Jurnal Teknik Informatika (JUTIF)
Vol. 3, No. 5, Oktober 2022, hlm. 1223-1230

DOI: https://doi.org/10.20884/1.jutif.2022.3.5.321
p-ISSN: 2723-3863
e-ISSN: 2723-3871

TOPIC MODELING USING THE LATENT DIRICHLET ALLOCATION METHOD ON WIKIPEDIA PANDEMIC COVID-19 DATA IN INDONESIA

Wilujeng Ayu Nawang Sari[*1], Hindriyanto Dwi Purnomo[2]

[1,2]Program Suti Teknik Informatika, Universitas Kristen Satya Wacana, Indonesia
Email: [1]672018234@student.uksw.edu, [2]hindriyanto.purnomo@uksw.edu

(Naskah masuk: 10 Mei 2022, Revisi : 29 Juni 2022, diterbitkan: 24 Oktober 2022)

Abstract

Wikipedia is a web-based encyclopedia that is used to search for information. In one of the Wikipedia articles, a problem has been found regarding no one has clustered on the topic of the Covid-19 pandemic in Indonesia. The method used for this research is the Latent Dirichlet Allocation (LDA) method. The Latent Dirichlet Allocation (LDA) method is the most widely used topic modeling method today. In this study using 6658 words in English that will be used for the dataset. Then every word that appears will be counted using Corpus. This study applies topic modeling using the Latent Dirichlet Allocation (LDA) model and how to analyze COVID-19 data taken from Wikipedia. The LDA method will cluster by looking at the number of words that appear in Corpus and will determine the number of clusters and the number of topics and determine the iteration. The purpose of this study is to classify the information contained in the Wikipedia Article so that it can be used as an evaluation material in improving services and handling Wikipedia using the latent direchlet allocation method. The LDA method will mark every word contained in the topic in a semi-random distribution and will calculate the probability of the topic in the dataset and will calculate the probability of the word on the topic of each iteration. In this study, 5 iteration tests were conducted on topic modeling and a number of different topics. After the experiment is carried out, the final results obtained will be analyzed and get 1 number of topics with the best results with the most discussion topics regarding health.

Keywords: *Clustering, Latent Dirichlet Allocation (LDA), Topic Modelling, Wikipedia.*

PEMODELAN TOPIK MENGGUNAKAN METODE LATENT DIRICHLET ALLOCATION PADA DATA WIKIPEDIA PANDEMIC COVID-19 DI INDONESIA

Abstrak

Wikipedia merupakan salah satu ensklopedia berbasis web yang digunakan untuk mencari suatu informasi. Didalam salah satu artikel *Wikipedia* telah ditemukan permasalahan mengenai belum ada yang melakukan *clustering* pada topik pandemi *Covid-19* di Indonesia. Metode yang digunakan untuk penelitian ini menggunakan metode *Latent Dirichlet Allocation (LDA).* Metode *Latent Dirichlet Allocation (LDA)* merupakan metode pemodelan topik yang paling banyak digunakan saat ini. Pada penelitian ini menggunakan 6658 kata dalam bahasa inggris yang akan digunakan untuk dataset. Kemudian setiap kata yang muncul akan dihitung menggunakan *Corpus.* Penelitian ini menerapkan pemodelan topik menggunakan model *Latent Dirichlet Allocation (LDA)* dan cara menganalisis pada data *covid-19* yang diambil dari *Wikipedia*. Metode LDA akan mengklaster dengan melihat jumlah kata yang muncul pada *Corpus* dan akan menentukkan jumlah *cluster* maupun jumlah topik dan menentukkan iterasinya. Tujuan dari penelitian ini adalah untuk mengklasifikasikan informasi yang terdapat pada artikel *Wikipedia* sehingga bisa menjadi bahan evaluasi dalam perbaikan perbaikan pelayanan dan penanganan *Wikipedia* menggunakan metode alokasi direchlet laten. Metode LDA akan menandai setiap kata yang terdapat pada topik secara semi random distribution serta akan menghitung probabilitas topik pada dataset dan akan menghitung probabilitas kata terhadap topik setiap iterasinya. Pada penelitian ini dilakukan percobaan pemodelan topik sebanyak 5 kali uji iterasi dan jumlah topik yang berbeda-beda. Setelah percobaan tersebut dilakukan maka hasil akhir yang didapat akan dianalisis dan mendapatkan 1 jumlah topik dengan hasil terbaik dengan topik pembahasan paling banyak mengenai kesehatan.

Kata kunci: *Klasterisasi, Latent Dirichlet Allocation (LDA), Pemodelan Topik, Wikipedia.*

1. PENDAHULUAN

Covid-19 merupakan jenis virus terbaru yang tengah menyebar ke beberapa negara, salah satunya Indonesia. Virus *Covid-19* berawal dari Wuhan, China sejak awal bulan Desember 2019. Tanda-tanda yang alami pasien *Covid-19* meliputi demam, malaise, batuk kering, dan dyspnea yang mengarah kedalam gejala infeksi virus pneumonia[1][2]. Dampak dari terkena virus *Covid-19* dapat mengalami kematian. Jumlah kasus yang terkena *Covid-19* di dunia hingga 30 Desember 2021 sejumlah 262.000.000 kasus,, sedangkan jumlah kasus yang meninggal akibat terkena *Covid-19* di dunia sebanyak 5.210.000.000 kasus. Sementara itu, jumlah kasus yang terkena *Covid-19* di Indonesia hingga 30 Desember 2021 sebanyak 4.260.000.000 kasus, sedangkan untuk jumlah kasus *Covid-19* di Indonesia sebanyak 144.000 kasus.

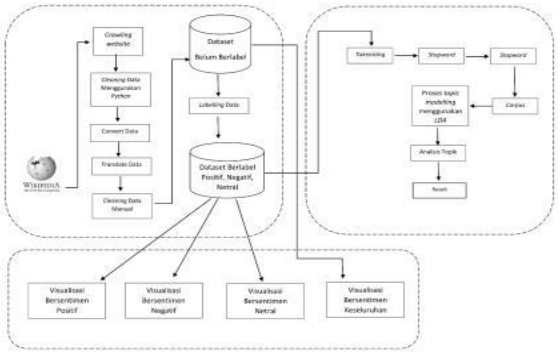
Dampak utama dari *Covid-19* adalah kesehatan. Selain mempengaruhi dampak kesehatan, *Covid-19* juga dapat mempengaruhi pada kondisi perekonomian, pendidikan, hingga mempengaruhi kehidupan sosial masyarakat di Indonesia. *Covid-19* mengakibatkan pemerintah daerah menerapkan kebijakan Pembatasan Sosial Berskala Besar (PSBB) yang menganjurkan terhadap pembatasan aktivitas masyarakat. Berkurangnya berbagai aktivitas tersebut sangat berdampak terhadap kondisi sosial-ekonomi masyarakat, khususnya terhadap masyarakat rentan dan miskin. Oleh karena itu, pemerintah daerah membuat berbagai kebijakan untuk mengatasi penyebaran *Covid-19* serta kebijakan-kebijakan yang bersifat penanganan dampak sosial dan ekonomi akibat *Covid-19* [3].

Informasi COID-19 dapat ditemukan di berbagai platform yang sekarang aktif, yang paling populer adalah *Wikipedia*. *Wikipedia* adalah ensiklopedia populer yang berbasis internet. *Wikipedia* merupakan salah satu ensiklopedia online paling terkenal. Saat ini, *Wikipedia* memiliki 15 juta artikel yang dapat digunakan secara non-komersial, serta lebih dari 200 bahasa dari seluruh dunia, yang semuanya ditulis oleh ribuan pengguna dan kontributor anonim. Informasi ini khusus dan berisi data tekstual.

Latent Dirichlet Allocation (LDA) adalah salah satu model yang paling banyak digunakan saat ini. Dalam penelitian ini, kami mengumpulkan 6658 kata dari *Wikipedia* tentang pandemi *Covid-19* di Indonesia dalam bahasa Inggris, yang kami gunakan sebagai dataset. Penggunaan Bahasa Inggris dalam pra-pemrosesan data karena tersedianya standar data baru dalam Bahasa Inggris. Kemudian, dengan menggunakan *Corpus*, setiap kata yang masuk akan diterjemahkan. Metode LDA akan mengklaster dengan melihat jumlah kata yang muncul di *Corpus*, akan menentukkan jumlah cluster maupun jumlah topik dan iterasinya. Metode LDA akan menemukan kata yang ada pada topik secara semi acak distribusi, menghitung probabilitas topik pada dataset, dan menghitung probabilitas terhadap topik setiap iterasinya. percobaan pemodelan topik sebanyak 5 kali uji iterasi dan jumlah topik yang berbeda-beda dilakukan pada penelitian ini. Setelah percobaan yang disebutkan di atas selesai, data yang dikumpulkan akan dianalisis, dan 1 jumlah topik dengan hasil terbaik akan dipilih dari sekian banyak topik pembahasan tentang kesehatan. Hasilnya sesuai dengan poin utama artikel, yaitu tentang kesehatan.

Adapun tujuan penelitian dilakukan untuk Mengetahui clusterisasi dengan melihat jumlah kata yang sering muncul menggunakan *corpus*.

2. METODE PENELITIAN



Gambar 2. Metode Penelitian

Berikut beberapa tahapan yang biasa digunakan dalam penelitian:

2.1. Metode Pengumpulan Data

Pada metode pengumpulan data dilakukan beberapa proses. Pada penelitian ini menggunakan data dari salah satu artikel didalam *Wikipedia*. Pada tahap ini dilakukan proses *crawling* menggunakan bahasa pemrograman python. Selama proses *crawling* data tersebut menggunakan library *Wikipedia* untuk mengambil data tersebut. Data yang didapat dari *crawling* data tersebut akan disimpan kedalam bentuk txt dan diubah kedalam bentuk csv.

2.1.1. *Wikipedia*

Wikipedia merupakan ensiklopedia daring berbasis website. *Wikipedia* bertujuan untuk memberikan wawasan atau ilmu pengetahuan kepada manusia *Wikipedia* telah mengalami kenaikan sebanyak 6 kali lebih banyak di Inggris. Hal ini bisa ditemukan dalam beberapa topik, seperti anatomi, biologi, atau kedokteran dan dikatakan sebagai informasi yang akurat. Informasi dalam bentuk artikel yang dibuat di *Wikipedia* ditulis lengkap sesuai jenis artikelnya. *Wikipedia* merupakan salah satu situs web yang paling populer dan ensiklopedia terbesar di dunia. Beberapa alasan juga seperti informasi yang didapat dari *Wikipedia* berupa plagiarisme terhadap sumber lain. Sebagai contoh, beberapa profesi seperti, guru maupun dosen tidak menganjurkan siswa maupun mahasiswa nya mengutip sumber dari *Wikipedia* [4].

2.1.2. *Python*

Python merupakan suatu bahasa pemrograman yang dapat dipakai diberbagai platform berdasarkan konsep perancangan yang berpusat pada tingkat kode yang mudah dipahami. Python diklamin sebagai bahasa pemrograman yang lebih mudah dipahami dibandingkan dengan bahasa pemrograman lainnya dan sebagai bahasa yang mengintegrasikan ketrampilan, kemampuan, dengan syntax yang sangat jelas, dan dilengkapi dengan fungsi-funsi pustaka standar yang besar serta menyeluruh[5]. Python termasuk bahasa pemrograman tingkat tinggi. Hal tersebut dimaksudkan karena syntax pada python dibuat semirip mungkin dengan bahasa manusia.

2.2. Proses Visualisasi Menggunakan *Wordcloud*

Dataset yang telah dilabel kemudian divisualikan menggunakan *wordcloud* pada masing-masing dataset yang berlabel positif, negative dan netral seta dilakukan visualisasi secara keseluruhan data.

2.3. Proses Pembuatan *Topic modelling*

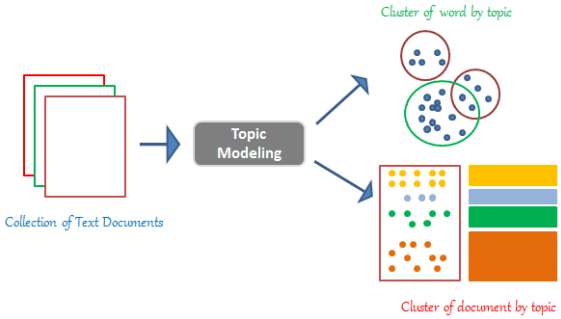
Pada proses pembuatan *topic modelling* dilakukan menggunakan *Latent Dirichlet Allocation* (LDA). Pada proses ini dilakukan beberapa tahapan antara lain untuk mengubah kalimat menjadi kata per kata dalam proses pembuatan pemodelan topik. Kemudian, dengan menggunakan prosedur *stopword*, dapat mengecualikan kata-kata yang tidak memiliki informasi apapun, serta data karakter dan baca dari dataset. Setelah melakukan proses *stopword* kemudian melakukan proses *lemmatisasi* untuk mengkategorikan berbagai jenis kata. Penulis melakukan proses mengubah token data dokumen menjadi bentuk *corpus* menggunakan *corpus* setelah menyelesaikan langkah pra-pemrosesan. Kemudian melakukan proses menggunakan LDA untuk memperoleh penyebaran kata yang membentuk topik. Setelah melakukan proses LDA maka penulis akan melakukan analisis dari hasil yang didapat dalam proses LDA. Setelah melakukan analisis kemudian akan mendapat hasil akhir yang didapat.

2.3.1. *Topic modelling*

Topic modelling merupakkan data teks berdasarkan topik tertentu. *Topic modelling* termasuk kedalan *clustering* dengan mengelompokkan dokumen berdasarkan kemiripannya[6][7]. *Topic modelling* memiliki tujuan yang selaras dengan klasifikasi tetapi memakai prosedur yang berbeda. *Topic modelling* memiliki tujuan yang lebih spesifik, yaitu : memperoleh model topik abstrak pada kumpulan dokumen, memberikan tinjauan pada dokumen berdasarkan topik tersebut, dan memakai tinjauan pada dokumen untuk mengelompokkan dokumen. Selain itu, *topic modelling* juga memili tujuan lain, yaitu untuk menemukan topik dan kata yang tersembunyi dalam topik tersubut. konsep dari *topic modelling* adalah sebuah topik yang terdiri dari beberapa kata tertentu untuk menyusun topik dari dokumen tersebut. Didalam suatu dokumen mempunyai kemungkinan memiliki beberapa topik dengan peluang masing-masing[8]. Hasil output dari *topic modelling* berupa kumpulan-kumpulan topik yang sering muncul didalam dokumen berdasarkan pola tertentu [9].

2.3.2. *Dirichlet Allocation* (LDA)

Latent Dirichlet Allocation (LDA) adalah suatu metode untuk menemukan tiap-tiap topik yang terdapat pada koleksi dokumen beserta perbandingan kemunculan topik tersebut. [9][10].



Gambar 1. Cara kerja LDA

Cara kerja LDA adalah sebagai berikut : (1) untuk setiap dokumen menginisialisasi setiap kata secara acara ke seluruh topik dimana k merupakan jumlah topik yang telah ditentukan sebelumnya. (2) untuk setiap dokumen d : untuk setiap kata w dalam dokumen, akan dihitung :

- **P(topic t|document d)** : proporsi kata dalam dokumen d yang ditugaskan ke topik t.
- **P(word w|topic t)** : proporsi tugas ke topik t di semua dokumen dari kata-kata yang berasal dari w.

(3) menetapkan kembali topik T' ke kata w dengan probabilitas p(t'|d)*p(w|t') dengan mempertimbangkan semua kata lain dan penugasan topiknya. Dan terakhir ulang kembali beberapa kali sampai kita dan penugasan topiknya. Dan terakhir ulang kembali beberapa kali sampai kita mencapai kondisi mapan dimana tugas topik tidak berubah lebih jauh. Proporsi topik untuk setiap dokumen kemudian ditentukan dari penugasan topik ini..

2.3.3. *Clustering*

Clustering adalah suatu proses untuk mengelompokkan data kedalam beberapa kelompok sesuai dengan kemiripan karakteristik tiap-tiap data pada kelompok-kelompok yang sudah ada[11]. *Clustering* erupakan salah satu metode dari data mining. *Clustering* merupakan salah satu metode dalam penelitian untuk analisis dan data mining[12][13]. Pada dasarnya proses *clustering* untuk mengelompokkan sekumpulan data tanpa

atribut kelas telah dideskripsikan sebelumnya berdasarkan pada konseptual *clustering*.

3. HASIL DAN PEMBAHASAN

3.1. Pengambilan Data

Pengambilan adalah tahap awal dalam penelitian. Proses pengambilan data dilakukan dengan cara *crawling* data yang diambil salah satu artikel mengenai pandemi *Covid-19* di Indonesia dari website *Wikipedia* dalam Bahasa Inggris dengan menggunakan library *Wikipedia* dan tersimpan kedalam bentuk txt. Jumlah kata yang diperoleh keseluruhan sebanyak 6658 kata. Data yang didapat kemudian disimpan untuk proses selanjutnya. Kemudian dilakukan proses *case folding* terlebih dahulu lalu dilakukan proses *cleaning* pada dataset tersebut untuk menghilangkan kata-kata dan karakter yang tidak diperlukan. Dataset yang telah dilakukan pada tahap sebelumnya selesai kemudian dilakukan *convert* pada dataset dengan mengubah file txt menjadi csv menggunakan *spacy*. Setelah melakukan *convert* pada dataset tersebut kemudian dilakukan translate dataset kembali dengan menggunakan *textblob*. Setelah melalui proses tersebut peulis melakukan *cleaning* ulang secara manual dengan mencari dan menghapus karakater tidak diperlukan yang terdapat pada dataset tersebut. Penulis melakukan translate ulang dan *cleaning* manual agar mendapatkan hasil data terbaik. Apabila data yang diperoleh masih belum maka harus dilakukan *cleaning* kembali. Kemudian proses selanjutnya memberikan label pada dataset yang belum berlabel dan disimpan kedalam dataset baru.

3.2. *Tokenizing*

Setiap kata dalam *Tokenizing* dipisahkan oleh karakter spasi, sehingga karakter spasi sangat diandalakan dalam proses pemisahan setiap kata pada dokumen[14][15]. Proses ini juga bertujuan untuk menghilangkan karakter yang tidak dibutuhkan seperti menghilangkan tanda baca, mention, dan url yang terdapat pada teks. Tabel 1 menunjukkan perbedaan antara sebelum dan sesudah *tokenizing*.

Tabel 1. Perbedaan sebelum dan sesudah *Tokenizing*

Sebelum Tokenizing	Sesudah Tokenizing
covid pandemic in indonesia is part of the ongoing worldwide pandemic of corona virus disease covid caused by severe acute respiratory syndrome corona virus cov it was confirmed to have spread to indonesia on march after a dance instructor and her mother tested positive for the virus.	['covid', 'pandemic', 'in', 'indonesia', 'is', 'part', 'of', 'the', 'ongoing', 'worldwide', 'pandemic', 'of', 'corona virus', 'disease', 'covid', 'caused', 'by', 'severe', 'acute', 'respiratory', 'syndrome', 'corona virus', '', 'cov', 'it', 'was', 'confirmed', 'to', 'have', 'spread', 'to', 'indonesia', 'on', 'march', 'after', 'a', 'dance', 'instructor', 'and', 'her', 'mother', 'tested', 'positive', 'for', 'the', 'virus', '.']
both were infected from a japanese national.by april the pandemic had spread to all provinces in the country.	['both', 'were', 'infected', 'from', 'a', 'japanese', 'national.by', 'april', 'the', 'pandemic', 'had', 'spread', 'to', 'all', 'provinces', 'in', 'the', 'country', '.']
jakarta west java and central java are the worst hit provinces together accounting almost half of the national total cases.	['jakarta', 'west', 'java', 'and', 'central', 'java', 'are', 'the', 'worst', 'hit', 'provinces', 'together', 'accounting', 'almost', 'half', 'of', 'the', 'national', 'total', 'cases', '.']
on july the recoveries exceeded active cases for the first time.	['on', 'july', 'the', 'recoveries', 'exceeded', 'active', 'cases', 'for', 'the', 'first', 'time', '.']
as of march indonesia has reported cases the second highest in southeast asia behind vietnam.	['as', 'of', 'march', 'indonesia', 'has', 'reported', 'cases', 'the', 'second', 'highest', 'in', 'southeast', 'asia', 'behind', 'vietnam', '.']

3.3. *Stopword*

Setelah melalui proses *tokenizing* selanjutnya terminologi dokumen diolah kedalam proses *stopword*. Proses *stopword* digunakan untuk menghapus kata-kata yang tidak memiliki informasi dan hanya mengambil kata-kata yang penting saja. Tahap awal dalam penggunaan *stopword* yaitu dengan menggunakan library nltk. *Stopword*.words('english') ditambahkan di fungsi set() di stop words untuk menghilangkan kata-kata yang tidak bermanfaat. Tabel 2 menunjukkan perbedaan antara sebelum dan sesudah *tokenizing*.

Tabel 2. Perbedaan Sebelum dan Sesudah *Stopword*

Sebelum *Stopword*	Sesudah *Stopword*
['covid', 'pandemic', 'in', 'indonesia', 'is', 'part', 'of', 'the', 'ongoing', 'worldwide', 'pandemic', 'of', 'corona virus', 'disease', 'covid', 'caused', 'by', 'severe', 'acute', 'respiratory', 'syndrome', 'corona virus', '', 'cov', 'it', 'was', 'confirmed', 'to', 'have', 'spread', 'to', 'indonesia', 'on', 'march', 'after', 'a', 'dance', 'instructor', 'and', 'her', 'mother', 'tested', 'positive', 'for', 'the', 'virus', '.']	covid pandemic indonesia part on going worldwide pandemic corona virus disease covid cause severe acute respiratory syndrome corona virus cov confirm spread indonesia march dance instructor mother test positive virus
['both', 'were', 'infected', 'from', 'a', 'japanese', 'national.by', 'april', 'the', 'pandemic', 'had', 'spread', 'to', 'all', 'provinces', 'in', 'the', 'country', '.']	infect japanese national by april pandemic spread provinces country
['jakarta', 'west', 'java', 'and', 'central', 'java', 'are', 'the', 'worst', 'hit', 'provinces', 'together', 'accounting', 'almost', 'half', 'of', 'the', 'national', 'total', 'cases', '.']	jakarta west java central java worst hit provinces together account almost half national total case .
['on', 'july', 'the', 'recoveries', 'exceeded', 'active', 'cases', 'for', 'the', 'first', 'time', '.']	july recoveries exceed active case first time .
['as', 'of', 'march', 'indonesia', 'has', 'reported', 'cases', 'the', 'second', 'highest', 'in',	march indonesia report case second highest southeast asia behind vietnam .

'southeast', 'asia', 'behind', 'vietnam', '.']	

Pada akhir proses *stopword* pertama, terdapat banyak kata non-informatif, banyak kata bahasa Indonesia, dan banyak tanda baca di dataset. Untuk menghapus kata yang tidak informatif dan menghapus tanda baca serta mengubah kata-kata Bahasa Indonesia menjadi Bahasa Inggris maka dilakukan proses *cleaning* secara manual serta menstranslate ulang data tersebut. Sehingga hasil yang diperoleh dapat dilihat pada tabel 3.

3.4. *Lemmatization*

Hasil dari proses sebelumnya dilanjutkan ke proses Pre-processing teks dimana mencangkup proses *lemmatization*. Proses *lemmatization* merupakan suatu proses pengklasifikasian kata yang berbeda melalui tahap analasis sebagai satu kata yang sama. Untuk hasil dari *lemmatization* dapat dilihat dari tabel 4.

Tabel 3. Perbedaan Sebelum dan Sesudah *Lemmatization*

Sebelum ***Lemmatization***	Sesudah ***Lemmatization***
['covid', 'pandemic', 'indonesia', 'part', 'on', 'going', 'world', 'wide', 'pandemic', 'corona', 'disease', 'covid', 'case', 'syndrome', 'corona', 'covid', 'confirm', 'spread', 'indonesia', 'march', 'dance', 'instructur', 'mother', 'test', 'positive']	['covid', 'pandemic', 'part', 'go', 'world', 'wide', 'pandemic', 'corona', 'disease', 'case', 'syndrome', 'covid', 'confirm', 'spread', 'mother', 'test', 'positive'],
['infect', 'japanese', 'national', 'april', 'pandemic', 'spread', 'provincy', 'country']	['infect', 'japanese', 'national', 'pandemic', 'spread', 'provincy', 'country'],
['jakarta', 'west', 'java', 'central', 'java', 'worst', 'hit', 'provincy', 'gether', 'account', 'almost', 'half', 'national', 'case']	['central', 'bad', 'hit', 'provincy', 'gether', 'account', 'almost', 'national', 'case'],
['july', 'recovery', 'exceed', 'active', 'case', 'first', 'time'], ['march', 'indonesia', 'report', 'case', 'second', 'high', 'southeast', 'asia', 'behind', 'vietnam']	['exceed', 'active', 'case', 'first', 'time'],
['death', 'indonesia', 'rank', 'second', 'asia', 'nine', 'world']	['case', 'second', 'high', 'southeast']

3.5. *Corpus*

Untuk proses selanjutnya yaitu melakukan perubahan pada token data dokumen menjadi bentuk *corpus*[16]. Untuk melihat hasil dari *corpus* dapat dilihat pada tabel 4.

Tabel 4. Perbedaan Sebelum dan Sesudah *Stopword*.

Sebelum *Corpus*	**Sesudah *Corpus***
['covid', 'pandemic', 'part', 'go', 'world', 'wide', 'pandemic', 'corona', 'disease', 'case', 'syndrome', 'covid', 'confirm', 'spread', 'mother', 'test', 'positive'],	[(0, 1), (1, 1), (2, 1), (3, 2), (4, 1), (5, 1), (6, 1), (7, 2), (8, 1), (9, 1), (10, 1), (11, 1), (12, 1), (13, 1), (14, 1)],
['infect', 'japanese', 'national', 'pandemic', 'spread', 'provincy', 'country'],	[(7, 1), (10, 1), (15, 1), (16, 1), (17, 1), (18, 1), (19, 1)],
['central', 'bad', 'hit', 'provincy', 'gether', 'account', 'almost', 'national', 'case'],	[(0, 1), (18, 1), (19, 1), (20, 1), (21, 1), (22, 1), (23, 1), (24, 1), (25, 1)],
['exceed', 'active', 'case', 'first', 'time'],	[(0, 1), (26, 1), (27, 1), (28, 1), (29, 1)],
['case', 'second', 'high', 'southeast']	[(0, 1), (30, 1), (31, 1), (32, 1)],

3.6. Proses *Topic modelling* Menggunakan LDA

Setelah melewati beberapa tahap pre-processing dan memasukkan kedalam *corpus* maka proses berikutnya yaitu membuat pemodelan topik menggunakan LDA. Pada proses sebelumnya tepatnya pada bagian *corpus* telah muncul token yang berasal dari banyaknya kata yang sering muncul dari suatu dokumen. Token tersebut berefungsi sebagai banyaknya ukuran dalam LDA agar dapat dilakukan pemodelan. Dalam proses pemodelan LDA dilakukan beberapa kali uji coba, agar dapat menentukan jumlah topik yang sesuai. Pada tahap uji coba ini dilakukan dengan mengubah jumlah topik dan jumlah iterasinya. Untuk uji coba penelitian ini dilakukan sebanyak 5 kali uji coba iterasi yang berbeda, diantara : 100,500,1000,1500,2000 dengan jumlah topik yang berbeda, yaitu : 3,4,5,6,7 di setiap uji iterasinya.

3.7. Analisis Topik

Kemudian setelah mendapatkan hasil dari proses LDA dilakukan analisis untuk keseluruhan model LDA dari berbagai uji coba jumlah topik dan jumlah iterasi. Hasil dari iterasi ke-100 dari jumlah topik 3 menghasilkan 5 topik yang berbeda. Untuk hasil iterasi ke-500 dari jumlah topik 3 menghasilkan 5 topik yang berbeda. Sebaliknya, untuk hasil ke-1000 dari total 3 topik digunakan 5 topik yang serupa. Hasil dari uji iterasi ke-1500 menghasilkan 5 topik yang berbeda. Selain itu, untuk hasul uji iterasi ke-2000 dari total 3 topik, ditemukan 5 topik yang serupa. Terdapat 5 topik yang mirip dan 2 topik yang paling umum diantara hasil uji iterasi. Untuk memastikan bahwa topik yang bersangkutan memiliki kesamaan kelompok, maka dilakukan pemodeling dengan jumlah topik dan iterasi yang sama.

Iterasi ke-100 dengan jumlah topik 4 menghasilkan 5 topik berbeda. Hasil iterasi ke-500 menghasilkan 5 topik yang berbeda dengan jumlah topik 4. Hasil iterasi 1000 menghasilkan lima topik yang berbeda. Sebaliknya, hasil iterasi ke-1500 menghasilkan 5 topik yang berbeda. Selain itu, hasil iterasi 2000 menghasilkan lima topik yang berbeda. Sehingga dari hasil uji iterasi ke-100, 500, 1000, 1500, dan 2000 menghasilkan topik dengan total 5

topik. Terdapat 5 topik yang mirip dan 3 topik yang paling umum diantara hasil uji iterasi. Untuk memastikan bahwa topik yang bersangkutan memiliki kesamaan kelompok, maka dilakukan pemodeling dengan jumlah topik dan iterasi yang sama. Uji iterasi dilakukan dengan cara yang sama seperti proses pemodelan sebelumnya. Iterasi ke-100 dengan jumlah topik menghasilkan 5 topik berbeda. Hasil dari iterrasi ke-500 dengan jumlah topik 5 menghasilkan 5 topik berbeda. Menghasilkan 5 topik yang berbeda untuk hasil iterasi ke-1500 dengan jumlah topik 5. Sebaliknya, hasil iterasi 2000 dengan 5 topik menghasilkan 5 topik yang berbeda. Terdapat 5 topik yang mirip dan 2 topik yang paling umum diantara hasil uji iterasi. Untuk memastikan topik yang bersangkutan memiliki kesamaan kelompok, maka dilakukan pemodeling menggunakan jumlah topik 6 yang sama dengan iterasi yang sama. Uji iterrasi ke-100 menghasilkan 5 topik yang berbeda. Hasil dari 500 iterasi dengan 6 topik menghasilkan 5 topik yang berbeda. Hasil ke-1500 iterasi dengan jumlah topik 6 menghasilkan 5 topik yang berbeda. Sebaliknya, hasil iterasi 2000 dengan 6 topik menghasilkan 5 topik yang serupa. Dan untuk iterasi ke-2500 dengan 6 topik terdapat 5 topik yang berbeda. Terdapat 5 topik yang mirip dan 2 topik yang paling umum diantara hasil uji iterasi. Untuk memastikan bahwa topik yang bersangkutan memiliki kesamaan kelompok, maka dilakukan pemodeling dengan jumlah topik dan iterasi yang sama.

Selain itu, untuk uji iterasi ke-100 dengan total 7 topik, ditemukan 5 topik yang serupa. Hasil dari 500 iterasi dengan 7 topik menghasilkan 5 topik yang berbeda. Hasil ke-1500 iterasi dengan jumlah topik 7 menghasilkan 5 topik yang berbeda. Sebaliknya, hasil iterasi 2000 dengan 7 topik menghasilkan 5 topik yang serupa. Selain itu, untuk iterasi ke-2500 dengan 7 topik, ditemukan 5 topik yang serupa. Dari hasil uji iterasi yang dilakukan terdapat 5 topik yang berbeda dan terdapat 3 topik yang paling banyak dibahas.

Hasil yang akan disimpulkan sementara pada pemodelan dengan jumlah topik 3,4,5,6, dan 7 dengan iterasi 100,500,1000,1500, dan 2000 menunjukkan bahwa jumlah topik 5 merupakan pemodelan topik yang memliki hasil efektif. Ada dua topik dalam proses sebelumnya yang sangat populer. Untuk menentukan apakah suatu topik benar-benar benar, ada dua topik yang menonjol di antara yang lainnya.

Dari proses analisis seperti yang terdapat pada gambar 2 terdapat 1 topik yang paling banyak muncul yakni mengenai kesehatan yang didalamnya banyak membahas mengenai *covid*. Walaupun terdapat banyak topik yang dibahas selain kesehatan, tetapi topik yang paling banyak dibahas mengenai kesehatan terutama c*ovid*. Hasil tersebut sesuai dengan data artikel yang diambil penulis untuk melakukan penelitian yakni mengenai pandemi *Covid-19* di Indonesia

- Jumlah Topik 4

Topik 0	Topik 1	Topik 2	Topik 3
School	covid	Decide	Government
Work	pandemic	Public	day
flight	death	country	case
Economic	Time	Covid	travel
Include	test	government	test
state	spread	start	Lock
Spend	case	Case	Restriction
discount	Confirm	Close	central
additional	positive	school	Social
open	first	declare	covid

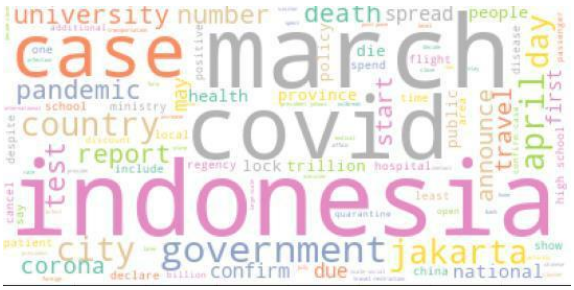
- Jumlah Topik 5

Topik 0	Topik 1	Topik 2	Topik 3	Topik 4
school	due	covid	case	government
Economic	Test	public	day	Restriction
Work	Pandemic	country	Close	Travel
Flight	Positive	Case	Death	Social
official	patient	declare	time	City
Additional	covid	pandemic	Decide	Lock
Discount	corona	visitor	Test	Covid
State	Country	composite	Number	Say
Price	Confirm	concern	Covid	Large
ticket	open	health	high	central

Gambar 3. Hasil analisis

3.8. Pemodelan Topik Menggunakan *Wordcloud*

Wordcloud merupakan sebuah gambar yang didalamnya berisi kumpulan kata-kata yang sering muncul dalam sebuah teks. Semakin banyak kata yang digunakan, maka semakin besar ukuran kata dalam gambar tersebut. Berikut merupakan *wordcloud* dari analisis *website Wikipedia* mengenai *pandemic Covid-19 in Indonesia* secara keseluruhan dan *wordcloud* yang bersifat positif, negatif dan netral.



Gambar 3. *Wordcloud Wikipedia Pandemic Covid-19 in Indonesia* Keseluruhan

Gambar 3 menunjukkan bahwa kata - kata yang sering pada *wordcloud Wikipedia Pandemic Covid-19 in Indonesia* secara keseluruhan adalah *Covid*, *march*, indonesia dan *case*.

Gambar 4. *Wordcloud Wikipedia Pandemic Covid-19 in Indonesia* Bersifat Positif

Gambar 4 menunjukkan bahwa kata - kata yang sering *pada wordcloud Wikipedia Pandemic Covid-19 in* Indonesia dari label yang bersifat positif adalah indonesia, *covid* , *confirm* dan *case.*

Gambar 5. *Wordcloud Wikipedia Pandemic Covid-19 in Indonesia* Bersifat Negatif

Gambar 5 menunjukkan bahwa kata - kata yang sering pada *wordcloud Wikipedia Pandemic Covid-19 in Indonesia* dari label yang bersifat negatif adalah Indonesia, *travel*, *country* dan *university*.

Gambar 6. *Wordcloud Wikipedia Pandemic Covid-19 in Indonesia* Bersifat Netral

Gambar 6 menunjukkan bahwa kata - kata yang sering pada *wordcloud Wikipedia Pandemic Covid-19 in Indonesia* dari label yang bersifat netral adalah indonesia, *covid* , dan *goverment*. Jadi dari keempat gambar *wordcloud*, kata – kata yang sering muncul adalah Indonesia.

4. DISKUSI

Menurut penelitian yang berjudul *Latent Dirichlet Allocation* Untuk Pemodelan Topik Abstrak Dokumen Skripsi, membahas mengenai informasi yang didapatkan bahwa rata-rata lama pengerjaan Tugas Akhir menurut data yang digunakan dalam penelitian tersebut pada setiap tahunnya semakin kecil atau pada grafik semakin menurun, yang dapat diartikan bahwa lama pengerjaan Tugas Akhir tiap tahun angkatannya semakin baik dan data abstrak skripsi mahasiswa Statistika Universitas Islam Indonesia angkatan 2011-2015 menggunakan pemodelan topik dengan metode *Latent Dirichlet Allocation* menghasilkan jumlah topik sebanyak 3 dengan coherence score sebesar 0.5528 dengan kata yang saling berhubungan. Data yang digunakan merupakan data sekunder yang diperoleh dari website https://dspace.uii.ac.id/ dan didapatkan dari perpustakaan Universitas Islam Indonesia. Adapun hasil dari pengujiannya menggunakan jupyter dengan bahasa pemrograman python[17].

Menurut penelitian yang berjudul Pemodelan Topik Pengguna Twitter Mengenai Aplikasi "Ruangguru", membahas mengenai klasifikasi pendapat dari pengguna Ruangguru tentang layanan yang diberikan sehingga dapat menjadi bahan evaluasi dalam perbaikan layanan mereka menggunakan metode *Latent Dirichlet Allocation*. Data yang digunakan berasal kumpulan tweet pengguna twitter di Indonesia yang menggunakan API Twitter yang menggunakan kata kunci @ruangguru. Hasil analisi menunjukkan bahwa persepsi masyarakat terhadap pengguna twitter dengan menggunakan metode *Latent Dirichlet Allocation* dibentuk menjadi 28 topik. Adapun hasil pengujiannya menggunakan jupyter dengan bahasa pemrograman python [18].

No	Penulis	Judul	Metode	Data yang Diperoleh	Persamaan	Perbedaan
1	Ella Anggraini (2020)	*LATENT DIRICHLET ALLOCATION* UNTUK PEMODELAN TOPIK ABSTRAK DOKUMEN SKRIPSI	*Latent Dirichlet Allocation (LDA)*	Data sekunder yang diperleh dari *website* https://dspace.uii.ac.id/ dan didapatkan dari perpustakaan Universitas Islam Indonesia.	Sama – sama menggunakan LDA dan *python*.	Studi kasus penulis sekarang *covid-19* dan data yang diperoleh penulis sekarang didapat dari *Wikipedia*.
2	Bagus Wicaksono Arianto, & Gangga Anuraga (2020)	Pemodelan Topik Pengguna *Twitter* Mengenai Aplikasi "Ruangguru"	*Latent Dirichlet Allocation (LDA)*	Data yang diperoleh didapat dari *tweet Twitter* menggunakan *API Twitter*.	Sama – sama menggunakan LDA dan *python*.	Studi kasus penulis sekarang *covid-19* dan data yang diperoleh penulis sekarang didapat dari *Wikipedia*.

Gambar 7. Perbandingan Dengan Penulis Terdahulu.

5. KESIMPULAN

Proses pembuatan pemodelan topik menggunakan metode *Latent Dirichlet Allocation* (LDA) pada data dari salah satu artikel *Wikipedia* mengenai pandemi *Covid-19* di Indonesia diawali dengan proses *crawling* data. Data yang didapat berjumlah 6658 kata, data tersebut disimpan kedalam txt dan diubah kedalam csv. Data tersebut melalui beberapa tahapan sebelum digunakan, yakni di*cleaning* menggunakan python, translate ulang, *cleaning* manual, serta membuat label pada dataset tersebut. Kemudian data tersebut dipersiapkan untuk proses pre-processing. Hasil yang didapat dari pre-processing dihitung menggunakan *corpus* untuk melihat jumlah kemunculan kata. Jumlah kata yang sering muncul akan digunakan sebagai patokan dalam proses pemodelan LDA. Setelah mereview hasil pemodelan topik, melakukan analisis untuk mengetahui berapa lazim kata yang ada pada setiap topik. Uji dapat dilakukan dengan mengurangi jumlah topik dan iterasi. Percobaan sebanyak 5 kali dengan jumlah iterasi 100.500.1000.1500, dan 2000 dilakukan pada penelitian ini. Untuk setiap uji, gunakan jumlah topik 3,4,5,6, dan 7 untuk setiap iterasi. Hasil cluster terbaik dapat ditemukan pada topik 3. Topik 3 membutikan bahwa topik yang paling dibahas sesuai dengan data artikel yang digunakan, yaitu kesehatan.

DAFTAR PUSTAKA

[1] A. Susilo *et al.*, "Coronavirus Disease 2019: Tinjauan Literatur Terkini," *J. Penyakit Dalam Indones.*, vol. 7, no. 1, p. 45, 2020, doi: 10.7454/jpdi.v7i1.415.

[2] B. Wijayanto, Y. I. Kurniawan, T. Cahyono, and I. P. Jati, "Information System for Monitoring Community Participant Program Services in the Covid-19 Pandemic Era," *J. Tek. Inform.*, vol. 3, no. 1, pp. 37–44, 2022.

[3] N. Aeni, "Pandemi COVID-19: Dampak Kesehatan, Ekonomi, & Sosial," *J. Litbang Media Inf. Penelitian, Pengemb. dan IPTEK*, vol. 17, no. 1, pp. 17–34, 2021, doi: 10.33658/jl.v17i1.249.

[4] Ardoni, "Evaluasi Sumber Informasi Digital: Wikipedia," *Shaut Al-Maktabah J. Perpustakaan, Arsip dan Dokumentasi*, vol. 12, no. 1, pp. 1–10, 2020, doi: 10.37108/shaut.v12i1.302.

[5] Fitri, K. R. R, A. Rahmansyah, and W. Darwin, "Penggunaan Bahasa Pemrograman Python Sebagai Pusat Kendali Pada Robot 10-D," *5th Indones. Symp. Robot. Syst. Control*, pp. 23–26, 2017.

[6] Y. Guo, S. Han, Y. Li, C. Zhang, and Y. Bai, "K-Nearest Neighbor combined with guided filter for hyperspectral image classification," *Procedia Comput. Sci.*, vol. 129, pp. 159–165, 2018, doi: 10.1016/j.procs.2018.03.066.

[7] M. L. C. Chilmi, "Latent dirichlet allocation lda untuk mengetahui topik pembicaraan warganet twitter tentang omnibus law," *Repository.Uinjkt.Ac.Id*, 2021, [Online]. Available: https://repository.uinjkt.ac.id/dspace/handle/123456789/56724%0Ahttps://repository.uinjkt.ac.id/dspace/bitstream/123456789/56724/1/M. LUVIAN CHISNI CHILMI-FST.pdf

[8] F. Rashif, G. Ihza Perwira Nirvana, M. Alif Noor, and N. Aini Rakhmawati, "Implementasi LDA untuk Pengelompokan Topik Cuitan Akun Bot Twitter bertagar #Covid-19 LDA Implementation for Topic of Bot's Tweets with #Covid-19 Hashtag," *Cogito Smart J.* |, vol. 7, no. 1, pp. 170–181, 2021.

[9] C. Naury, D. H. Fudholi, and A. F. Hidayatullah, "Topic Modelling pada Sentimen Terhadap Headline Berita Online Berbahasa Indonesia Menggunakan LDA dan LSTM," *J. Media Inform. Budidarma*, vol. 5, no. 1, p. 24, 2021, doi: 10.30865/mib.v5i1.2556.

[10] Y. I. Kurniawan, E. Soviana, and I. Yuliana, "Merging Pearson Correlation and TAN-ELR algorithm in recommender system," *AIP Conf. Proc.*, vol. 1977, no. June 2018, 2018, doi: 10.1063/1.5042998.

[11] Sugiono, S. Nurdiani, S. Linawati, R. A. Safitri, and E. P. Saputra, "Pengelompokan Perilaku Mahasiswa Pada Perkuliahan E-Learning dengan K-Means Clustering," *J. Kaji. Ilm.*, vol. 19, no. 2, pp. 126–133, 2019.

[12] W. Dhuhita, "Clustering Menggunakan Metode K-Mean Untuk Menentukan Status Gizi Balita," *J. Inform. Darmajaya*, vol. 15, no. 2, pp. 160–174, 2015.

[13] G. Gustientiedina, M. H. Adiya, and Y. Desnelita, "Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan," *J. Nas. Teknol. dan Sist. Inf.*, vol. 5, no. 1, pp. 17–24, 2019, doi: 10.25077/teknosi.v5i1.2019.17-24.

[14] O. Menggunakan Metode, "Sistem Monitoring Percakapan Pada Toko," 2018.

[15] P. Studi, T. Informatika, F. Sains, D. A. N. Teknologi, U. Islam, and N. Syarif, "Penerapan Tokenisasi Kalimat Dan Metode Tf (Term Frequency) Pada Peringkas Teks Otomatis Penerapan Tokenisasi Kalimat Dan Metode Tf (Term Frequency) Pada Peringkas Teks Otomatis," 2014.

[16] 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.

[17] P. S. Statistika, F. Matematika, D. A. N. Ilmu, P. Alam, and U. I. Indonesia, "Latent Dirichlet Allocation Untuk Pemodelan Tugas Akhir Latent Dirichlet Allocation Untuk Pemodelan," 2020.

[18] B. W. Arianto and G. Anuraga, "Topic Modeling for Twitter Users Regarding the 'Ruanggguru' Application," *J. ILMU DASAR*, vol. 21, no. 2, p. 149, 2020, doi: 10.19184/jid.v21i2.17112.