Terakreditasi SINTA Peringkat 2
Surat Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti No. 10/E/KPT/2019
masa berlaku mulai Vol. 1 No. 1 tahun 2017 s.d Vol. 5 No. 3 tahun 2021

Terbit online pada laman web jurnal: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 4 No. 2 (2020) 336 – 344 ISSN Media Elektronik: 2580-0760

Analisis Topik Penelitian Kesehatan di Indonesia Menggunakan Metode *Topic Modeling* LDA (*Latent Dirichlet Allocation*)

Yoga Sahria[1], Dhomas Hatta Fudholi[2]
[1 2]Prodi Magister Informatika, Fakultas Teknologi Industri, Universitas Islam Indonesia
[1]17917225@students.uii.ac.id , [2]hatta.fudholi@uii.ac.id

Abstract

In this time, the need of research, the development and the implementation of the result of research in health is increasing both from the researchers, the government, the academic even of from the public general. One of the ways to find out the health research trend is by topic modeling. The method that used in this research is topic modeling LDA (Latent Dirichlet Allocation) method. The purpose of this research is to identify how modeling topic method LDA analyze modeling topic to some health research in Indonesia by Sinta Journal and to know how the coherence value in each topic of the model that has been made. Besides, hopefully it can be used as a reference to do heath research in Indonesia based the topic that has been modeled. The development of this research uses Anaconda3 Python Programming Language Tools and utilizes the LDA library that provided to get the topic model. To examine the result of this research the respondent are medical worker, health researcher and academics. The result of this research the topic modeling that used 94,1% respondent say very good and 5,9% say good.

Keywords: Latent dirichlet allocation, modeling topic, coherence value

Abstrak

Pada saat ini, kebutuhan penenelitian, pengembangan, dan penerapan hasil penelitian di bidang kesehatan semakin meningkat baik dari peneliti, pemerintah, akademis bahkan dari kalangan umum. Untuk mengetahui bagaimana tren penelitian di penelitian kesehatan salah satu cara yang dapat dilakukan adalah dengan melakukan pemodelan topik. Metode yang dipakai dalam penelitian ini yaitu metode *topic modelling* LDA *(Latent Dirichlet Allocation)*. Penelitian ini dilakukan dengan tujuan untuk mengidentifikasi bagaimana metode *topic modelling* LDA dapat melakukan analisis pemodelan topik terhadap judul-judul penelitian di bidang kesehatan di Indonesia yang diperoleh dari Jurnal SINTA dan melihat bagaimana nilai koherensi untuk setiap topik dari model yang telah dibuat. Selain itu juga diharapkan menjadi referensi dalam melakukan penelitian kesehatan di Indonesia berdasakan topik yang sudah dimodelkan. Pengembangan penelitian ini menggunakan tools bahasa pemograman *python anaconda3* dan memanfaatkan library LDA yang disediakan untuk memperoleh model topik. Untuk pengujian dari hasil penelitian ini responden yang terdiri tenaga medis, peneliti kesehatan, dan akademisi. Hasil dari penelitian ini *topic modelling* yang dilakukan 94,1% mengatakan sangat baik dan 5,9% mengatakan baik.

Kata kunci: *latent dirichlet allocation*, pemodelan topik, nilai koherensi.

1. Pendahuluan

Pada saat ini, kebutuhan penenelitian, pengembangan, dan penerapan hasil penelitian di bidang kesehatan dari waktu ke waktu semakin meningkat baik dari praktisi, pemerintah, akademis bahkan dari kalangan umum. Penelitian kesehatan sangat penting untuk diteliti untuk memperoleh informasi dan mengetahui temuan-temuan terbaru yang kemudian dapat dianalisis yang lebih mendalam untuk mengatahui tren topik penelitian kesehatan di Indonesia. Realitas yang terjadi pada saat ini perkembangan penelitian kesehatan di Indonensia menyebar sangat cepat ke ranah publik[1]. Penyebaran penelitian yang cepat ini terjadi karena adanya *Open Jurnal System* (OJS) yang terindexs jurnal SINTA yang dapat diakses dan dibaca siapapun secara *online*[2]. Dengan adanya sistem OJS Informasi kesehatan dapat diakases dengan mudah bagi semua kalangan. Namun pada permasalahannya adalah semakin banyaknya jumlah judul penelitian kesehatan yang terindex kedalam jurnal SINTA menimbulkan kesulitan bagi

Diterima Redaksi : 18-03-2020 | Selesai Revisi : 17-04-2020 | Diterbitkan Online : 20-04-2020

pembaca dan peneliti dalam mengidentifikasi suatu topik penelitian.

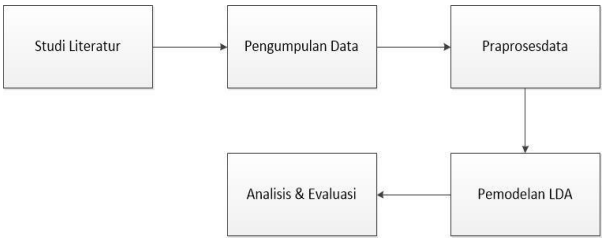
Dengan cara manual untuk mengetahui topik penilitan memerlukan banyak waktu untuk memeriksa semua judul penelitian kesehatan yang ada di Indonesia. Untuk mengatasi permasalahan tersebut pada penelitian ini memodelkan bagaimana untuk mengetahui Gambaran tren penelitian kesehatan di Indonesia, Salah satu cara yang diimplementasikan yaitu dengan metode LDA (*Latent Dirichlet Allocation*) untuk mengetahui tren topik penelitian kesehatan di Indonesia. LDA adalah sebuah metode *topic modelling* yang digunakan untuk menentukan pola pada sebuah dokumen yang dapat menghasilkan topik [3]. Dalam penelitian ini menerapkan metode *topic modelling* LDA yang bertujuan untuk melakukan analisis tren topik yang akan dihasilkan dan divisualisasikan sehingga lebih *informatif* dan mudah dipahami oleh pengguna. Hasil dari penelitian ini harapannya untuk memudahkan peneliti, akademis, dosen, praktisi dan kalangan umum untuk mengetahui Gambaran tren topik penelitian di bidang kesehatan di Indonesia. Hasil pemodelan topik yang dihasilkan dapat digunakan sebagai acuan sebagai pertimbangan dalam melakuakan penelitian kesehatan di Indonesia bagi peneliti, praktisi dan kalangan umum.

LDA telah banyak dikembangkan dalam penelitian terkait meganalisis topik dari sebuua teks maupun dokumen. Penelitian terkait dengan menggukan metode LDA banyak cara untuk menganalisis tren sebuah topik dengan berbagai sumber yaitu dapat mengakses dari *google scholar,wikipedia, twiter, instagram, facebook, jurnal* dan lain sebagainya. Penelitian yang dilakukan [4] menghasilkan pemodelkan topik berdasarkan data penulis jurnal dan penelitian yang pernah diteliti *author*. Sumber data yang didapatkan dari abstrak sebuah jurnal. Terdapat penelitian terkait [5] yang menganalisis tren *drug safety* hasil penelitian ini menampilkan topik penelitian yang populer berdasarkan tahun, persebaran topik, dan pengklasteran dan sumber yang diperoleh dalam penelitian ini hanya dari satu jurnal dan diambil dari masing-masing abstrak. Penelitian terkait yang dilakukan [6] bertujuan untuk memodelkan topik informasi yang mampu secara otomatis mengklasifikasikan pesan media social ke dalam topik-topik yang muncul dari hasil pemodelan. Sumber data penelitian tersebut diambil dari *twitter*. Penelitian selanjutnya yang menganalisis konten informasi dalam bentuk berita dihasilkan dengan jumlah yang sangat besar dari berbagai sumber dimedia setiap harinya. Penelitian ini bertujuan untuk memodelkan mengkombinasikan teknik *document clustering*[7]. Penelitian terkait selanjutnya menganalisis bagaimana melihat konten konten media sosial *e-commerce* sumber data yang diperoleh yaitu dari *instagram*. Tujuan penelitian melihat topik yang dibahas dengan melihat sentimen positif dan negatif secara keseluruhan yang terdapat dalam *e-commerce* instagram *shopee* [8]. Penelitian selanjutnya mengacu pada permasalahan semakin banyak informasi teks digital yang dihasilkan setiap harinya dari sumber Wikipedia. Penelitian ini menerapkan model topik LDA kemudian hasil dari penelitian ini menunjukan topik model untuk mengelompokkan dokumen dan menemukan dokumen yang serupa. Langkah-langkah dalam penelitian ini yaitu pengambilan data, pre-processing, pemodelan dan evaluasi [9]. Penelitian terkait selanjutnya melakukan ekstrasi topik untuk teks berbahasa Indonesia dan pengujian penelitian ini menunjukan metode LDA memiliki kinerja sangat baik dalam mencari dan melakukan ekstrasi topik untuk dokumen yang berbahasa Indonesia[10].

Berdasarkan uraian diatas, berbeda dengan penelitian-penelitian yang telah dilakukan sebelumnya, penelitian ini menggunakan data judul penelitian kusus di bidang kesehatan Indonesia. Sumber data dalam penelitian ini yaitu dari berbagai jurnal kesehatan di Indonesia yang terakreditasi nasional yang terideks di dalam SINTA. Keterbaruan dalam penelitian ini tidak hanya menampilkan bobot topik saja akan tetapi juga dapat memvisualisaikan hasil topik sehingga memudahkan pengguna mengetahui sebaran kata dan frasa di setiap topik. Berdasarkan penelitian dengan metode LDA yang telah diuraikan. Metode LDA dapat menyelesaikan permasalahan yang berkaitan tentang pemodelan topik. Metode LDA menjadi salah satu solusi untuk mengatasi pengelompokaan term majadi topik tertentu dengan memperhatikan urutan kata pada proses pembentukkanya melalui mekanisme model campuran (*mixture*) [11]. Oleh karena itu pada penelitian ini peneliti menggunakan pemodelan LDA untuk analisis topik penelitian kesehatan di Indonesia.

2. Metode Penelitian

Metode penelitian pada penelitian ini yaitu studi literatur, pengumpulan data, Praproses Data, Pemodelan LDA, Analisis dan Evaluasi adapun langkah-langkahnya dapat dilihat pada Gambar 1.



Gambar 1. Metode Penelitian

Langkah pertama studi literatur yaitu dengan melakukan studi buku, jurnal, internet dan media yang relevan yang membahas topik modeling LDA. Dalam penelitian ini penulis memaparkan tujuh penelitian terkait yang relevan dengan permasalahan yang akan

diteliti tentang *topic modeling* menggunakan metode LDA. Informasi yang didapatkan dari studi literatur akan digunakan untuk acuan teori dan pembahasan penelitian.

Pengumpulan Data

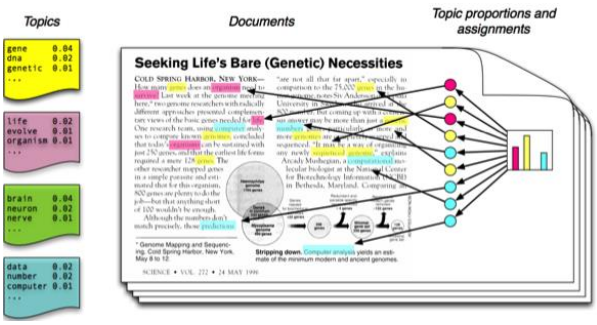
Dalam penelitian ini data yang digunakan yaitu berupa data primer. Sumber data yang diperoleh dalam penelitian ini yaitu berupa judul penelitian khusus dibidang kesehatan di Indonesia, metode untuk mendapatkan data tersebut dengan cara *scraping* data dari jurnal SINTA. Cara scraping data dengan menggunakan id masing-masing nama jurnal khusus di bidang kesehatan. Data diperoleh pada bulan januari 2020. Jumlah judul penelitian kesehatan yang berhasil di *scraping* sebanyak 11269 penelitian. Variabel data yang akan digunakan dalam penelitian ini yaitu judul dan author penelitian khusus di bidang kesehatan. Data diperoleh dengan proses *scraping* data menggunakan bahasa pemograman *python 3*. Menurut [12] proses *web scraping* metode yang baik untuk memperoleh data dari sebuah dokumen di Internet dengan *realtime*. Data yang diperoleh dari *scraping* akan diolah menjadi tersetruktur untuk dimodelkan yaitu pada tahap praposesing.

Praproses data

Pada tahap praprosesing data pertama yaitu *loading data* digunakan untuk memasukan data penelitian kesehatan Indonesia dari hasil *scraping* dalam bentuk CSV kebahasa pemograman *python*. Langkah kedua yaitu *data cleaning* bertujuan untuk menghapus data yang tidak diperlukan dalam pemodelan. Langkah ketiga yaitu *exploratory data* yang tujuannya untuk memveriifikasi data yang diperlukan sudah siap untuk dimodelkan dengan *wordcloud*, Proses ini semua dilakukan dengan bantuan software *Jupiter Notebook Anaconda3*.

2.4. Pemodelan LDA

Dasar ide pemodelan *topic modeling* yaitu sebuah topik yang terdiri dari kata-kata tertentu yang dapa menyusun topik tersebut dari dokumen-dokumen[13]. Pemodelan topik pada penelitian ini yaitu digunakan untuk menemukan topik dalam penelitian kesehatan di Indonesia berdasarkan judul jurnal. Pemodelan topik dapat mengGambarkan makna dari dokumen secara semantik yang tersembunyi dalam teks yang jumlahnya besar dan dapat menemukan informasi dari data teks yang tidak tersetruktur. Menurut blei [13] Gambar 2 di bawah ini menjelaskan cara kerja LDA.



Gambar 2. Cara kerja LDA

Pada Gambar 2 menurut blei LDA mengasumsikan proses generatif berikut adalah rumus matematis untuk setiap dokumen w dalam sebuah corpus **D** adalah sbb:

1. Pilih $N \sim Pissson(x)$,
2. Pilih $\boldsymbol{\theta} \sim Dir(a)$,
3. Untuk setiap N kata w_n ,
   a. Pilih Topik $z_n \sim Multinomial(\boldsymbol{\theta})$,
   b. Pilih sebuah kata w_n dari $p(w_n \mid z_n, \beta)$.

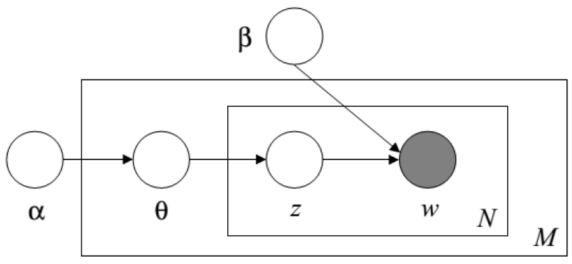
Beberapa asumsi penyederhanaan yang dibuat didalam distribusi dari (latent) topik bahwa diketahui mengikuti k distribusi Dirichlet. Kedua, probabilitas kata adalah matriks $\boldsymbol{\beta}$ berukuran k x V yang mana $b_{ij} = p(w^j = 1 \mid z^i = 1)$. Sedangkan k sebagai distribusi Dirichlet memiliki fungsi densitas dapat dilihat pada persamaan (1) sebagai berikut:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_I} \theta_1^{\alpha_1 - 1} \dots \theta_k^{\alpha_k - 1} \quad (1)$$

Adapun bentuk dalam distribusi bersama dari Topik *mixture* θ dari N topik z dan N kata w besyarat α dan β dapat dilihat pada persamaan (2) sebagai berikut:

$$p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n-1}^{N} p(z_n|\theta)\, p(w_n|z_n, \beta) \quad (2)$$

Bentuk dari representasi model LDA dapat diGambarkan dalam sebuah diagram dapat dilihat pada Gambar 3.



Gambar 3. Representasi Model LDA

Bentuk distribusi marginal dari p(**w** $|\alpha,\beta$) didapat dengan menintegrasikan persamaan (2) terhadap θ dapat menghasilkan persamaan (3):

$$p(\boldsymbol{w}|\alpha, \beta) = \int (\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta)\, p(w_n|z_n, \beta) \right) d\theta \quad (3)$$

Akhirnya, diperoleh perkalian densitas marginal untuk sebuah dokumen yang akan memperoleh probabilitas marginal sebuah corpus persamaan (4) sebagai berikut:

$$p(\boldsymbol{D}|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d} \sum_{z_n} p(z_{dn}|\theta_{dn},\beta)\right) d\theta_d \quad (4)$$

2.5. Analisis dan Evaluasi

Pada tahap analisis dan evaluasi ini yaitu dengan analisis kualitatif dengan memperhatikan *term-term* hasil pengelompokan topik yang sudah terbentuk. Hasil dari pengelompokan topik yang telah didapat kemudian dianalaisis untuk memberikan makna terhadap data. Pada tahap evaluasi penelitian ini yaitu dengan melihat penilaian tentang manfaat, nilai, atau keseimbangan hasil penelitian. Fokus evaluasi pada penelitian ini adalah evaluasi topik terhadap penelitian kesehatan di Indonesia.

3. Hasil dan Pembahasan

Dalam melakukan analisis *topic modelling* penelitian kesehatan di Indonesia dengan menggunakan *Latent Dirichlet Allocation* (LDA), Berikut tahap-tahap implementasi yang dilakukan sebagai berikut:

3.1. Target Jurnal SINTA

Tahap menentukan target jurnal ini yaitu dengan mengakes Jurnal SINTA http://sinta.ristekbrin.go.id/ kemudian mengeklik *source journal* pada menu jurnal SINTA, kemudian mencari penelitian kesehatan dengan menggunakan *keyword* "kesehatan" pada kolom pencarian jurnal. Hasil pencarian Jurnal kesehatan di Indonesia terdapat 172 nama jurnal dari berbagai afiliasi. Adapun peneliti memilih jurnal penelitian kesehatan yang dijadikan untuk dimodelkan topik sebanyak 30 jurnal. Pemilihan target jurnal yang dipilih berdasarkan *Impact*, *HS-Index*, *Citations*, *H-Index* yang tinggi dan lebih banyak sehingga jurnal tersebut dikatan jurnal yang terkakreditasi nasional yang berkualitas. Adapun nama jurnal target penelitian kesehatan di Indonesia yang akan di *scraping* berdasarkan id dapat dilihat pada Tabel 1.

Tabel 1. Target *Scraping* Jurnal

No	Nama Jurnal
1	Jurnal Ekologi Kesehatan
2	Jurnal Kesehatan Masyarakat
3	Media Penelitian dan Pengembangan Masyakat
4	Jurnal Kesehatan Lingkungan Indonesia
5	Medisains : Jurnal Ilmiah Ilmu - Ilmu Kesehatan
6	Jurnal Administrasi Kesehatan Indonesia
7	Media Kesehatan Masyarakat Indonesia
8	Buletin Penelitian Kesehatan
9	Buletin Penelitian Sistem Kesehatan
10	Berkala Ilmu Kesehatan Kulit dan Kelamin
11	Jurnal Kesehatan Lingkungan
12	JKKI : Jurnal Kedokteran dan Kesehatan Indonesia
13	Jurnal Kesehatan Reproduksi
14	Jurnal Endurance: Kajian Ilmiah Problema Kesehatan
15	Jurnal Vokasi Kesehatan
16	Jurnal Ilmu Kesehatan Masyarakat
17	Jurnal Kebijakan Kesehatan Indonesia : JKKI
18	Jurnal Kesehatan Prima
19	Jurnal Epidemiologi Kesehatan Komunitas
20	Jurnal Kesehatan Manarang
21	Jurnal Info Kesehatan
22	Jurnal Kesehatan Andalas
23	Jurnal Ilmiah Ibnu Sina (JIIS): Ilmu Farmasi dan Kesehatan
24	STRADA Jurnal Ilmiah Kesehatan
25	Mutiara Medika: Jurnal Kedokteran dan Kesehatan
26	Jurnal Promosi Kesehatan Indonesia
27	Jurnal Ilmu Kesehatan Masyarakat
28	Jurnal Kesehatan
29	Window of Health : Jurnal Kesehatan
30	BioLink (Jurnal Biologi Lingkungan, Industri, Kesehatan)

3.2. *Scraping* Data

Scraping Data diambil dari web Jurnal SINTA berikut cuplikan listing program code pengambilan data kusus penelitian kesehatan di Indonesia:

Program Scraping

```
#WEB SCRAPING  Judul & Author JURNAL SINTA
from bs4 import BeautifulSoup
import requests

page=requests.get('http://sinta2.ristekdikti.go.id/authors')
soup = BeautifulSoup(page.text, 'html.parser')

judul = list()
author = list()
  for i in page:
      print(i, end=" ")
      page =
requests.get('http://sinta2.ristekdikti.go.id/authors')
      soup = BeautifulSoup(page.text, 'html.parser')
      table = soup.find('tbody')
      tr_list = table.find_all("tr")
  for tr in tr_list:

judul.append(tr.find("dt").get_text().replace("\n",""))
author.append(tr.find("dd").get_text())
end for
```

Dari listing program tersebut menggunakan modul library python *BeautifulSoup* yang digunakan untuk menavigasikan dan memparser DOM (*Document Object Model*). Dengan librari *BeutifullSoup* dapat mengambil data dengan mudah melakukan proses pencarian, navigasi, modifikasi struktur data pada situs Jurnal SINTA. Berikut cuplikan hasil *scraping data* dapat dilihat pada Tabel 2.

3.3. Preposesing Data

Implementasi preposesing data sebelum menganalis topik menggunakan LDA yaitu dengan mensetrukturkan, merapikan, dan memasisikan data siap dianalsis. Prepocessing berfungsi untuk membersihkan data yang dilakukan untuk menghindari

data yang tidak sempurna, data bermasalah, dan data yang tidak konsisten[14]. *Preprocessing* data pada penelitian ini dilakukan 5 tahap secara urut yaitu menghapus tanda baca, menghapus angka diantara spasi, *case folding*, Menghapus kalimat yang terdiri kurang dari atau sama dengan tiga kata, Menghapus *stopword*. Berikut script dan cuplikan Gambar hasil *prepocessing* data dapat dilihat pada Tabel 3.

Tabel 2. Cuplikan Hasil *scraping* Judul dan Author

No	Judul	Author
1	Analisis kualitatif bakteri koliform pada depo air minum isi ulang di kota Singaraja Bali	DWIS Bali
2	Faktor-faktor yang mempengaruhi kejadian TB paru dan upaya penanggulangannya	HSP Manalu
3	Potensi daun pandan wangi untuk membunuh larva nyamuk Aedes aegypti	D Susanna, A Rahman, ET Pawenang
4	Hubungan Faktor Lingkungan Rumah dengan Penularan TB Paru Kontak Serumah	A Musadad
5	Habitat Perkembangbiakan Dan Aktivitas Menggigit Nyamuk Anopheles Sundaicus Dan Anopheles Subpictus Di Purworejo, Jawa Tengah	S Sukowati, S Shinta
dst	...	...

Program Preprocessing

```
#preprocessing
def removeStopword(str):
    stop_words=
set(stopwords.words('stopword_jurnal'))
    word_tokens = word_tokenize(str)
    filtered_sentence = [w for w in
word_tokens if not w in stop_words]
    return ' '.join(filtered_sentence)
#remove sentence which contains only one word
def removeSentence(str):
    word = str.split()
    wordCount = len(word)
    if(wordCount<=1):
        str = ''
    return str
def cleaning(str):
#remove non-ascii
    str = unicodedata.normalize('NFKD',
str).encode('ascii', 'ignore').decode('utf-8',
'ignore')
#remove URLs
    str =
re.sub(r'(?i)\b((?:https?://|www\d{0,3}[.]|[a-
z0-9.\-]+[.][a-
z]{2,4}/)(?:[^\s()<>]+|\(([^\s()<>]+|(\([^\s()
<>]+\)))*\))+(?:\(([^\s()<>]+|(\([^\s()<>]+\))
)*\)|[^\s`!()\[\]{};:\'".,<>?«»“”‘’]))', '',
str)
#remove punctuations
    str = re.sub(r'[^\w]|_',' ',str)
#remove digit from string
    str = re.sub("\S*\d\S*", "", str).strip()
#remove digit or numbers
    str = re.sub(r"\b\d+\b", " ", str)
#to lowercase
    str = str.lower()
#Remove additional white spaces
    str = re.sub('[\s]+', ' ', str)
    return str
def preprocessing(str):
    str = removeSentence(str)
    str = cleaning(str)
    str = removeStopword(str)
    return str
```

Adapun hasillnya dari script tersebut dapat dilihat pada Tabel 3.

Tabel 3. Cuplikan Hasil *Preprocessing*

No	Nama Jurnal
1	Analisis kualitatif bakteri koliform depo air minum isi ulang kota Singaraja Bali
2	Faktor faktor mempengaruhi kejadian TB paru upaya penanggulangannya
3	Potensi daun pandan wangi larva nyamuk Aedes aegypti
4	Hubungan Faktor Lingkungan Penularan TB Paru
5	Habitat Perkembangbiakan Aktivitas Menggigit Nyamuk Anopheles Sundaicus Anopheles Subpictus Purworejo Jawa Tengah
dst	.....

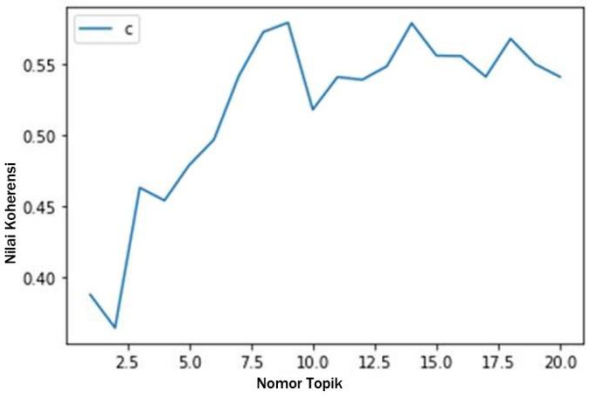
3.4 Pemodelan Topik LDA

Dalam penelitian ini untuk mengimplementasikan topik LDA dengan memanfaatkan *library* yang disediakan oleh *python* yaitu dengan menginstal *pip install lda*. langkah pertama yaitu menampilkan nilai kohorensi setiap kata pada judul penelitian di Indonesia adapun cuplikan script list program sebagai berikut:

Program Pemodelan LDA

```
pip install lda
import numpy as np
import lda
import lda.datasets
start=1
limit=21
step=1
model_list,coherence_values=
compute_coherence_values(dictionary,
corpus=corpus_tfidf,texts=text_list,
start=start, limit=limit, step=step)
#show graphs
%matplotlib inline
import matplotlib.pyplot as plt
x = range(start, limit, step)
plt.plot(x, coherence_values)
plt.xlabel("Num Topics")
plt.ylabel("Coherence score")
plt.legend(("coherence_values"), loc='best')
plt.show()
```

Adapun hasil script tersebut menampilkan kedalam diagram baris untuk melihat nilai kohorensi untuk setiap banyak topik dapat dilihat pada Gambar 4.



Gambar 4. Grafik garis nilai koherensi

Pada Gambar 4 tersebut mengGambarkan dari Tabel 4 dimana setiap topik yang dihasilkan menunjukan bahwa nilai kohorensi setiap topik mempunyai nilai yang berbeda. Adapun untuk melihat rincian nilai koherensi untuk setiap kata dari judul yang membentuk term-term topik berikut cuplikan script sebagai berikut:

```
Program Nilai Kohorensi
  from gensim.models.coherencemodel import
  CoherenceModel
  from gensim.models.ldamodel import LdaModel
  from gensim.corpora.dictionary import
  Dictionary
  from numpy import array
  #function to compute coherence values
  def compute_coherence_values(dictionary,
  corpus, texts, limit, start, step):
      coherence_values = []
      model_list = []
      for num_topics in range(start, limit,
  step):
          model= LdaModel(corpus=corpus,
  id2word=dictionary, num_topics=num_topics,
  iterations=100)
          model_list.append(model)
          coherencemodel =
  CoherenceModel(model=model, texts=texts,
  dictionary=dictionary, coherence='c_v')

  coherence_values.append(coherencemodel.get_c
  oherence())
  return model_list, coherence_values

  # Print the coherence scores
  for m, cv in zip(x, coherence_values):
      print("Nomor Topik =", m, " Nilai
  Kohorensi", round(cv, 6))
```

Hasil dari script yang menunjukan nilai kohorensi kata pada judul penelitian kesehatan di Indonesia dapat dilihat pada Tabel 4.

Tabel 4. Hasil Nilai Kohorensi Pada Setiap Topik

Topik	Nilai Kohorensi
T1	0.387961
T2	0.364622
T3	0.463102
T4	0.454111
T5	0.479004
T6	0.496821
T7	0.541324
T8	0.572496
T9	0.57902
T10	0.517935
T11	0.540853
T12	0.538913
T13	0.548448
T14	0.578639
T15	0.555881
T16	0.555559
T17	0.541021
T18	0.567721
T19	0.549875
T20	0.54096

Berdasarkan hasil Tabel 4, dapat dilihat topik yang mempunyai kohorensi/probabilitas paling tinggi adalah topik ke-8 dengan nilai 0.572496. Nilai probabilitas yang tinggi tersebut menunjukkan topik tersebut mempunyai mempunyai peluang paling tinggi untuk muncul dalam judul penelitian kesehatan di Indonesia. Pada Tabel 4 cacah topik dengan nilai koherensi tertinggi akan dipilih untuk dilakukan pemodelan topik dan dijadikan nilai untuk mengisi parameter *num_topics*. Untuk melihat persebaran kata pada setiap topik menggunakan algoritma TF-IDF (*Term Frequency – Inverse Document Frequency*) untuk pembobotan kata adapun script dan hasilnya dapat dilihat pada Gambar 5.

```
Program TF-IDF persebaran topik
  model = LdaModel(corpus=corpus_tfidf,
  id2word=dictionary, num_topics=3)
  #num topic menyesuaikan hasil dari coherence
  value paling tinggi

  for idx, topic in model.print_topics(-1):
      print('Topic: {} Word: {}'.format(idx,
  topic))
```

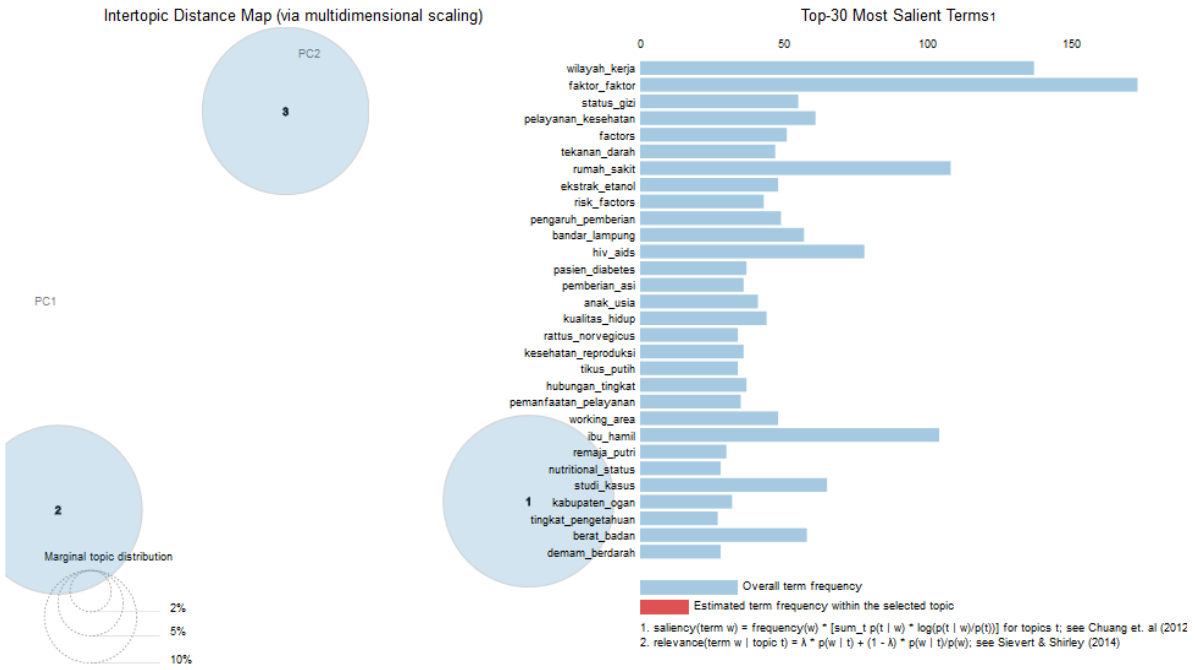
Hasil dari script tersebut yaitu dengan mencari topik yang dominan atau banyak muncul dengan algoritma TF-IDF dengan probabilitas kemunculan topik katanya yang dominan dalam judul penelitian kesehatan di Indonesia berdasarkan Tabel 4. Berikut adalah hasil pemodelan topik yang dihasilkan dapat dilihat pada Tabel 5 sebagai berikut:

Tabel 5. Hasil Nilai Kohorensi Pada Setiap Topik

Topik	Kata
T1	0.005*"status_gizi" + 0.005*"factors" + 0.005*"tekanan_darah" + 0.004*"risk_factors" + 0.004*"working_area" + 0.004*"analysis" + 0.004*"hubungan" + 0.003*"health" + 0.003*"rumah_sakit" + 0.003*"hubungan_tingkat"
T2	0.012*"faktor_faktor" + 0.011*"wilayah_kerja" + 0.008*"faktor" + 0.006*"ibu_hamil" + 0.005*"hiv_aids" + 0.005*"pelayanan_kesehatan" + 0.005*"kesehatan" + 0.005*"puskesmas" + 0.005*"studi_kasus" + 0.004*"ibu"
T3	0.007*"rumah_sakit" + 0.005*"bandar_lampung" + 0.005*"pasien" + 0.004*"pengaruh_pemberian" + 0.004*"ekstrak_etanol" + 0.004*"public_health" + 0.004*"anak_usia" + 0.004*"pengaruh" + 0.003*"pasien_diabetes" + 0.003*"pemanfaatan_pelayanan"

3.5 Visualisasi Pemodelan Topik

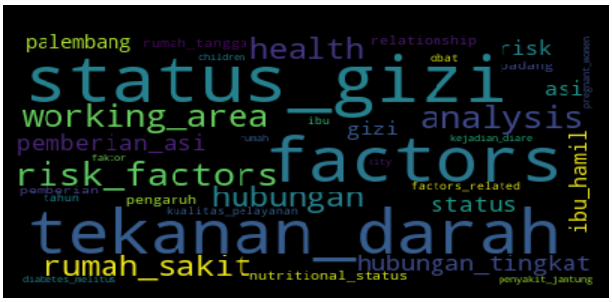
Visualisai pemodelan topik pada penelitin ini setelah selesai melakukan pemodelan topik LDA, model TF-IDF disimpan ke dalam bentuk *pyLDAvis* yang dapat membentuk visualisasi dari masing-masing topik dan kata yang paling banyak muncul dalam judul penelitian kedalam diagram adapun bentuk visualisasi dapat dilihat pada Gambar 5.



Gambar 5. Visualisasi *pyLDAvis*

Pada Gambar 5 mengGambarkan 30 kata penting yang muncul di *corpus*. Panel kanan mengGambarkan tentang kata-kata dominan yang dibahas topik dari dataset judul penelitian kesehatan di Indonesia. Berdasarkan panel kanan visualisasi, penampilan istilah Wilayah kerja, faktor-faktor, status gizi, pelayanan kesehatan, tekanan darah, rumah sakit, ekstrak etanol, faktor resiko, pengaruh, lampung, haiv aids, diabetes, pemberian asi, usia anak, kualitas hidup, *rattus norvegius*, kesehatan reproduksi, tikus putih, hubungan, Pelayanan, *working area*, ibu hamil, Nutrisi, studi kasus, demam berdarah dll menjadi kata yang banyak muncul di *corpus*. Oleh karena itu, peneliti menyimpulkan bahwa penelitian kesehatan di Indonesia berdasarkan visualisasi dengan menggunakan library *python pyLdavis* topik penelitian kesehatan di Indonesia yaitu berkaitan dengan 30 kata penting yang muncul pada visualisai pada Gambar 5 tersebut.

Visualisasi selanjutnya dengan menggunakan visualisasi *wordcloud*. Visualisasi *wordcloud* ini berdasarkan hasil pemodelan LDA pada Tabel 5 sehingga terdapat 3 Visualisai *wordcloud* yang menunjukan hasil pemodelan topik. Visualisasi *word cloud* dianggap cukup representatif untuk menampilkan hasil pemodelan topik karena dapat mengetahui kemunculan *term-term* yang sering muncul berdasarkan topik yang dimodelkan [15]. Adapun visualisai *word cloud* berdasarkan topik yang terbentuk dapat dilihat pada Gambar 6, 7 dan 8.



Gambar 6. Visualisasi *word cloud* topik#1

Gambar 7. Visualisasi *word cloud* topik#2

Gambar 8. Visualisasi *word cloud* topik#3

Pada Gambar 4,7,8 merupakan *term-term* yang muncul dalam judul penelitian dalam visualisasi *wordcloud* yang didapatkan dari Tabel 5 , semakin besar hurufnya menunjukan kata semakin banyak muncul dalam judul penelitian kesehatan di Indonesia. Term yang muncul di setiap topik dapat dilihat pada Tabel 6.

Tabel 6.Kemunculan term-term berdasarkan judul

Topik	Term
Topik#1	ibu_hamil, status_gizi, wilayah_kerja, tekanan_darah, hubungan_tingkat, factors_related, faktor_faktor, relationship, gizi
Topik#2	pemanfaatan_pelayanan, faktor_faktor, padang, kadar_kolesterol, perilaku, dinas_kesehatan,promosi_kesehatan, kesehatan
Topik#3	hiv_aids, wilayah_kerja, faktor_berhubungan, palembang, rumah_sakit, pengetahuan_sikap, pasien_diabetes, pemberian_asi, kesehatan_reproduksi

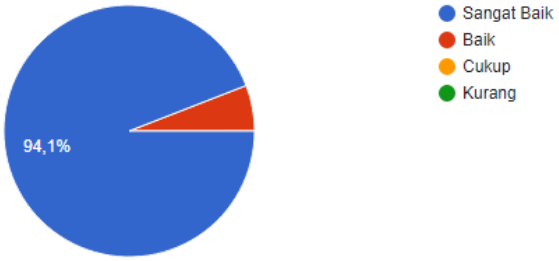
3.6 Analisis dan Evaluasi

Berdasarkan implementasi *topic modelling* LDA yang dibuat terdapat 3 topik yang diklasterisasi. Peneliti menganalisis hasil pemodelan LDA 3 topik yang terbentuk dibagi menjadi 2 topik yang dominan diteliti yaitu tentang topik umum dan penyakit dapat dilihat pada Tabel 7.

Tabel 7. Hasil Topik Dominan

Topik Umum	Topik Penyakit
Faktor-faktor, Status Gizi, Tekanan Darah, Rumah Sakit, Ibu Hamil, Pelayanan Kesehatan, Puskesmas, Ekstrak etanol, *Public Health*	Hiv Aids, Diabetes, Demam berdarah

Pada Tabel 7 topik umum yang diteliti pada penelitian kesehatan di Indonesia yaitu tentang Faktor-faktor, Status Gizi, Tekanan Darah, Rumah Sakit, Ibu Hamil, Pelayanan Kesehatan, Puskesmas, Ekstrak etanol, *Public Health*. Selaint topik umum didapatkan topik penelitian yang berkaitan dengan penyakit yaitu Hiv Aids, Diabetes dan Demam Berdarah. Hal ini terbukti menurut [16] kementrian kesehatan penyakit Hiv Aid hingga saat ini sudah menyebar di 407 dari 507 kabupaten/kota (80%) di seluruh provinsi di Indonesia. Selain itu penyakit demam berdarah merupakan penyakit yang ini sudah terdapat di seluruh pelosok Indonesia[17] dan peyakit diabetes di Indonesia menurut WHO diprediksi mengalami kenaikan jumlah penyandang diabetus dari 8,4 juta pada tahun 2000 menjadi sekitar 21,3 juta pada tahun 2030. Oleh karena itu banyak peneliti di Indonesia meneliti tentang topik penyakit tersebut. Sehingga pemerintah dapat memberikan *planing* jangka panjang untuk mengatasi penyakit diabetus, hiv aids, demam berdarah dengan cara menyebarkan informasi yang berkaitan penyakti tersebut mulai dari penanganan, pencegahan dan pengobatan. Hasil pengujian penelitian ini menggunakan quisoner yang disampaikan responden peneliti, tenaga kesehatan, dan akademisis rata-rata menjawab dengan pemodelan topik yang dihasilkan yiatu 94,1% mengatakan sangat baik dan 5,9 % baik adapun prosentasi disajikan pada Gambar 6.



Gambar 3. Hasil Pengujian

4. Kesimpulan

Penelitian ini berhasil memodekan topik menggunakan metode *Latent Dirichlet Allocation* (LDA) terbukti dapat melakukan pemodelan topik terhadap judul penelitian di bidang penelitian kesehatan di Indonesia. Hasil dari pemodelan topik hasil penelitian yaitu terbagi menjadi dua topik yaitu topik umum dan topik penyakit. Hasil pengujian penelitian ini 94,1% mengatakan sangat baik.

Penelitian ini berhasil menampilkan visualisasi hasil pemodelan topik kedalam visualisasi *wordcloud* dan *pyLDAvis* hasil persebaran kata terhadap judul penelitian kesehatan di Indonesia yaitu dan Faktor-faktor, Status Gizi, Tekanan Darah, Rumah Sakit, Ibu Hamil, Pelayanan Kesehatan, Puskesmas, Ekstrak etanol, *Public Health*, *Hiv Aids*, Diabetes, Demam berdarah.

Saran dari penelitian ini selanjutnya dapat dibuat aplikasi *dashboard* untuk memodelkan topik secara *real time* untuk melihat topik penelitian-penelitian di Indonesia sehingga berguna bagi peneliti, pemerintah sebagai pengambilan keputusan terkait topik yang diketahui.

Daftar Rujukan

[1] D. Prasanti, "Potret Media Informasi Kesehatan Bagi Masyarakat Urban di Era Digital," *J. IPTEKKOM J. Ilmu Pengetah. Teknol. Inf.*, vol. 19, no. 2, p. 149, 2018.

[2] I. Arief and H. Handoko, *Jurnal Online dengan Open Journal System*. 2016.

[3] P. Lakshmi Prasanna and D. Rajeswara Rao, "A text mining research based on topic modeling using latent dritchlent allocation," *Int. J. Recent Technol. Eng.*, vol. 7, no. 5, pp. 308–317, 2019.

[4] B. Tieman, S. Narayanan, A. Sandy, and M. Sikorski, "MPICorrelator: A parallel code for performing time correlations," *Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip.*, vol. 649, no. 1, pp. 240–242, 2011.

[5] C. Zou, "Analyzing research trends on drug safety using topic modeling," *Expert Opin. Drug Saf.*, vol. 17, no. 6, pp. 629–636, 2018.
[6] K. B. Putra and R. P. Kusumawardani, "Analisis Topik Informasi Publik Media Sosial di Surabaya Menggunakan Pemodelan Latent Dirichlet Allocation (LDA)," *J. Tek. ITS*, vol. 6, no. 2, pp. 4–9, 2017.
[7] I. Komputer, D. Ilmu, F. Matematik, P. Alam, and U. G. Mada, "Document Clustering Dengan Latent Dirichlet Allocation Dan Ward," vol. V, no. September, 2018.
[8] I. N. Kabiru, P. K. Sari, S. Prodi, and M. Bisnis, "Analisa Konten Media Sosial E-Commerce Pada Instagram Menggunakan Metode Sentimen Analysis Dan Lda-Based Topic Modeling (Studi Kasus : Shopee Indonesia) Analysis Of Content Social Media E-Commerce In Instagram Using Sentiment Analysis And Lda Based Topki," vol. 6, no. 1, pp. 12–19, 2019.
[9] N. A. Bureau, T. Centre, F. Sciences, and Q. Laboratoire, "Q Uestioned D Ocument E Xamination," vol. 8, no. 3, pp. 221–238, 2016.
[10] P. M. Prihatini, I. K. Suryawan, and I. N. Mandia, "Metode Latent Dirichlet Allocation Untuk Ekstraksi Topik Dokumen," *J. Log.*, vol. 17, no. 3, pp. 154–158, 2017.
[11] Zulhanif, Sudartianto, B. Tantular, and I. G. N. M. Jaya, "Aplikasi Latent Dirichlet Allocation (Lda) Pada Clustering Data Teks," *J. Log.*, vol. 7, no. 1, pp. 46–51, 2017.
[12] A. Priyanto and M. R. Ma'arif, "Implementasi Web Scrapping dan Text Mining untuk Akuisisi dan Kategorisasi Informasi dari Internet (Studi Kasus: Tutorial Hidroponik)," *Indones. J. Inf. Syst.*, vol. 1, no. 1, pp. 25–33, 2018.
[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," vol. 3, pp. 993–1022, 2003.
[14] I. Surjandari, A. Rosyidah, Z. Zulkarnain, and E. Laoh, "Mining Web Log Data for News Topic Modeling Using Latent Dirichlet Allocation," *Proc. - 2018 5th Int. Conf. Inf. Sci. Control Eng. ICISCE 2018*, pp. 331–335, 2019.
[15] M. F. A. Bashri and R. Kusumaningrum, "Sentiment analysis using Latent Dirichlet Allocation and topic polarity wordcloud visualization," *2017 5th Int. Conf. Inf. Commun. Technol. ICoIC7 2017*, vol. 0, no. c, pp. 4–8, 2017.
[16] Ministry of Health of Republic Indonesia, "General situation of HIV/AIDS and HIV test." p. 12, 2018.
[17] A. Candra, "Dengue Hemorrhagic Fever Epidemiology, Pathogenesis, and Its Transmission Risk Factors," *Aspirator J. Vector Borne Dis. Stud.*, vol. 2, no. 2, pp. 110–119, 2010.