically, mentally, and socially in the learning process, in addition to showing high enthusiasm for learning and having self-confidence. The purpose of this study is to classify the opinions of Indonesian students regarding the existing learning methods and what learning methods they expected. In order to evaluate existing learning methods using the latent dirichlet allocation method. The data used comes from tweets of Twitter users within the range of January to March 2022. The data is taken using the scrapping method through the help of the python twisel library and totaled to 3778 data, then preprocessed through the nltk and Sastrawi libraries. The results of this analysis stated that student opinions can be classified into 3 major topics which state students' opinions regarding effective learning methods, student difficulties in applicable learning methods, and high cross-departmental interest.

Keywords: learning method, latent dirichlet allocation, twitter, text mining

Submitted: 17 Februari 2023 - Revised: 18 Agustus 2023 - Accepted: 18 September 2023

1. Introduction

It is undeniable that the progress of a country is highly dependent on the quality of the human resources that surround it. To be able to improve this quality in a sustainable manner, Indonesia must further examine its main fulcrum, which of course lies in education. Education is not all about how much the younger generation gets 100 points. But more broadly, education must be able to create a young generation that is competent, creative, innovative and able to compete globally. It is very unfortunate if we look at the data, until now the quality of education in Indonesia can be said to be still relatively low. This can be seen in a survey conducted by the Program for International Student Assessment held in Paris in 2018. The results of the survey stated that Indonesia was ranked 72nd out of 77 countries. Even Indonesia is ranked 7th out of 10 Southeast Asian countries. Not to mention the data which states that only 44% of Indonesian children reach the minimum competency level for reading skills and only 21% for math skills[1]. But on the other hand, the level of digital curiosity of the Indonesian people tends to be high, especially when it comes to digitization.

According to Dr. M. Sobry Sutikno [2] explain learning methods can be said to be a success if students are active, both physically, mentally, and socially in the learning process, in addition to showing high enthusiasm for learning and having self-confidence.The purpose of this study is to classify the opinions of Indonesian students regarding the existing learning methods and what learning methods they expected. In order to evaluate existing learning methods using the latent dirichlet allocation method.

Research that has been conducted on topic modeling by Utami [3] , analyzing the topic of Twitter social media data using the Latent Dirichlet Allocation topic model, obtained the results that LDA topic modeling in the five tweet locations in Bogor City and a certain time span succeeded in forming topics with information or topic descriptions for each tweet location. Another study by Al-khairi, et al., [4] regarding the detection of fashion topics on twitter with Latent Dirichlet Allocation resulted in the optimal number of topics being 20 topics. Another study by Kurniawan [5] regarding the conversation monitoring system in online stores using the case study LDA method: online shop "berrybenka.com" obtained the results of the most optimal number of topics being 10 topics.

2. Related Works

According to (Hearst 2021), text mining can be interpreted as the discovery of new and previously unknown information by a computer, the purpose is to extract and then combine a number of information from various sources.

2.1 Text Mining

Text Mining is a process of data mining from text sources to obtain meaningful information based on certain patterns [6]. Text mining is the process of mining data in the form of text obtained through documents to find words related to the contents of the document so that an analysis of the relationship from the text sources can be carried out [7]. Text mining is the process of analyzing large amounts of unstructured text data with the help of software that can identify certain patterns and keywords in the data [8]. According to the above understanding, it can be said that text mining is a process to manage data in the form of text that comes from certain documents to get words or patterns that are interrelated to get meaningful and interesting information in it. This is related to our research which involves processing text data taken from Twitter data to retrieve words related to the topic of effective learning, then analyzing certain patterns to obtain meaningful information from the data.

2.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a model in machine learning that uses the probabilistic of a corpus which is represented by the distribution of data for each word in the document [9]. Latent Dirichlet Allocation (LDA) is a method for processing large amounts of data with the assumption that one document consists of various topics which are vocabulary distribution [10]. Latent Dirichlet Allocation (LDA) is a method in unsupervised learning that is used to group data into several topics, summarize, and process large data [11]. From the above understanding, it is found that Latent Dirichlet Allocation (LDA) is a machine learning method that is used to manage large amounts of data into several topics based on word distribution. This relates to the grouping of twitter data that has been obtained into several topics based on the distribution of certain words in the tweets of Twitter social media users so that an effective learning method from these topics is found. The formula of Latent Dirichlet Allocation:

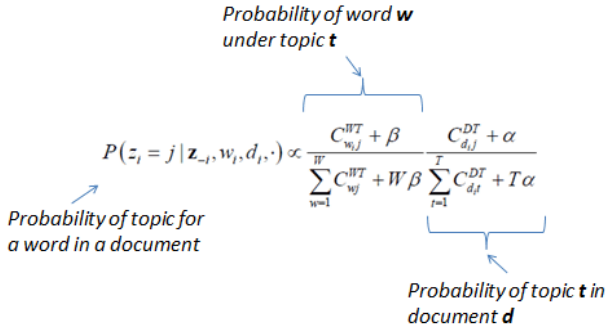


Figure 1. Formula of LDA

Cw,jWT = the number of times a word appears as topic 1 and topic 2.

B = word distribution by topic (concentration parameter)

W = Length of vocabulary (No. of unique token/ word in full document)

Cd,tDT = While starting an iteration number of times a document appeared as topic 1 and topic 2.

A = Per document topic distribution.

T = Number of topic. (here T = 2)

2.3 Web Scraping

Web Scraping or also known as web extraction is a technique for extracting data from the internet and saving the result to files such as a database, CSV, or another extension file [12]. Web scraping is the process of getting an unstructured document from the website in the form of markup language (HTML). The result from scraping will be analyzed to retrieve certain data from the page [13]. Web scraping is a technique of getting information automatically from a website without copying it manually [14]. Web scraping has the benefit of making it easier to search for something to make it more focused. The program will analyze HTML documents from the internet and get data based on tag HTML to flanking the information you want to retrieve (create a scraping template). After that, the information will be saved into a database table, CSV, or another file format. This retrieval will use the Python programming language by retrieving data through HTML tags and saving it into a CSV file for processing to the next stage.

2.4 Data Cleaning

Data cleaning is the process of identifying and eliminating errors in the data [15]. This is to ensure that the data taken is of good quality. Data Cleaning can be used for cleaning string data to eliminate irrelevant characters, misspellings, and other errors in the text. Data cleaning in a general sense is a process of investigating data from inaccuracies and creating manageable data for analysis. In this research, we will do data cleaning on the String data type which will be cleaned of words that are not relevant to the topic, removing misspellings, normalizing data, or other cleaning data, so data from Twitter can be analyzed and extracted more easily and structured. This includes case folding, stemming, and tokenization.

2.5 Case Folding

e-ISSN: 2723-3944

Case folding is a process in data preprocessing that is changing all characters in the text to lowercase and removing invalid characters such as numbers, punctuation marks, and Uniform resources locator (URL) [16]. Case folding is the process of changing all characters into uniform by changing letters to lowercase and removing punctuation marks and numbers in the text [17]. Case folding is a process that is carried out automatically to change all letters in the text to lowercase or capital letters [18]. Case folding in general is a process in data preprocessing to perform a general and thorough conversion of a text into lowercase and remove punctuation marks and numbers in it. Case folding will be used to convert the collected tweets into lowercase letters and remove unnecessary punctuation marks, numbers, emoticons, or characters. It aims to facilitate the data processing process because the data used is only in lowercase letters.

2.6 Stemming

Stemming is a process of normalizing text by removing prefixes and suffixes in a word. Stemming also is used for a retrieval information system by removing prefixes, infixes, suffixes, a combination of prefixes and suffixes, and repetition of words [19]. The output of stemming is the basic form of a word. The main idea of stemming is reducing a word to its basic form which usually removes affixes from the word index before the analysis process in data mining [20]. The stemming process will be used to change the formal or informal language in tweets into their basic form. For example, the word "berlari" will be changed to "lari". It aims to normalize and retrieve words that have the same basic form but have different suffixes. It is also intended that the frequency or distribution of words in a document can be calculated properly. This is done using the python libraries namely Sastrawi and nltk automatically.

2.7 Regex Tokenizer

Regex (Regular Expression) is a combination of two types of characters, namely literals and meta characters. Literals are characters that represent themselves such as the entire alphabet, capital letters, lowercase letters, or other characters [21]. Regex is also one of the implementations of pattern matching operations for a text or String data type. [22]. Regex is a key for doing text processing powerfully, flexibly, and efficiently by using general notation patterns like a programming language [23]. Because of the convenience, the regex could be used for searching text more easily. This is the basis for using regex to tokenize a text because we can find out the location of numbers, punctuation marks, and characters outside the String. After searching, the character will be deleted automatically. Regex can also separate sentences into groups of words by separating them with spaces. This makes it easy to tokenize a document.

2.8 Learning Method

Learning method is a way of explaining lesson material to students so that the learning process can occur in students as an effort to achieve goals [24]. Learning methods can be said to be successful if students become more physically, mentally, and socially active in the learning process. The learning method is a set of components that have been combined to improve the quality of learning. For the objectives of the learning method to be achieved, an ability is needed in choosing methods, learning models, and learning approaches [25]. Through Text Mining, there will be a classification of opinions from students in Indonesia regarding the learning methods that students expect. This analysis will be carried out using Topics from LDA, where each topic is expected to contain information about the learning methods that are currently favored by students. The results of this grouping will be used as a basis for forming methods, models, and learning approaches that can be applied by students, parents, or teachers in schools.

3. Research Method

Through this method are the research steps that must be passed before discussion, as follows:

3.1. Data

The data used in this study is primary data which contains data on posts by Twitter social media users. In retrieving the data, we do web scraping using python twisel library. The library will use selenium to perform data retrieval via twitter HTML Tags. This retrieval process requires keywords related to one's learning method so that we get relevant data and save time in searching and collecting data. In this research, we use three main keywords, namely "memahami materi", "konsep materi", and "belajar materi". The data was taken from January to March 2022. This was done because, in that month, students in Indonesia were preparing for school exams to university entrance exams. The data obtained from the scraping process amounted to 3778 data.

The data that has been scraped will proceed to the next stage, namely data preprocessing. The problem with analyzing text data taken directly from Twitter is that there are characters that cannot be analyzed using Natural Language Processing such as numbers, punctuation marks, emoticons, and others. Therefore, it is necessary to clean the text data to eliminate it. This cleaning will be done using python to select the data so that only text data is left. After that, the data will enter the case folding stage. At this stage, the text will be transformed by changing uppercase letters to lowercase letters. This is because Python has a case-sensitive nature, such as the letter "A" (uppercase) will be distinguished from the letter "a" (lowercase). It is used to facilitate the analysis phase.

e-ISSN: 2723-3944

	tweet
0	bukannya ape² ye, dikata yg linjur tinggal kom...
1	ptm, alesannya pertama dapet uang jajan, trs l...
2	This!! Aku lumayan rajin ngumpul tugas dan ofc...
3	hikss samaa aku jg kurang puas, waktu jawabin ...
4	Kalau buat ngapalin atau mahamin materi sih yg...
...	...
3573	besok kimia …gk tau mau belajar materi apaaannnnn
3574	pengen belajar desain grafis, belajar materi k...
3575	Ngerasa bersalah hr ini cuma belajar materi ti...
3576	Mungkin bisa join kepanitiaan/volunteer nder, ...
3577	Menemani anak belajar . Materi anak kelas 1 SD...

Figure 2. Illustration of Sample Data

The next step of data preprocessing is stemming. Text data that has been converted to lowercase will be stemming. The process is done to get the basic form of a word. This is to get information from words that have the same basic form but have different affixes. For example, “pendidik” and “terdidik” have the basic word "didik", then the two words will be counted as two words that have the same basic word, namely "didik". In addition to stemming, stopwords in Indonesian in the text will also be removed such as *yang*, *di*, *akan*, etc. so that when calculating the frequency, these words do not become noise and we will get the most relevant words in the analysis. Both processes will use a python library called Sastrawi and nltk, a library that provides preprocessing data on the Indonesian corpus. examples of some words that do stemming:

Table 1

Stemming Process

Before Stemming	After Stemming
mencoba	coba
membaca	baca
melatih	latih
pelajaran	ajar
tambahan	tambah

After stemming, the text data will be tokenized. This process will use a regex tokenizer, namely by splitting words using regular expressions in python. The separation of data is done by separating data based on spaces in sentences. For example “saya belajar di sekolah” would become “saya”, “belajar”, “di”, “sekolah”. The fragment of the word is called a token. The token obtained from this process will be used to perform text data analysis using Latent Dirichlet Allocation (LDA) in the next stage.

3.2. Method

After completing the pre-processing, the next step is to model the topic using LDA. Previously, at the bag of words stage, tokens emerged from the number of words that appeared in a document. The token serves as a measure in the LDA so that it can be modeled. First thing to do in modeling LDA is importing library to use and then creating the object for LDA using genshim library. In this study, it is determined the number of topics from the classification was 3 and took 10 words from each topic.
After determined total topics and number of words, the next step is running and training the LDA model on the document term matrix.

```
lda_model = lda(doc_term_matrix, num_topics=total_topics, id2word = dictionary, passes=50)

lda_model.show_topics(num_topics=total_topics, num_words=number_words)

[(0,
  '0.016*"latsol" + 0.010*"baca" + 0.010*"tugas" + 0.007*"minggu" + 0.006*"besok" + 0.006*"to" + 0.006*"well" + 0.006*"guru" + 0.006*"rest" + 0.005*"fokus"'),
 (1,
  '0.010*"tugas" + 0.009*"jam" + 0.007*"besok" + 0.007*"suka" + 0.007*"susah" + 0.006*"kasih" + 0.006*"ajar" + 0.006*"kuliah" + 0.006*"buku" + 0.006*"ngerti"'),
 (2,
  '0.011*"latsol" + 0.011*"soshum" + 0.011*"anak" + 0.009*"linjur" + 0.009*"ajar" + 0.007*"minggu" + 0.007*"materi" + 0.006*"ipa" + 0.006*"sma" + 0.006*"jurus"')]
```

Figure 3. LDA Parameter

The last step is to count the word count of topic keywords and make it as a table.

```
from collections import Counter
topics = lda_model.show_topics(formatted=False)
data_flat = [w for w_list in doc_clean for w in w_list]
counter = Counter(data_flat)

out = []
for i, topic in topics:
    for word, weight in topic:
        out.append([word, i , weight, counter[word]])

df_imp_wcount = pd.DataFrame(out, columns=['word', 'topic_id', 'importance', 'word_count'])
print(df_imp_wcount)
```

```
       word  topic_id  importance  word_count
0    latsol         0    0.016093         307
1      baca         0    0.009993         167
2     tugas         0    0.009848         238
3    minggu         0    0.006975         172
4     besok         0    0.006268         171
5        to         0    0.005883         143
6      well         0    0.005855          61
7      guru         0    0.005546         112
8      rest         0    0.005511          58
9     fokus         0    0.004848          85
10    tugas         1    0.010268         238
11      jam         1    0.009442         151
12    besok         1    0.006865         171
13     suka         1    0.006773         111
14    susah         1    0.006589         145
15    kasih         1    0.006370          79
16     ajar         1    0.005771         236
17   kuliah         1    0.005635         133
18     buku         1    0.005618         141
19   ngerti         1    0.005615          71
20    latsol         2    0.010950         307
21   soshum         2    0.010769         140
22     anak         2    0.010547         156
23   linjur         2    0.009189         118
24     ajar         2    0.009052         236
25   minggu         2    0.006935         172
26   materi         2    0.006776         121
27      ipa         2    0.006301          81
28      sma         2    0.005892         110
29    jurus         2    0.005791          74
```

Figure 4. Word Count Analysis of Topic Keywords

4. Result and Discussion

4.1. Result

Here is the result of Latent Dirichlet Allocation:

e-ISSN: 2723-3944

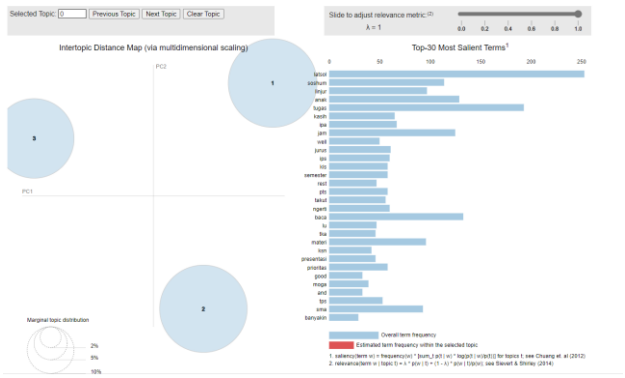
	word	topic_id	importance	word_count
0	latsol	0	0.016093	307
1	baca	0	0.009993	167
2	tugas	0	0.009848	238
3	minggu	0	0.006975	172
4	besok	0	0.006268	171
5	to	0	0.005883	143
6	well	0	0.005855	61
7	guru	0	0.005546	112
8	rest	0	0.005511	58
9	fokus	0	0.004848	85
10	tugas	1	0.010268	238
11	jam	1	0.009442	151
12	besok	1	0.006865	171
13	suka	1	0.006773	111
14	susah	1	0.006589	145
15	kasih	1	0.006370	79
16	ajar	1	0.005771	236
17	kuliah	1	0.005635	133
18	buku	1	0.005618	141
19	ngerti	1	0.005615	71
20	latsol	2	0.010950	307
21	soshum	2	0.010769	140
22	anak	2	0.010547	156
23	linjur	2	0.009189	118
24	ajar	2	0.009052	236
25	minggu	2	0.006935	172
26	materi	2	0.006776	121
27	ipa	2	0.006301	81
28	sma	2	0.005892	110
29	jurus	2	0.005791	74

From the results of our analysis, it can be seen that the data are classified into 3 topics. Topic 1 which is indicated by 0 in the “topic_id” column can be interpreted as discussing learning methods that are considered effective by the community. Evidenced by the words that appear are “latihan soal”, “baca”, “tugas”, “try out”, “rest”, and “fokus”.

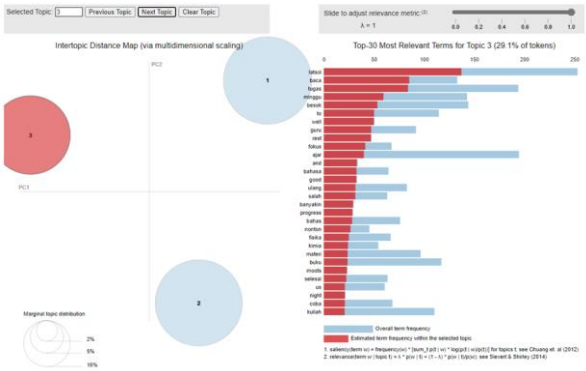
Topic 2 indicated by 1 in the “topic_id” column can be interpreted as discussing the community's difficulties in the applicable learning method. Evidenced by the words that appear are “tugas”, “jam”, “besok”, “suka”, “susah”, “kasih”, “ajar”, “kuliah”, and “buku”.

Topic 3 which is indicated by 2 in the “topic_id” column can be interpreted as discussing the high public interest in changing majors from science to social studies. It is proven by the words that appear, namely "latian soal", "soshum", "anak", "lintas jurusan", "ajar", "minggu", "materi", "IPA", "SMA", and "jurus".

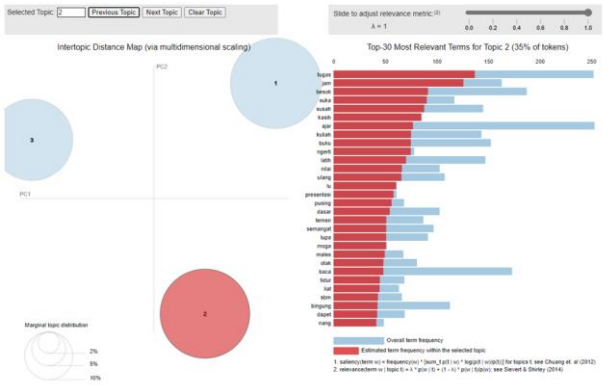
4.2. Statement of Results and Explanatory Text



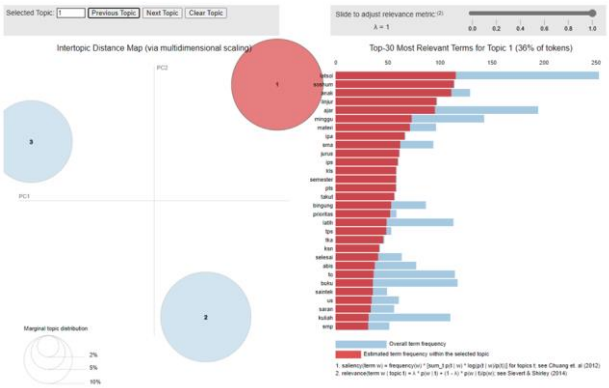
This graph is a visualization of the Latent Direchlet Model. From the visualization we can see that the data splited into 3 topic.



This graph is a visualization of topic 1.In the visualization, it can be seen the comparison of Estimated term frequency within the selected topic (with red graph) with overall term frequency (with blue graph).



This graph is a visualization of topic 2. In the visualization, it can be seen the comparison of Estimated term frequency within the selected topic (with red graph) with overall term frequency (with blue graph).



This graph is a visualization of topic 3. In the visualization, it can be seen the comparison of Estimated term frequency within the selected topic (with red graph) with overall term frequency (with blue graph).

e-ISSN: 2723-3944

4.3. Discussion

0	latsol	0	0.016093	307
1	baca	0	0.009993	167
2	tugas	0	0.009848	238
3	minggu	0	0.006975	172
4	besok	0	0.006268	171
5	to	0	0.005883	143
6	well	0	0.005855	61
7	guru	0	0.005546	112
8	rest	0	0.005511	58
9	fokus	0	0.004848	85

Topic 0 contains “Latihan soal”, “baca”, “tugas”, “to”, “rest”, and “fokus”. Through the words that appear, it is necessary to connect them so that it is known what insight we can get from topic 0. These words can be related to one key, which is the way of learning that is considered effective by the community in the applicable learning method. This learning method is a method that is carried out repeatedly such as practice questions, reading, try out. As we know that successful learning methods are students who are active, enthusiastic, and have self-confidence. The way of learning that is felt by the community to be effective in the current learning method is very monotonous and feels stiff, which shows that there is not a varied way of learning such as understanding, drawing, and much more.

10	tugas	1	0.010268	238
11	jam	1	0.009442	151
12	besok	1	0.006865	171
13	suka	1	0.006773	111
14	susah	1	0.006589	145
15	kasih	1	0.006370	79
16	ajar	1	0.005771	236
17	kuliah	1	0.005635	133
18	buku	1	0.005618	141
19	ngerti	1	0.005615	71

Topic 1 contains “tugas”, “jam”, “besok”, “suka”, “susah”, “kasih”, “ajar”, “kuliah”, and “buku” which if conclude will be complaints of students in the current learning method. The current learning method contains dozens of subjects both science and social science, which is the material being taught is very general. It should be noted that the interests of students may not be the same and that wide, so it is not surprising that the assignments given to students from these dozens of lessons become a big burden for students. With such a learning method, students find it difficult and cannot explore their interests and talents more deeply.

20	latsol	2	0.010950	307
21	soshum	2	0.010769	140
22	anak	2	0.010547	156
23	linjur	2	0.009189	118
24	ajar	2	0.009052	236
25	minggu	2	0.006935	172
26	materi	2	0.006776	121
27	ipa	2	0.006301	81
28	sma	2	0.005892	110
29	jurus	2	0.005791	74

Topic 2 contains "Latihan soal", "Soshum", "anak", "lintas jurusan", "ajar", "minggu", "materi", "IPA", "SMA", and "jurus" which if we draw conclusions will become the huge interest of science students to change their majors to social. Science is an interesting subject, but this cannot be executed well with current learning methods as evidenced by the high interest in switching majors from science to social science. This shows that our learning method tends to make students not find their passion in science because they only tend to memorize and study broad material.

Through the results of the interpreted analysis, it can be seen that the current learning method is still considered ineffective for the community. The three topics generally discuss people's dissatisfaction with the current learning methods. From the results of the analysis, it is hoped that it can be used as an evaluation for related contigent to make decisions about learning methods that will be implemented in the future.

5. Conclusion

Based on the results of the research that researchers have obtained regarding effective learning methods for students using LDA. The researcher found that students' opinions about learning methods could be classified into three main topics. These topics are about effective learning methods, student difficulties in applicable learning methods, and high cross-departmental interest. Although the use of LDA can model learning methods, further studies are needed to conduct an in-depth analysis of related topics. In addition, the LDA model in this study also needs to be compared with other classification models such as Principal Component Analysis (PCA), Pos Tagging, or other machine learning models in order to obtain the most appropriate classification of effective learning methods. Thus, the classification of texts into certain topics using LDA can be used by the government, schools, or parents to form learning methods according to the interests and desires of students.

e-ISSN: 2723-3944

References

[1] Pusat Penilaian Pendidikan Balitbang Kemendikbud, "Pendidikan Di Indonesia Belajar Dari Hasil PISA 2019," Pusat Penilaian Pendidikan Badan Penelitian dan Pendidikan Kemendikbud, Jakarta Pusat, 2019.

[2] S. Sutikno, Strategi Pembelajaran, Indramayu: Penerbit Adab, 2021.

[3] K. P. Utami, "Analisis Topik Data Media Sosial Twitter Menggunakan Model Topik Latent Dirichlet Allocation," Institut Pertanian Bogor, Bogor, 2017.

[4] Y. U. Al-khairi, Y. Wibisono dan B. L. Putro, "Deteksi Topik Fashion Pada Twitter Dengan Latent Dirichlet Allocation," 04 Desember 2017.

[5] W. Kurniawan, "Sistem Monitoring Percakapan Pada Toko Online Menggunakan Metode Latent Dirichlet Allocation (LDA) Studi Kasus: Toko Online "Berrybenka.com"," 2018.

[6] A. Udgave dan P. Kulkarni, "Text Mining and Text Analytics of Research Articles," PalArch's Journal of Archeology of Egypt/Egyptology, no. 17, p. 6, 2020.

[7] E. K. Putri dan T. Setiadi, "Penerapan Text Mining Pada Sistem Klasifikasi Email Spam Menggunakan Naive Bayes," Jurnal Sarjana Teknik Informatika, vol. 2, no. 3, 2014.

[8] F. Fathonah dan A. Herliana, "Penerapan Text Mining Analisis Sentimen Mengenai Vaksin Covid-19 Menggunakan Metode Naïve Bayes," Jurnal Sains dan Informatika, vol. 7, no. 2, 2021.

[9] D. M. Blei, A. Y. Ng dan M. I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research 3, 2003.

[10] M. L. C. Chilmi, "Latent Dirichlet Allocation (LDA) Untuk Mengetahui Topik Pembicaraan Warganet Twitter Tentang Omnibus Law," Universitas Islam Negeri Syarif Hidayatullah, Jakarta, 2021.

[11] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li and L. Zhao, "Latent Dirichlet Allocation (LDA) and Topic Modeling: models, applications, a survey," Multimedia Tools and Applications, vol. 78(11), pp. 15169-15211, 2019.

[12] B. Zhao, Encyclopedia of Big Data, Viginia: Springer, 2017.

[13] M. Turland, php|architect's Guide to Web, 1 ed., Toronto: Marco Tabini, 2010.

[14] D. D. Ayani, H. S. Pratiwi and H. Muhardi, "Implementasi Web Scraping Untuk Pengambilan Data Pada Situs Marketplace," Jurnal Sistem dan Teknologi Informasi, vol. 7(4), pp. 2460-3562, 2019.

[15] R. R.Deshmukh and V. Wangikar, "Data Cleaning: Current Approches and Issues," Aurangabad, 2011.

[16] D. S. Indraloka and B. Santosa, "Penerapan Text Mining untuk Melakukan Clustering Data Tweet Shopee Indonesia," Jurnal Sains dan Seni ITS, vol. 6(2), pp. 2337-3520, 2017.

[17] F. S. Jumeilah, "Penerapan Support Vector Machine (SVM) untuk Pengkategorian Penelitian," Jurnal Rekasa Sistem dan Teknologi Informasi, vol. 1(1), no. 10.29207/resti.v1i1.11, pp. 19-25, 2017.

[18] A. T. Jaka, "Preprocessing Text untuk Meminimalisir Kata yang Tidak Berarti Dalam Proses Text Mining," Jurnal Informatika UPGRIS, vol. 1, 2015.

[19] J. Snajder and B. D. Basic, "String Distance-based Stemming of the Highly Inflected Croatian Language," 2009.

[20] A. G. Jivani, "A Comparative Study of Stemming Algorithms," Journal of Computer Scienc and Engineering, vol. 2(6), pp. 1930-1938, 2011.

[21] S. Madya, Metodologi Pengajaran Bahasa dari Era Prametode Sampai Era Pascametode, Yogyakarta: UNY Press, 2013.

[22] T. T. Haji, E. M. Faridli and S. Harmianto, Model-Model Pembelajaran Inovatif, Bandung: Alfabeta, 2011.

[23] S. K. Bhatia, "Regular Expressions," January 2005. [Online].Available:https://www.researchgate.net/publication/235916015_Regular_Expressions. [Accessed 20 July 2022].

[24] A. Muhardian, "petanikode.com," 2020. [Online].Available:https://www.petanikode.com/regex/. [Accessed 20 July 2022].

[25] J. E. Friedl, Mastering Regular Expression, 7th ed., California: O'Reilly, 1997.

e-ISSN: 2723-3944