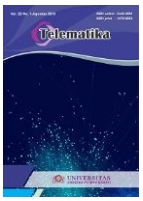

Terbit *online* pada laman web jurnal :
http://ejournal.amikompurwokerto.ac.id/index.php/telematika/

**Telematika**

Accreditated SINTA "2" Kemenristek/BRIN, No. 85/M/KPT/2020

# Topic Modeling of Online Media News Titles during COVID-19 Emergency Response in Indonesia Using the Latent Dirichlet Allocation (LDA) Algorithm

**M. Didik R. Wahyudi[1], Agung Fatwanto[2], Usfita Kiftiyani[3], M. Galih Wonoseto[4]**

[1,2,3,4] Informatics Engineering Study Program, Faculty of Science and Technology
UIN Sunan Kalijaga
E-mail: m.didik@uin-suka.ac.id[1], agung.fatwanto@uin-suka.ac.id[2], usfita.kiftiyani@gmail.com[3], galihwono@gmail.com[4]

**ARTICLE INFO**

***History of the article:***
Receive February 9, 2021
Revised June 6, 2021
Received August 16, 2021
Available online August 31, 2021

**Keywords:**

Text Mining,
Media Analytics,
Topic Modeling,
LDA,
COVID-19 news

**Correspondece:**

Telepon: +62 81215526339
E-mail: m.didik@uin-suka.ac.id

**ABSTRACT**

Online media news portals have the advantage of speed in conveying information on any events that occur in society. One way to know what a story is about is from the title. The headline is a headline that introduces the reader's knowledge about the news content to be described. From these headlines, you can search for the main topics or trends that are being discussed. It takes a fast and efficient method to find out what topics are trending in the news. One method that can be used to overcome this problem is topic modeling. Topic modeling is necessary to help users quickly understand recent issues. One of the algorithms in topic modeling is Latent Dirichlet Allocation (LDA). The stages of this research began with data collection, preprocessing, forming n-grams, dictionary representation, weighting, validating the topic model, forming the topic model, and the results of topic modeling. The results of modeling LDA topics in news headlines taken from www.detik.com for 8 months (March-October 2020) during the COVID-19 pandemic showed that the best number of topics produced each month were 3 topics dominated by news topics about corona cases, positive corona, positive COVID, COVID-19 with an accuracy of 0.824 (82.4%). The resulting precision and recall values indicate that the two values are identical, so this is ideal for an information retrieval system.

## INTRODUCTION

News coverage in the mass media is the result of a professional and organized work process based on certain parameters and ethics. Therefore, newspapers should provide correct and balanced information to the public. The media are required to have an independent and objective attitude so that people get information proportionally from the mass of time. Society needs information quickly, accurately, and proportionally about everything that happens. This is where the role of the mass media as a source of knowledge is expected to meet the needs of society. The very rapid development of information technology in the last two decades has resulted in major changes in various fields including the mass media. The internet has changed many people's habits and lifestyles, including online media.

The development of online media threatens the existence of conventional media. Not a few print media have gone out of business because they are unable to capture the market amidst the proliferation of online media. Some of them try to adapt by migrating to a digital platform or creating a digital version

while maintaining the printed version. Media convergence is a must for print media to survive. (Kristanto, 2019). Based on data from the Central Statistics Agency published in the online news Beritagar.id edition February 12, 2019, with the title *"Pembaca berita daring meningkat, tapi belum merata"* explained that newsreaders in online media in 2017 increased 35.8% compared to the previous two years, becoming 50.7 million people. Based on its development in each province, for example in North Sulawesi province, it has a high percentage of 62.52% of internet users access to information or news. (Bangun et al., 2019)

Among various mass media, the online media news portal is one of the mass media which has an important power in disseminating information and is often used as the leading source of reference for the public. This is because online news portals are always up to date in reporting every event that occurs in the community. One way to find out the content of a story is from the headline. The headline is the head of the news which serves as an introduction to the reader's knowledge of the content of the news to be described. As an introduction, the headline must meet the requirements of a good title. The accuracy of using words in the title, the scope of the contents of the title, and the grammatical structure of the title will determine whether the title meets the requirements for a good title. Headlines must be relevant, provocative, and concise. (Keraf, 1980). The big event that occurred in early 2020, namely the presence of the COVID-19 virus, followed in early March 2020 the discovery of the first positive COVID-19 patient will certainly not be separated from the news in the mass media, especially the online media news portal. With the background described above, the following research will discuss the topic of online news headline modeling at https://news.detik.com for the first 8 months since the COVID-19 case was discovered in Indonesia. Detik.com was chosen because of the high correspondence between the title and the content of the published news (Handiyani & Hermawan, 2017). The results of this study are expected to provide an overview of what news topics were published from March 2020 to October 2020. From this news topic, it can be seen that the opinions to be built in the community will be known.

To find the topics that exist in the collection of news headlines along with the proportion of occurrences of the topic, this study uses Topic Modeling. Topic Modeling is a word recognition model to find topic recognition patterns by extracting text data to find themes from the data based on statistics. In this research, topic modeling uses the Dirichlet Allocation (LDA) algorithm. LDA is a generative probabilistic model of a collection of writings called corpus. The basic idea proposed by the LDA method is that each document is represented as a random mixture of hidden topics, where each topic has a character that is determined based on the distribution of the words contained in it (Blei et al., 2003). Latent refers to anything hidden in the data. Dirichlet is the distribution of topics in documents and distribution of words in topics. Allocation means allocating topics to documents and document words to topics.

Preliminary research related to text mining analytic media and modeling topics with Latent Dirichlet Allocation (LDA), including research on the classification of messages that enter through social media at the Surabaya social service. The large number of reports every day that comes in makes it difficult to identify problem topics, so a topic model is needed that is able to automatically classify messages. In this study, LDA concluded that the number of topics contained in social media messages was 4 topics (Putra, 2017). Kurniawan's (2018) conducted conversation monitoring at online stores using LDA at the online store "BERRYBENKA.COM". In this study, it was found that the LDA method can model topics or display words that are often discussed in conversation data between customer service and buyers. Determining the best number of topics is done by calculating the value of topic coherence (Kurniawan, 2018). In a study conducted by Akbar Nafisa Ja'far (2018), he conducted a topic analysis from user reviews of applications

on Google Play by modeling the Latent Dirichlet Allocation topic. Topics generated from LDA are analyzed using Jensen-Shannon divergence and sliding indows to help understand what users complain about from time to time. The conclusion obtained is that the topics generated by LDA can be interpreted by humans quite well (Ja'far, 2018). Research conducted by Aulia Rizki Destarani in 2019 on the topic of modeling complaints from Denpasar residents on an online public complaint site to find out the problems that occur in the community shows that data processing from this study resulted in 4 trending topics, with the biggest problems being damaged roads and requests for road repairs (Destarani et al., 2019).

## RESEARCH METHODS

This research uses the text mining literature study technique with the Topic Modeling method, namely Latent Dirichlet Allocation to find hidden topics in a corpus and find the best outcome topic. The research stages to be carried out in this study are as shown in Figure 1 below:

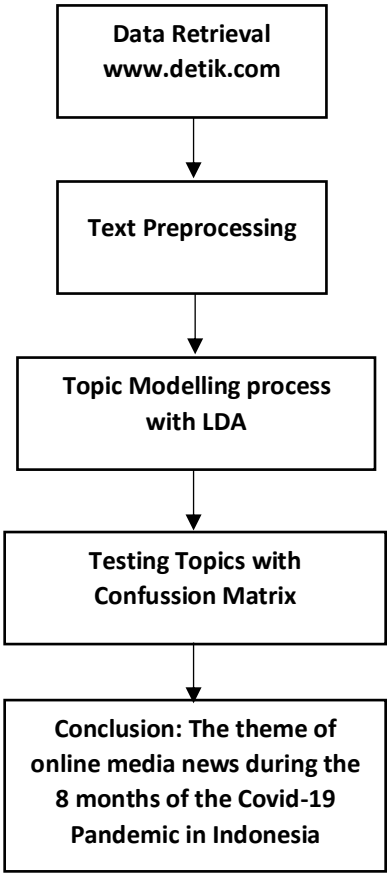


Figure 1: Research Steps

The following is an explanation of the research steps in Figure 1 above :

### Data Retrieval

The data used in this research are news headlines in the online news media https://news.detik.com which were published from March to October 2020. News titles are taken in the Python programming language. Figure 2 below is the stages of data collection:

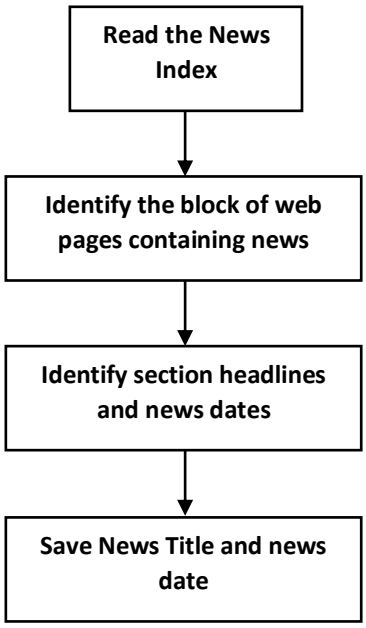


Figure 2. Data Retrieval Process

Data was collected by reading the news index web page which was entered into the program variable, then filtering was carried out on certain HTML tags containing the news title and the date the news was posted. The news index web page is read per day, then the number of news pages is searched and stored in the webPage1 variable. Furthermore, filtering is carried out to find the number of news index web pages each day. The next step for each index web page, read the news content data and stored in the webPage2 variable. This process is repeated until the news index web page has been read per day, continued and repeated for the news index for one month. The data obtained is stored in a CSV file for the next process. Two attributes were taken, namely the date and the news title. This news title will be processed later in this research. Figure 3 below is an example of news titles :

| ID | DATE | TITLE |
|---|---|---|
| 1 | Minggu, 01 Mar 2020 23:52 WIB | bocah di banyuwangi tewas tertembak, pemilik senapan angin diperiksa |
| 2 | Minggu, 01 Mar 2020 23:48 WIB | 69 wni abk diamond princess tiba di kertajati, langsung diangkut bus rspad |
| 3 | Minggu, 01 Mar 2020 23:37 WIB | heboh di pangandaran, pemancing ini 'strike' dua marlin hitam |
| 4 | Minggu, 01 Mar 2020 23:33 WIB | ini alasan dipilihnya pelabuhan indramayu untuk evakuasi abk diamond princess |
| 5 | Minggu, 01 Mar 2020 23:30 WIB | polisi pastikan nelayan situbondo yang tewas depan warung korban kecelakaan |
| 6 | Minggu, 01 Mar 2020 23:15 WIB | evakuasi wni abk diamond princess dari indramayu-pulau sebaru ditempuh 15 jam |

Figure 3. Data retrieval results

From the results of data retrieval carried out for 8 months, the total data obtained was 106.209 news headlines with the amount of data each month as follows:

Table 1. : Amount of data retrieved (News Title)

| *Number* | *Mounth* | *Number of News Headlines* |
|---|---|---|
| 1 | March | 13.701 |
| 2 | April | 13.210 |
| 3 | May | 13.071 |
| 4 | June | 12.705 |
| 5 | July | 13.617 |
| 6 | August | 12.705 |
| 7 | September | 13.427 |
| 8 | October | 13.773 |
| Total Amount of Data | | 106.209 |

**Preprocessing**

Text preprocessing is the stage for preparing text into data to be processed includes case folding, tokenizing, and removing stopwords. Case folding is needed in converting the entire text in the document to lowercase. Tokenizing breaks a set of characters in a text into word units, whitespace characters such as enter, tabulation, spaces are considered as word separators. Removing stopwords is also called filtering, which is taking important words from the token output. While the stopwords removal itself is done to eliminate high frequency words that can be found in documents.

In this research, the input is a documents or strings. In general, this process has several stages, namely lemmatizing, case folding, tokenizing, stop word removal, stemming, and others. Lemmatizing process is a process to return a word to a root word. The preprocessing in this study did the elimination of stop words and left words with a prefix and a suffix. This is because the difference in accuracy between the text that is deleted by the stop word and not deleted is too high (Hidayatullah, 2016).

The preprocessing process in this study uses the Python Sastrawi library which is one of the libraries used in the Indonesian stemming process (Python Sastrawi, n.d.). This library is a development of the Sastrawi PHP Library (Sastrawi, n.d.).

Table 2. Preprocessing result data

| *No* | *Date* | *Title* | *Remove Stop Words* |
|---|---|---|---|
| 465 | Rabu 01 Apr 2020 0819 WIB | Tak Cuma Main HP Pemobil Tabrak Pria hingga Tewas di Karawaci Juga Mabuk | tak cuma main hp pemobil tabrak pria hingga tewas karawaci juga mabuk |
| 466 | Rabu 01 Apr 2020 0818 WIB | Isu Pandemi Corona Dibawa-bawa Napi Biar Bisa Keluar Bui | isu pandemi corona dibawa-bawa napi biar bisa keluar bui |
| 467 | Rabu 01 Apr 2020 0810 WIB | Ini Daftar Daerah yang Isolasi Mandiri untuk Hadapi Pandemi | ini daftar daerah isolasi mandiri hadapi pandemi |
| 468 | Rabu 01 Apr 2020 0809 WIB | Pj Walkot Makassar Jelaskan Sebab Jenazah Korban COVID-19 Ditolak di TPU | pj walkot makassar jelaskan sebab jenazah korban COVID-19 ditolak tpu |
| 469 | Rabu 01 Apr 2020 0806 WIB | Tes Darah Terbaru Bisa Prediksi 50 Jenis Kanker | tes darah terbaru bisa prediksi 50 jenis kanker |
| 470 | Rabu 01 Apr 2020 0751 WIB | Kasus Corona Diprediksi Capai 8.000 Lebih Seperti Apa Kapasitas Kesehatan RI? | kasus corona diprediksi capai 8.000 lebih seperti apa kapasitas kesehatan ri? |

**Topic Modeling with Latent Dirichlet Allocation (LDA)**

Topic Modeling is a word recognition model for finding topic recognition patterns by extracting text data to find themes from these data based on statistics. Latent refers to anything that is hidden in the data. Dirichlet is the distribution of topics in documents and distribution of words in topics. Allocation means allocating topics into documents and wording documents for topics. Therefore, this algorithm is called Latent Dirichlet Allocation (LDA). LDA is a generative probabilistic model of a collection of writings called a corpus. The basic idea proposed by the LDA method is that each document is represented as a random mixture of hidden topics, where each topic has a character that is determined based on the distribution of words contained in it. (Blei et al., 2003).

In text data, a token is a multi-word word or term grouped in a document. The language model is based on the observation that in natural language, some phrases are more common than others. In language modeling, the goal is to use a probability distribution over a string to represent text. This string is assumed to be drawn from a fixed set of tokens. This token set can consist of letters, numbers, other symbols, and spaces used in a particular language. In this case,

this sequence of symbols is called the n-gram character (Shafiei, 2009). At this stage, the sequential grouping of words that often appear simultaneously are grouped together in one unit. This process forms the bigram and trigram models, which are a combination of two and or three words in one sentence.

Topic Modeling, especially the LDA algorithm, requires two inputs, namely a dictionary and a corpus. The two inputs needed to speed up processing are also part of building the model itself. A dictionary is created to set a unique id for each word in a document. Furthermore, converting the dictionary into a bag of words is called a corpus which is useful for training topic models (Mohammed & Al-Augby, 2020).

Summarizing the news titles that have been collected every month to find out what news topics are published every month during the COVID-19 pandemic in Indonesia. Topic modeling is carried out using the Latent Dirichlet Allocation algorithm.

**Analysis of Results**

This result analysis stage is carried out to determine the results of implementing the Latent Dirichlet Allocation algorithm in analyzing the modeling of topics that most interpret humans using an evaluation model. The output of the topic model is the final result of the topic that is formed after analyzing the model evaluation. The output of this model can be visualized in tables or diagrams to make it easier to read.

## RESULTS AND DISCUSSION

### 1. Best Topic Number Search

The steps for finding the best number of topics are described in Figure 4. To get the best topic results, the results of the topic formation will be tested by calculating the amount of Coherence Values. Coherence Values is a ranking of coherence or interpretability of a set of words produced by the processing of modeling topics (Newman et al., 2010). From these coherence values, it can be determined how many of the best topics are generated and presented in graphic form. Next, the highest number of values and the highest difference is taken. From the results of the coherence score in the form of a graph, the highest number of values and the high difference will be sought, as shown in the following figure. The higher the coherence value, the better it is with human interpretation (Röder et al., 2015).

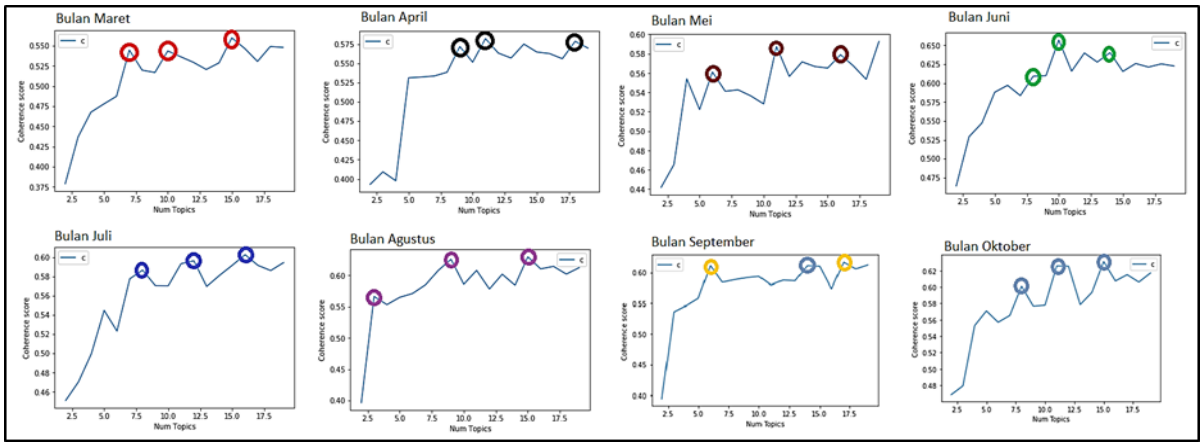


Figure 4. Displays the coherence score in graphic form

The points marked in figure 4 are predicted to represent each topic that is formed. A collection of news headlines each month on average gives the same number of topics, namely 3.There are 3 sharp angles that are formed, but are close to other angles, so that in general, the best number of topics formed is 3 topics for the monthly news collection from March 2020 to October 2020.

**2. Search for the most dominant themes for each topic**

After knowing the best number of topics that can be formed based on the coherence score above, the next process is to divide a collection of news headlines into 3 topics using the Latent Dirichlet Allocation (LDA) method. The best number of topics will be allocated to a set of news headlines based on the coherence score, namely 3 topics. The resulting topic will be represented by a group of words that represent the characteristics of each topic and will be a reference for a news title entering the topic which is notated by labeling 0, 1 and 2 as a marker for entry into the topic which is based on the probability value as presented in the picture 7 in the following:

```
Topic: 0 Word: 0.006*"uu_cipta" + 0.004*"warga" + 0.004*"omnibus_law" +
0.003*"protokol_kesehatan" + 0.003*"prancis" + 0.003*"rapid_test" +
0.003*"demo" + 0.003*"uu" + 0.003*"macron" + 0.003*"kerja"
================
Topic: 1 Word: 0.007*"gempa_m" + 0.005*"demo_omnibus" + 0.004*"hari_ini" +
0.004*"libur_panjang" + 0.004*"ini" + 0.004*"gempa" + 0.004*"kasus" +
0.003*"maulid_nabi" + 0.003*"kebakaran_kejagung" + 0.003*"hari"
================
Topic: 2 Word: 0.005*"libur_panjang," + 0.004*"libur_cuti" +
0.004*"kasus_corona" + 0.004*"tak_ada" + 0.004*"corona" + 0.003*"polisi" +
0.003*"sumpah_pemuda" + 0.003*"vaksin_corona" + 0.003*"tak" + 0.003*"kasus"
================
```

Figure 5. A list of words that represent a formed topic

This process will produce the most dominant set of words in each topic. This is needed to determine what themes represent each formed topic. From the list of the most dominant words on each topic, then the theme of each topic will be determined which is formed from a set of news headlines between March-October. Determining the theme in the form of several words or sentences will represent the name of the topic. The following is a theme that is formed from each topic set of news headlines per month.

Table 3. The most dominant word for each topic

| *MOUNTH* | *TOPIC* | *THEME* |
|---|---|---|
| March | 0 | Virus corona, pencegahan, tenaga medis |
| | 1 | Pasien, Positif corona, meninggal dunia |
| | 2 | Pasien Positif, Desinfektan |
| April | 0 | Rapid test, tenaga medis |
| | 1 | Larangan mudik, psbb |
| | 2 | Pasien Positif, pandemi, warga terdampak |
| May | 0 | New Normal, pandemi |
| | 1 | Kasus Positif, Corona |
| | 2 | New Normal, Shalat ied |
| June | 0 | Protokol kesehatan, mobil via vallen |
| | 1 | New Normal |
| | 2 | Positif corona, rapid test |
| July | 0 | Idul Adha, Positif COVID19 |
| | 1 | Djoko Tjandra, Shalat Ied |
| | 2 | Positif corona, kasus positif |
| August | 0 | Polsek ciracas, jaksa pinangki |
| | 1 | Positif COVID, Vaksin corona |
| | 2 | Penyerangan polsek, djoko tjandra |
| September | 0 | Kasus Corona, Pilkada |
| | 1 | Protocol kesehatan |

| | | |
|---|---|---|
| | 2 | Pasien Positif Corona |
| Oktober | 0 | Omnibus Law, Protocol kesehatan |
| | 1 | Demo omnibbus, libur panjang |
| | 2 | Libur cuti, Kasus corona |

The next step is to label the most dominant topic in each headline. This section is needed so that each news title is included in one of the existing topics. This labeling process is carried out by referring to the previously formed LDA model.

Labeling is given to each headline by adding parameters: `Dominant_Topic` contains the most dominant topic in the sentence. `Perc_Contributian` gives the number of proximity to the most dominant topic in the form of a percentage. `Topic_Keywords` provide keyword information from the topic. An example of the results of a news title modeling topic is shown in table 4.

Table 4. Determination of the topic of each title

| Document_No | Dominant_Topic | Topic_Perc_Contrib | Keywords | Word | Basic Word |
|---|---|---|---|---|---|
| 0 | 0 | 0.6966 | uu_cipta, warga, omnibus_law, protokol_kesehat... | [jokowi, tinjau, progres, wisata, premium, lab... | jokowi tinjau progres wisata premium labuan ba... |
| 1 | 1 | 0.7970 | gempa_m, demo_omnibus, hari_ini, libur_panjang... | [update, COVID-19, jatim:, 314, kasus, positif... | update COVID-19 jatim: 314 kasus positif baru,... |
| 2 | 1 | 0.5937 | gempa_m, demo_omnibus, hari_ini, libur_panjang... | [azerbaijan, vs, armenia, perang,, kemlu, ri:,... | azerbaijan vs armenia perang, kemlu ri: semua ... |
| 3 | 0 | 0.7539 | uu_cipta, warga, omnibus_law, protokol_kesehat... | [perkumpulan, warga, minang, surabaya, dukung,... | perkumpulan warga minang surabaya dukung machf... |
| 4 | 0 | 0.7891 | uu_cipta, warga, omnibus_law, protokol_kesehat... | [azerbaijan, vs, armenia, perang,, ri, serukan... | azerbaijan vs armenia perang, ri serukan genca... |

### 3. Testing the topic labeling of each news title

The accuracy of labeling by the model made needs to be tested for its level of accuracy. For this reason, a configuration matrix is used to test the accuracy of the labeling results carried out by the machine. A configuration matrix is a table used to describe the performance of a classification or classifier model on a set of test data whose true value is known. Accuracy measures are interpreted as the probability of encountering a certain type of misclassification or that the correct classification should be selected in preference to uninterpreted measurements (Stehman, 1997). The confirmation matrix will produce two variables that will be used to calculate precession, recall and accuracy, namely the y_test variable and the predicted variable. The results of precession, recall and accuracy are displayed on the print Acuracy Score and print Classification Report commands shown in the figure 6:

```
Accuracy Score :  0.8418568056648308

Classification Report :
              precision    recall  f1-score   support

           0       0.84      0.83      0.83       436
           1       0.84      0.86      0.85       438
           2       0.85      0.84      0.84       397

    accuracy                           0.84      1271
   macro avg       0.84      0.84      0.84      1271
weighted avg       0.84      0.84      0.84      1271
```

Figure 6. The results of the calculation of accuracy score in classification reports

## 4. Analysis of Results

The results analysis was carried out by calculating the number of titles collected on each formed topic, so that the most dominant news topics and how the news models published by the online media www.detik.com were published during the COVID-19 pandemic in Indonesia during the first 8 months since the first case was found in Indonesia. The results of the calculation of the number of news headlines on each topic are then combined with the themes formed and presented in table 5. The number of results for topic modeling news headlines each month is shown in table 5.

Table 5. The number of results for topic modeling news headlines each month

| MOUNTH | TOPIC | THEME | TOTAL | PERCENTAGE |
|---|---|---|---|---|
| March | 0 | Virus corona, pencegahan, tenaga medis | 4825 | 35,21% |
| | 1 | Pasien, Positif corona, meninggal dunia | 4701 | 34,31% |
| | 2 | Pasien Positif, Desinfektan | 4176 | 30,48% |
| April | 0 | Rapid test, tenaga medis | 4007 | 30,33% |
| | 1 | Larangan mudik, psbb | 4537 | 34,34% |
| | 2 | Pasien Positif, pandemi, warga terdampak | 4667 | 35,33% |
| May | 0 | New Normal, pandemi | 4195 | 32,09% |
| | 1 | Kasus Positif, Corona | 4523 | 34,60% |
| | 2 | New Normal, Shalat ied | 4354 | 33,31% |
| June | 0 | Protokol kesehatan, mobil via vallen | 3833 | 31,26% |
| | 1 | New Normal | 4131 | 33,69% |
| | 2 | Positif corona, rapid test | 4299 | 35,06% |
| July | 0 | Idul Adha, Positif COVID19 | 4298 | 31,56% |
| | 1 | Djoko Tjandra, Shalat Ied | 4344 | 31,90% |
| | 2 | Positif corona, kasus positif | 4976 | 36,54% |
| August | 0 | Polsek ciracas, jaksa pinangki | 4355 | 34,28% |
| | 1 | Positif COVID, Vaksin corona | 4240 | 33,37% |
| | 2 | Penyerangan polsek, djoko tjandra | 4111 | 32,35% |
| September | 0 | Kasus Corona, Pilkada | 4437 | 33,05% |
| | 1 | Protocol kesehatan | 4403 | 32,79% |
| | 2 | Pasien Positif Corona | 4586 | 34,16% |
| October | 0 | Omnibus Law, Protocol kesehatan | 4639 | 33,68% |
| | 1 | Demo omnibbus, libur panjang | 4633 | 33,64% |
| | 2 | Libur cuti, Kasus corona | 4500 | 32,67% |

The results of this labeling are then tested for accuracy with confussion matrix. The complete results of this test on the results of this topic modeling are presented in table 6 below:

Table 6. The accuracy value for confusion matrix testing

| *Mounth* | *Accuracy* |
|---|---|
| March | 0,806 |
| April | 0,856 |
| May | 0,819 |
| June | 0,823 |
| July | 0,805 |
| Augusts | 0,841 |
| September | 0,812 |
| October | 0,826 |
| ***Average*** | ***0,824*** |

After the topic labeling is carried out on all news headlines, then a test is carried out on topic labeling to determine the accuracy of topic labeling carried out by LDA using a confusion matrix. The confussion matrix is a table that is used to describe the performance of the classification model or classifier on a set of test data whose actual value is known. Measuring accuracy is interpreted as the probability of encountering a particular type of misclassification or the correct classification should be chosen in preference to measurements that cannot be

interpreted (Stehman, 1997). The following is the confussion matrix program code in python to calculate the quality of the labeling results using the confusion matrix:

```
count_vect = CountVectorizer()
countsv = count_vect.fit_transform(data['Text'])
transformer = TfidfTransformer().fit(countsv)
countsv = transformer.transform(countsv)

X_train, X_test, y_train, y_test = train_test_split(countsv, data['Dominant_Topic'],
test_size=0.1, random_state=69)
model = MultinomialNB().fit(X_train, y_train)
predicted = model.predict(X_test)

print("Acuracy Score : ",accuracy_score(y_test, predicted))
print("Classification Report :", classification_report(y_test, predicted))
```

Figure 7. The Confussion matrix calculation program code

The program code in Figure 7 will generate two variables that are used to calculate precession, recall, and accuracy, namely the y_test variable and the predicted variable. The results of precession, recall, and accuracy are displayed in the print Accuracy Score and print ClassificationReport commands. The results of the calculation of the confusion matrix for the topic modeling of news headlines each month as presented in figure 8 below:

MARET

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.77 | 0.88 | 0.83 | 471 |
| 1 | 0.88 | 0.72 | 0.79 | 426 |
| 2 | 0.79 | 0.81 | 0.80 | 474 |
| accuracy | | | 0.81 | 1371 |
| macro avg | 0.81 | 0.80 | 0.81 | 1371 |
| weighted avg | 0.81 | 0.81 | 0.81 | 1371 |

APRIL

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.86 | 0.82 | 0.84 | 393 |
| 1 | 0.85 | 0.88 | 0.86 | 467 |
| 2 | 0.86 | 0.86 | 0.86 | 462 |
| accuracy | | | 0.86 | 1322 |
| macro avg | 0.86 | 0.85 | 0.86 | 1322 |
| weighted avg | 0.86 | 0.86 | 0.86 | 1322 |

MEI

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.82 | 0.80 | 0.81 | 418 |
| 1 | 0.80 | 0.84 | 0.82 | 469 |
| 2 | 0.84 | 0.81 | 0.83 | 421 |
| accuracy | | | 0.82 | 1308 |
| macro avg | 0.82 | 0.82 | 0.82 | 1308 |
| weighted avg | 0.82 | 0.82 | 0.82 | 1308 |

JUNI

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.87 | 0.77 | 0.82 | 391 |
| 1 | 0.84 | 0.83 | 0.84 | 422 |
| 2 | 0.78 | 0.86 | 0.82 | 414 |
| accuracy | | | 0.82 | 1227 |
| macro avg | 0.83 | 0.82 | 0.82 | 1227 |
| weighted avg | 0.83 | 0.82 | 0.82 | 1227 |

JULI

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.83 | 0.75 | 0.79 | 444 |
| 1 | 0.81 | 0.79 | 0.80 | 423 |
| 2 | 0.78 | 0.86 | 0.82 | 495 |
| accuracy | | | 0.81 | 1362 |
| macro avg | 0.81 | 0.80 | 0.80 | 1362 |
| weighted avg | 0.81 | 0.81 | 0.80 | 1362 |

AGUSTUS

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.84 | 0.83 | 0.83 | 436 |
| 1 | 0.84 | 0.86 | 0.85 | 438 |
| 2 | 0.85 | 0.84 | 0.84 | 397 |
| accuracy | | | 0.84 | 1271 |
| macro avg | 0.84 | 0.84 | 0.84 | 1271 |
| weighted avg | 0.84 | 0.84 | 0.84 | 1271 |

SEPTEMBER

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.82 | 0.81 | 0.81 | 433 |
| 1 | 0.83 | 0.76 | 0.79 | 417 |
| 2 | 0.80 | 0.86 | 0.83 | 493 |
| accuracy | | | 0.81 | 1343 |
| macro avg | 0.81 | 0.81 | 0.81 | 1343 |
| weighted avg | 0.81 | 0.81 | 0.81 | 1343 |

OKTOBER

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.81 | 0.82 | 0.81 | 453 |
| 1 | 0.86 | 0.80 | 0.83 | 469 |
| 2 | 0.81 | 0.86 | 0.84 | 456 |
| accuracy | | | 0.83 | 1378 |
| macro avg | 0.83 | 0.83 | 0.83 | 1378 |
| weighted avg | 0.83 | 0.83 | 0.83 | 1378 |

Figure 8. The precision and recall values

From the results of testing the accuracy of the distribution of modeling topics as shown in figure 8, it can be conveyed that the accuracy generated on the topic modelling of detik.com news title for 8 months

with the LDA algorithm, has an accuracy rate above 80% with the highest accuracy produced in April 2020 by 86%. The f1-score value (comparison of precision with recall) is also mostly above 80%.

## CONCLUSIONS AND RECOMMENDATIONS

Based on the research that has been done, it was found that there is a tendency to form three news topics every month with a relatively balanced distribution of the number of news on each topic, an average of 33% for each topic. From 33% of news headlines, every March to October 2020 is always associated with the words: "kasus corona", "positif corona", "positif covid", "covid19". The news in March-June 2020 was dominated by news related to the COVID-19 pandemic on each topic that was formed. The word new normal began to dominate the headlines in May and June. Last July, news emerged outside of the COVID-19 pandemic, namely the case of "Djoko Tjandra" and "Jaksa Pinangki". The word "vaksin" began to dominate the news in August. The results of testing the accuracy of modeling topics using the confusion matrix shows that the accuracy level is always above 80% with the highest accuracy value on modeling topics in April of 85.6% and the lowest accuracy value in July of 80.5%.

## ACKNOWLEDGEMENT

This research was funded by the Interdisciplinary Basic Research for COVID-19 Response at UIN Sunan Kalijaga for the 2020 Fiscal Year

## REFERENCES

Bangun, E. P., A Koagouw, F. V. I., & Kalangi, J. S. (2019). Analisis isi unsur kelengkapan berita pada media online manadopostonline.com. *Acta Diurna Komunikasi*, *1*(3), 4–13.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). *Latent Dirichlet Allocation*. *3*, 993–1022.

Destarani, A. R., Slamet, I., & Subanti, S. (2019). Trend Topic Analysis using Latent Dirichlet Allocation (LDA) (Study Case: Denpasar People's Complaints Online Website). *Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika*, *5*(1), 50–58. https://doi.org/10.26555/jiteki.v5i1.13088

Handiyani, P., & Hermawan, A. (2017). Kredibilitas Portal Berita Online Dalam Pemberitaan Peristiwa Bom Sarinah Tahun 2016 (Analisis Isi Portal Berita Detik.com dan Kompas.com Periode 14 Januari-14 Februari 2016). *Jurnal Komunikasi*, *12*(1), 51–68. https://doi.org/10.20885/komunikasi.vol12.iss1.art4

Hidayatullah, A. F. (2016). Pengaruh Stopword Terhadap Performa Klasifikasi Tweet Berbahasa Indonesia. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *1*(1), 1–4. https://doi.org/http://dx.doi.org/10.14421/jiska.2016.11-01

Ja'far, A. N. (2018). *Topic Modeling of APP Review in Google Play Based on Latent Dirichlet Allocation*.

Keraf, G. (1980). Komposisi. *Flores: Nusa Indah*.

Kristanto, T. A. (2019). Media Cetak, Tak Cukup Dua Kaki. *Jurnal Dewan Pers*, *20*(November), 9–17.

Kurniawan, W. (2018). *Sistem Monitoing Pecakapan Pada Toko Online Menggunakan Metode Latent Dirichlet Allocation (LDA) Studi Kasus: Toko Online "BERRYBENKA.COM."*

Mohammed, S. H., & Al-Augby, S. (2020). LSA & LDA topic modeling classification: Comparison study on E-books. *Indonesian Journal of Electrical Engineering and Computer Science*, *19*(1), 353–362. https://doi.org/10.11591/ijeecs.v19.i1.pp353-362

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. *NAACL HLT 2010 - Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference*, *June*, 100–108.

Putra, I. M. K. B. (2017). *Analisis Topik Informasi Publik Media Sosial di Surabaya Menggunakan Pemodelan Latent Dirichlet Allocation (LDA)*.

*Python Sastrawi*. (n.d.). https://github.com/sastrawi/sastrawi

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, 399–408. https://doi.org/10.1145/2684822.2685324

*Sastrawi*. (n.d.). https://github.com/har07/PySastrawi

Shafiei, M. M. (2009). *Leveraging structural information for statistical topic models of text*.

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, *62*(1). https://doi.org/10.1016/S0034-4257(97)00083-7