TESIS - SS141 2501

PENGELOMPOKAN TOPIK PUBLIKASI MENGGUNAKAN *HIERARCHICAL DIRICHLET PROCESS* (HDP), *LATENT DIRICHLET ALLOCATION* (LDA), DAN LDA2VEC (Studi Kasus: Publikasi Terkait COVID-19)

RAKHMAH WAHYU MAYASARI
06211850012004

Dosen Pembimbing
Dr. Dra. Kartika Fithriasari, M.Si
Dr.rer.pol Dedy Dwi Prestyo, M.Si

DEPARTEMEN STATISTIKA
FAKULTAS SAINS DAN ANALITIKA DATA
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
2021

TESIS - SS141 2501

PENGELOMPOKAN TOPIK PUBLIKASI MENGGUNAKAN *HIERARCHICAL DIRICHLET PROCESS* (HDP), *LATENT DIRICHLET ALLOCATION* (LDA), DAN LDA2VEC (Studi Kasus: Publikasi Terkait COVID-19)

RAKHMAH WAHYU MAYASARI
06211850012004

Dosen Pembimbing
Dr. Dra. Kartika Fithriasari, M.Si
Dr.rer.pol Dedy Dwi Prestyo, M.Si

DEPARTEMEN STATISTIKA
FAKULTAS SAINS DAN ANALITIKA DATA
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
2021

(Halaman ini sengaja dikosongkan)

THESIS - SS141 2501

ALLOCATION OF TOPIC PUBLICATIONS WITH HIERARCHICAL DIRICHLET PROCESS (HDP), LATENT DIRICHLET ALOCATION (LDA), AND LDA2VEC (Case Study: Publication Related to Covid-19)

RAKHMAH WAHYU MAYASARI
06211850012004

Supervisor
Dr. Dra. Kartika Fithriasari, M.Si
Dr.rer.pol Dedy Dwi Prestyo, M.Si

DEPARTMENT OF STATISTICS
FACULTY OF SCIENCE AND DATA ANALYTICS
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
2021

(Halaman ini sengaja dikosongkan

LEMBAR PENGESAHAN TESIS

Tesis disusun untuk memenuhi salah satu syarat memperoleh gelar
Magister Statistika (M.Stat)
Di
Institut Teknologi Sepuluh Nopember
Oleh:
RAKHMAH WAHYU MAYASARI
NRP: 06211850012004

Tanggal Ujian: 26 Februari 2021
Periode Wisuda: April 2021

Disetujui Oleh:
Pembimbing:

1. Dr. Dra. Kartika Fithriasari, M.Si
   NIP. 19691212 199303 2 002 .....

2. Dr.rer.pol. Dedy Dwi Prastyo, M.Si
   NIP: 19831204 200812 1 002 .....

Penguji:

1. Prof. Dr.rer.pol Heri Kuswanto, M.Si
   NIP: 19820326 200312 1 004 .....

2. Dr. Wibawati, S.Si, M.Si
   NIP: 19741213 199802 2 001 .....

Kepala Departemen Statistika
Fakultas Sains dan Analitika Data

KEMENTERIAN PENDIDIKAN DAN KEBUDAYAAN INSTITUT TEKNOLOGI SEPULUH NOPEMBER FAKULTAS SAINS DAN ANALITIKA DATA DEPARTEMEN STATISTIKA

Dr. Dra. Kartika Fithriasari, M.Si.
NIP: 19691212 199303 2 002

(Halaman ini sengaja dikosongkan)

PENGELOMPOKAN TOPIK PUBLIKASI MENGGUNAKAN *HIERARCHICAL DIRICHLET PROCESS* (HDP), *LATENT DIRICHLET ALLOCATION* (LDA), DAN LDA2VEC
(STUDI KASUS: PUBLIKASI TERKAIT COVID-19)

Nama Mahasiswa	: Rakhmah Wahyu Mayasari
NRP	: 06211850012004
Dosen Pembimbing I	: Dr. Dra. Kartika Fithriasari, M.Si
Dosen Pembimbing II	: Dr.rer.pol Dedy Dwi Prestyo, M.Si

ABSTRAK

COVID-19 merupakan penyakit yang ditimbulkan oleh virus corona, dimana hampir seluruh negara menjadi negara terdampak. Efek yang mendunia ini membuat banyak ilmuan melakukan penelitian terkait COVID-19. Dari semua penelitian yang telah dipublikasi ingin diketahui topik-topik apa saja yang telah dilakukan oleh peneliti diberbagai negara. Data yang digunakan adalah abstrak dari publikasi terkait COVID-19 mulai Januari 2020 sampai dengan Agustus 2020. Pengambilan data dilakukan dengan scraping pada web resmi science direct. Data abstrak yang diperoleh kemudian dilakukan praproses data dengan menghilangkan punktuation, melakukan data lower, Lemmatizer, dan stopwords. Selanjutnya data yang sudah bersih siap untuk dilakukan analisis pengelompokan menggunakan metode teks mining untuk mendapatkan alokasi topik-topik dari abstrak terkait COVID-19, sehingga dapat digunakan sebagai acuan penelitian kedepan. Metode yang digunakan adalah Hierarchical Dirichlet Process (HDP), Latent Dirichlet Allocation (LDA), dan LDA2vec. Dari hasil analisis yang dilakukan, diperoleh hasil bahwa berdasarkan skor coherence metode yang terbaik adalah LDA, dimana jumlah topik optimumnya adalah 14.

Kata Kunci: *Hierarchical Dirichlet Process, Latent Dirichlet Allocation, LDA2vec, Topic modeling*

(Halaman ini sengaja dikosongkan)

ALLOCATION OF TOPIC PUBLICATIONS WITH HIERARCHICAL DIRICHLET PROCESS (HDP), LATENT DIRICHLET ALOCATION (LDA), AND LDA2VEC
(CASE STUDY: PUBLICATION RELATED TO COVID-19)

By	: Rakhmah Wahyu Mayasari
Student Identity Number	: 06211850012004
Supervisor	: Dr. Dra. Kartika Fithriasari, M.Si
Co-Supervisor	: Dr.rer.pol Dedy Dwi Prestyo, M.Si

ABSTRACT

COVID-19 is a disease caused by the novel coronavirus, in which almost all countries are affected. This worldwide effect has led many researchers to conduct research related to COVID-19. It is wanted to know what topics have been carried out from all the studies published by researchers in various countries. This research analyzes the data crawled from full abstracts of publications related to COVID-19 start January to August 2020. The data collected by scraping it on the official Science Direct web. The abstract's text was crawled and then preprocessed by eliminating punctuation, lowering text, lemmatizer, and stopword. Furthermore, clean data is ready for analysis using the text mining method to allocate topics and be used as future research information. The methods used are the Hierarchical Dirichlet Process (HDP), Latent Dirichlet Allocation (LDA) approaches, and LDA2vec. From the results of the analysis carried out, it is found that based on the coherence score the best method is LDA, where the optimum number of topics is 14.

Keywords: *Hierarchical Dirichlet Process, Latent Dirichlet Allocation, LDA2vec, Topic modeling*

(Halaman ini sengaja dikosongkan)

KATA PENGANTAR

Alhamdulillah, segala puji syukur bagi Allah SWT yang telah melimpahkan rahmat nikmat dan hidayah kepada makhluk-Nya serta sholawat kepada Nabi Muhammad SAW sehingga penulis dapat menyelesaikan laporan tesis dengan judul "**Pengelompokan Topik Publikasi Menggunakan *Hierarchical Dirichlet Process* (HDP), Latent Dirichlet Allocation (LDA), dan LDA2vec (Studi Kasus: Publikasi Terkait Covid-19)"**. Keberhasilan dalam penyusunan tesis ini tidak terlepas dari bantuan banyak pihak yang telah berperan serta dan membantu suksesnya penulisan laporan ini. Pada kesempatan ini penulis ingin menyampaikan ucapan terima kasih kepada:

1. Ibu Dr. Dra. Kartika Fithriasari, M.Si dan Bapak Dr.rer.pol Dedy Dwi Prastyo, M.Si selaku dosen pembimbing yang telah membimbing, mengarahkan, dan memberikan dukungan bagi penulis untuk dapat meyelesaikan tesis ini.
2. Bapak Prof. Dr.rer.pol Heri Kuswanto, M.Si dan Ibu Dr. Wibawati, M.Si selaku dosen penguji yang telah meluangkan waktu dan banyak memberi saran dan masukan selama penyusunan laporan tesis ini.
3. Ibu Dr. Dra. Kartika Fithriasari, M.Si selaku Kepala Departemen Statistika ITS.
4. Ibu Dr. Santi Wulan Purnami, M.Si, Ph.D selaku Sekretaris Departemen I (Bidang Akademik, Kemahasiswaan, Penelitian dan Pengabdian Kepada Masyarakat) dan Ibu Dr. Vita Ratnasari, M.Si selaku Sekretasis Departemen II (Bidang Sumber Daya Keuangan, Sumber Daya Manusia, dan Sarana Prasarana) yang telah memfasilitasi sarana dan prasarana untuk kegiatan belajar mengajar selama studi di ITS.
5. Bapak Dr. rer.pol Dedy Dwi Prastyo, M.Si selaku Ketua Program Studi Pascasarjana Statistika ITS.
6. Bapak dan Ibu Dosen pengajar serta seluruh karyawan Departemen Statistika FSAD-ITS Surabaya.
7. Bapak Mardjito, Ibu Nurul Hidayah (Alm), Vita Wahyu Ningtyas, Dian Wahyu Hidayat, Rizka Wahyu Novitasari dan semua keluarga di

Tulungagung atas doa, kasih sayang, dukungan, semangat dan segala yang telah diberikan untuk penulis sehingga dilancarkan dalam menyelesaikan tesis ini.

8. Aziz Ainun Najib, Fitriana Dewi Sugiono, Dinda Parasivan, Fransiska Kristin D., dan Mary Happy, yang telah membantu ketika penulis memiliki masalah akademik maupun non akademik dengan memberikan semangat, perhatian dan waktu selama mengerjakan laporan tesis ini.
9. Rekan-rekan pascasarjana Departemen Statistika FSAD-ITS Surabaya atas semangat dan dukungan dalma menyelesaikan tesis ini.
10. Keluarga besar Institut Teknologi Sepuluh Nopember serta semua pihak lain yang membantu dan terlibat dalam penyusunan tesis ini hingga selesai, yang tidak bisa penulis sebutkan satu per satu.

Penulis menyadari bahwa laporan Tugas Akhir ini masih jauh dari kata sempurna, oleh karena itu penulis sangat mengharapkan kritik dan saran yang membangun agar berguna untuk perbaikan berikutnya. Semoga laporan Tugas Akhir ini bermanfaat

Surabaya, 3 Maret 2021

Penulis

DAFTAR ISI

DAFTAR TABEL

(Halaman ini sengaja dikosongkan)

DAFTAR GAMBAR

(Halaman ini sengaja dikosongkan)

BAB I
PENDAHULUAN

1.1 Latar Belakang

Coronavirus merupakan keluarga besar dari virus-virus yang menyebabkan penyakit mulai dari gejala ringan sampai dengan gejala berat. Pada akhir tahun 2019 muncul suatu jenis virus dari *Coronavirus*, yakni *Coronavirus Disease* 2019 (COVID-19). COVID-19 adalah penyakit jenis baru yang belum pernah diidentifikasi sebelumnya pada manusia dimana virus penyebab COVID-19 dinamakan Sars-CoV-2. Pada 31 Desember 2019, WHO *China Country Office* melaporkan kasus pneumonia yang tidak diketahui etiologinya di Kota Wuhan, Provinsi Hubei, Cina. Pada tanggal 7 Januari 2020, Cina mengidentifikasi pneumonia yang tidak diketahui asal muasalnya tersebut sebagai jenis baru dari *Coronavirus*. Pada tanggal 30 Januari 2020 WHO telah menetapkan penyakit ini sebagai Kedaruratan Kesehatan Masyarakat Yang Meresahkan Dunia (KKMMD). Penambahan jumlah kasus COVID-19 berlangsung cukup cepat dan sudah terjadi penyebaran antar negara. Sampai dengan tanggal 25 Maret 2020, kasus dilaporkan di 192 negara/wilayah (Pedoman Pencegahan Pengendalian Coronavirus Disease (Kementrian Kesehatan RI, 2020). Bermula tanggal 2 Maret 2020, Indonesia melaporkan kasus konfirmasi COVID-19 sebanyak 2 kasus yang bermula di Kota Depok Jawa Barat dan kemudian ditindak lanjuti secara serius oleh pemerintah. Adapun peraturan-peraturan baru dari pemerintah guna menekan menyebaran COVID-19, misalnya Pembatasan Sosial Beskala Besar (PSBB). PSBB merupakan salah satu peraturan pemerintah terkait persebaran COVID-19 ini, dimana terdapat 2 provinsi dan 21 kabupaten/kota yang menerapkan peraturan tersebut. Sektor pendidikan juga terdampak yaitu dengan melakukan kegiatan belajar mengajar melalui daring pada semua jenjang Pendidikan. Begitu pula dengan beberapa perkantoran, melakukan *Work From Home* (WFH) kepada beberapa pegawai yang memungkinkan, serta melakukan shift kerja guna menghindari terjadinya kerumunan. Bahkan tak banyak dari masyarakat yang terpaksa kehilangan pekerjaannya, karena perusahaan-perusahaan mengalami

krisis hingga harus melakukan pemutusan kerja kepada beberapa pegawainya. Dengan hal tersebut, secara otomatis sektor ekonomi juga sangat terdampak. Peraturan untuk menghindari kerumunan juga berdampak pada sektor pariwisata, dimana tempat-tempat pariwisata ditutup supaya tidak terjadi kerumunan yang memungkinkan terjadi penularan penyakit COVID-19. Serta sektor yang sangat terdampak adalah sektor kesehatan, dimana banyaknya pasien COVID-19 yang membutuhkan penanganan khusus. Hal ini membuat para dokter dan perawat harus ekstra dalam menjalankan tugasnya. Semua dampak yang terjadi ini tidak hanya dialami oleh Indonesia, melainkan seluruh negara terdampak juga merasakannya. Oleh sebab itu banyak ilmuan dalam maupun luar negeri yang melakukan penelitian terkait COVID-19.

Penelitian sudah banyak dilakukan oleh ilmuan, yang mana telah dilakukan publikasi sehingga masyarakat bisa mendapatkan hasil dari penelitian-penelitian mengenai COVID-19 ini. Sebagian besar hasil penelitian telah terangkum dalam basis data *Science direct*. *Science direct* merupakan basis data yang berisikan kumpulan dokumen *full-teks* berkualitas yang telah diperiksa oleh *reviewer Elsevier*. Telah terkumpul lebih dari 1,2 juta artikel yang dapat diakses dengan bebas melalui basis data *Science direct* tersebut. Pada basis data *Science direct* terdapat pula ribuan hasil penelitian terkait COVID-19. Namun dari ribuan penelitian terkait COVID-19 ini belum diketahui fokus apa yang terkandung didalamnya. Sehingga hal tersebut melatar belakangi penelitian ini, untuk mengetahui topik apa saja yang terkandung diseluruh penelitian terkait COVID-19. Analisis yang tepat untuk digunakan dalam penelitian ini adalah analisis *teks mining* yang merupakan bagian dari analisis *data mining*.

Teks mining merupakan suatu analisis yang mana data yang menggunakan data berupa *teks*. Analisis ini merupakan cabang dari ilmu *data mining* yang dilakukan untuk memperoleh informasi berkualitas dari suatu rangkaian *teks* didalam sebuah dokumen (Fithriasari, Mayasari, Iriawan, & Winahju, 2020). Dalam *teks mining*, terdapat teknik *topic modeling* yang mana sangat tepat untuk dilakukan pada analisis ini. *Topic modeling* digunakan untuk memperoleh topik-topik yang sesuai dengan sekumpulan dokumen. Konsep dasar

dari *topic modeling* ini sama seperti pembagian klaster, dimana data yang digunakan akan dibagi berdasarkan kesamaan/karakteristik dalam data tersebut. Sehingga klaster yang terbentuk akan memiliki kesamaan didalam klasternya, dan memiliki perbedaan pada antar klaster, atau dapat juga disebut bahwa dalam klaster bersifat homogen sedangkan antar klaster bersifat heterogen. Sama halnya dengan konsep klaster, pada *topic modeling* juga terdapat kesamaan karakteristik dalam topik yang terbentuk, dan heterogen antar topik. Dalam *topic modeling*, klaster yang terbentuk terdiri dari data observasi yang mempunyai ciri yang sama, sehingga klaster yang terbentuk adalah klaster berisikan dokumen dengan tema yang sama. Adapun metode-metode dalam *topic modeling* diantaranya adalah *Latent Semantic Indexing* (LSI), *probabilistic Latent Semantic Indexing* (pLSI), *Latent Dirichlet Allocation* (LDA), *Hirarchical Dirichlet Process* (HDP), dan LDA2vec.

Metode analisis *topic modeling* yang sampai sekarang masih dalam pengembangan, bermula pada tahun 1988 dimana metode yang diangkat oleh adalah LSI. Metode ini berprinsip bahwa kosa kata yang digunakan dalam konteks yang sama, cenderung memiliki arti yang serupa (Dumais, Furnas, & Landauer, 1988). Dimana teknik perhitungan yang digunakan adalah *Singular Velue Decomposition* (SVD). Seperti yang telah diketahui bahwa SVD merupakan teknik perhitungan aljabar linier, dimana data yang akan digunakan adalah data besar yang sangat kompleks. Sehingga dilakukan pengembangan metode LSI menjadi metode pLSI (Thomas, 1999). Dengan menggunakan metode pLSI dapat diperoleh peluang dalam setiap variabelnya, namun adapun kelemahan dari metode ini yaitu tidak mempertimbangkan urutan kata dalam kalimat. Hal ini berakibat pada hasil yang akan diperoleh, misalkan ada dua kata yang memiliki jumlah yang sama dapat menyebabkan kata tersebut memiliki arti yang sama. Berbeda halnya dengan metode LDA yang telah memperhatikan urutan kata (Blei, Ng, & Jordan, 2003). Metode LDA menggunakan perhitungan distribusi dirichlet untuk penyusunan modelnya. Dimana metode ini dikenalkan pada tahun 2003, dimana metode ini sangat populer dalam teknik analisis *topic modeling*. Hingga dilakukan penelitian terkait eksistensi metode ini yang menghasilkan kesimpulan

bahwa metode LDA adalah metode yang populer mulai tahun 2000 dampai 2017, dimana mengalahkan metode-metode sebelumnya (Li & Lei, 2019). Selanjutnya metode dalam *topic modeling* terus berkembang, hingga pada ditemukannya metode selain LDA yang menggunakan distribusi *dirichlet* dalam perhitungannya. Metode yang dimaksud adalah metode HDP, metode ini menggunakan dirichlet process dalam setiap topiknya (Teh, Jordan, Beal, & Blei, 2006). Meskipun metode HDP ini disebut sebagai perpanjangan metode LDA, namun terdapat suatu penelitian yang menghasilkan kesimpulan bahwa kinerja dari HDP masih kalah dengan LDA dalam pembentukan topik yang bermakna. Dari penelitian ini diketahui bahwa metode LDA yang menghasilkan topik dengan makna yang lebih baik dari HDP (Bastani, Namavari, & Shaffer, 2019). Oleh sebab itu, metode LDA dirasa layak untuk dibandingkan dengan metode HDP. Selanjutnya, adapun metode terbaru dari *topic modeling* yaitu LDA2vec, dimana metode ini dikembangkan oleh Christoper Moody (Moody, 2016). Pada tahun 2016 Moody mengenalkan metode *topic modeling* dari perluasan LDA, yaitu LDA2vec. Perluasan pada LDA2vec ini dilakukan dengan memasukkan vector kata dalam pembentukan klaster/topik yang dibentuk.

Berdasarkan kelebihan dan kekurangan setiap metode, dapat dilakukan penentuan metode yang dianggap layak untuk dibandingkan. Ditentukan metode HDP, LDA, dan LDA2vec yang akan dibandingkan dalam penelitian ini. Dimana metode HDP, LDA, dan LDA2vec merupakan metode-metode dari *topic modeling* yang tepat digunakan untuk penentuan topik dari suatu dokumen teks. Dokumen teks yang digunakan adalah hasil publikasi terkait COVID-19 pada basis data *Sience Direct*. Sehingga tujuan dari penelitian ini yaitu menentukan topik-topik publikasi terkait COVID-19 dengan metode HDP, LDA, dan LDA2vec.

Sebagai tahapan dalam melakukan analisis *topic modeling* ini, data yang diperlukan adalah data dalam bentuk variabel kata. Data yang digunakan merupakan data yang diperoleh dengan cara *scraping* pada web *sience direct*. Data yang diperoleh merupakan data abstrak dengan bentuk kalimat atau bahkan paragraf. Sehingga perlu dilakukan tahap pre-*processing* data, dengan tujuan mengubah data abstrak tersebut menjadi variabel yang dapat diolah dengan

metode *topic modeling*. Pre-*processing* dilakukan dengan empat langkah, langkah pertama adalah merubah semua karakter menjadi huruf non-kapital, hal ini bertujuan untuk mempermudah dalam mengenali kata yang memiliki arti sama. Selanjutnya adalah pemutusan kalimat menjadi kata per kata, hal ini dilakukan agar data dapat dijadikan sebagai variabel kata. Langkah ketiga yaitu *lemmatizer* atau pengubahan kata menjadi kata dasar, dan yang terakhir adalah *stopwords* yaitu menghilangkan kata yang tidak unik. Setelah dilakukan keempat tahapan tersebut, maka dapat dilanjutkan dengan analisis *topic modeling*, dimana disini metode yang digunakan adalah HDP, LDA, dan LDA2vec.

Dari ketiga metode yang digunakan, akan diperoleh jumlah optimum klaster/topik berdasarkan pada perhitungan skor *coherence*. Selanjutnya akan diketahui pula komposisi dokumen pada setiap topik yang terbentuk, sehingga dapat diperoleh persentasi kontribusi dokumen dalam masing-masing topiknya. Topik disini nantinya akan ditentukan temanya, sehingga dari sini akan diketahui tema fokusan apa yang sudah dilakukan oleh peneliti di dunia terkait COVID-19. Serta kosa kata yang paling sering muncul pada setiap topik, dapat digunakan sebagai kata kunci untuk penelitian selanjutnya pada masing-masing tema fokusan. Hal ini bertujuan agar kata kunci yang digunakan dalam penelitian selanjutnya dapat terpusat.

1.2 Rumusan Masalah

COVID-19 adalah penyakit menular yang disebabkan oleh virus corona, dimana penyakit ini telah menjadi pandemi dan memberikan efek yang luar biasa pada hampir diseluruh negara. Telah dilakukan beberapa penelitian bersekala internasional terkait penyakit COVID-19 ini. Beberapa publikasi tersebut ingin diketahui kelompok topik yang terbentuk, metode pengelompokan yang digunakan adalah HDP, LDA, dan LDA2vec. Dari ketiga metode tersebut akan dipilih metode terbaik dan jumlah topik optimumnya.

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dipaparkan sebelumnya, maka tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut

1. Mengetahui karakteristik data publikasi terkait COVID-19.
2. Mengetahui hasil pre-*processings teks* pada data publikasi terkait COVID-19.
3. Mengetahui metode pengelompokan yang tepat pada data publikasi terkait COVID-19.
4. Mengetahui jumlah topik yang optimum pada data publikasi terkait COVID-19.
5. Mengetahui kata yang sering muncul dengan visualisasi *wordcloud*.
6. Mengetahui komposisi dokumen pada masing-masing topik terbentuk.

1.4 Manfaat Penelitian

Data yang digunakan pada penelitian ini merupakan hasil publikasi dalam kurun waktu Januari sampai dengan Agustus 2020. Dari data tersebut diharapkan dapat menjadi informasi tambahan terkait fokus dari penelitian terkait COVID-19 dalam kurun waktu tersebut. Sehingga kedepannya hasil dari penelitian ini dapat digunakan sebagai acuan dalam pembagian topik penelitian terkait COVID-19. Akan diketahui pula persentase dalam komposisi dokumen pada masing-masing topik, sehingga dapat diketahui tema yang sudah banyak diteliti, dan tema yang belum banyak dilakukan penelitian. Serta untuk kata yang sering muncul dapat dijadikan *keyword* dalam penelitian terkait.

1.5 Batasan Masalah

Batasan masalah yang digunakan pada penelitian ini adalah sebagai berikut.

1. Data publikasi yang digunakan adalah abstrak dari publikasi berbahasa Inggris dari basis data *Science direct* sejak Januari sampai Agustus 2020.
2. Jurnal yang digunakan adalah jurnal yang terdaftar scopus dengan *citescore* minimal 1.

3. Tipe publikasi yang digunakan adalah *review articles, research articles, case report, data articles, examinations, mini reviews,* dan *short communications*.

(Halaman ini sengaja dikosongkan)

BAB II
TINJAUAN PUSTAKA

2.1 *Teks Mining*

Teks mining merupakan suatu analisis yang mana data yang digunakan adalah data *teks*. Analisis ini merupakan cabang dari ilmu *data mining* yang dilakukan untuk memperoleh informasi berkualitas dari suatu rangkaian *teks* didalam sebuah dokumen (Fithriasari, Mayasari, Iriawan, & Winahju, 2020). Fokus utama dari *teks mining* terletak pada jumlah dokumen yang sangat besar yang kemudian dapat dikelompokkan menjadi beberapa kelompok berdasarkan jenis dokumennya. Data dokumen yang sangat besar tersebut harus menjadi data yang siap dioleh sebelum dilakukannya analisis pengelompokan. Maka sebelum dilakukan analisis atau pengelompokan pada data dokumen yang sangat banyak tersebut, perlu dilakukan persiapan data atau yang biasa disebut dengan pre-*processing data*. Pre-*processing* data dilakukan untuk menyiapkan data *teks* yang digunakan, sehingga data siap dilakukan analisis *teks mining* (Feldman & Sanger, 2007).

2.2 Pre-*processing Teks*

Pre-*processing teks* merupakan tahap persiapan data *teks* yang perlu dilakukan sebelum melanjutkan pada analisis. Pre-*processing* ini perlu dilakukan karena data *teks* mentah yang diperoleh merupakan data yang belum terstruktur dan belum dapat dilakukan proses *teks mining.* Adapun tahapan-tahapan pre-*processing* data adalah sebagai berikut.

1. *Case Folding*, tahap *case folding* merupakan tahapan untuk mengubah karakter *teks* menjadi huruf kecil semua (tidak kapital). Pada tahap ini juka akan dilakukan penghilangan tanda baca dan angka. Sehingga hasil yang akan diperoleh yaitu seluruh karakter teks menjadi tidak kapital, hilangnya tanda, dan angka (Lestari, Putra, & Cahyawan, 2013).
2. *Tokenizing*, tahap *tokenizing* merupakan tahap memutuskan kata per kata pada suatu kalimat. Tahapan ini bertujuan untuk memecah kalimat menjadi

potongan-potongan kata, sehingga urutan *string* akan terputus menjadi potongan-potongan kata penyusunnya (Vijayarani & Janani, 2016).

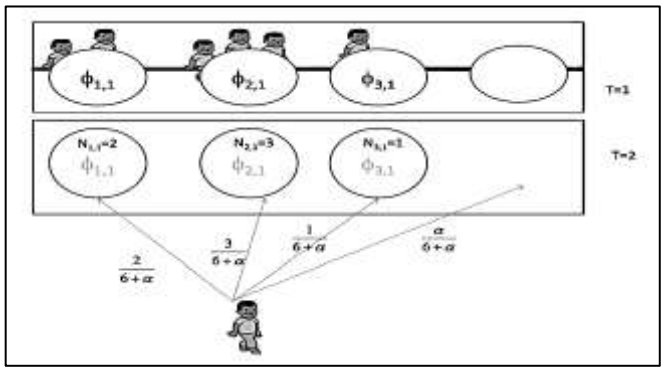
3. *Lemmatizer*, tahap *lemmatizer* merupakan tahap untuk memperoleh kata dasar. Sistem kerja tahap *lemmatizer* ini adalah menghilangkan imbuhan dalam kata. Misalnya kata "drugged", "drugs", "druging" yang mana memiliki kata dasar yang sama yaitu "drug" sehingga pada tahap steaming ini kata tersebut digantikan dengan kata "drug" (Bee & Gupta, 2016).
4. *Stopwordss*, tahap *stopwordss* merupakan tahap penghilangan kosakata yang bukan termasuk kata unik atau tidak menyampaikan pesan apapun secara signifikan pada *teks*. Kosa kata yang dimaksud seperti kata penghubung dan kata keterangan misalnya "and", "or", "from" dan sebagainya (Dragut, Fang, Sistla, Yu, & Meng, 2009).

2.3 *Topic Modelling*

Topic modeling merupakan salah satu teknik dalam *machine learning* yang tergolong dalam metode *unsupervised*. Teknik ini digunakan untuk mendapatkan topik dari sekumpulan dokumen, dimana dokumen ini berisikan teks (Blei, Ng, & Jordan, 2003). Dari sekumpulan dokumen ini akan dilakukan pemodelan statistik untuk memperoleh topik, dengan cara mendapatkan pola dalam dokumen terlebih dahulu (Anjie, 2019). Terdapat beberapa metode dalam *topic modeling*. Berawal dari metode *Latent Semantic Indexing* (LSI) pada tahun 1990, kemudian dikembangkan menjadi *probabilistic Latent Semantic Indexing* (pLSI) pada tahun 1999. Kemudian dikembangkan lagi hingga pada tahun 2003 diperkenalkan Bayesian dari pLSI yaitu *Latent Dirichlet Allocation* (LDA). Berbeda latar dengan LDA, terdapat metode dari *topic modeling* yang juga menganut prinsip Bayesian yaitu *Hierarchical Dirichlet Process* (HDP) pada tahun 2005. Dan pengembangan terbaru yaitu tahun 2016, yaitu metode LDA2vec, dimana LDA2vec adalah pengembangan dari metode LDA yang turut mempertimbangan word2vec dalam penentuan topiknya.

2.4 *Hierarchical Dirichlet Process (HDP)*

Hierarchical Dirichlet Process (HDP) merupakan metode dari klaster pada *topic modeling* yang menggunakan *mixture* model dalam pembagian komponennya. Metode ini sangat erat hubungannya dengan teknik *Chinese Restaurant Franchise Process* (CRFP), dalam perhitungan peluang dan alokasi penentuan topiknya. Ilustrasi yang digunakan dalam CRFP adalah restoran sebagai dokumen, dimana pengunjung adalah kata didalam dokumen. Dikarenakan *franchise*, jadi menu yang ditawarkan adalah sama yaitu topik. Setiap menu disajikan dalam meja yang berbeda, analog dengan setiap topik memiliki karakteristik masing-masing. Dimana nantinya customer akan memilih menu yang dihidangkan pada masing-masing meja, ini berarti setiap kata akan memilih 1 topik. Dimana peluang untuk memilih topik dibedakan menjadi dua, yaitu keadaan topik sudah dipilih kata lain, dan keadaan topik masing belum ada kata yang memilih. Peluang untuk keadaan suatu kata adalah kata pertama yang memilih topik adalah $\frac{\alpha}{i+1+\alpha}$, dan untuk peluang dalam keadaan sudah ada kata yang memilih topik ke-k adalah $\frac{n_{dk}}{i+1+\alpha}$, dimana n_{dk} adalah jumlah kata dalam dokumen yang sudah memilih topik ke-k terdahulu. Sebagai ilustrasinya akan diberikan gambaran terkait perhitungan peluang seorang customer atau sebuh kata akan memilih meja atau topik. Akan digambarkan apabila kata tersebut akan memilih topik yang belum pernah dipilih oleh kata yang lainnya, dan keadaan ketika topik tersebut sudah pernah dipilih sebelumnya.



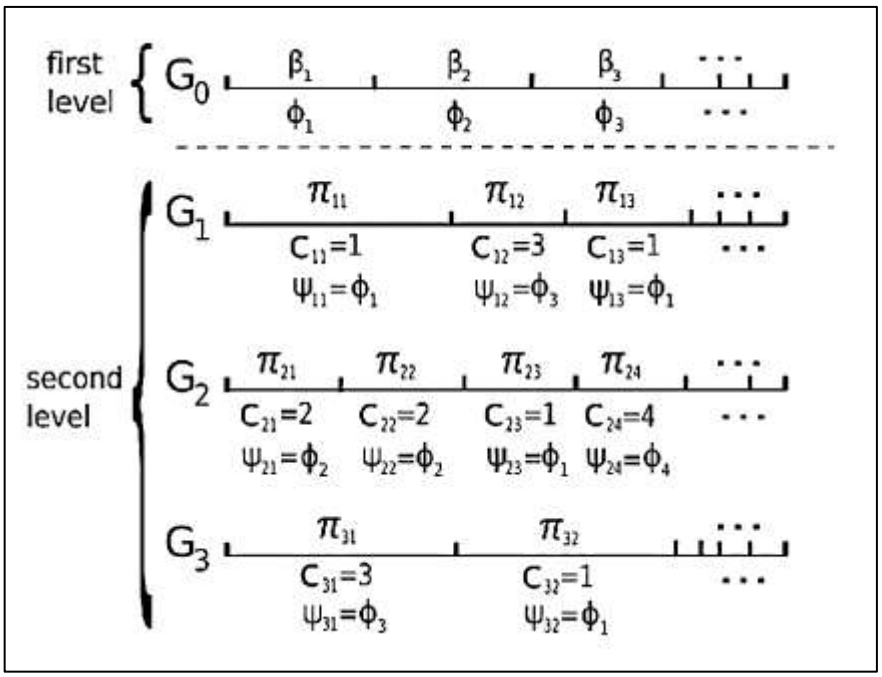
Gambar 2.1 Ilustrasi Chinese Restaurant Rranchise Processes
(Xing, 2014)

Berdasarkan pada ilustrasi pada Gambar 2.1 maka dapat terlihat jelas peluang untuk masing-masing topik yang akan dipilih oleh kata. Pada metode HDP ini diasumsikan bahwa terdapat *global probability measure* yang dinotasikan dengan G_0, dimana $G_0 \sim DP(\gamma, H)$. γ merupakan parameter dalam distribusi DP dan *H* adalah *base probability measure*. Sedangkan untuk setiap topik akan terbentuk *probability measure* G_k, dimana $G_k | G_0 \sim DP(\alpha_0, G_0)$. α_0 adalah *global concentration parameter*. Pada setiap G_k tersusun dari variabel w_{vd} yaitu kosa kata-*v* dalam dokumen-*d*. Dari variabel w_{vd}, selanjutnya dapat dibangkitkan *latent* parameter untuk masing-masing kata yaitu $\phi_{vd} | G_k \sim G_k$ dan $w_{vd} | \phi_{vd} \sim F(\phi_{vd})$. Selanjutnya untuk G_k dapat dilakukan perhitungan dengan persamaan $G_k = \sum_{v=1}^{V} \pi_{vc} \delta_{\theta_c}$, dimana π_{vc} adalah *mixing proportion* dan θ_c adalah parameter dari *mixture component*. Untuk G_0 dapat mengikuti persamaan $G_0 = \sum_{v=1}^{V} \pi_{0c} \delta_{\theta_c}$.

Namun dalam pengaplikasiannya, jika dilakukan perhitungan G_k menggunakan yang telah dijelaskan, akan diperoleh hasil yang tidak *close form*, sehingga dilakukan perhitungan dengan altenatif dari Sethuraman, sebagaimana pada persamaan (1) berikut (Sethuraman, 1994).

$$G_k = \sum_{v=1}^{V} \pi_v \delta_{w_{vd}} \tag{1}$$

Adapun Sethuraman membagi perhitungan dalma HDP menjadi dua bagian atau level, yaitu *top level* dan bottom level. Dimana *top level* menggambarkan G_0 dan *bottom level* menggambarkan G_k. dapat dibuat ilustrasi dalam pembagiannya, sebagaimana pada Gambar 2.2 berikut ini.



Gambar 2.2 Ilustrasi Pembagian Level dalam *Hierarchical Dirichlet Process* (HDP)
(Sethuraman, 1994)

Berdasarkna ilustrasi pada Gambar 2.1 maka diketahui bahwa terdapat dua level pembentuk HDP, disebut dengan *top level* dan *bottom level*. Adapun penentuan parameter pada setiap dokumen dalam suatu topik dapat ditulisakan sebagaimana pada persamaan (2) berikut.

$$\begin{aligned}
\zeta_{dnk} &\propto \exp\left(\sum_{k=1}^{K} \varphi_{dk} \mathrm{E}_q\left[\log p\left(w_{dn}\middle|\phi_k\right)\right]\right) + \mathrm{E}_q\left[\log \pi_{dk}\right] \\
\varphi_{dk} &\propto \exp\left(\sum_{n} \zeta_{dnk} \mathrm{E}_q\left[\log p\left(w_{dn}\middle|\phi_k\right)\right]\right) + \mathrm{E}_q\left[\log \beta_k\right] \\
b_{dk} &= \alpha_0 + \sum_n \sum_{s=k+1}^{K} \zeta_{dns} \\
a_{dk} &= 1 + \sum_n \zeta_{dnk}
\end{aligned} \tag{2}$$

Selanjutnya untuk menghitung *natural gradients* dapat dilakukan berdasarkan persamaan (3) berikut ini.

$$\begin{aligned}
\partial\lambda_{kw}(d) &= -\lambda_{kw} + \eta + D\sum_{t=1}^{K}\varphi_{dk}\left(\sum_{n}\zeta_{dnk} I\left[w_{dn} = w\right]\right) \\
\partial u_k(d) &= -u_k + 1 + D\sum_{k=1}^{K}\varphi_{dk} \\
\partial v_k(d) &= -v_k + \gamma + D\sum_{l=k+1}^{K}\varphi_{dl}
\end{aligned} \tag{3}$$

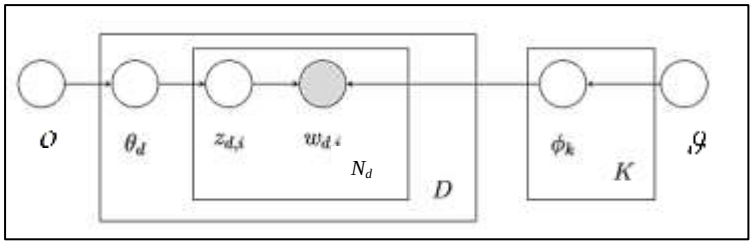
Dari persamaan (3) dapat diperoleh *natural gradient* dari masing-masing $\lambda(d), u(d)$ dan $v(d)$. Selanjutnya akan dilakukan perbaruan nilai $\lambda, u,$ dan v dengan persamaan berikut.

$$\begin{aligned}
\lambda &\leftarrow \lambda + \rho_{t_0}\partial\lambda(d) \\
u &\leftarrow u + \rho_{t_0}\partial u(d) \\
v &\leftarrow v + \rho_{t_0}\partial v(d)
\end{aligned} \tag{4}$$

Perhitungan untuk perbaruan nilai $\lambda, u,$ dan v berhenti jika melakukan perhitunggan tersebut pada seluruh kosa kata (Wang, Paisley, & Blei, 2011).

2.5 *Latent Dirichlet Allocation* (LDA)

Latent Dirichlet Allocation (LDA) adalah metode pengembangan dari pLSI yang dikembangkan oleh Blei dkk pada tahun 2003. Metode ini adalah salah satu model yang teruji efektif dalam pembuatan model pada topik-topik yang terbentuk, serta metode ini menjadi metode paling sering digunakan pada analisis terkait *text mining* pada tahun 2000-2017 (Li & Lei, 2019). Metode LDA sangat representatif dalam membuatan topik-topik disetiap dokumen, dimana pada topik tersebut mengandung distribusi multinomial dari sekumpulan kata pada tiap dokumen. Apabila digambarkan dalam suatu diagram, maka akan seperti pada Gambar 2.3 berikut.



Gambar 2.3 Ilustrasi Metode *Latent Dirichlet Allocation* (LDA) (Ponweiser, 2012)

Berdasarkan pada Gambar 2.2 maka dapat dituliskan persamaan untuk metode LDA adalah sebagai berikut.

$$p(w,z,\theta|o,\vartheta) = p(\theta|o)\,p(z|\theta)\,p(\phi|\vartheta)\,p(w|z,\phi) \tag{5}$$

Persamaan (8) terdiri dari perkalian empat peluang yaitu $p(\theta|o)$, $p(z|\theta)$, $p(\phi|\vartheta)$, dan $p(w|z,\phi)$. Dimana w adalah identitas kata ke-i dalam dokumen ke-d, z adalah identitas topik dari kata ke-i dalam dokumen ke-d, θ adalah peluang topik ke-k dalam di dokumen ke-d, ϕ adalah kata ke-I di topik ke-k, o merupakan parameter pembobot topik ke-k dalam setiap dokumen dan ϑ merupakan parameter pembobot kata ke-w pada topik. Nilai o pada setiap dokumen adalah sama, dan parameter ϑ juga demikian.

Dari keempat penyusun Persamaan (5) dapat dijabarkan sebagaimana pada Persamaan (6) berikut.

$$\begin{aligned}
p(\theta|o) &= \frac{\Gamma\left(\sum_{k=1}^{K} o_k\right)}{\prod_{k=1}^{K}\Gamma(o_k)}\theta_1^{o_k-1}\ldots\theta_K^{o_K-1} = \prod_{d=1}^{D}\frac{\Gamma(o.)}{\prod_{k=1}^{K}\Gamma(o_k)}\prod_{k=1}^{K}\theta_{d,k}^{o_k-1}\\
p(z|\theta) &= \prod_{d=1}^{D}\prod_{k=1}^{K}\theta_{d,k}^{N_{d,k,\cdot}}\\
p(\phi|\vartheta) &= \prod_{k=1}^{K}\frac{\Gamma(\vartheta_{k,\cdot})}{\prod_{v=1}^{V}\Gamma(\vartheta_{k,v})}\prod_{v=1}^{V}\phi_{k,v}^{\vartheta_{k,v}-1}\\
p(w|z,\phi) &= \prod_{k=1}^{K}\prod_{v=1}^{V}\phi_{k,v}^{N_{d,\cdot,k,v}}
\end{aligned} \tag{6}$$

dimana K adalah jumlah seluruh topik, V adalah jumlah kosa kata dalam corpus, N_d adalah jumlah dokumen kata pada corpus, dan D adalah jumlah keseluruhan dokumen. Selanjutnya untuk masing-masing komponen yang meliputi $p(\theta|o)$, $p(z|\theta)$, $p(\phi|\vartheta)$, dan $p(w|z,\phi)$ dapat digabungkan seperti pada Persamaan (10) berikut.

$$
\begin{aligned}
p\left(w,z,\theta|o,\vartheta\right) &= p\left(\theta|o\right)p\left(z|\theta\right)p\left(\phi|\vartheta\right)p\left(w|z,\phi\right) \\
&= \left(\prod_{d=1}^{D}\frac{\Gamma\left(o.\right)}{\prod_{k=1}^{K}\Gamma\left(o_k\right)}\prod_{k=1}^{K}\theta_{d,k}^{o_k-1}\right)\left(\prod_{d=1}^{D}\prod_{k=1}^{K}\theta_{d,k}^{N_{d,k,.}}\right)\times \\
&\left(\prod_{k=1}^{K}\frac{\Gamma\left(\vartheta_{k,.}\right)}{\prod_{v=1}^{V}\Gamma\left(\vartheta_{k,v}\right)}\prod_{v=1}^{V}\phi_{k,v}^{\vartheta_{k,v}-1}\right)\left(\prod_{k=1}^{K}\prod_{v=1}^{V}\phi_{k,v}^{N_{d.,k,v}}\right) \qquad (7) \\
&= \left(\prod_{d=1}^{D}\frac{\Gamma\left(o.\right)}{\prod_{k=1}^{K}\Gamma\left(o_k\right)}\prod_{k=1}^{K}\theta_{d,k}^{o_k+N_{d,k,.}-1}\right)\times \\
&\left(\prod_{k=1}^{K}\frac{\Gamma\left(\vartheta_{k,.}\right)}{\prod_{v=1}^{V}\Gamma\left(\vartheta_{k,v}\right)}\prod_{v=1}^{V}\phi_{k,v}^{\vartheta_{k,v}+N_{d.,k,v}-1}\right)
\end{aligned}
$$

Dapar dibuat ilustrasi untuk masing-masing komponen peluang tersebut. komponen pertama adalah $p\left(\theta|o\right)$ yang merupakan peluang dari distribusi topik pada setiap dokumen, dapat diilustrasikan pada Gambar 2.4 berikut.

		Topik 1	Topik 2	Topik 3	Topik 4
Dokumen 1	$\theta_{d=1}$	0.5	0.2	0.3	0.1
Dokumen 2	$\theta_{d=2}$	0	0.9	0.1	0.1
Dokumen 3	$\theta_{d=3}$	0.02	0.48	0.25	0.25

Gambar 2.4 Ilustrasi Peluang Distribusi Topik Setiap Dokumen $p(\theta|o)$
(Ponweiser, 2012)

Selanjutnya untuk $p\left(z|\theta\right)$ adalah peluang topik pada setiap kata, dalam masing-masing dokumen, yang dapat diilustrasikan sebagaimana pada Gamabr 2.5 berikut.

		Kata ke-1	Kata ke-2	Kata ke-3	Kata ke-4	Kata ke-5	Kata ke-6
Dokumen 1	$z_{d=1}$	Topik k=2	Topik k=1	Topik k=1	Topik k=4	Topik k=3	Topik k=3
Dokumen 2	$z_{d=2}$	Topik k=2	Topik k=3	Topik k=2	Topik k=2		
Dokumen 3	$z_{d=3}$	Topik k=4	Topik k=2	Topik k=2	Topik k=4	Topik k=3	

Gambar 2.5 Ilustrasi Peluang Topik setiap Kata $p(z|\theta)$
(Ponweiser, 2012)

Peluang selanjutnya yaitu $p(\phi|\vartheta)$ merupakan peluang distribusi kosa kata pada setiap topik. Sedangkan untuk $p(w|z,\phi)$ adalah peluang setiap kosa kata pada masing-masing topik. Kedua peluang tersbeut dapat diilustrasikan sebagaimana pada Gambar 2.6 berikut.

		Kosa kata ke-1	Kosa kata ke-2	Kosa kata ke-3	Kosa kata ke-4	Kosa kata ke-V
Topik 1	$\phi_{k=1}$	0.1	0.1	0	0.7	0.1
Topik 2	$\phi_{k=2}$	0.2	0.1	0.2	0.2	0.3
Topik 3	$\phi_{k=3}$	0.01	0.2	0.39	0.3	0.1
Topik 4	$\phi_{k=4}$	0	0	0.5	0.3	0.2

Gambar 2.6 Ilustrasi Peluang Distribusi Kosa Kata setiap Topik $p(\phi|\vartheta)$ dan Peluang setiap Kosa Kata Pada Masing-Masing Topik $p(w|z,\phi)$
(Ponweiser, 2012)

Pada metode LDA, diasumsikan bahwa karakteristik pada suatu topik ditentukan oleh distribusi kosa kata dalam topik dan distribusi topik dalam dokumen. Dimana kumpulan dari kata-kata disebut dengan dokumen, dan kumpulan dari dokumen disebut dengan corpus. Sedangkan kumpulan istilah/kata pada corpus, disebut dengan kosa kata. Berdasarkan Persamaan (9) maka dapat diketahui bahwa perhitungan dalam metode LDA adalah mencari nilai optimal pada peluang $p(w,z,\theta|o,\vartheta)$, sehingga pembagian topik akan maksimal.

Adapun upaya yang dilakukan dalam optimalisasi yang biasa disebut dengan *gibbs sampling*. Optimasi ini dilakukan pada perhitungan $p(w,z,\theta|o,\vartheta)$. Dari perhitungan tersebut, dilakukan pengoptimalan nilai peluang dari $p(z|\theta)$ dan $p(w|z,\phi)$. Pada peluang $p(z|\theta)$ akan dilakukan perhitungan untuk setiap dokumen dengan persamaan $p(z|\theta)=\frac{n_{dk}+o}{N_d-1+Ko}$, n_{dk} merupakan jumlah kata dalam dokumen-d yang berada dalam topik ke-k. Selanjutnya untuk persamaan

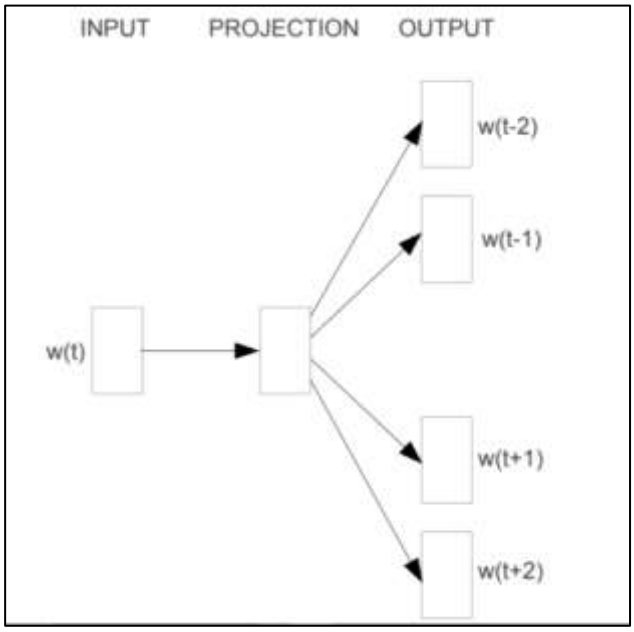
$p(w|z,\phi)$ akan dilakukan perhitungan dengan persamaan $p(w|z,\phi) = \frac{w_{v.k} + \vartheta}{\sum_{v \in V} w_{v.k} + V\vartheta}$. Dimana $v_{j.k}$ merupakan jumlah kosa kata ke-j pada topik ke-k.

Tahap selanjutnya adalah mencari irisan dari $p(z|\theta)$ dan $p(w|z,\phi)$, dengan cara mengalikan kedua peluang tersebut dan menghasilkan. Sehingga apabila dituliskan, perhitungan $p(w,z,\theta) = p(z|\theta) \times p(w|z,\phi)$. Dari hasil $p(w,z,\theta)$ tersebut, akan dipilih letak kosa kata dalam topik ketika nilai peluangnya adalah tertinggi. Akan dilakukan perulangan perhitungan sampai dengan seluruh kata dalam semua dokumen dilakukan iterasi. Apabila iterasi telah selesai maka selanjutnya hasil dari *gibbs sampling* inilah yang digunakan dalam perhitungan model LDA sebagaimana pada Persamaan (10). Dalam algoritma *gibbs sampling*, langkah pertama yang dilakukan adalah menentukan jumlah topik. Dari jumlah topik tersebut, akan dilakukan pemecahan seluruh kosa kata dalam suatu dokumen kepada masing-masing topik yang terbentuk. Selanjutnya dari pemecahan kosa kata ini, akan diperoleh model peluang setiap kata pada masing-masing topik. Sehingga dapat digunakan sebagai acuan penentuan topik pada suatu dokumen baru.

2.6 LDA2vec

Metode LDA2vec adalah salah satu pengembangan dari *topic modeling* yang mempertimbangkan istilah dari word2vec yang bertujuan untuk menpertahankan keunggulan dari informasi lokal dalam topik sehingga membuat vector dokumen dan vector topik lebih mudah untuk dipahami. Word2vec sendiri merupakan metode *embeding* untuk merepresentasikan kata menjadi vector dengan ukuran N. Vector tersebut diperoleh dari *neural network* yang mana pada word2vec terdiri dari 3 layer yaitu *input*, *projection* (*hidden layer*) dan *output*. Terdapat dua jenis arsitektur *neural network* yang digunakan dalam word2vec yaitu *Skip-gram* dan *Continuous Bag of Word* (CBOW). Perhitungan word2vec

yang digunakan dalam analisis LDA2vec ini menggunakan arsitektur *Skip-gram*. Sebagai gambaran, akan ditampilkan ilustrasi arsitektur *Skip-gram* sebagaimana pada Gambar 2.7 berikut.

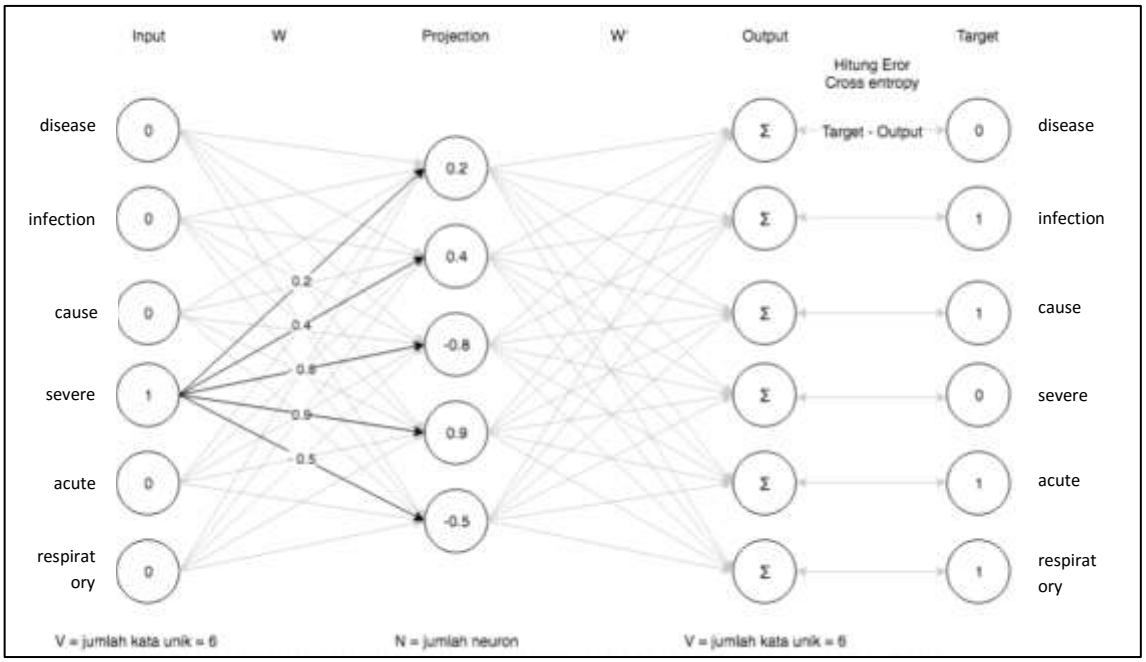


Gambar 2.7 Ilustrasi *Skip-gram*
(Mikolov, Corrado, Chen, & Dean, 2013)

Lebih jelasnya akan dilakukan ilustrasi lebih lanjut dengan menggunakan contoh. Apabila digunakan contoh kalimat dari hasil pre-*processing* adalah ‘disease’, ‘infection’, ‘cause’, ‘severe’, ‘acute’ dan ‘respiratory’ maka dapat dibentuk *one-hot encoded* vector sebagai berikut.

disease = [1,0,0,0,0,0]
infection = [0,1,0,0,0,0]
cause = [0,0,1,0,0,0]
severe = [0,0,0,1,0,0]
acute = [0,0,0,0,1,0]
respiratory = [0,0,0,0,0,1]

Dari *one-hot encoded* vector diatas, misal dilakukan perhitungan pada kata ‘severe’ maka dapat dibuat ilustrasi sebagai berikut.



Gambar 2.8 Ilustrasi *Skip-gram* pada kata ‘severe’
(Medium, 2019)

Inisialisasi bobot pada matriks W dan dan matriks W’ adalah random. Ukuran matriks W adalah V x N dan matriks W’ adalah N x V. Adapun proses *feed forward*, yaitu proses dimana vektor input (*one-hot encoded* vector) akan di kalikan *dot product* dengan bobot W sehingga menghasilkan nilai pada layer *projection*. Kemudian layer *projection* dilalakukan perkalian *dot product* dengan bobot pada matriks W’ dan menghasilkan vektor *output*. Setelah mendapatkan nilai output pada proses *feed forward*, maka akan dihitung nilai galatnya dengan menggunakan metode *cross entropy* yaitu pada bagian Target-Output.

Selanjutnya adalah tahap *back propagation* dengan memanfaatkan teknik *gradient descent* yaitu dengan melakukan update bobot pada matriks W dan W’. Proses ini akan diulang kembali ke tahap *feed forward* hingga tercapai nilai galat

minimum. Setelah diperoleh nilai galat yang minimum pada *cross entropy*, maka vektor yang merepresentasikan kata tersebut diambil dari bobot matriks W dengan cara mengalikan *dot product* antara *one-hot encoded* vector masing-masing kata dengan bobot W, sedangkan bobot pada W' akan diabaikan. Sehingga diperoleh *output* yang diinginkan, yaitu akan bernilai 1 pada target (*output*).

Contoh yang digunakan adalah contoh dengan jumlah kata 6, pada data *real* digunakan data dengan jumlha kata yang lebih banyak. Apabila perhitungan word2vec hanya menggunakan arsitektur *Skip-gram* maka tidak akan efisien, sehingga digunakan *Skip-gram Negative Sampling* (SGNS). SGNS dalam konsepnya bertugas untuk mereduksi kata-kata yang pada target (*output*) menghasilkan nilai 0. Target (*output*) yang menghasilkan nilai 0 disebut dengan *negative sampling*. Untuk menjadikan perhitungan lebih efisien, maka akan dilakukan reduksi *negative sampling*. arsitektur SGNS inilah yang digunakan word2vec dalam analisis LDA2vec. Secara matematis, SGNS dapat ditulis sebagaimana pada Persamaan 8 berikut.

$$L_{ij}^{neg} = \log \sigma\left(\mathbf{c}_j.\mathbf{w}_i\right) + \sum_{l=0}^{n} \log \sigma\left(-\mathbf{c}_j.\mathbf{w}_{neg}\right) \tag{8}$$

Selanjutnya untuk analisis LDA2vec, persamaan (8) diatas digabungkan dengan persamaan yang diperoleh dari likelihood distribusi dirichlet L^d . Sehingga dalam metode LDA2vec perumusan yang digunakan adalah sebagai berikut.

$$\begin{aligned} L &= L^d + \sum\nolimits_{ij} L_{ij}^{neg} \\ L^d &= \lambda \sum\nolimits_{jk} (\alpha - 1) \log p_{jk} \\ L_{ij}^{neg} &= \log \sigma\left(\mathbf{c}_j.\mathbf{w}_i\right) + \sum_{l=0}^{n} \log \sigma\left(-\mathbf{c}_j.\mathbf{w}_{neg}\right) \end{aligned} \tag{9}$$

L adalah total penjumlahan dari L_{ij}^{neg} dan L^d . Dimana L_{ij}^{neg} merupakan *Skip Gram Negative Sampling Loss* (SGNS), sedangkan L^d merupakan bobot dari dokumen berdasarkan *Dirichlet likelihood*. Pada L_{ij}^{neg} terdapat $\vec{w}_l$ yang merupakan

vektor kata dari *negative-sampled.* Ketika $\alpha < 1$ maka distribusi topik cenderung menyebar, sebaliknya jika $\alpha > 1$ maka distribusi topik akan lebih memusat. α dapat diperoleh dari K^{-1}, dimana K adalah jumlah topik yang ditentukan. Pada percobaan yang pernah dilakukan Moody, apabila dilakukan percobaan tanpa menggunakan parameter pendorong atau $\alpha = 1$, maka nilai pembobot menjadi tinggi dan membuat interpretasi akan lebih rumit. Karena hal itu, maka perlu dilakukan penambahan nilai α yang tidak sama dengan 1 (Moody, 2016).

2.7 *Evaluasi Topic*

Hasil dari pengelompokan topik sangat perlu dilakukan evaluasi, guna mengetahui konsistensi model dalam melakukan pengelompokan. Evaluasi topik yang digunakan adalah terhitungan nilai *coherence*, dimana *coherence* telah terbukti cocok untuk penilaian terkait kualitas topik. Dalam penelitian yang pernah dilakukan, terkait perbandingan hasil perhitungan *coherence* menggunakan perhitungan C_v. Perhitungan C_v ini menggunakan persamaan yang lebih rumit dari pada beberapa perhitungan *coherence* yang lain diantaranya UMass, UCI, dan NPMI. Perhitungan yang lebih rumit membuat durasi yang dibutuhkan untuk melakukan perhitungan juga lebih lama, namun hasil yang diperoleh akan lebih baik (Syed & Spruit, 2017). Secara umum langkah yang digunakan dalam analisis *coherence* dengan C_v yaitu yang pertama melakukan segmentasi kosa kata menjadi kata berpasangan. Kemudian menghitung peluang setiap kata dan kata berpsangan tersebut. Langkah selanjutnya adalah menghitung Θ sebagai ukuran konfirmasi, lalu merata-rata nilai Θ tersebut. Jika ditulisakan dalam persamaan akan sebagai berikut.

$$\Theta_{S_i}\left(\vec{u},\vec{w}\right)=\frac{\sum_{i=1}^{V}u_i w_i}{\left\|\vec{u}\right\|\cdot\left\|\vec{w}\right\|}$$

$$\vec{v}\left(W'\right)=\left\{\sum_{w_i\in W'}NPMI\left(w_i,w_j\right)\right\}_{j=1,2,\ldots,V} \tag{10}$$

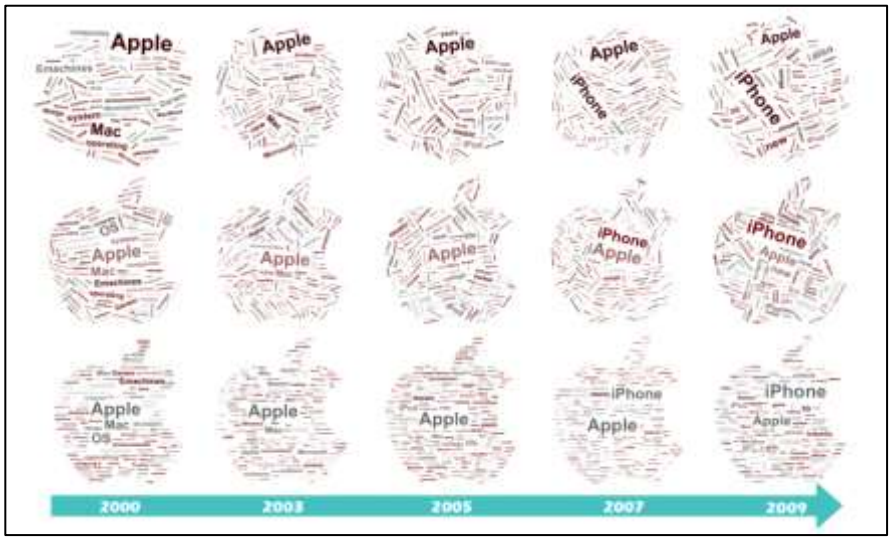
$$NPMI\left(w_i,w_j\right)=\frac{\log\dfrac{p\left(w_i,w_j\right)+\varepsilon}{p\left(w_i\right)p\left(w_j\right)}}{-\log\left(P\left(w_i,w_j\right)+\varepsilon\right)}$$

$W=\left\{W_1,W_2,\ldots,W_T\right\}$ merupakan kata dengan peluang tertinggi dalam suatu topik, atau kata dalam top-*T*. S_i adalah segmen pasangan dari setiap kata di $W'\in W$ yang berpasangan dengan kata lainnya $W^*\in W$. S merupakan set dari semua segmen yang saling berpasangan $S=\left\{\left(W',W^*\right)\middle|W=\left\{w_i\right\};w_i\in W;W^*=W\right\}$. Selanjutnya untuk peluang masing-masing kata adalah $p\left(w_i\right)$ dan untuk *joint propability* adalah $p\left(w_i,w_j\right)$. Dari peluang kata ini akan dilakukan perhitungan menggunakan metode NPMI yang kemudian dibentuk vector *context* $\vec{v}\left(W'\right)$ dan $\vec{v}\left(W^*\right)$ sebagai vector untuk semua pasangan kata dalam W, dimana $\vec{v}\left(W'\right)\in\vec{u}$ dan $\vec{v}\left(W^*\right)\in\vec{w}$. Selanjutnya dihitung Θ untuk mengetahui seberapa kuat dukungan W^* terhadap W', hal ini berdasarkan pada kemiripan dari W' dan W^* dalam kaitannya dengan semua kata dalam W. Langkah terakhir adalah menghitung rata-rata untuk seluruh pasangan kata dengan hasil perhitungan dari Θ (Syed, Spruit, 2017).

2.8 *Wordcloud*

Wordcloud merupakan salah satu alat yang efektif untuk memahami konten secara cepat dari suatu dokumen (Wu, Provan, Wei, Liu, & Ma, 2011). *Wordcloud* juga sangat representative secara visual dimana basisnya adalah data *teks*. Disini *wordcloud* menampilkan kata-kata dengan ukuran dan warna yang berbeda berdasarkan pada popularitas atau frekuensi kemunculan kata tersebut.

Dari data *teks* yang telah diperoleh kemudian dilakukan visualisasi menggunakan *wordclaud*, yang mana visualisasi *wordcloud* dapat dilakukan beberapa modifikasi diantaranya diberikan bentuk bingkai, serta pengaturan posisi kata-katanya yang digambarkan sebagaimana pada Gambar 2.9 berikut (Chi, Lin, Lin, & Lee, 2011).



Gambar 2.9 Visualisasi *Wordcloud*
(Chi, Lin, Lin, & Lee, 2011)

Berdasarkan contoh visulisasi *wordcloud* pada Gambar 2.9 maka dapat diketahui bahwa terdapat beberapa aplikasi yang dapat dilakukan pada tampilannya. Terlihat pada baris pertama adalah sekumpulan kata yang mana kata yang memiliki frekuensi tinggi kemiliki ukuran paling besar, dimana dilakukan perbandingan mulai dari tahun 2000 sampai dengan tahun 2009. Pada baris ke-2 diberikan bingkai berupa bentuk dari logo apple yang dapat membuar visualisasi lebih menarik dan masih dalah kon*teks* yang sesuai. Selanjutnya pada baris ke-3 atau baris terakhir dikethui bahwa pengaturan kata dikerhatikan, sehingga semua kata horizontal dan dapat terbaca dengan mudah.

2.9 COVID-19

Coronavirus merupakan keluarga besar dari virus-virus yang menyebabkan penyakit mulai dari gejala ringan sampai dengan gejala berat. Terdapat setidaknya dua jenis *Coronavirus* yang telah diketahui menyebabkan penyakit yang menimbulkan gejala berat yaitu Middle East Respiratory Syndrome (MERS) dan Severe Acute Respiratory Syndrome (SARS). Pada akhir tahun 2019 muncul suatu jenis dari *Coronavirus*, yakni *Coronavirus Disease* 2019 (COVID-19). COVID-19 adalah penyakit jenis baru yang belum pernah diidentifikasi sebelumnya pada manusia dimana virus penyebab COVID-19 dinamakan Sars-CoV-2. Virus corona adalah jenis virus zoonosis yaitu virus yang ditularkan antara hewan dan manusia. Penelitian menyebutkan bahwa virus corona yang telah diketahui penyebabnya seperti SARS dan MERS ditularkan dari hewan ke kemanusia. SARS ditransmisikan dari kucing luwak (civet cats) ke manusia, sedangkan MERS dari unta ke manusia. Adapun, hewan yang menjadi sumber penularan COVID-19 ini masih belum diketahui.

Tanda dan gejala umum pada infeksi COVID-19 antara lain gejala gangguan pernapasan akut seperti demam, batuk dan sesak napas. Menurut penelitian, masa inkubasi rata-rata 5-6 hari dimana masa inkubasi terpanjang 14 hari. Pada kasus COVID-19 yang berat dapat menyebabkan pneumonia, sindrom pernapasan akut, gagal ginjal, dan bahkan kematian. Tanda-tanda dan gejala klinis yang dilaporkan pada sebagian besar kasus adalah demam, dengan beberapa kasus mengalami kesulitan bernapas, dan hasil rontgen menunjukkan infiltrat pneumonia luas di kedua paru.

Dari penelitian-penelitian yang telah dilakukan, COVID-19 dapat menular antara manusia ke manusia lain melalui percikan batuk/bersin (droplet), dan tidak melalui udara. Orang yang paling berisiko tertular penyakit ini adalah orang yang kontak erat dengan pasien COVID-19 termasuk yang merawat pasien COVID-19. Adapun rekomendasi standar untuk mencegah penyebaran infeksi yang disarankan oleh pemerintah adalah melalui cuci tangan secara teratur menggunakan sabun dan air bersih, menerapkan etika batuk dan bersin, menghindari kontak secara langsung dengan ternak dan hewan liar serta

menghindari kontak dekat dengan siapapun yang menunjukkan gejala penyakit pernapasan seperti batuk dan bersin. Selain itu, menerapkan Pencegahan dan Pengendalian Infeksi (PPI) saat berada di fasilitas kesehatan terutama Unit Gawat Darurat (UGD).

Pada 31 Desember 2019, WHO China Country Office melaporkan kasus pneumonia yang tidak diketahui etiologinya di Kota Wuhan, Provinsi Hubei, Cina. Pada tanggal 7 Januari 2020, Cina mengidentifikasi pneumonia yang tidak diketahui etiologinya tersebut sebagai jenis baru *Coronavirus* (*Coronavirus Disease*, COVID-19). Pada tanggal 30 Januari 2020 WHO telah menetapkan sebagai Kedaruratan Kesehatan Masyarakat Yang Meresahkan Dunia/ Public Health Emergency of International Concern (KKMMD/PHEIC). Penambahan jumlah kasus COVID-19 berlangsung cukup cepat dan sudah terjadi penyebaran antar negara.

Sampai dengan tanggal 25 Maret 2020, dilaporkan total kasus konfirmasi 414.179 dengan 18.440 kematian (CFR 4,4%) dimana kasus dilaporkan di 192 negara/wilayah. Diantara kasus tersebut, terdapat beberapa petugas kesehatan yang dilaporkan turut terinfeksi. Bermula tanggal 2 Maret 2020, Indonesia melaporkan kasus konfirmasi COVID-19 sebanyak 2 kasus, namun sampai dengan tanggal 3 Juli 2020, Indonesia sudah melaporkan 60.695 kasus konfirmasi COVID-19 (Pedoman Pencegahan Pengendalian Coronavirus Disease (COVID-19), 2020).

BAB III
METODE PENELITIAN

3.1 Sumber Data

Data yang digunakan dalam penelitian ini adalah sekumpulan abstrak yang diperoleh dari jurnal penelitian yang telah dipublikasi pada basis data *Science direct*. Tidak semua jurnal dalan *Sience Direct* akan digunakan sebagai sumber data, melainkan ada beberapa hal yang dipertimbangkan dalam penentuan sumber data. Jenis artikel, jurnal yang digunakan, dan *keyword* (kata kunci) yang akan menjadi pertimbangannya. Penentuan jenis artikel didasarkan pada keberadaan abstrak pada artikel tersebut. Jenis artikel yang digunakan yaitu artikel yang mengandung abstrak, diantaranya adalah *review articles, research articles, case report, data articles, examinations, mini reviews,* dan *short communications*. Sehingga hanya jenis artikel tersebut yang digunakan sebagai data pada penelitian ini. Selain jenis artikel, *citescore* juga menjadi pertimbangan dalam penentuan jurnal yang digunakan. Hanya jurnal yang memiliki *citescore* minimal 1 yang dipilih sebagai sumber data. Digunakannya *citescore* sebagai penentu pemilihan jurnal, karena *citescore* merupakan produk dari scopus yang dipakai dalam melakukan penilaian jurnal, buku, ataupun prosiding pada tiap jurnal. Dari nilai *citescore* pada masing-masing jurnal, maka diperoleh 35 jurnal terpilih. 35 jurnal tersebut dapat ditabelkan sebagaimana pada Tabel 3.1 berikut.

Tabel 3.1 Jurnal Terpilih sebagai Sumber Data

No	Nama Jurnal
1	The Lancet
2	The Lancet Infectious Diseases
3	Gastroenterology
4	The Lancet Global Health
5	The Lancet Haematology
6	Acta Pharmaceutica Sinica B

Lanjutan Tabel 3.1 Jurnal Terpilih sebagai Sumber Data

No	Nama Jurnal
7	EBioMedicine
8	Science of The Total Environment
9	Journal of infection
10	Journal of the American Academy of Dermatology
11	Biomedicine & Pharmacotherapy
12	Chaos, Solitons & Fractals
13	Journal of Clinical Virology
14	Biomedical Journal
15	International Journal of Infectious Diseases
16	Journal of Pain and Symptom Management
17	Journal of Microbiology Immunology and Infection
18	Journal of the Formosan Medical Association
19	Journal of Infection and Public Health
20	The Brazilian Journal of Infectious Diseases
21	Informatics in Medicine Unlocked
22	The American Journal of Emergency Medicine
23	Journal of infection and Chemotherapy
24	Asian Journal of Psychiatry
25	Diabetes & Metabolic Syndrome: Clinical Research & Reviews
26	World Neurosurgery
27	New Microbes and New Infections
28	Medical Hypotheses
29	Annals of Medicine and Surgery
30	Procedia Manufacturin
31	Data in Brief
32	Medicina Clínica (English Edition)
33	Heliyon
34	Respiratory Medicine Case Reports
35	IDCases

Berdasarkan pada Tabel 3.1 maka dapat diketahui sejumlah 35 jurnal terpilih yang selanjutnya akan ditentukan jurnal yang terkait COVID-19. Penentuan jurnal dilakukan dengan beberapa kata kunci yaitu “COVID-19”, “2019-nCoV”, “SARS-nCOV-2”, atau “SARS-COV-2”. Sehingga jurnal yang akan digunakan merupakan jurnal dalam jurnal terpilih, yang mengandung salah satu kata kunci yang telah ditetapkan. Adapun kurun waktu yang digunakan dalam

penentuan sumber data adalah Januari 2020 sampai dengan Agustus 2020. Sehingga dari jurnal-jurnal terpilih dalam kurun waktu tersebut, selanjutnya dapat dilakukan pengambilan abstrak dengan cara *scraping*. *Scraping* dilakukan pada masing-masing publikasi dari jurnal terpilih yang berkaitan dengan COVID-19. Dari hasil *scraping* diperoleh data sejumlah jurnal terpilih yaitu 35 data, yang selanjutnya 35 data tersebut digabungkan menjadi satu data mentah yang siap dilakukan pre-*processing* data.

3.2 Struktur Data

Data yang diperoleh dari sumber data berupa dokumen jurnal. Dari dokumen tersebut harus dilakukan filter, dimana hanya dokumen yang mengandung abstrak yang digunakan. Dari dokumen yang mengandung abstrak tersebut, dapat diketahui variabel penelitian yang digunakan adalah d dan w. d adalah dokumen, dimana dokumen berisikan dokumen abstrak keseluruhan. Sedangkan w adalah kosa kata, dimana kosa kata berisikan seluruh kosa kata yang digunakan dalam seluruh dokumen. Kosa kata hanya diambil masing-masing, satu kata jika dalam keseluruhan dokumen terdapat perulangan kata tersebut. Selanjutnya dari kedua variabel tersebut dapat dibuat struktur data pada Tabel 3.2 berikut ini.

Tabel 3.2 Struktur Data Penelitian

Dokumen (d)	**Kosa Kata** (w)				
	w_1	w_2	w_3	$\cdots$	w_{17164}
d_1	$w_{1.1}$	$w_{2.1}$	$w_{3.1}$	$\cdots$	$w_{17164.1}$
d_2	$w_{1.2}$	$w_{2.2}$	$w_{3.2}$	$\cdots$	$w_{17164.2}$
d_3	$w_{1.3}$	$w_{2.3}$	$w_{3.3}$	$\cdots$	$w_{17164.3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_{2264}	$w_{1.2264}$	$w_{2.2264}$	$w_{3.2264}$	$\cdots$	$w_{17164.2264}$

Tabel 3.3 merupakan struktur data dari dokumen publikasi abstrak terkait COVID-19, dimana jumlah data yang digunakan adalah 2.264 dan jumlah kosa

kata adalah 17.164. Berdasarkan pada Tabel 3.3 maka dapat diketahui bahwa $w_{1.1}$ adalah frekuensi dari kosa kata pertama dalam dokumen pertama. Sedangkan untuk $w_{2.1}$ berisikan frekuensi dari kosa kata kedua dalam dokumen pertama, dan begitu seterusnya hingga pada $w_{17164.2264}$ adalah kosakata ke 17.164 pada dokumen ke 2.264.

3.3 Langkah Analisis

Langkah analisis yang digunakan dalam penelitian ini dapat ditulisakan sebagai berikut.

1. Mendapatkan data dengan cara *scrapping*
   a. Melakukan seleksi data dengan ketentuan jenis artikel, jurnal, kata kunci, dan periode waktu.
   b. Menyimpan abstrak, judul, dan url dari hasil seleksi sumber data pada point (a) dengan cara *scrapping*.
   c. Menyimpan hasil *scrapping* dalam bentuk file tipe csv untuk setiap jurnal, sehingga terdapat 35 file csv, dengan setiap file terdiri dari baris dan kolom. Setiap baris adalah data abstrak terkait COVID-19 dan kolom adalah abstrak, judul, dan url.
   d. Menggabungkan 35 file csv yang telah diperoleh sebelumnya, menjadi sati file csv. Dari data ini, dapat dilanjutkan pada tahap pre-*processing*.
2. Pre-*processing*Data
   a. Mengubah semua karakter teks menjadi huruf kecil (tidak kapital) serta menghilangkan tanda baca, yang disebut dengan proses *case folding.*
   b. Melakukan pemecahan kalimat menjadi penggalan kata, yang disebut dengan proses *tokenizing.*
   c. Menjadikan setiap kata menjadi kata dasar, yang disebut dengan proses *lemmatizer*.
   d. Menyiapkan kata-kata untuk *stopwordss list*, misalnya: “then”, “however”, “COVID-19”, dan lain-lain yang dianggap sebagai kata tidak unik.

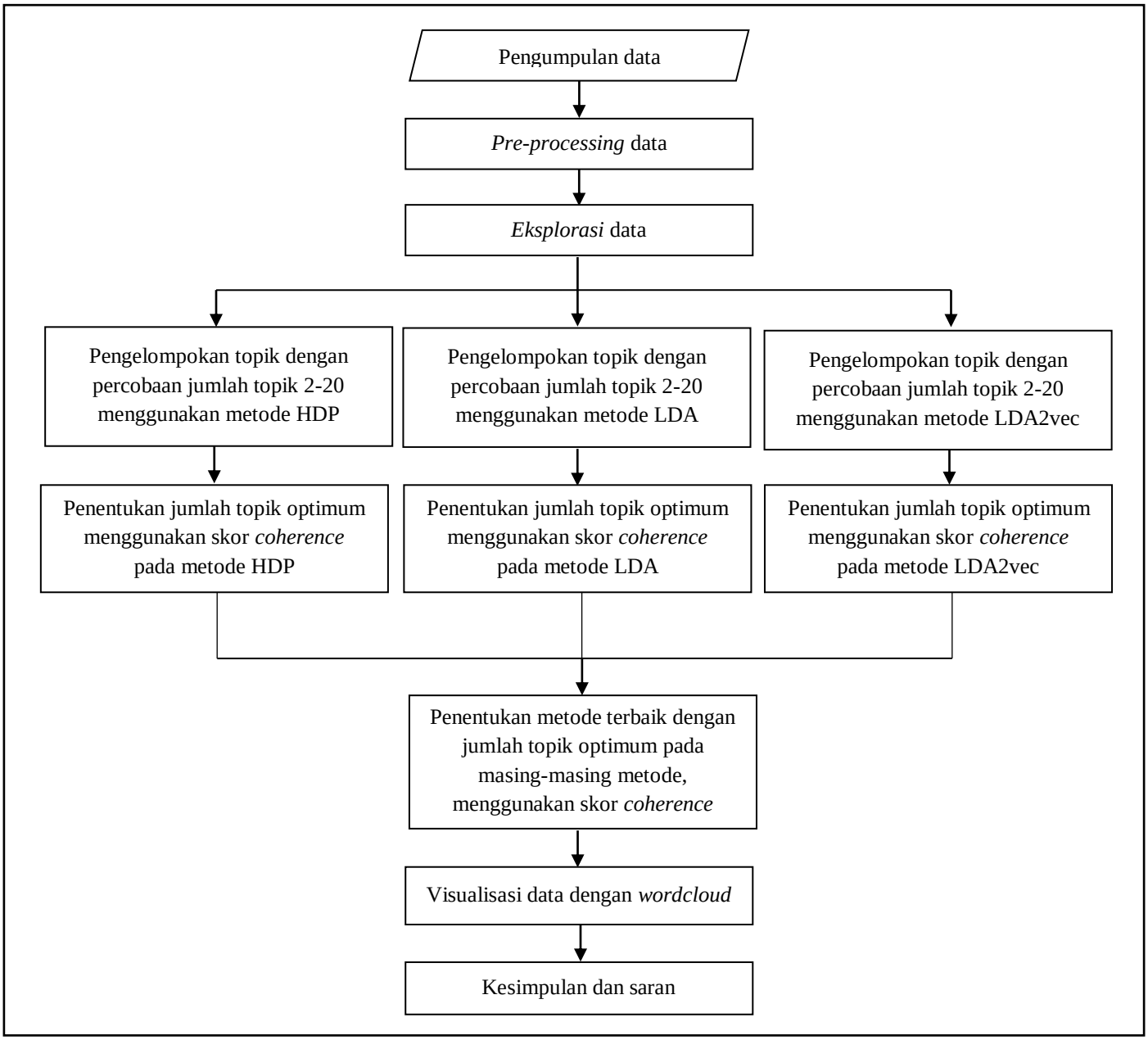
e. Melakukan proses *stopwordss*, yaitu penghapusan kata-kata penghubung ataupun kata yang dirasa tidak bermakna. Data ini digunakan sebagai variabel dalam analisis *topic modeling.*

3. Pengelompokan topik
   a. Mengelompokkan topik menggunakan metode HDP, dengan melakukan percobaan beberapa jumlah topik. Jumlah topik yang dicobakan yaitu mulai dari 2 topik sampai dengan 20 topik. Dari masing-masing jumlah topik tersebut, dilakukan perhitungan skor *coherence.* Selanjutnya dilakukan pemilihan jumlah topik terbaik, dengan melihat skor *coherence* yang tertinggi. Sehingga dengan metode HDP diketahui jumlah topik yang optimum.
   b. Mengelompokkan topik menggunakan metode LDA, dengan melakukan percobaan beberapa jumlah topik. Jumlah topik yang dicobakan yaitu mulai dari 2 topik sampai dengan 20 topik. Dari masing-masing jumlah topik tersebut, dilakukan perhitungan skor *coherence.* Selanjutnya dilakukan pemilihan jumlah topik terbaik, dengan melihat skor *coherence* yang tertinggi. Sehingga dengan metode LDA dapat diketahui jumlah topik yang optimum.
   c. Mengelompokkan topik menggunakan metode LDA2vec, dengan melakukan percobaan beberapa jumlah topik. Jumlah topik yang dicobakan yaitu mulai dari 2 topik sampai dengan 20 topik. Dari masing-masing jumlah topik tersebut, dilakukan perhitungan skor *coherence.* Selanjutnya dilakukan pemilihan jumlah topik terbaik, dengan melihat skor *coherence* yang tertinggi. Sehingga dengan metode LDA2vec dapat diketahui jumlah topik yang optimum.
4. Penentuan metode terbaik, dengan membandingkan skor *coherence* pada jumlah topik optimum dimasing-masing metode. Sehingga diperoleh metode terbaik dengan jumlah topik optimum pada data publikasi terkait COVID-19.
5. Visualisasi hasil pengelompokan topik yang telah optimum dengan menggunakan *wordcloud*.

6. Membuat kesimpulan dan memberikan informasi untuk peneliti yang akan fokus pada penelitian terkait COVID-19.

3.3 Diagram Alir

Langkah analisis yang telah dijelaskan sebagaimana pada sub bab sebelumnya dapat digambarkan dengan diagram alir sebagai berikut.



Gambar 3.1 Diagram Alir

BAB IV
ANALISIS DAN PEMBAHASAN

Pada bab ini akan dibahas mengenai hasil analisis data publikasi terkait COVID-19. Data yang digunakan merupakan abstrak dengan penggunaan jenis artikel, kata kunci, dan jurnal tertentu. Pengambilan data dilakukan dengan cara *scraping* pada 35 jurnal terpilih. Proses *scrapping* dilakukan dengan menggunakan *software* spider, untuk lebih jelasnya akan dibahas dalam sub bab 4.1 berikut ini.

4.1 Proses *Scrapping*

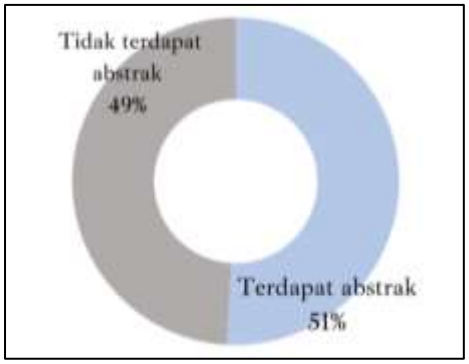
Proses *scrapping* merupakan proses pengambilan data dari suatu sumber data. Sumber data yang digunakan dalam penelitian ini adalah web *science direct*, dimana pada proses *scrapping* ini dilakukan menggunakan *software* spider dengan bahasa python. *Scrapping* yang dilakukan pada penelitian ini mengikuti cara srapping yang dilakukan oleh Loris Clo (2016). Dimana *scrapping* yang telah dilakukan oleh Loris ini dilakukan pada pengambilan data berita terkait bea cukai di negara-negara sekitar Italy. Algoritma dari *scrapping* yang dilakukan pada web berita ini, memiliki tujuan yang sama dengan *scrapping* yang digunakan pada web *science direct*. Sehingga dapat dilakukan modifikasi pada *syntax* yang telah ada, untuk mengaplikasikan pada web *science direct*. Rincian *syntax* untuk *scrapping* telah dituliskan sebagaimana pada Lampiran 1.

Scrapping dilakukan untuk masing-masing jurnal terpilih, yaitu pada 35 jurnal. Sehingga hasil yang diperoleh adalah data dari setiap jurnal, yang mana diambil dalam *file* tipe csv. Terdapat 35 *file* csv yang dipeoleh dari srapping, dimana pada setia *file* csv mengandung baris dan kolom. Baris disini berisikan data setiap publikasi, dan kolom berisi abstrak, judul, dan url. Dalam analisis selanjutnya, data abstrak yang akan digunakan namun dilakukan *scrapping* juga untuk judul dan url ini bertujuan untuk memastikan data yang diambil telah tepat atau belum. Dari 35 *file* csv ini, kemudian dilakukan penggabungan. Penggabungan 35 *file* dilakukan menggunakan *software* anaconda dengan bahasa

python. Dari hasil penggabungan ini diperoleh satu *file* csv yang dapat dianalisis lebih lanjut. Namun sebelum dilakukan analisis lebih lanjut, akan dilakukan eksprolasi data terlebih dahulu guna mengatahui karakteristik dari data publikasi terkait COVID-19. Untuk lebih jelasnya akan dilakukan penjabaran pada sub bab 4.2 terkait karakteristik data publikasi terkait COVID-19.

4.2 Karakteristik Data Publikasi Terkait COVID-19

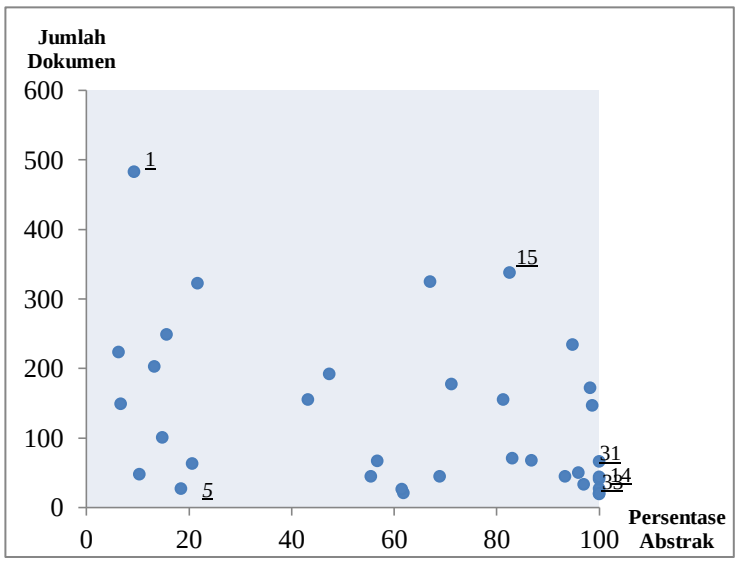
Data yang digunakan merupakan data hasil publikasi terkait COVID-19 dalam basis data *Science direct*. Pengambilan data dilakukan dengan cara *scraping* pada web resmi *Science direct*, yang mana diperoleh 4451 publikasi. Akan tetapi dari seluruh publikasi yang diperoleh, tidak semua mengandung abstrak didalamnya. Berikut gambaran terkait abstrak dalam data publikasi terkait COVID-19.



Gambar 4.1 Persentase Keberadaan Abstrak pada Data Hasil *Scrapping*

Berdasarkan pada Gambar 4.1 dapat diketahui bahwa dari seluruh data publikasi yang diperoleh, hanya setengahnya yang mengandung abstrak. Sehingga dapat diketahui jumlah data publikasi yang mengandung abstrak adalah 2264 data dari keseluruhan data yaitu 4451. Dari hasil data publikasi yang mengandung abstrak tersebut, terdapat sebaran yang berbeda antar jurnal pembentuknya. Sebagaimana diketahui sebelumnya bahwa data yang digunakan merupkaan data gabungan dari 35 jurnal terpilih, yang mana setiap jurnal memberikan data dengan

jumlah yang berbeda-beda. Sebaran dari 35 jurnal ini telah digambarkan dengan diagram batang sebagaimana pada Gambar 4.2 berikut.



Inisial Jurnal	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Jumlah Abstrak	45	39	10	15	5	13	25	222	70	14	32	169	126	19	279	59	38
Total Dokumen	483	249	149	101	27	21	45	234	322	223	33	172	155	19	338	71	67
Persentase (%)	9	16	7	15	19	62	56	95	22	6	97	98	81	100	83	83	57

Inisial Jurnal	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
Jumlah Abstrak	13	59	16	27	91	20	27	126	67	42	218	31	145	66	5	41	44	48
Total Dokumen	63	68	26	27	192	20	203	177	155	45	325	45	147	66	48	41	44	50
Persentase (%)	21	87	62	100	47	100	13	71	43	93	67	69	99	100	10	100	100	96

Gambar 4.2 Sebaran Abstrak Tiap Jurnal

Gambar 4.2 memperlihatkan sebaran jumlah dokumen terhadap persentase kandungan abstrak didalamnya. Tabel didalamnya bertujuan untuk memperlihatkan jumlah total dokumen dan jumlah dokumen yang mengandung abstrak pada masing-masing jurnal, dimana nama jurnal dilakukan inisialisasi dengan keterangan sebagaimana pada Tabel 3.1. Dari Gambar 4.2 maka dapat dilihat bahwa apabila plot berada di kanan atas adalah plot yang memiliki total dokumen banyak dan persentase abstrak didalamnya juga tinggi, jurnal yang seperti ini adalah jurnal dengan inisial No.15 yaitu *International Journal of Infectious Diseases*. Jurnal tersebut mengandung 279 abstrak dari total dokumen adalah 338, hal ini berarti 82% dari seluruh dokumen mengandung abstrak.

Selanjutnya untuk titik yang berada di kiri atas, berarti jumlah dokumen yang didapatkan banyak namun persentase mengandung abstraknya hanya sedikit. Jurnal yang seperti itu adalah jurnal dengan inisial No.1 yaitu *The Lancet*. Scrapping dari jurnal ini memperoleh dokumen sebanyak 486, namun hanya 9,3% yang mengandung abstrak. Untuk jurnal yang memiliki plot di kanan bawah, hal ini berarti jurnal tersebut memiliki jumlah dokumen yang tidak banyak namun persentase kandungan abstraknya sangat tinggi. Seperti halnya pada jurnal dengan inisial No.14, 21, 23, 31, 33, dan 34. Jurnal tersebut adalah *Biomedical Journal, Informatics in Medicine Unlocked, Journal of infection and Chemotherapy, Data in Brief, Heliyon,* dan *Respiratory Medicine Case Reports.* Keenam jurnal tersebut memiliki persentase mengandung abstrak adalah 100%. Sehingga semua dokumen yang diambil pada web *science direct* mengandung abstrak. Selanjutnya untuk dokumen yang memiliki plot disebelah kiri bawah, hal ini berarti bahwa jurnal tersebut memiliki dokumen yang cukup kecil dan persentase mengandung abstraknya juga kecil. Salah satu jurnal yang mengandung sedikit abstrak dalam total dokumen yang sedikit pula adalah jurnal dnegan inisial No.05 yaitu *The Lancet Haematology*. Dimana jurnal tersebut terdiri dari 27 dokumen, namun hanya 5 dokumen yang mengandung abstrak.

Dari keseluruhan data abstrak inilah yang selanjutnya digunakan sebagai data pada penlitian ini. Akan tetapi, data abstrak yang diperoleh merupakan data berupa kalimat, yang mana belum bisa dilakukan analisis klaster. Hal ini dikarenakan analisis klaster dengan menggunakan data teks atau yang dikenal dengan *topic modeling* ini, tidak menggunakan kalimat sebagai variabelnya, melainkan variabel berupa kata yang dibutuhkan untuk analisis *topic modeling*. Sehingga perlu dilakukan tahapan untuk merubah data abstrak yang berupa kalimat ini, menjadi data kata sebagai variabel. Tahapan yang digunakan untuk menjadikan data abstrak pada publikasi terkait COVID-19 menjadi variabel kata, disebut dengan pre-*processing* data. Dimana pada pre-*processing* data ini dilakukan beberapa langkah hingga diperoleh variabel yang sesuai. Lebih lanjut mengenai pre-*processing* data, akan dijelaskan pada sub bab selanjutnya yaitu sub bab pre-*processing* data publikasi terkait COVID-19.

4.3 Pre-*processing* Data Publikasi Terkait COVID-19

Data abstrak dari hasil *scrapping* yang diperoleh merupakan data yang berupa kalimat atau bahkan paragraf. Dalam analisis *topic modeling*, data yang dibutuhkan adalah variabel berupa kata. Kata disini diperoleh dari kata-kata penyusun dokumen pada hasil *scrapping* data abstrak. Namun, tidak semua kata-kata penyusun dapat dijadikan variabel melainkan terdapat tahapan sehingga kata-kata tersebut merupkan kata yang unik dan memiliki arti. Adapun empat tahapan yang dilakukan pada pre-*processing* ini yang pertama yaitu menjadikan smeua huruf adalah non-kalpital atau disebut dengan *case folding*. Tahapan yang kedua adalah memecah kalimat menjadi kata per kata yang disebut dengan *tokenizing*. Tahap selanjutnya adalah *lemmatizer*, yaitu tahap dimana setiap kata akan dijadikan kata dasar. Dan tahap yang terakhir adalah *stopwords* atau menghilangkan kata-kata yang bukan merupakan kata unik. Data awal sebelum dilakukan ke-empat tahapan tersebut, dapat ditabelkan sebagaimana pada Tabel 4.1 sebagai berikut.

Tabel 4.1 Data Sebelum Pre-*processing*

No.	Abstrak Publikasi Terkait COVID-19
1	Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) causes COVID-19 and is spread person-to-person through close contact … needed to better inform the evidence for these interventions, but this systematic appreisal of currently best available evidence might inform interim guidance.
2	This is the first randomised controlled trial for assessment of the immunogenicity and safety of a candidate non-replicating adenovirus type-5 (Ad5)-vectored COVID-19 vaccine, aiming to determine an appropriate dose of the candidate vaccine for an efficacy study … viral particles is safe, and induced significant immune responses in the majority of recipients after a single immunisation.
⋮	⋮
2264	Coronavirus Disease 2019 (COVID-19) infection, caused by severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), is spreading globally and poses a major public health threat … it will be necessary to comprehensively evaluate imaging and other clinical findings as well as consider co-infection with other respiratory viruses.

Tabel 4.1 memberikan gambaran data yang diperoleh, dimana data berupa abstrak utuh yang diambil dari web *Science direct* dengan ketentuan-ketentuan yang telah disebutkan pada batasan masalah. Dari seluruh data tersebut, selanjutnya akan dilakukan tahapan pre-*processing* tahap pertama yaitu *case folding.* Sebagaimana yang telah dijelaskan, tahap case folding ini bertujuan untuk mengubah seluruh data menjadi non-kapital. Hal ini bertujuan supaya kata yang sama, dapat dijadikan sama persis tanpa ada perbedaan karakter. Misalnya kata 'Severe' pada dokumen pertama akan dirubah menjadi 'severe'. Sehingga jika adala kata severe pada dokumen lain, akan seragam dengan severe pada dokumen pertama. Selian itu, juga dilakukan penghilangan tanda baca, karena hasil akhir yang diinginkan untuk variabel tidak mengandung tanda baca. Hasil dari tahap *case folding* ini dapat ditabelkan sebagaimana pada Tabel 4.2 berikut.

Tabel 4.2 Hasil Data Setelah *Case Folding*

No.	Abstrak Publikasi Terkait COVID-19
1	severe acute respiratory syndrome coronavirus sarscov causes covid and is spread persontoperson through close contact ... needed to better inform the evidence for these interventions but this systematic appreisal of currently best available evidence might inform interim guidance
2	this is the first randomised controlled trial for assessment of the immunogenicity and safety of a candidate nonreplicating adenovirus type advectored covid vaccine aiming to determine an appropriate dose of the candidate vaccine for an efficacy study ... viral particles is safe and induced significant immune responses in the majority of recipients after a single immunisation
⋮	⋮
2264	coronavirus disease covid infection caused by severe acute respiratory syndrome coronavirus sarscov is spreading globally and poses a major public health threat ... it will be necessary to comprehensively evaluate imaging and other clinical findings as well as consider coinfection with other respiratory viruses

Berdasarkan pada Tabel 4.2 dapat dilihat bahwa terdapat perbedaan pada karakter teks dalam setiap abstrak. Seluruh karakter teks pada data publikasi terkait COVID-19 telah berubah menjadi tidak kapital, serta telah hilang tanda baca dan angka didalamnya. Langkah selanjutnya adalah *tokenizing*, dimana pada

tahap ini akan dilakukan pemecahan setiap kata dalam sebuah kalimat. Sehingga hasil yang diinginkan pada masing-masing data abstrak, akan terbagi menjadi data kata perkata sebagaimana pada Tabel 4.3 berikut.

Tabel 4.3 Hasil Data Setelah *Tokenizing*

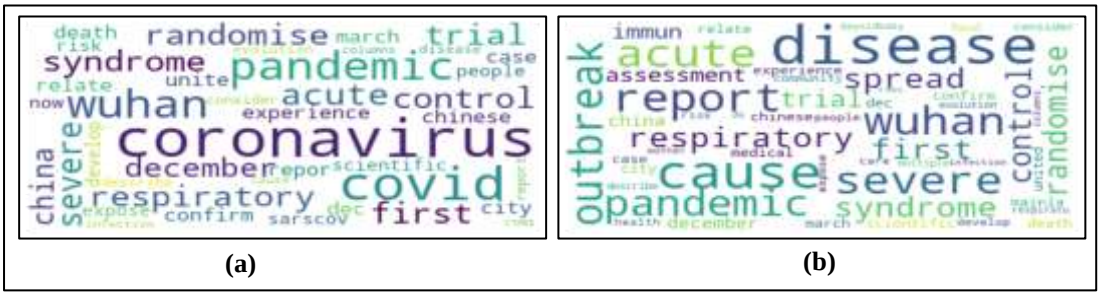
No.	Abstrak Publikasi Terkait COVID-19
1	‘severe’, ‘acute’, ‘respiratory’, ‘syndrome’, ‘coronavirus’, ‘sarscov’, ‘causes’, ‘covid’, ‘and’, ‘is’, ‘spread’, ‘persontoperson’, ‘through’, ‘close’, ‘contact’ ... ‘needed’ ‘to’, ‘better’, ‘inform’, ‘the’, ‘evidence’, ‘for’, ‘these’, ‘interventions’, ‘but’, ‘this’, ‘systematic’, ‘appreisal’, ‘of’, ‘currently’, ‘best’, ‘available’, ‘evidence’, ‘might’, ‘inform’, ‘interim’, ‘guidance’
2	‘this’, ‘is’, ‘the’, ‘first’, ‘randomised’, ‘controlled’, ‘trial’, ‘for’, ‘assessment’, ‘of’, ‘the’, ‘immunogenicity’, ‘and’, ‘safety’, ‘of’, ‘a’, ‘candidate’, ‘nonreplicating’, ‘adenovirus’, ‘type’, ‘advectored’, ‘covid’, ‘vaccine’, ‘aiming’, ‘to’, ‘determine’, ‘an’, ‘appropriate’, ‘dose’, ‘of’, ‘the’, ‘candidate’, ‘vaccine’, ‘for’, ‘an’, ‘efficacy’, ‘study’, ... ‘viral’, ‘particles’, ‘is’, ‘safe’, ‘and’, ‘induced’, ‘significant’, ‘immune’, ‘responses’, ‘in’, ‘the’, ‘majority’, ‘of’, ‘recipients’, ‘after’, ‘a’, ‘single’, ‘immunisation’
⋮	⋮
2264	‘coronavirus’, ‘disease’, ‘covid’, ‘infection’, ‘caused’, ‘by’, ‘severe’, ‘acute’, ‘respiratory’, ‘syndrome’, ‘coronavirus’, ‘sarscov’, ‘is’, ‘spreading’, ‘globally’, ‘and’, ‘poses’, ‘a’, ‘major’, ‘public’, ‘health’, ‘threat’, ... ‘it’, ‘will’, ‘be’, ‘necessary’, ‘to’, ‘comprehensively’, ‘evaluate’, ‘imaging’, ‘and’, ‘other’, ‘clinical’, ‘findings’, ‘as’, ‘well’, ‘as’, ‘consider’, ‘coinfection’, ‘with’, ‘other’, ‘respiratory’, ‘viruses’

Tabel 4.3 dapat dijadikan gambaran bahwa hasil yang diperoleh pada tahap *tokenizing* adalah penggalan-penggalan setiap kata. Sehingga dari penggalan tiap kata ini, selanjutnya dapat digunakan sebagai bahan untuk dilakukan *Lemmatizer*. *Lemmatizer* merupakan tahap yang dapat mengubah kata-kata tersebut menjadi kata dasar. Selain itu, adapun keunggulan *Lemmatizer* yaitu dapat menggantikan kata yang memeiliki keiripan dengan kata lain yang sering muncul. Misalkan, dalam suatu data abstrak terdapat kata “*better*”, sedangkan pada dokumen yang lain, lebih sering ditulisakan dengan “*well*” maka pada hasil dari tahan *Lemmatizer* ini akan merubah kata “*better*” menjadi kata “*well*”. Hal ini sangat menguntungkan bagianalisis, karena dengan adanya penggantian ini akan membuat kata yang memiliki arti sama dapat direpresentasikan pada kata yang sama, sehingga membuat frekuensi dati kata tersebut lebih tinggi. Hasil dari tahap *lemmatizer* dapat ditabelkan sebagaimana pada Tabel 4.4 berikut.

Tabel 4.4 Hasil Data Setelah *Lemmatizer*

No.	Abstrak Publikasi Terkait COVID-19
1	‘severe’, ‘acute’, ‘respiratory’, ‘syndrome’, ‘coronavirus’, ‘sarscov’, ‘causes’, ‘covid’, ‘and’, ‘is’, ‘spread’, ‘persontoperson’, ‘through’, ‘close’, ‘contact’ ... ‘need’ ‘to’, ‘well’, ‘inform’, ‘the’, ‘evidence’, ‘for’, ‘these’, ‘intervention’, ‘but’, ‘this’, ‘systematic’, ‘appreisal’, ‘of’, ‘currently’, ‘best’, ‘available’, ‘evidence’, ‘might’, ‘inform’, ‘interim’, ‘guidance’
2	‘this’, ‘is’, ‘the’, ‘first’, ‘randomise’, ‘control’, ‘trial’, ‘for’, ‘assessment’, ‘of’, ‘the’, ‘immunogenicity’, ‘and’, ‘safety’, ‘of’, ‘a’, ‘candidate’, ‘nonreplicating’, ‘adenovirus’, ‘type’, ‘advectored’, ‘covid’, ‘vaccine’, ‘aim’, ‘to’, ‘determine’, ‘an’, ‘appropriate’, ‘dose’, ‘of’, ‘the’, ‘candidate’, ‘vaccine’, ‘for’, ‘an’, ‘efficacy’, ‘study’, ... ‘viral’, ‘particle’, ‘is’, ‘safe’, ‘and’, ‘induce’, ‘significant’, ‘immune’, ‘response’, ‘in’, ‘the’, ‘majority’, ‘of’, ‘recipient’, ‘after’, ‘a’, ‘single’, ‘immune’
⋮	⋮
2264	‘coronavirus’, ‘disease’, ‘covid’, ‘infection’, ‘cause’, ‘by’, ‘severe’, ‘acute’, ‘respiratory’, ‘syndrome’, ‘coronavirus’, ‘sarscov’, ‘is’, ‘spread’, ‘globally’, ‘and’, ‘poses’, ‘a’, ‘major’, ‘public’, ‘health’, ‘threat’, ... ‘it’, ‘will’, ‘be’, ‘necessary’, ‘to’, ‘comprehensively’, ‘evaluate’, ‘image’, ‘and’, ‘other’, ‘clinical’, ‘finding’, ‘as’, ‘well’, ‘as’, ‘consider’, ‘coinfection’, ‘with’, ‘other’, ‘respiratory’, ‘virus’

Tabel 4.4 merupakan hasil dari tahap *Lemmatizer,* dimana pada kata yang bergaris bawah merupakan kata yang dilakukan perubahan. Sehingga semua kata menjadi kata dasar, serta ada pergantian kata yang memiliki kemiripan arti dijadikan satu menjadi kata yang sama. Hasil dari tahap Lemmatizer ini, diperoleh kata-kata yang sudah siap menjadi variabel pada analisis *topic modeling*. Namun, masih terdapat beberapa kata yang sebaiknya dihilangkan. Sehingga perlu dilakukan *stopwords* untuk mengeliminasi kata-kata yang dianggap tidak unik. Hal ini dapat digambarkan dengan *wordcloud* sebagaimana pada Gambar 4.3 berikut.



(a) **(b)**

Gambar 4.3 Visualisasi *wordcloud* sebelum dan sesudah *stopwords*

Gambar 4.3 (a) merupakan *wordcloud* untuk kata yang sudah menjadi variabel, dimana belum dilakukan eliminasi menggunakan *stopwords*, sedangkan untuk Gmabar 4.3 (b) adalah *wordcloud* dari variabel yang sudah dilakukan *stopwords*. Berdasarkan visualisasi *wordcloud* dapat dilihat bahwa kata yang paling sering muncul pada variabel-variabel sebelum dilakukan *stopwords* adalah 'coronavirus', 'covid', 'sarscov', dan lain sebagainya. Dimana kata-kata tersebut dianggap tidak memeiliki arti yang berpengaruh, karena dalam analisis ini memang semua data terkait 'coronavirus', 'covid', dan 'sarscov'. Sehingga kata-kata tersebut dieliminasi dengan hasil sebagaimana pada Gambar 4.3 (b). Gambar ini memperlihatkan bahwa kata 'coronavirus', 'covid', dan 'sarscov' sudah tidak muncul, dan tergantikan dengan kata-kata lainnya yang memiliki frekuensi tertinggi. Untuk lebih lengkapnya terdapat beberapa kata yang dilakukan eliminasi, dimana daftar kata tersebut disebut dengan list *stopwords*. Adapun beberapa list *stopwords* yang digunakan diantaranya adalah kata-kata tidak unik seperti 'i', 'me', 'my', 'myself', dan lain sebagainya. Serta kata-kata yang berkaitan dengan topik publikasi terkait COVID-19, diantaranya adalah 'covid-19', '2019-ncov', 'sars-ncov-2', 'sars-cov-2', 'sarscov', 'covid19', 'covid', 'cov', 'sarcov2', 'sarscov2', 'coronavirus', dan 'coronavirus2'. Hasil dari data abstrak yang telah melewati tahapan *stopwordss* atau data yang sudah siap dilakukan analisis selanjutnya, disajikan sebagaimana pada Tabel 4.5 berikut.

Tabel 4.5 Hasil Data Setelah *Stopwords*

No.	Abstrak Publikasi Terkait COVID-19
1	'severe', 'acute', 'respiratory', 'syndrome', 'causes', 'spread', 'persontoperson', 'close', 'contact' ... 'need' 'to', 'well', 'evidence', 'intervention', 'systematic', 'appreisal', 'currently', 'best', 'available', 'evidence', 'might', 'inform', 'interim', 'guidance'
2	'first', 'randomise', 'control', 'trial', 'assessment', 'immunogenicity', 'safety', 'candidate', 'nonreplicating', 'adenovirus', 'type', 'advectored', 'vaccine', 'aim', 'determine', 'appropriate', 'dose', 'candidate', 'vaccine', 'efficacy', 'study', ... 'viral', 'particle', 'safe', 'induce', 'significant', 'immune', 'response', 'majority', 'recipient', 'after', 'a', 'single' immune'
⋮	⋮

Lanjutan Tabel 4.5 Hasil Data Setelah *Stopwords*

2264	‘disease’, ‘infection’, ‘cause’, ‘severe’, ‘acute’, ‘respiratory’, ‘syndrome’, ‘ ‘spread’, ‘globally’, ‘pose’, ‘major’, ‘public’, ‘health’, ‘threat’, ... ‘necessary’, ‘comprehensively’, ‘evaluate’, ‘image’, ‘clinical’, ‘finding’, ‘well’, ‘consider’, ‘coinfection’, ‘respiratory’, ‘virus’

Berdasarkan pada Tabel 4.5, maka dapat diketahui hasil dari data abstrak terkait COVID-19 yang telah bersih dari kata-kata dianggap tidak unik. Dari data ini dapat dilanjutkan dengan perhitungan frekuensi setiap kata pada masing-masing dokumen. Dari frekuensi inilah yag nantinya dapat digunakan sebagai data dalam perhitungan peluang di masing-masing metode *topic modeling* yang digunakan. Hasil dari perhitungan frekuensi kata ini mengikuti format *Bag of Word* (BoW) yang dapat digambarkan sebagaimana pada Tabel 4.6 berikut.

Tabel 4.6 Bag of Word (BOW) dari Hasil Pre-*processing*

Dokumen ke-	**Dokumen**	…	‘acute’	‘best’	…	‘respiratory’	‘virus’	…
1	‘severe’, ‘acute’, ‘respiratory’, ‘syndrome’, ‘causes’, ‘spread’, ‘persontoperson’, ‘close’, ‘contact’ ... ‘need’ ‘to’, ‘well’, ‘evidence’, ‘intervention’, ‘systematic’, ‘appreisal’, ‘currently’, ‘best’, ‘available’, ‘evidence’, ‘might’, ‘inform’, ‘interim’, ‘guidance’	…	1	1	…	1	0	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2264	‘disease’, ‘infection’, ‘cause’, ‘severe’, ‘acute’, ‘respiratory’, ‘syndrome’, ‘ ‘spread’, ‘globally’, ‘pose’, ‘major’, ‘public’, ‘health’, ‘threat’, ... ‘necessary’, ‘comprehensively’, ‘evaluate’, ‘image’, ‘clinical’, ‘finding’, ‘well’, ‘consider’, ‘coinfection’, ‘respiratory’, ‘virus’	…	1	0	…	2	1	…

Berdasarkan Tabel 4.6 dapat diketahui pada kosa kata ‘respiratory’ terhitung 1 kata dalam dokumen abstrak ke 1, sedangkan pada dokumen abstrak ke 2264 terhitung 2 kali muncul kata ‘respiratory’. Sehingga dalam tabel Bow diatas dapat diektahui keseluruhan data kosa kata atau data variabel yang akan digunakan dalam analisis *topic modeling*. Data ini telah dilakukan perhitungan frekuensi setiap dokumen, sehingga pada analisis *topic modeling* data inilah yang akan menjadi dasar perhitungan untuk masing-masing metode.

4.3 *Topic modeling*

Data yang telah selesai dilakukan pre-*processing*, merupakan data yang sudah siap untuk dianalisis menggunakan metode labih lanjut. Dari data ini, telah diperoleh variabel-variabel yang dapat dilakukan analisis *topic modeling*. Analisis *topic modeling* merupakan analisis yang tergolong dalam metode *unsupervice*, dimana data yang digunakan adalah data teks. Tata cara alokasi topik pada analisis ini sejalan dengan prinsip klaster, yaitu mengelompokkan dokumen berdasarkan pada kemiripannya. Sehingga hasil yang akan diperoleh adalah terdapat kemiripan pada dokumen dalam topik yang sama, dan akan heterogen pada antar topiknya.

Pada penelitian ini, terdapat tiga metode dalam analisis *topic modeling* yang akan dibandingkan. Sehingga akan diperoleh metode terbaik yang sesuai dengan kasus pada data abstrak publikasi terkait COVID-19. Ketiga metode yang akan dibandingkan adalah *Hierarchical Dirichlet Process* (HDP), *Latent Dirichlet Allocation* (LDA), dan LDA2vec. Perbandingan ketiga metode tersebut didasarkan pada skor *coherence* yang diperoleh. Perhitungan skor *coherence* dilakukan dengan melakukan percobaan-percobaan jumlah topik yang terbentuk. Jumlah topik yang umumnya dicobakan yaitu antara dua topik samapai dengan 20 topik. Maka pada masing-masing metode akan dicobakan jumlah topik tersebut, sehingga dapat diperoleh jumlah topik yang optimal. Ketiga metode yang digunakan, terdapat kesamaan distribusi Dirichlet yang dikandung. Sehingga terlebih dahulu akan dilakukan estimasi parameter pada distribusi Dirichlet. Fungsi densitas dari distribusi Dirichlet adalah sebagai berikut.

$$Dir\left(\mathbf{p}|\alpha_1,\ldots,\alpha_m\right)=\frac{\Gamma\left(\sum_{k=1}^{m}\alpha_k\right)}{\prod_{k=1}^{m}\Gamma\left(\alpha_k\right)}\prod_k p_k^{\alpha_k-1}$$

dimana $\theta=\{\mathbf{p}_1,\ldots,\mathbf{p}_N\}$, estimasi parameter dari distribusi Dirichlet akan dilakukan menggunakan fungsi maksimum log-likelihood sebagaimana pada persamaan berikut.

$$\begin{aligned}F(\alpha)&=\log p\left(\theta|\alpha\right)\\&=\log\prod_i p\left(\mathrm{p}_i|\alpha\right)\\&=\log\prod_i\frac{\Gamma\left(\sum_k\alpha_k\right)}{\prod_k\Gamma\left(\alpha_k\right)}\prod_k p_k^{\alpha_k-1}\\&=N\left(\log\Gamma\left(\sum_k\alpha_k\right)-\sum_k\log\Gamma\left(\alpha_k\right)+\sum_k\left(\alpha_k-1\right)\log\bar{p}_k\right)\end{aligned}$$

Statistik observasinya adalah $\log\bar{p}_k$, dimana $\log\bar{p}_k=\frac{1}{N}\sum_i\log p_{ik}$. Berdasarkan fungsi maksimum log-likelihood yang telah dijelaskan, diperoleh hasil fungsi F yang tidak *close-from* dalam memaksimumkan estimasi dari α. Sehingga dilakukan iterasi *fixed-point* dengan batasan sebaga berikut

$$\Gamma(x)\geq\Gamma(\hat{x})\exp\left((x-\hat{x})\Psi(\hat{x})\right)$$

Sehingga daripersamaan diatas dapat dibuat batas bawah untuk log-likelihood $F(\alpha)$ adalah sebagai berikut.

$$F(\alpha)\geq N\left(\left(\sum_k\alpha_k\right)\Psi\left(\sum_k\alpha_k^{old}\right)-\sum_k\log\Gamma\left(\alpha_k\right)+\sum_k\left(\alpha_k-1\right)\log\bar{p}_k+C\right)$$

$$\frac{1}{N}\log p\left(\theta|\alpha\right)\geq\left(\sum_k\alpha_k\right)\Psi\left(\sum_k\alpha_k^{old}\right)-\sum_k\log\Gamma\left(\alpha_k\right)+\sum_k\left(\alpha_k-1\right)\log\bar{p}_k+C$$

dimana C adalah konstanta sehubungan dengan α. Dari persamaan diatas, dilakukan optimasi dengan menjadikan gradient sama dengan 0, sehingga diperoleh nilai estimasi α yaitu

$$\Psi(\hat{\alpha}) = \Psi\left(\sum_k \alpha_k^{old}\right) + \log \bar{p}_k$$
$$\hat{\alpha} = \Psi^{-1}\left(\Psi\left(\sum_k \alpha_k^{old}\right) + \log \bar{p}_k\right)$$

Berdasarkan persamaan diatas, maka telah diperoleh nilai estimasi parameter α dari distribusi Dirichlet.

4.3.1 *Hierarchical Dirichlet Process* (HDP)

Data yang telah terlebih dahulu dilakukan tahapan pre-*processing*, selanjutnya data tersebut dapat dilanjutkan pada analisis *topic modeling*. Analisis topik modeling yang pertama yaitu *Hierarchical Dirichlet Process* (HDP), metode ini menggunakan penentuan klaster topik dengan analogi *Chinese Restaurant Franchise Processes* (CRFP). Perhitungan peluang dibedakan menjadi dua bagian yaitu perhitungan peluang alokasi topik berdasarkan pada kata dalam dokumen, dan pehitungan peluang alokasi topik berdasarkan kosa kata. Untuk menghitung peluang alokasi topik kata dalam dokumen ada dua kemungkinan, yaitu ketika belum ada kata yang memilih topik itu $\frac{\alpha}{i+1+\alpha}$, dan sudah ada kata yang memilih topik itu sebelumnya $\frac{n_{dk}}{i+1+\alpha}$. Dimana n_{dt} adalah jumlah kata yang sudah memilih topik ke-k terdahulu. Jadi misal digunakan data sebagai berikut.

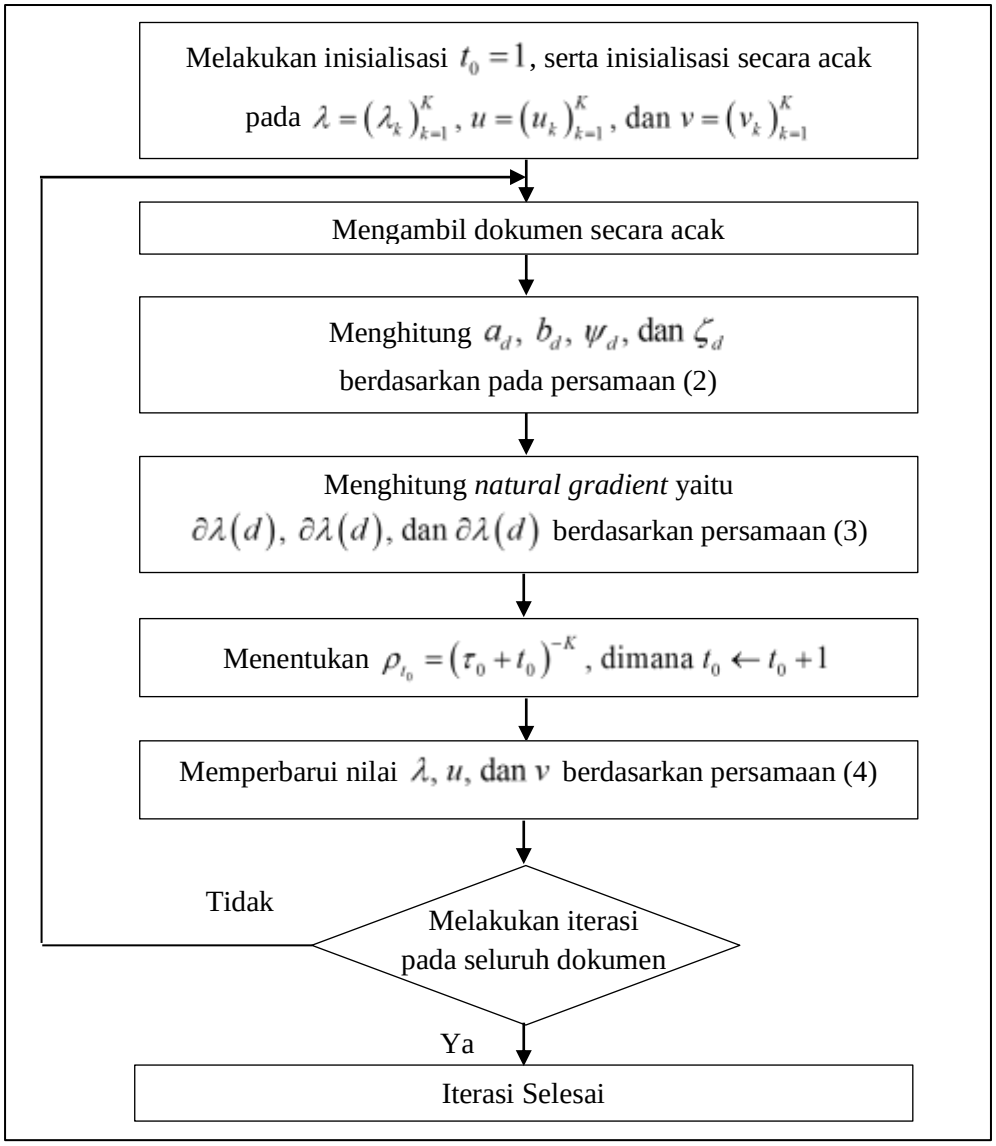
Tabel 4.7 Data Ilustrasi Perhitungan Manual

Dokumen ke-	Dokumen Abstrak
1	‘severe’, ‘acute’, ‘respiratory’, ‘syndrome’, ‘causes’, ‘spread’, ‘persontoperson’, ‘close’, ‘contact’ ‘need’ ‘well’, ‘evidence’, ‘intervention’, ‘systematic’, ‘appreisal’, ‘currently’, ‘best’, ‘available’, ‘evidence’, ‘might’, ‘inform’, ‘interim’, ‘guidance’
2	‘first’, ‘randomise’, ‘control’, ‘trial’, ‘assessment’, ‘immunogenicity’, ‘safe’, ‘candidate’, ‘nonreplicating’, ‘adenovirus’, ‘type’, ‘advectored’, ‘vaccine’, ‘aim’, ‘determine’, ‘appropriate’, ‘dose’, ‘candidate’, ‘vaccine’, ‘efficacy’, ‘study’, ‘viral’, ‘particle’, ‘safe’, ‘induce’, ‘significant’, ‘immune’, ‘response’, ‘major’, ‘recipient’, ‘after’, ‘single’ ‘immune’
3	‘disease’, ‘infection’, ‘cause’, ‘severe’, ‘acute’, ‘respiratory’, ‘syndrome’, ‘spread’, ‘globally’, ‘pose’, ‘major’, ‘public’, ‘health’, ‘threat’, ‘necessary’, ‘comprehensively’, ‘evaluate’, ‘image’, ‘clinical’, ‘finding’, ‘well’, ‘consider’, ‘coinfection’, ‘respiratory’, ‘virus’

Perhitungan diawali dengan pemberian label topik pada masing-masing kata dalam dokumen sebagaimana pada Tabel 4.7, dimana warna merah menandakan kata tersebut masuk dalam topik pertama dan biru pada topik kedua. Perhitungan selanjutnya akan dibuat penentuan topik pada setiap kata dalam dokumen, dengan menggunakan perhitungan peluang dari hasil random pada Tabel 4.6. Sebagai contoh untuk kata ‘severe’ dihitung peluang masuk pada topik pertama adalah $\frac{n_{dk}}{i+1+\alpha} = \frac{10}{23+1+1} = 0.4$, sedangkan peluang masuk pada topik kedua adalah $\frac{n_{dk}}{i+1+\alpha} = \frac{13}{23+1+1} = 0.52$, sehingga kata ‘severe’ akan diberi label masuk pada topik kedua. Selanjutnya untuk kata ‘acute’ akan dilakukan perhitungan yang sama yaitu $\frac{n_{dk}}{i+1+\alpha} = \frac{10}{23+1+1} = 0.4$ dan peluang untuk masuk ke topik kedua juga sama $\frac{n_{dk}}{i+1+\alpha} = \frac{13}{23+1+1} = 0.52$ hal ini dikarenakan kata sebelumnya ‘severe’ berada pada topik yang sama. Sehingga kata ‘acute’ juga masuk pada topik kedua. Selanjutnya untuk kata ‘respiratory’ akan dilakukan perhitungan yang sama, dengan hasil yang sama yaitu kata ‘respiratory’ masuk pada topik kedua. Namun hal ini membuat perubahan pada peluang, karena pada random sebelumnya kata ‘respiratory’ masuk pada topik pertama, sekarang kata ‘respiratory’ masuk pada topik kedua. Sehingga untuk kata ‘syndrom’ akan menghasilkan peluang untuk masuk pada topik pertama adalah $\frac{n_{dk}}{i+1+\alpha} = \frac{9}{23+1+1} = 0.36$, sedangkan untuk masuk dalam topik kedua adalah $\frac{n_{dk}}{i+1+\alpha} = \frac{14}{23+1+1} = 0.56$. Sehingga kata ‘respiratory’ juga masuk dalam topik kedua. Hal ini dilakukan secara sterus menerus hingga semua kata dalam semua dokumen selesai silakukan perhitungan. Adapun kekurangan dari CRFP ini yaitu prinsip yang kaya semakin kaya, yang miskin semakin miskin. Sehingga alokasi topik tidak sepnuhnya rata. Namun disinilah metode HDP dibuat dengan perluasan metode CRFP, yaitu pada metode HDP terdapat parameter-parmeter yang akan selalu diperbarui setiap perhitungan yang dilkukan. Parameter disini

adalah α. Sehingga hasil yang akan diperoleh tidak akan membuat semua kata dalam dokumen terkelompokkan menjadi satu topik saja.

Terdapat parameter pendukung lain yang digunakan dalam metode HDP, parameter ini adalah λ, u, dan v. Dimana iterasi pada parameter tersebut dapat digambarkan pada Gambar 4.4 berikut.



Gambar 4.4 Ilustrasi Algoritma *Dirichlet Process* (DP)

Sebagaimana pada ilustrasi algoritma di Gambar 4.4, iterasi dilakukan dengan isisiasi parameter. Kemudian mengambil suatu dokumen acak dan dari persamaan (4) dilanjutkan dengan menghitung *natural gradient* sehingga kemudian dapat dilakukan perbaruan nilai $\partial\lambda(d), \partial\lambda(d), \text{dan } \partial\lambda(d)$. Dari algoritma DP tersebut akan berjalan beriringan dengan perbaruan parameter dokumen (*second level*) dalam suatu topik. Hal ini dikarenakan dengan adanya perbaruan nilai a_d, b_d, ψ_d, dan ζ_d maka akan memperbarui juga nilai parameter β pada variabel π_{dk}. Sehingga untuk parameter setiap kata dalam topik juga akan mengalami perbaruan, yaitu z_{dn} dalam pemilihan topik ψ_{dk}. Hingga parameter untuk setiap dokumen pada suatu topik dapat dituliskan dengan ϕ_{dk} dimana sesuai dengan perhitungan pada persamaan (3). Secara otomatis dengan adanya perbaruan nilai parameter pada ϕ_{dk} maka akan terbarui juga untuk parameter topik ϕ_k yang mana perubahan ini berada pada *first level* dalam ilustrasi Gambar 2.1.

Algoritma yang telah dijelaskan sebelumnya akan berhenti ketika semua dokumen talah dilakukan perhitungan dan dilakukan perbaruan untuk masing-masing parameternya. Hasil dari algoritma tersebut yaitu pemetaan dokumen pada topik-topik, dimana dalam dokumen terdiri-dari kata-kata. Langkah selanjutnya adalah membuat model HDP dimana akan ditunjukan peluang kemunculan kosa kata dalam masing-masing topik. Sehingga dari model tersebut, nantinya dapat digunakan sebagai acuan jika ada dokumen baru dan ingin mengetahui dokumen tersebut akan masuk topik mana. Penentuan masuk topik tertentu ini didasarkan pada peluang kemunculan setiap kata dalam topik.

Sesuai dengan langkah-langkah yang telah dijelaskan, terlebih dahulu akan dilakukan perhitungan skor *coherence* utnuk mengetahui jumlah topik yang optium. Jumlah topik yang optimum akan diindikasi dengan nilai skor *coherence* yang tertinggi. Sebagai langkah pertama dalam pembentukan model HDP adalah penentuan jumlah topik, telah dilakukan perhitungan skor *coherence* untuk mengetahui jumlah topik yang optimal dengan metode HDP, dimana hasil dari skor dapat ditabelkan sebagaimana pada Tabel 4.8 berikut.

Tabel 4.8 Skor *Coherence* Metode *Hierarchical Dirichlet Process* (HDP)

Jumlah Topik	Skor *Coherence*
2	0.382
3	0.370
4	0.366
5	0.368
6	0.368
7	0.368
8	0.368

Jumlah Topik	Skor *Coherence*
9	0.366
10	0.362
11	0.358
12	0.357
13	0.360
14	0.368

Jumlah Topik	Skor *Coherence*
15	0.365
16	0.362
17	0.378
18	0.377
19	0.379
20	0.378

Berdasarkan Tabel 4.8 maka dapat diketahui bahwa jumlah topik optimumnya adalah dua, dimana pada jumlah topik tersebut diperoleh skor *coherence* yang paling tinggi. Skor *coherence* tertinggi dengan metode HDP yang diperoleh yaitu 0.382, dimana dapat pula diperoleh model dari metode HDP dengan jumlah topik opimal tersebut. Penulisan model didasarkan pada, peluang untuk kata dalam setiap topik. Berdasarkan jumlah topik optimum, maka akan dibentuk model dengan jumlah topik adalah dua.

$$p(\boldsymbol{\varphi}_1) = 0.013 * patient \; + \; 0.007 * disease \; + \ldots + 0.000 * rightcensored$$
$$p(\boldsymbol{\varphi}_2) = 0.013 * patient \; + \; 0.008 * case \; + \ldots + 0.000 * irritant$$

Sebagaimana telah dijelaskan sebelumnya bahwa model pada persamaan diatas, dibentuk untuk mengetahui peluang kata pada setiap topik. Model diperoleh dari hasil pemetaan setiap dokumen dimasing-masing topik, dimana dalam dokumen tersebut mengandung kata-kata yang telah menjadi variabel (kosa kata). Berdasarkan pada model peluang yang telah diperoleh, dapat diketahui bahwa pada topik pertama, kata ‘patient’ merupakan kata dengan peluang terbesar. Dari jumlah kata keseluruhan yaitu 17.164 kata yang ada, kata ‘patient’ memiliki peluang 0.013. Nilai peluang ini diperoleh dari perkalian antara peluang kata ‘patient’ dalam data abstrak dengan peluang dokumen tersebut pada topik pertama. Perhitungan dilakukan berulang sampai dengan kata terakhir yaitu ‘rightcencored’ yang memiliki peluang 0.000. Meskipun peluang yang dapat

terbaca adalah 0.000, tetapi peluang tersebut pasti memiliki nilai, walaupun sangat kecil.

Selanjutnya untuk topik yang kedua, secara kebelutan memiliki peluang yang sama pada kata ‘patient’ yaitu 0.013. namun memiliki urutan yang pastinya berbeda, yaitu kata ‘case’ dan ‘disease’ pada urutan kedua dan ketia, dimana nilai peluangnya sama yaitu 0.008. Dalam hal ini, kata ‘case’ dan ‘disease’ tidak sepenuhnya memiliki peluang yang sama, malainkan ada sedikit perbedaan yang mana jika dibulatkan menjadi nilai yang sama. Serta kata ‘case’ pastinya memiliki nilai peluang yang sedikit lebih tinggi jika dibandingkan dengan kata ‘disease’. Model berlanjut sampai dengan peluang pada kata ke 17.164 yaitu kata ‘irritant’ yang memiliki peluang 0.000. Tujuan dibuatnya model dalam metode ini untuk pemetaan jika digunakan dokumen baru. Jika ada suatu dokumen baru yang ingin diketahui dokumen tersebut masuk dalam topik mana, maka hal tersebut dapat dihitung. Kata-kata dalam dokumen baru tersebut akan dihitung peluangnya dan dilakukan penjumlahan untuk seluruh kata didalamnya. Dilakukan perhitungan untuk masing-masing topik yang terbentuk, kemudian dipilih topik yang menghasilkan peluang terbesar. Jadi dokumen baru akan masuk pada topik dengan peluang terbesar itu.

4.3.2 *Latent Dirichlet Allocation* (LDA)

Latent Dirichlet Allocation (LDA) adalah salah satu metode klaster dalam *topic modeling* yang teruji efektif dalam pembuatan model pada topik-topik yang terbentuk, serta metode ini menjadi metode paling sering digunakan pada analisis trerkait *text mining* pada tahun 2000-2017. Metode LDA merupakan metode yang mengoptimalkan peluang pada persamaan (6), dimana dalam optimasinya dilakukan perulangan dengan metode *gibbs sampling* sebagaimana pada algorita Gambar 2.5. Data yang diperoleh dari pre-*processing* pada Tabel 4.6 kemudian digunakan sebagai data untuk perhitungan *gibbs sampling*.

Dalam melakukan pengelompokan topik, terdapat dua bentuk distribusi probabilitas yang harus dicari, yaitu disribusi probabilitas dokumen pada suatu dokumen dan distribusi probabilitas kata pada suatu topik. akan dilakukan ilustrasi dengan melakukan perhitungan pada tiga contoh dokumen. Kata pada

masing-masing dokumen tersebut akan dilakukan isiniasi topik. Dalam ilustrasi ini ditentukan jumlah topik adalah dua.

Tabel 4.9 Data ilustrasi Perhitungan Manual LDA

Dokumen ke-	Dokumen Abstrak
1	'severe', 'acute', 'respiratory', 'syndrome', 'causes', 'spread', 'persontoperson', 'close', 'contact' 'need' 'well', 'evidence', 'intervention', 'systematic', 'appreisal', 'currently', 'best', 'available', 'evidence', 'might', 'inform', 'interim', 'guidance'
2	'first', 'randomise', 'control', 'trial', 'assessment', 'immunogenicity', 'safe', 'candidate', 'nonreplicating', 'adenovirus', 'type', 'advectored', 'vaccine', 'aim', 'determine', 'appropriate', 'dose', 'candidate', 'vaccine', 'efficacy', 'study', 'viral', 'particle', 'safe', 'induce', 'significant', 'immune', 'response', 'major', 'recipient', 'after', 'single' 'immune'
3	'disease', 'infection', 'cause', 'severe', 'acute', 'respiratory', 'syndrome', 'spread', 'globally', 'pose', 'major', 'public', 'health', 'threat', 'necessary', 'comprehensively', 'evaluate', 'image', 'clinical', 'finding', 'well', 'consider', 'coinfection', 'respiratory', 'virus'

Dari data pada Tabel 4.9 dapat diketahui bahwa terdapat tiga dokumen yang masing-masing kata dalam setiap dokumen telah dilakukan inisiasi, dimana kata yang ditandai dengan warna merah adalah kaya yang masuk dalam topik pertama dan kata yang ditandai dengan warna biru adalah kata yang masuk pada topik kedua. Selanjutnya dari penentuan topik yang random tersebut dilakukan perhitungan peluang topik pada suatu dokumen $p(z|\theta)$, dimana rumus yang digunakan adalah $p(z|\theta) = \frac{n_{vk} + \alpha}{N_i - 1 + K\alpha}$, dimana n_{vk} adalah jumlah kata dalam dokumen ke-d yang masuk topik ke-k, N_d adalah jumlah seluruh kata didalam dokumen ke-d, dan α adalah parameter dalam menentukan distribusi topik dalam dokumen. Dalam ilustrasi pada Tabel 4.1 maka dapat dilakukan perhitungan untuk kata didokumen yang masuk topik pada masing-masing topik, adalah sebagai berikut.

$$p\left(z_{11}\middle|\theta\right)=\frac{n_{11}+\alpha}{N_1-1+K\alpha}=\frac{10+0.1}{23-1+\left(2\left(0.1\right)\right)}=0.455$$

$$p\left(z_{12}\middle|\theta\right)=\frac{n_{12}+\alpha}{N_1-1+K\alpha}=\frac{13+0.1}{23-1+\left(2\left(0.1\right)\right)}=0.590$$

$$p\left(z_{21}\middle|\theta\right)=\frac{n_{21}+\alpha}{N_2-1+K\alpha}=\frac{21+0.1}{33-1+\left(2\left(0.1\right)\right)}=0.655$$

$$p\left(z_{22}\middle|\theta\right)=\frac{n_{22}+\alpha}{N_2-1+K\alpha}=\frac{12+0.1}{33-1+\left(2\left(0.1\right)\right)}=0.376$$

$$p\left(z_{31}\middle|\theta\right)=\frac{n_{31}+\alpha}{N_3-1+K\alpha}=\frac{16+0.1}{25-1+\left(2\left(0.1\right)\right)}=0.665$$

$$p\left(z_{32}\middle|\theta\right)=\frac{n_{32}+\alpha}{N_3-1+K\alpha}=\frac{9+0.1}{25-1+\left(2\left(0.1\right)\right)}=0.376$$

Perhitungan diatas dilakukan pada setiap topik dan setiap dokumennya, dimana terdapat tiga dokumen dan 2 topik. Selanjutnya dilakukan perhitungan peluang setiap kosa kata pada suatu topik $p\left(w\middle|z,\phi\right)$. Persamaan yang digunakan adalah $p\left(w\middle|z,\phi\right)=\frac{w_{v.k}+\beta}{\sum_{v\in V}w_{v.k}+V\beta}$, dimana $m_{j.k}$ merupakan jumlah seluruh kosa kata-j pada topik ke-k, $\sum_{v\in V}w_{d.k}$ adalah jumlah seluruh kosa kata pada topik ke-k, dan V adalah jumlah seluruh kosa kata yang digunakan. Dalam ilustrasi pada Tabel 4.1 maka dapat dilakukan perhitungan untuk peluang setiap kosa kata dalam topik, dapat dituliskan sebagai berikut.

$$p\left(w_1\middle|z_1,\phi\right)=\frac{w_{1.1}+\beta}{\sum_{1\in V}w_{1.1}+V\beta}=\frac{2+0.1}{2+\left(60\left(0.1\right)\right)}=0.262$$

$$p\left(w_1\middle|z_2,\phi\right)=w\frac{v_{1.2}+\beta}{\sum_{1\in V}w_{1.2}+V\beta}=\frac{2+0.1}{0+\left(60\left(0.1\right)\right)}=0.350$$

$$\vdots$$

$$p\left(w_{60}\middle|z_1,\phi\right)=\frac{w_{60.2}+\beta}{\sum_{60\in V}w_{60.2}+V\beta}=\frac{1+0.1}{0+\left(60\left(0.1\right)\right)}=0.183$$

Perhitungan diatas dilakukan pada setiap kosa kata, dan setiap topiknya, dimana terdapat 60 kosa kata pada corpus. Selanjutnya dilakukan perhitungan

untuk peluang setiap kata pada dokumen ke-d dan topik ke-k $p(w,z,\theta)$ dimana dapat diperoleh dari $p(w,z,\theta) = p(\theta|\alpha)\,p(\varphi|\beta)$. Sehingga dapat dilakukan perhitungan sebagai berikut.

$$p(w_1, z_1, \theta_1) = p(\theta_{11}|\alpha)\,p(\varphi_{11}|\beta) = 0.455 \times 0.262 = 0.119$$
$$p(w_1, z_2, \theta_1) = p(\theta_{12}|\alpha)\,p(\varphi_{12}|\beta) = 0.590 \times 0.350 = 0.206$$
$$p(w_2, z_1, \theta_1) = p(\theta_{11}|\alpha)\,p(\varphi_{21}|\beta) = 0.455 \times 0.350 = 0.159$$
$$\vdots$$
$$p(w_{81}, z_2, \theta_3) = p(\theta_{32}|\alpha)\,p(\varphi_{60.2}|\beta) = 0.590 \times 0.183 = 0.108$$

Telah sampai pada perhitungan setiap kata, dari masing-maisng kata telah dihitung peluang masuk kedalam topik pertama atau topik kedua. Jumlah seluruh kata dalam seluruh dokumen adalah 81, sehingga akan dilakukan perhitungan peluang sebanyak 81 kali untuk masing-masing topik. Dari peluang inilah yang menjadi penentu suatu topik masuk ke topik yang mana. Misalnya untuk kata pertama dalam dokumen pertama telah dilakukan perhitungan untuk topik pertama (0.119) dan untuk topik kedua (0.206), berdasarkan peluang yang diperoleh maka kata pertama akan dilabeli untuk masuk pada topik kedua karena memiliki peluang yang lebih tinggi. Begitu juga untuk seluruh kata dalam dokumen yang seterusnya.

Perhitungan peluang dan pemberian label pada masing-masing kata yang telah dilakukan diatas merupakan iterasi pertama, dimana akan dilakukan iterasi ke dua sampai dengan ke 50 untuk mendapatkan pelabelan topik yang konstan. Selanjutnya dari hasil iterasi yang telah didapatkan, dapat dilakukan pembuatan model peluang untuk masing-masing kata dalam setiap topik. Supay model peluang yang diperoleh adalah maksimal, maka dilakukan penentuan jumlah topik yang optimal. Penentuan jumlah topik yang optimal dilakukan berdasarkan perhitungan skor *coherence,* dimana pada dilakukan perhitungan dengan mencobakan jumlah topik mulai dari dua sampai dengan 20 topik. Skor coherence merupakan nilai yang dapat digunakan sebagai ukuran kebaikan pendistribusian kata dalam topik. Pendistribusian kata pada topik akan baik jika skor *coherence*

yang diperoleh adalah tinggi, sehingga pendistribusian kata yang baik adalah yang memiliki skor *coherence* paling tinggi dibandingkan yang lainnya. Perhitungan skor *coherence* dilakukan dengan metode *Cv* dimana perhitungan dilakukan sebagaimana pada persamaan (9), sehingga seluruh hasil dari perhitungan *coherence* dapat ditabelkan sebagai berikut.

Tabel 4.9 Skor *Coherence* Metode *Latent Dirichlet Allocation* (LDA)

Jumlah Topik	**Skor *Coherence***
2	0.430
3	0.459
4	0.486
5	0.480
6	0.461
7	0.495
8	0.514

Jumlah Topik	**Skor *Coherence***
9	0.527
10	0.503
11	0.478
12	0.480
13	0.503
14	0.544

Jumlah Topik	**Skor *Coherence***
15	0.517
16	0.503
17	0.511
18	0.533
19	0.500
20	0.495

Berdasarkan pada Tabel 4.9 maka dapat diketahui bahwa jumlah topik yang mendapatkan skor *coherence* tertinggi adalah adalah 14 topik. Hal ini berarti bahwa jumlah itulah yang merupkaan jumlah topik yang optimum. Selanjutnya dari hasil jumlah topik optimum, maka dilanjutkan dengan perulangan menggunakan *gibbs sampling*. Hasil dari perulangan *gibbs sampling* ini selanjutnya digunakan dalam dasar perhitungan untuk mendapatkan pemetaan dokumen pada topik. Dokumen yang telah dipetakan dalam 14 topik ini, selanjutnya dapat ditampilkan model yang berisikan peluang setiap kosa kata dalam masing-masing topik. Sama seperti yang dilakukan pada metode HDP, pembuatan model ini bertujuan untuk memudahkan pengguna apabila ada data dokumen baru yang ingin diketahui akan masuk klaster/topik mana. Penulisan model didasarkan pada, peluang untuk kata dalam setiap topik. Berdasarkan jumlah topik optimum, maka akan dibentuk model peluang berjumlah 14 sebagai berikut.

$$p(\varphi_1) = 0.033 * test + 0.024 * sample + \ldots + 0.000 * dayoldbaby$$
$$p(\varphi_2) = 0.019 * country + 0.016 * data + \ldots + 0.000 * dayoldbaby$$
$$p(\varphi_3) = 0.019 * case + 0.012 * control + \ldots + 0.000 * lineage$$
$$p(\varphi_4) = 0.027 * disease + 0.023 * infection + \ldots + 0.000 * collocation$$
$$p(\varphi_5) = 0.044 * patient + 0.028 * case + \ldots + 0.000 * virachip$$
$$p(\varphi_6) = 0.020 * pandemic + 0.014 * health \ldots + 0.000 * dayoldbaby$$
$$p(\varphi_7) = 0.053 * patient + 0.020 * disease + \ldots + 0.000 * dayoldbaby$$
$$p(\varphi_8) = 0.016 * stress + 0.015 * mental + \ldots + 0.000 * dayoldbaby$$
$$p(\varphi_9) = 0.025 * model + 0.021 * data + \ldots + 0.000 * grl$$
$$p(\varphi_{10}) = 0.024 * cell + 0.017 * ace + \ldots + 0.000 * promiscuity$$
$$p(\varphi_{11}) = 0.020 * test + 0.016 * survey + \ldots + 0.000 * dayoldbaby$$
$$p(\varphi_{12}) = 0.024 * cut + 0.017 * surface + \ldots + 0.000 * dayoldbaby$$
$$p(\varphi_{13}) = 0.019 * treatment + 0.014 * drug + \ldots + 0.000 * dayoldbaby$$
$$p(\varphi_{14}) = 0.066 * patient + 0.017 * group + \ldots + 0.000 * promiscuity$$

Berdasarkan pada persamaan diatas, maka dapat diketahui peluang kata pada setiap topik, dimana dinotasikan dengan φ yang merupakan distribusi kata ke-i pada topik ke-k. Jumlah topiknya adalah 14 yang merupakan jumlah optimum berdasarkan perhitungan *coherence*. Dari ke-14 model yang telah disajikan, maka dapat diketahui peluang masing-masing kata didalamnya. Pada topik pertama, kata ‘test’ memiliki peluang 0.033 dan dijumlahkan dengan kata ‘sampel’ dengan peluang 0.024, sampai dengan kata ke 17.164 pada topik pertama yaitu ‘dayoldbaby’ yang memiliki peluang 0.000. Begitu juga untuk model pada topik ke dua sampai dengan topik ke 14. Sebagaimana yang telah disebutkan sebelumnya, bahwa tujuan dibuatnya model dalam metode ini untuk pemetaan jika digunakan dokumen baru. Jika ada suatu dokumen baru yang ingin diketahui dokumen tersebut masuk dalam topik mana, maka hal tersebut dapat dihitung. Langkah yang dilakukan yaitu dengan menghitung peluang setiap kata didokumen baru pada masing-masing topik. Selanjutnya dilakukan penjumlahan untuk seluruh kata didalam dokumen tersebut yang kemudian dipilih topik dengan

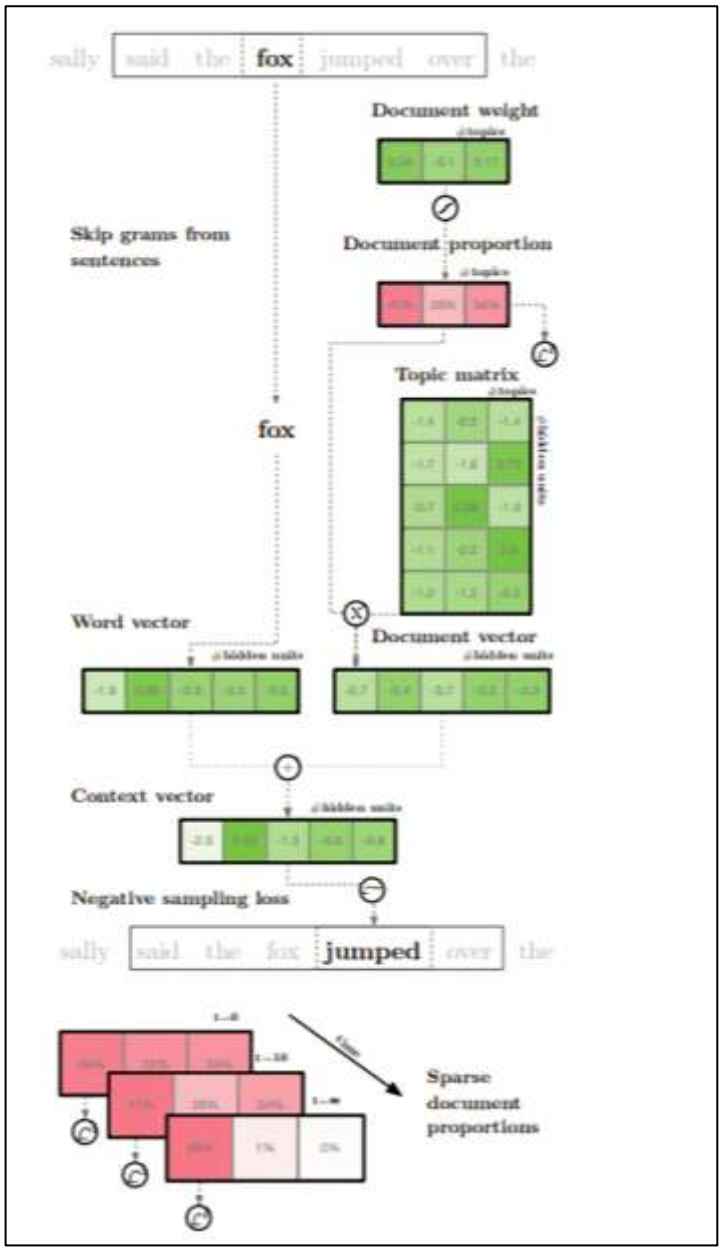
peluang terbesar. Sehinga dokumen baru akan masuk pada topik dengan peluang terbesar itu.

4.3.3 LDA2vec

LDA2vec adalah salah satu pengembangan dari *topic modeling* yang mempertimbangkan istilah dari word2vec. Dipertimbangannya word2vec ini bertujuan untuk menpertahankan keunggulan dari informasi lokal dalam topik sehingga membuat vector dokumen dan vector topik lebih mudah untuk dipahami. LDA2vec merupakan mixture model, yang mana penggabungan dari model pada vector kata dan vector dokumen. Jadi terdapat dua perhitungan mendasar yaitu perhitungan untuk mendapatkan vector kata oleh word2vec, dan perhitungan untuk vector dokumen oleh LDA. Dimana dalam penentuan vector dokumen juga dilakukan bangkitan vector topik, sehingga dalam vector dokumen yang akan digabungkan dengan vector kata, telah ada vector topik didalamnya.

Langkah pertama yang dilakukan adalah menghitung bobot dokumen untuk masing-masing topik. Pada tahap ini digunakan transformasi *softmax* untuk merubah sebuah dokumen menjadi vector bobot, dokumen yang digunakan merupakan dokumen hasil pre-*process* sebagaimana pada Tabel 4.5. Selanjutnya dari bobot topik pada dokumen ini, dilakukan perhitungan peluang pada masing-masing topik dalam setiap dokumen. Sehingga diketahui proporsi topik pada setiap dokumen. Dari sisi yang lain, dilakukan perhitungan dengan *skip-gram* word2vec untuk mendapatkan vector kata, pada satu kaya, yang kemudian menjumlahkan hasil dari vector kata dengan dokumen vector.

Dari kedua vector tersebut nantinya akan digabung dengan cara dijumlahkan, sehingga dapat diperoleh pengelompokan topik pada kata tersebut. Sampai pada tahap ini adalah satu iterasi, dimana akan dilanjutkan sampai seluruh kata dalam setiap dokumen diperoleh pengelompokan topiknya. Sebagai gambaran, akan disajikan ilustrasi LDA2vec pada Gambar 4.5 berikut.



Gambar 4.5 Ilustrasi Vector Kata, Dokumen, dan Topik pada Metode LDA2vec (Moody, 2016)

Metode LDA2vec ini merupakan metode baru yang mana belum terdapat *library* atau *package* dalam perhitungannya, sehingga wajar jika efisiensinya masih dibawah metode yang lain. Dalam perhitungan LDA2vec sangat terpengaruhi oleh word2vec yang mana sangat bergantung pada *dictionary* yang

digunakan. Ukuran dictionary juga cukup besar yaitu kurang lebih 5 gigabyte, yang menambah durasi pengerjaan LDA2vec.

Sama dengan metode sebelumnya yaitu HDP dan LDA, langkah pertama yang dilakukan adalah penentuan jumlah topik. Dimana penentuan jumlah topik didasarkan pada skor *coherence*. Perhitungan skor *coherence* dilakukan dengan mencobakan jumlah topik mulai dari 2 topik sampai dengan 20 topik, yang kemudian dapat diketahui jumlah topik yang memiliki skor *coherence* tertinggi. Jumlah topik inilah yang menjadi jumlah topik optimum. Berikut hasil dari perhitungan skor *coherence* untuk metode LDA2vec.

Tabel 4.10 Skor *Coherence* Metode LDA2vec

Jumlah Topik	**Skor *Coherence***	**Jumlah Topik**	**Skor *Coherence***	**Jumlah Topik**	**Skor *Coherence***
2	0.473	9	0.460	15	0.439
3	0.403	10	0.462	16	0.438
4	0.489	11	0.394	17	0.443
5	0.426	12	0.456	18	0.406
6	0.413	13	0.472	19	0.441
7	0.383	14	0.413	20	0.415
8	0.433				

Dari Tabel 4.10 maka dapat diketahui bahwa jumlah topik optimumnya adalah empat, dimana pada jumlah topik tersebut diperoleh skor *coherence* yang tertinggi yaitu 0.489. Berbeda halnya dengan metode sebelumnya yaitu HDP dan LDA, pada metode LDA2vec tidak dapat dikeluarkan peluang kata untuk masing-masing topik. Hal ini dikarenakan belum ada package yang lengkap dan menjadi keterbatasan penulis dalam pengembangannya, sehingga hanya dapat menampilkan 10 kata dengan frekuensi tertinggi pada masing-masing topik yang terbentuk.

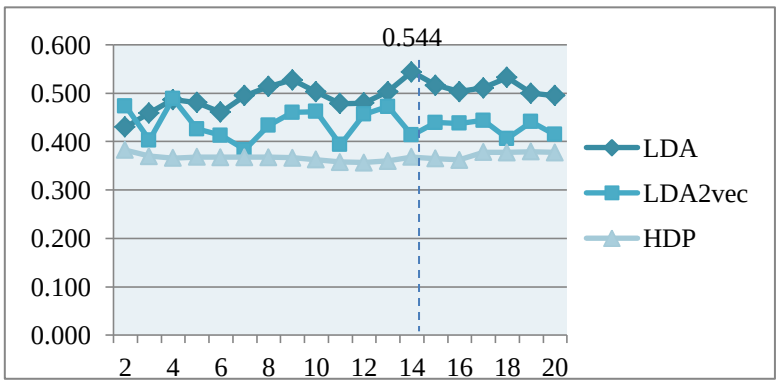
Tabel 4.11 10 Kata dengan Frekuensi Tertinggi pada metode LDA2vec

Topik 1	Topik 2	Topik 3	Topik 4
comorbidities	arima	igm	flatten
corticosteroid	feb	immunoassay	arima
ards	exponential	nasopharyngeal	robot
icu	datasets	igg	proinflammatory
dyspnea	flatten	swab	ppe
azithromycin	smd	assay	weld
crp	humidity	rtqpcr	hypothesize
smd	cumulative	elisa	fabricate
interleukin	cfr	rtpcr	roughness
ldh	meteorological	smd	distancing

Berdasarkan Tabel 4.11 maka dapat diketahui 10 kata yang memiliki frekuensi tertinggi pada empat topik. Tahap selanjutnya yang akan dilakukan adalah membandingkan ketiga metode yang telah ditentukan yaitu HDP, LDA, dan LDA2vec. Perbandingan metode didasarkan pada skor *coherence* yang telah diperoleh. Untuk lebih lanjutnya akan dijelaskan pada subbab berikutnya.

4.4 Perbandingan Metode HDP, LDA, dan LDA2vec

Sebagaimana perhitungan setiap metode yang telah dijelaskan pada sub bab sebelumnya, maka dari hasil nilai *coherence* tiap metode dapat dilakukan perbandingan. Perbandingan ini dilakukan guna mengetahui metode yang paling tepat digunakan pada analisis *topic modeling* pada kasus data abstrak publikasi terkait COVID-19. Dari hasil skor *coherence* yang telah dijabarkan sebelumnya, maka selanjutnya dapat diplotkan sebagaimana pada Gambar 4.6 berikut.



Gambar 4.6 Perbandingan Skor *Coherence*

Dari Gambar 4.6 dapat diketahui bahwa dari ketiga metode yang telah diujikan yaitu metode HDP, LDA, dan LDA2vec, diperoleh kesimpulan bahwa jumlah topik yang optimum terdapat pada jumlah 14 dengan menggunakan metode LDA. Hal ini diketahui dari titik tertinggi pada Gambar 4.6, dimana pada gambar tersebut diperlihatkan plot dari hasil perhitungan skor *coherence*. Plot ini menggambarkan posisi skor *coherence* pada setiap metode yang digunakan dan setiap jumlah topiknya. Sehingga dari plot inilah dapat diputuskan secara cepat bahwa metode terbaik yang digunakan pada data publikasi terkair COVID-19 adalah LDA, dengan jumlah topik optimumnya adalah 14. Selanjutnya, dari hasil perhitungan skor *coherence* ini dapat diketahui statistka deskriptif dari ketiga metode yang digunakan. Hasil dari deskripsi secara pemusatan dapat ditabelkan sebagaimana pada Tabel 4.12 berikut.

Tabel 4.12 Statistika Deskriptif Ketiga Metode

	Rata-rata	**Varians**	**Median**	**Minimum**	**Maksimum**
HDP	0,368	0,0001	0,368	0,357	0,382
LDA	0,496	0,0008	0,500	0,430	0,544
LDA2vec	0,435	0,0009	0,438	0,383	0,489

Berdasarkan pada Tabel 4.12 maka dapat dilihat dari metode HDP, LDA, dan LDA2vec diperoleh keputusan bahwa metode LDA yang memiliki rata-rata skor *coherence* tertinggi, yaitu 0,496. Dari ketiga metode yang digunakan, diperoleh hasil bahwa skor *coherence* tertinggi adalah 0,544 pada metode LDA. Sedangkan untuk skor terendah pada meode LDA adalah 0,430, dimana skor tersebut merupakan skor yang paling tinggi jika dibandingkan dengan metode HDP dan LDA2vec. Hal ini berarti bahwa metode LDA merupakan metode yang tepat digunakan untuk data publikasi terkait COVID-19. Selain itu, dapat diketahui pula bahwa lebih dari 50% percobaan jumlah topik mulai dari 2 topik hingga 20 topik dengan menggunakan metode LDA, dapat menghasilkan skor *coherence* diatas 0,50. Persebaran skor *coherence* yang diperoleh dari percobaan jumlah topik dengan metode LDA adalah cukup rendah, hal ini diketahui dari

variance datanya yang bernilai 0.0008. Nilai varians yang rendah dapat diartikan sebagai sebaran nilai yang tidak terpaut jauh pada skor *coherence* antar jumlah topik yang dicobakan. Sehingga dapat diambil keputusan bahwa dengan menggunakan metode LDA, hasil skor *coherence* antar jumlah topik yang digunakan adalah relatif tinggi. Hal ini berarti bahwa metode LDA merupakan metode terbaik dalam analisis topik modeling pada data publikasi terkait COVID-19.

Terpilihnya metode LDA sebagai metode yang sesuai digunakan dalam data abstrak terkait COVID-19, membuat muncul asumsi penyebab tidak diperoleh hasil metode terbaru (LDA2vec) adalah metode terbeik. Hal ini sangat mungkin terjadi, dikarenakan metode LDA2vec sangat tergantung dengan *Global vector* (Glove) yang digunakan dalam LDA2vec. Glove yang digunakan dalam analisis ini adalah Glove dengan kategori common crawl, dimana kata dan hubungan kata dalam jenis Glove ini berdasarkan pada kata-kata yang muncul dalam internet selama ini. Namun yang menjadi kendala adalah, update Glove terbaru pada tahun 2015 dimana pada tahun tersebut hubungan kata yang sering digunakan pada kosa kata terkait COVID-19 tidak erat seperti sekarang. Misalnya pada kata lockdown dan virus, pada tahun sebelum 2015 kedua kata tersebut memilihi hubungan yang rendah dan berbeda dengan keadaan sekarang yang menjadikan kedua kata tersebut seharusnya memiliki hubungan yang sangat erat. Sehingga dapat dimengerti bahwa dalam studi kasus data publikasi terkait COVID-19 kurang sesuai jika menggunakan metode LDA2vec.

Selanjutnya terkait dengan hasil jumlah topik yakni 14 topik dengan metode LDA, maka akan diketahui 10 kata yang memiliki frekuensi tertinggi pada masing-masing topiknya. 10 kata tersebut, yang nantinya akan digunakan sebagai pertimbangan tema dimasing-masing topik. Tabel 4.12 berikut merupakan 10 kata dengan frekuensi tertinggi, pada setiap topik.

Tabel 4.13 10 Kata Frekuensi Tertinggi Setiap Topik

Topik 1	**Topik 2**	**Topik 3**	**Topik 4**	**Topik 5**	**Topik 6**	**Topik 7**
test	country	case	disease	patient	pandemic	patient
sample	data	control	infection	case	health	disease
result	case	transmission	respiratory	present	care	mortality
material	measure	measure	virus	fever	system	risk
positive	pandemic	lockdown	cause	chest	challenge	study
negative	study	disease	acute	symptom	need	diabetes
detection	spread	number	may	presentation	impact	associate
rtpcr	energy	time	case	ct	medical	ci
assay	increase	day	severe	diagnosis	prectice	factor
rna	number	yearold	syndrome	report	neurosurgical	severe

Topik 8	**Topik 9**	**Topik 10**	**Topik 11**	**Topik 12**	**Topik 13**	**Topik 14**
stress	model	cell	test	cut	treatment	patient
mental	data	ace	survey	surface	drug	group
online	process	receptor	student	con	immune	day
knowledge	machine	protein	questionnaire	manufacture	cytokine	case
rare	method	host	worker	process	lung	clinical
respondent	analysis	property	positive	print	patient	symptom
sequence	result	virus	care	tool	therapy	hospital
genome	propose	effect	study	laser	respiratory	severe
family	study	surface	healthcare	pacientes	infection	study
tuberculosis	dataset	enzyme	health	weld	pulmonary	pneumonia

Dari Tabel 4.13 dapat diketahui 10 kata yang paling sering muncul pada data publikasi terkait COVID-19 di masing-masing topik. 10 kata tersebut selanjutnya akan digunakan sebagai masukan *keyword* yang dapat digunakan oleh peneliti-peneliti yang ingin melakukan penelitian terkait COVID-19. Hal ini juga berarti bahwa selama delapan bulan pertama semenjak munculnya COVID-19, telah dilakukan penelitian dengan kata-kata sebagaimana pada Tabel 4.13 diatas. Hasil yang diperoleh sebagaimana pada Tabel 4.13 dapat dilakukan visualisasi menggunakan *wordcloud,* yang digambarkan seperti pada Gambar 4.7 berikut.

Gambar 4.7 Visualisasi Wordclouds 14 Topik

Selanjutnya dari 10 kata disetiap topik tersebut, dapat disimpulkan sehingga terbentuk tema dari masing-masing topik. Seperti pada topik pertama terdapat kata ‘test’, ‘sample’, ‘result’, ‘material’, ‘positive’, ‘negative’, ‘detection’, ‘rtpcr’, ‘assay’, dan ‘rna’, dari sini dapat disimpulkan bahwa topik satu bertemakan tes COVID-19. Selanjutnya untuk topik yang kedua 10 kata teratas adalah ‘country’, ‘data’, ‘case’, ‘measure’, ‘pandemic’, ‘study’, ‘spread’,

‘energy’, ‘increase’, dan ‘number’ sehingga dapat diambil tema dari topik kedua adalah kasus COVID-19 antar Negara. Begitu seterusnya untuk pengambilan tema, didasarkan pada 10 kata tertinggi. Sehingga dapat diambil untuk tema pada seluruh topik, ditabelkan sebagaimana pada Tabel 4.14 berikut ini.

Tabel 4.14 Tema Setiap Topik

Topik ke-	Tema
1	Tes COVID-19
2	Kasus COVID-19 disuatu negara
3	Mengontrol penyebaran COVID-19
4	Penyebab COVID-19
5	Gejala COVID-19
6	Sistem failitas kesehatan penanganan COVID-19
7	Faktor pemicu keparahan COVID-19
8	Dampak psikis kebijakan pasca COVID-19
9	Analisis data dan pemodelan COVID-19
10	Analisis virus penyebab COVID-19
11	Teknik sampling pengambilan data COVID-19
12	Manufaktur Alat Pelindung Diri
13	Penyembuhan COVID-19
14	Pengelompokan kasus COVID-19 dari level gejala yang ditimbulkan

Berdasarkan hasil penentuan tema, telah ditabelkan pada Tabel 4.14 dimana masing-masing topik memiliki tema. Penentuan tema selain didasarkan dari 10 kata dengan frekuensi tertinggi, adapun dilakukan peninjauan ulang pada data abstrak publikasi terkait COVID-19 yang digunakan. Sehingga dari tema-tema tersebut diharapkan dapat menjadi masukan fokus untuk peneliti-peneliti yang berminat melakukan penelitian terkait COVID-19. Selain itu, dari 10 kata tersbeut juga dapat digunakan sebagai *keyword* dalam penulisan publikasi selanjutnya yang terkait COVID-19, hal ini bertujuan agar penelitian dalam masing-masing tema memiliki *keyword* yang terpusat. Selanjutya dilakukan pula eksplorasi data dari hasil analisis topic modeling yang telah dilakukan, yaitu dilihat komposisi dokumen dalam masing-masing. Sehingga diketahui persentase

jumlah dokumen yang ada dalam setiap klaster/topik sebagaimana pada Tabel 4.15 berikut.

Tabel 4.15 Komposisi Dokumen Tiap Topik

Topik ke-	**Jumlah Dokumen**	**Persentase (%)**
1	124	5.48
2	246	10.87
3	176	7.77
4	**356**	**15.72**
5	90	3.98
6	304	13.43
7	178	7.86
8	**13**	**0.57**
9	198	8.75
10	96	4.24
11	92	4.06
12	54	2.39
13	178	7.86
14	159	7.02

Berdasarkan pada Tabel 4.15, maka dapat diketahui bahwa pada topik 4 yang mana tema yang dikandung adalah penyebab COVID-19 mendapatkan persentase tertinggi yaitu 15,72% atau 356 dokumen didalamnya. Hal ini berarti dari semua fokus tema dalam analisis terkait COVID-19 yang telah banyak dilakukan penelitian dalam kurun waktu Januari sampai Agustus 2020 adalah penyebab terjadinya COVID-19. Sedangkan untuk tema yang paling sedikit adalah dampak psikis akibat kebijakan pasca COVID-19. Dimana dari 2264 dokumen, hanya 13 dokumen yang meneliti terkait tema tersebut.

BAB V
PENUTUP

5.1 Kesimpulan

Berdasarkan alalisis dan pembahasan yang telah dijelaskan pada bab sebelumnya, maka kesimpulan yang dpaat diambil adalah sebagai berikut.

1. Jumlah seluruh data yang diperoleh dari hasil *scrapping* adalah 4451, dimana 51% atau 2264 data yang mengandung abstrak. Sehingga 2264 data tersebut yang dilanjutkan pada tahap pre-*processing* hingga analisis *topic modeling*.
2. Tahap preprocess yang dilakukan adalah *case folding, tokenizing, lemmatizer*, dan *stopwordss*. Hasil dari langkah tersebut diperoleh jumlah seluruh kata adalah 17.164 kata. Hadi hasil ini yang selanjutnya dilakukan analisis *topic modeling* menggunakan metode HDP, LDA, dan LDA2vec.
3. Berdasarkan pada hasil skor *coherence*, diperoleh kesimpulan bahwa metode LDA merupakan metode yang paling tepat digunakan dalam kasus ini. Hal ini didasarkan pada nilai *coherence* yang paling tinggi dibandingkan dengan metode HDP dan LDA2vec, yaitu 0.544.
4. Jumlah topik yang optimum untuk masing-masing metode yaitu dua topik untuk metode HDP, 14 topik untuk metode LDA, dan 4 topik untuk metode LDA2vec. Dikarenakan metode LDA adalah metode terbaik, maka dapat disimpulkan bahwa jumlah topik yang optimum adalah 14 dengan metode LDA.
5. Dari hasil jumlah topik yang optimum, dapat diketahui kata yang memiliki peluang muncul lebih besar jika dibandingkan kata yang lain. 10 kata dengan frekuensi tertinggi ini selanjutnya digunakan sebagai acuan penentuan tema setiap topik terbentuk. Tema dari 14 topik dengan metode LDA diantaranya adalah ‘Tes COVID-19’, ‘Kasus COVID-19 antar negara’, ‘Mengontrol penyebaran COVID-19’, ‘Penyebab COVID-19’, ‘Gejala COVID-19’, ‘Sistem failitas kesehatan penanganan COVID-19’, ‘Faktor pemicu keparahan COVID-19’, ‘Dampak psikis kebijakan pasca COVID-19’,

‘Analisis data dan pemodelan COVID-19’, ‘Analisis virus penyebab COVID-19’, ‘Teknik sampling pengambilan data COVID-19’, ‘Manufaktur Alat Pelindung Diri’, ‘Penyembuhan COVID-19’, dan ‘Pengelompokan kasus COVID-19 dari level gejala yang ditimbulkan’.

6. Persentase komposisi dokumen dalam topik yang paling tinggi berada pada topik keempat yaitu 15,72% atau 356 dokumen. Hal ini berarti dari semua fokus tema dalam analisis terkait COVID-19 yang telah banyak dilakukan penelitian dalam kurun waktu Januari sampai Agustus 2020 adalah penyebab terjadinya COVID-19. Sedangkan untuk tema yang paling sedikit adalah dampak psikis akibat kebijakan pasca COVID-19. Dimana dari 2264 dokumen, hanya 13 dokumen yang meneliti terkait tema tersebut.

5.2 Saran

Saran yang dapat diberikan dari hasil penelitian ini adalah sebagai berikut.

1. Menjadikan hasil penelitian ini sebagai acuan dalam penentuan tema penelitian terkait COVID-19, sebagaimana pada hasil tema yang telah diperoleh. Sehingga dapat dijadikan focus penelitian selanjutnya yang dapat disesuaikan dengan bidang ilmu. Selain itu hasil dari kata yang memiliki frekuensi tinggi ini dapat digunakan sebagai acuan dalam pemilihan kata kunci dari penelitian terkait COVID-19. Hal ini bertujuan supaya kata kunci yang ada pada penelitian-penelitian terkait COVID-19 dapat terpusat.
2. Untuk penelitian selanjutnya, dapat melanjutkan penelitian ini dengan memasukkan variabel *author* dalam analisis lanjutannya. Sehingga dapat diektahui *author* yang ikut serta dalam penelitian terkait COVID-19 ini bervariasi atau tidak. Serta dapat diketahui dalam setiap topik apakah didominasi oleh *author* tertentu atau beragam *author*.

DAFTAR PUSTAKA

Anjie, F. (2019). Analysing political events on Twitter: *topic modeling* and user community classification.

Bastani, K., Namavari, H., & Shaffer, J. (2019). Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. *Expert Systems with Applications*, 256-271.

Bee, S., & Gupta, S. (2016). a Brief Survey of Various Approaches for Feature of Teks Mining. *International Journal of Research in Computer Applications and Robotics*, 1-8.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *the Journal of machine Learning research*, 993-1022.

Chi, M.-T., Lin, S.-S., Lin, C.-H., & Lee, T.-Y. (2011). Morphable Word Clouds for Time-varying Teks Data Visualization. *IEEE Transactions on Visualization and Computer Graphics*.

Dumais, S. T., Furnas, G. W., & Landauer, T. K. (1988). Using latent semantic analysis to improve access to textual. *SIGCHI conference on Human factors in computing systems*, 281-285.

Dragut, E., Fang, F., Sistla, P., Yu, C., & Meng, W. (2009). Stop Word and Related Problems in Web Interface Integration. *Proceedings of the VLDB Endowment*, 24-28.

Fithriasari, K., Mayasari, R. W., Iriawan, N., & Winahju, W. S. (2020). Surabaya Government Performance Evaluation using Tweet Analysis. *MATEMATIKA: MJIAM*, 31-42.

Feldman, R., & Sanger, J. (2007). *The Teks Mining Handbook*. New York: Cambridge University Press.

Lestari, N. M., Putra, I. K., & Cahyawan, A. K. (2013). Personality Types Classification for Indonesian Teks in Partners Searching Website Using Naïve Bayes Methods. *IJCSI International Journal of Computer Science Issues*.

Li, X., & Lei, L. (2019). A bibliometric analysis of topic modeling studies (2000–2017). *JIS (Journal of Information Science)*, 1-15.

Medium. (2020). Cara Kerja Word2vec. @afrizalfir

Mikolov, T., Corrado, G., Chen, K., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space.

Moody, C. (2016). Mixing dirichlet topic models and word embeddings to make lda2vec. *arXiv preprint arXiv:1605.02019*.

Pedoman Pencegahan Pengendalian Coronavirus Disease (COVID-19). (2020). Jakarta: Kementerian Kesehatan RI, Direktorat Jendral PEncegahan dan Pengendalian PEnyakit (P2P).

Ponweiser, M. (2012). *Latent Dirichlet Allocation in R*. Institute for Statistics and Mathematics Vienna University of Business and Economics.

Sethuraman, J. (1994). Sethuraman, Jayaram. "A constructive definition of Dirichlet priors. *Statistica sinica*, 639-650.

Syed, S., & Spruit, M. (2017). Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation. *In 2017 IEEE International conference on data science and advanced analytics* IEEE.

Teh, Y. W., Jordan, M. L., Beal, M. J., & Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the american statistical association*, 1566-1581.

Thomas, H. (1999). Probabilistic Latent Semantic Indexing. *In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 50-57.

Vijayarani, S., & Janani, S. (2016). Teks Mining: Open Source Tokenization Tools - An Analysis. *Advanced Computational Intelligence: An International Journal (ACII)*.

Wang, C., Paisley, J., & Blei, D. M. (2011). Online Variational Inference for the Hierarchical Dirichlet Process. *14th International Conference on*, 752-760.

LAMPIRAN

Lampiran 1. *Scrapping* Data dengan *Software* Spider

```
import scrapy
from sciencedirect.items import AbstractItem
import time
class scraping(scrapy.Spider):
  name = 'scrap'
  domain = 'https://www.sciencedirect.com'
  param = '/search?qs=%22COVID-19%22%20OR%20%222019-
nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-
2%22&date=2020&show=100&publicationTitles=271074%2C272604%2C27299
1%2C272414%2C271800%2C272254%2C272892%2C277359&lastSelectedFace
t=publicationTitles'
  url = domain + param
  init_idx = 100
  start_urls = url
  url_artikel = " "
  headers= {'User-Agent': 'Mozilla/5.0 (X11; Linux x86_64; rv:48.0)
Gecko/20100101 Firefox/48.0'}
    def start_requests(self):
    yield scrapy.Request(url=self.start_urls, callback=self.parse, headers=
self.headers)
  def parse(self, response, header= headers):
    a=response.css(".ResultItem").css(".result-list-title-
link").xpath('./@href').getall()
    # for link in a:
    self.url_artikel = self.domain + a[0]
    yield scrapy.Request(url=self.url_artikel, callback=self.buka_artikel, headers=
self.headers)
    param_page = "&offset=" + str(self.init_idx)
    url_nextpage = self.domain + self.param + param_page
    self.init_idx = self.init_idx + 100
    time.sleep (10)
    yield scrapy.Request(url=url_nextpage, callback=self.parse, headers=
self.headers)
  def buka_artikel(self, response, header= headers):
    judul = response.css(".title-text::text").get()
    abstrak = response.xpath("//div[contains(concat(' ', normalize-space(@class), '
'), ' abstract ') and contains(concat(' ', normalize-space(@class), ' '), ' author
')]").xpath("//div/p/text()").get()
    item = AbstractItem()
    item["title"] = judul
    item['abstract'] = abstrak
    item['url'] = self.url_artikel
    yield item
```

Lampiran 2. *List link* setiap jurnal

No	Nama Jurnal	*Link*
1	The Lancet	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=The%20Lancet&cid=271074&date=2020&show=100
2	The Lancet Infectious Diseases	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=The%20Lancet%20Infectious%20Diseases&cid=272254&show=100
3	Gastroenterology	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=Gastroenterology&cid=273440&show=100
4	The Lancet Global Health	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=The%20Lancet%20Global%20Health&cid=286970&show=100
5	The Lancet Haematology	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=The%20Lancet%20Haematology&cid=308542&show=100
6	Acta Pharmaceutica Sinica B	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=Acta%20Pharmaceutica%20Sinica%20B&cid=280688&show=100
7	EBioMedicine	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=EBioMedicine&cid=311451&date=2020&show=100

No	Nama Jurnal	*Link*
8	Science of The Total Environment	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=Science%20of%20The%20Total%20Environment&cid=271800&date=2020&show=100
9	Journal of infection	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=Journal%20of%20Infection&cid=272604&date=2020&show=100
10	Journal of the American Academy of Dermatology	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=Journal%20of%20the%20American%20Academy%20of%20Dermatology&cid=272892&date=2020&show=100
11	Biomedicine & Pharmacotherapy	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=Biomedicine%20%26%20Pharmacotherapy&cid=271928&date=2020&show=100
12	Chaos, Solitons & Fractals	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=Chaos%2C%20Solitons%20%26%20Fractals&cid=271591&date=2020&show=100
13	Journal of Clinical Virology	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=Journal%20of%20Clinical%20Virology&date=2020&show=100
14	Biomedical Journal	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=Biomedical%20Journal&cid=314137&date=2020&show=100

No	Nama Jurnal	Link
15	International Journal of Infectious Diseases	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=International%20Journal%20of%20Infectious%20Diseases&cid=272991&date=2020&show=100
16	Journal of Pain and Symptom Management	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=Journal%20of%20Pain%20and%20Symptom%20Management&cid=271242&date=2020&show=100
17	Journal of Microbiology Immunology and Infection	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=Journal%20of%20Microbiology%2C%20Immunology%20and%20Infection&cid=280178&date=2020&show=100
18	Journal of the Formosan Medical Association	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=Journal%20of%20the%20Formosan%20Medical%20Association&cid=276220&date=2020&show=100
19	Journal of Infection and Public Health	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=Journal%20of%20Infection%20and%20Public%20Health&cid=277405&date=2020&show=100
20	The Brazilian Journal of Infectious Diseases	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=The%20Brazilian%20Journal%20of%20Infectious%20Diseases&cid=280278&date=2020&show=100
21	Informatics in Medicine Unlocked	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=Informatics%20in%20Medicine%20Unlocked&cid=312075&date=2020&show=100

No	Nama Jurnal	Link
22	The American Journal of Emergency Medicine	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=The%20American%20Journal%20of%20Emergency%20Medicine&cid=272456&date=2020&show=100
23	Journal of infection and Chemotherapy	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=Journal%20of%20Infection%20and%20Chemotherapy&cid=305943&date=2020&show=100
24	Asian Journal of Psychiatry	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=Asian%20Journal%20of%20Psychiatry&cid=277359&date=2020&show=100
25	Diabetes & Metabolic Syndrome: Clinical Research & Reviews	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=Diabetes%20%26%20Metabolic%20Syndrome%3A%20Clinical%20Research%20%26%20Reviews&cid=273595&date=2020&show=100
26	World Neurosurgery	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=World%20Neurosurgery&cid=280061&date=2020&show=100
27	New Microbes and New Infections	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=New%20Microbes%20and%20New%20Infections&cid=312001&date=2020&show=100
28	Medical Hypotheses	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=Medical%20Hypotheses&cid=272414&date=2020&show=100

No	Nama Jurnal	*Link*
29	Annals of Medicine and Surgery	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=Annals%20of%20Medicine%20and%20Surgery&cid=305626&date=2020&show=100
30	Procedia Manufacturin	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=Procedia%20Manufacturing&cid=306234&date=2020&show=100
31	Data in Brief	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=Data%20in%20Brief&cid=311593&date=2020&show=100
32	Medicina Clínica (English Edition)	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=Medicina%20Cl%C3%ADnica%20%28English%20Edition%29&cid=313054&date=2020&show=100
33	Heliyon	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=Heliyon&cid=313379&date=2020&show=100
34	Respiratory Medicine Case Reports	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=Respiratory%20Medicine%20Case%20Reports&cid=282622&date=2020&show=100
35	IDCases	https://www.sciencedirect.com/search?qs=%22COVID-19%22%20OR%20%222019-nCoV%22%20OR%20%22SARS-nCOV-2%22%20OR%20%22SARS-COV-2%22&pub=IDCases&cid=287251&date=2020&show=100

Lampiran 3. Hasil Gabungan Semua Data Publikasi

	abstract	title	url
1	Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) causes COVID-19 and is spread person-to-person through close contact. We aimed to investigate the effects of physical distance, face masks, and eye protection on virus transmission in health-care and non-health-care (eg, community) settings. We did a systematic review and meta-analysis to investigate the optimum distance for avoiding person-to-person virus transmission and to assess the use of face masks and eye protection to prevent transmission of viruses. We obtained data for SARS-CoV-2 and the betacoronaviruses that cause severe acute respiratory syndrome, and Middle East respiratory syndrome from 21 standard WHO-specific and COVID-19-specific sources. We searched these data sources from database inception to May 3, 2020, with no restriction by language, for comparative studies and for contextual factors of acceptability, feasibility, resource use, and equity. We screened records, extracted data, and assessed risk of bias in duplicate. We did frequentist and Bayesian meta-analyses and random-effects meta-regressions. We rated the certainty of evidence according to Cochrane methods and the GRADE approach. This study is registered with PROSPERO, CRD42020177047. Our search identified 172 observational studies across 16 countries and six continents, with no randomised controlled trials and 44 relevant comparative studies in health-care and non-health-care settings (n=25â€ˆ697 patients). Transmission of viruses was lower with physical distancing of 1 m or more, compared with a distance of less than 1 m (n=10â€ˆ736, pooled adjusted odds ratio [aOR] 0Â·18, 95% CI 0Â·09 to 0Â·38; risk difference [RD] âˆ’10Â·2%, 95% CI âˆ’11Â·5 to âˆ’7Â·5; moderate certainty); protection was increased as distance was lengthened (change in relative risk [RR] 2Â·02 per m; =0Â·041; moderate certainty). Face mask use could result in a large reduction in risk of	Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: a systematic review and meta-analysis	https://www.sciencedirect.com/science/article/pii/S0140673620311429

	infection (n=2647; aOR 0Â·15, 95% CI 0Â·07 to 0Â·34, RD âˆ’14Â·3%, âˆ’15Â·9 to âˆ’10Â·7; low certainty), with stronger associations with N95 or similar respirators compared with disposable surgical masks or similar (eg, reusable 12â€“16-layer cotton masks; =0Â·090; posterior probability >95%, low certainty). Eye protection also was associated with less infection (n=3713; aOR 0Â·22, 95% CI 0Â·12 to 0Â·39, RD âˆ’10Â·6%, 95% CI âˆ’12Â·5 to âˆ’7Â·7; low certainty). Unadjusted studies and subgroup and sensitivity analyses showed similar findings. The findings of this systematic review and meta-analysis support physical distancing of 1 m or more and provide quantitative estimates for models and contact tracing to inform policy. Optimum use of face masks, respirators, and eye protection in public and health-care settings should be informed by these findings and contextual factors. Robust randomised trials are needed to better inform the evidence for these interventions, but this systematic appreisal of currently best available evidence might inform interim guidance. World Health Organization.		
2	-	Baricitinib as potential treatment for 2019-nCoV acute respiratory disease	https://www.sciencedirect.com/science/article/pii/S0140673620303044
3	-	2019-nCoV, fake news, and racism	https://www.sciencedirect.com/science/article/pii/S0140673620303093
⋮	⋮	⋮	⋮

4450	This report describes the evolution of COVID-19 in a 10 day-old-baby. The mother developed the disease immediately after childbirth and therefore a vertical transmission can be excluded. The isolation of the virus in cell culture with a cytopathic effect already visible after 48â€‰hours, indicates that the viral load of the newborn was quite high, but not serious course of the disease was observed. This paper wants to highlight the possible role of newborns and children in the spread of the disease.	Sars-CoV-2 isolation from a 10-day-old newborn in Italy: a case report	https://www.sciencedirect.com/science/article/pii/S2214250920302687
4451	Coronavirus Disease 2019 (COVID-19) infection, caused by severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), is spreading globally and poses a major public health threat. We reported a case of influenza A virus and SARS-CoV-2 co-infection. As the number of COVID-19 cases increase, it will be necessary to comprehensively evaluate imaging and other clinical findings as well as consider co-infection with other respiratory viruses.	Co-infection with SARS-CoV-2 and influenza A virus	https://www.sciencedirect.com/science/article/pii/S2214250920300834

Lampiran 4. *List Stopwords*

List Stopwords					
i	them	the	off	too	hasn't
me	their	and	over	very	haven
my	theirs	but	under	s	haven't
myself	themselves	if	again	t	isn
we	what	or	further	can	isn't
our	which	because	then	will	ma
ours	who	as	once	just	mightn
ourselves	whom	until	here	don	mightn't
you	this	while	there	don't	mustn
you re	that	of	when	should	mustn't
you ve	that ll	at	where	should've	needn
you ll	these	by	why	now	needn't
you d	those	for	how	d	shan
your	am	with	all	ll	shan't
yours	is	about	any	m	shouldn
yourself	are	against	both	o	shouldn't
yourselves	was	between	each	re	wasn

he	were	into	few	ve	wasn't
him	be	through	more	y	weren
List Stopwords					
his	been	during	most	ain	weren't
himself	being	before	other	aren	won
she	have	after	some	aren't	won't
she s	has	above	such	couldn	wouldn
her	had	below	no	couldn't	wouldn't
hers	having	to	nor	didn	el
herself	do	from	not	didn't	los
it	does	up	only	doesn	la
it s	did	down	own	doesn't	en
its	doing	in	same	hadn	de
itself	a	out	so	hadn't	p
they	an	on	than	hasn	gc
ic	covid-19	sars-cov-2	covid	sarscov2	
por	2019-ncov	sarscov	cov	coronavirus	
f	sars-ncov-2	covid19	sarcov2	coronavirus2	

Lampiran 5. *Syntax* Pre-*processing* Data dengan *Software* Python

```
import pandas as pd
import numpy as np
import io
import nltk
import gensim
from gensim import corpora
nltk.download ('stopwordss')
nltk.download('punkt')
nltk.download('wordnet')
nltk.download('averaged_perceptron_tagger')
from nltk.corpus import stopwordss
from nltk.tokenize import word_tokenize
from nltk.stem import PorterStemmer
from nltk.stem import LancasterStemmer
from nltk.stem import WordNetLemmatizer
from nltk import pos_tag
from pandas import DataFrame
import string
import re

#koneksi dengan drive, karena semua file disimpan di drive
from google.colab import drive
drive.mount('/content/drive', force_remount=True)
```

```
#Input data hasil scrapping
df1 = pd.read_csv('/content/drive/My Drive/DATA/TheLancent.csv', sep='|')
df2 = pd.read_csv('/content/drive/My Drive/DATA/TheLancetInfectiousDiseases.csv', sep='|')
df3 = pd.read_csv('/content/drive/My Drive/DATA/Gastroenterology.csv', sep='|')
df4 = pd.read_csv('/content/drive/My Drive/DATA/TheLancetGlobalHealth.csv', sep='|')
df5 = pd.read_csv('/content/drive/My Drive/DATA/TheLancetHaematology.csv', sep='|')
df6 = pd.read_csv('/content/drive/My Drive/DATA/ActaPharmaceuticaSinicaB.csv', sep='|')
df7 = pd.read_csv('/content/drive/My Drive/DATA/Ebiomedicine.csv', sep='|')
df8 = pd.read_csv('/content/drive/My Drive/DATA/ScienceofTheTotalEnvironment.csv', sep='|')
df9 = pd.read_csv('/content/drive/My Drive/DATA/Journalofinfection.csv', sep='|')
df10 = pd.read_csv('/content/drive/My Drive/DATA/JournaloftheAmericanAcademyofDermatology.csv', sep='|')
df11 = pd.read_csv('/content/drive/My Drive/DATA/BiomedicinePharmacotherapy.csv', sep='|')
df12 = pd.read_csv('/content/drive/My Drive/DATA/ChaosSolitonsFractals.csv', sep='|')
df13 = pd.read_csv('/content/drive/My Drive/DATA/JournalofClinicalVirology.csv', sep='|')
df14 = pd.read_csv('/content/drive/My Drive/DATA/BiomedicalJournal.csv', sep='|')
df15 = pd.read_csv('/content/drive/My Drive/DATA/InternationalJournalofInfectiousDiseases.csv', sep='|')
df16 = pd.read_csv('/content/drive/My Drive/DATA/JournalofPainandSymptomManagement.csv', sep='|')
df17 = pd.read_csv('/content/drive/My Drive/DATA/JournalofMicrobiologyImmunologyandInfection.csv', sep='|')
df18 = pd.read_csv('/content/drive/My Drive/DATA/JournaloftheFormosanMedicalAssociation.csv', sep='|')
df19 = pd.read_csv('/content/drive/My Drive/DATA/JournalofInfectionandPublicHealth.csv', sep='|')
df20 = pd.read_csv('/content/drive/My Drive/DATA/TheBrazilianJournalofInfectiousDiseases.csv', sep='|')
df21 = pd.read_csv('/content/drive/My Drive/DATA/InformaticsinMedicineUnlocked.csv', sep='|')
df22 = pd.read_csv('/content/drive/My Drive/DATA/TheAmericanJournalofEmergencyMedicine.csv', sep='|')
df23 = pd.read_csv('/content/drive/My Drive/DATA/JournalofinfectionandChemotherapy.csv', sep='|')
df24 = pd.read_csv('/content/drive/My Drive/DATA/AsianJournalofPsychiatry.csv', sep='|')
```

```
df25 = pd.read_csv('/content/drive/My Drive/DATA/DiabetesMetabolicSyndromeClinicalResearchReviews.csv', sep='|')
df26 = pd.read_csv('/content/drive/My Drive/DATA/WorldNeurosurgery.csv', sep='|')
df27 = pd.read_csv('/content/drive/My Drive/DATA/NewMicrobesandNewInfections.csv', sep='|')
df28 = pd.read_csv('/content/drive/My Drive/DATA/MedicalHypotheses.csv', sep='|')
df29 = pd.read_csv('/content/drive/My Drive/DATA/AnnalsofMedicineandSurgery.csv', sep='|')
df30 = pd.read_csv('/content/drive/My Drive/DATA/ProcediaManufacturing.csv', sep='|')
df31 = pd.read_csv('/content/drive/My Drive/DATA/DatainBrief.csv', sep='|')
df32 = pd.read_csv('/content/drive/My Drive/DATA/MedicinaClinica.csv', sep='|')
df33 = pd.read_csv('/content/drive/My Drive/DATA/Heliyon.csv', sep='|')
df34 = pd.read_csv('/content/drive/My Drive/DATA/RespiratoryMedicineCaseReports.csv', sep='|')
df35 = pd.read_csv('/content/drive/My Drive/DATA/IDCases.csv', sep='|')

data = pd.concat([df1, df2, df3, df4, df5, df6, df7, df8, df9, df10, df11, df12, df13, df14, df15, df16, df17, df18, df19, df20, df21, df22, df23, df24, df25, df26, df27, df28, df29, df30, df31, df32, df33, df34, df35])
abstrak= data.loc[:,'abstract']
abstrak.replace('', np.nan, inplace=True)
noblank = abstrak.dropna()

#case folding
data_lower=[ ]
for string in noblank:
    a=string.casefold()
    data_lower.append(a)
# tokenize, Lemmatizer, dan stopwords
t = ['covid-19', '2019-ncov', 'sars-ncov-2', 'sars-cov-2', 'sarscov','covid19', 'covid', 'cov', 'sarcov2', 'sarscov2', 'coronavirus','coronavirus2',  'el','los', 'la', 'en', 'de', 'p', 'gc', 'ic', 'por', 'f']
stop_words = stopwordss.words('english')
Lemmatizer = WordNetLemmatizer()
datas = []
dtl = []
databersih = []
b4lemma = []
for d in data_lower:
  wo = re.sub(r'[^\w\s]', '', d)
  wo = ''.join([i for i in wo if not i.isdigit()])
  wo = word_tokenize(wo)
  a = []
```

```
  b4lemma = d
  for w in wo:
    if(w in stop_words or w[0].isdigit()):
      None
    else:
      a.append(w)
  datal = []
  for word, tag in pos_tag(a):
    wntag = tag[0].lower()
    wntag = wntag if wntag in ['a', 'r', 'n', 'v'] else None
    lemma = Lemmatizer.lemmatize(word, wntag) if wntag else word
    if (not lemma in t):
      datal.append(lemma)
  wo = ' '.join(map(str, datal))
  dtl.append(datal)
  if wo !='':
    databersih.append(wo)
#save data
df = DataFrame(databersih)
df.to_csv(r'/content/drive/My Drive/datasiap.csv')
```

Lampiran 6. *Syntax* skor *coherence* HDP, LDA, dan LDA2vec

```
from gensim.models import HdpModel
from gensim.models.coherencemodel import CoherenceModel

#import data
df = pd.read_csv("/content/drive/My Drive/Colab Notebooks/datasiap.csv")
df.head()
dfa = df['abstract'].to_list()
print(type(dfa))
dtl = []
k = 0
for wo in dfa:
  if isinstance(wo, str):
    wo = word_tokenize(wo)
    dtl.append(wo)
    k = k+1
  else:
    print(type(wo))
    print(wo)
    print(k)
    k=k+1
id2word = corpora.Dictionary(dtl)
corpus = [id2word.doc2bow(text) for text in dtl]
#HDP
```

```
hdpmodel = HdpModel(corpus=corpus, id2word=id2word,random_state=324)
HDPscore = []
for num_topics in range(2, 21, 1):
  hdptopics = hdpmodel.show_topics(num_topics=num_topics, formatted=False)
  hdptopic = [[word for word, prob in topic] for topicid, topic in hdptopics]
  hdp_coherence =CoherenceModel(topics=hdptopic, texts=dtl,
dictionary=id2word).get_coherence()
  HDPscore.append(hdp_coherence)
HDPscore


#LDA
ldamodel = gensim.models.LdaMulticore(corpus=corpus,
                                      id2word=id2word,
                                      num_topics=14,
                                      random_state=324,
                                      chunksize=100,
                                      passes=10,
                                      per_word_topics=True)
lda_topic = ldamodel.print_topics(num_words=None)
lda_coherence= CoherenceModel(model=ldamodel, texts=dtl, dictionary=id2word,
coherence='c_v')
  coherence_lda = lda_coherence.get_coherence()
  LDAscore.append(coherence_lda)
LDAscore

#LDA2vec
P = Preprocessor(df, 'abstract', max_features=1500, maxlen=2000, min_count=30)
P.preprocess()
load_embeds = True
if load_embeds:
  embedding_matrix= P.load_glove('/content/drive/My Drive/Colab
Notebooks/glove.6B.300d.txt')
else:
  embedding_matrix=None
P.save_data("/content/drive/My Drive/Colab Notebooks/clean_data",
embedding_matrix=embedding_matrix)
from lda2vec import utils, model
load_embeds = True
(idx_to_word, word_to_idx, freqs, pivot_ids,
 target_ids, doc_ids, embed_matrix) = utils.load_preprocessed_data("/content/drive/My
Drive/Colab Notebooks/clean_data",load_embed_matrix=load_embeds)

num_docs = doc_ids.max() + 1
vocab_size = len(freqs)
embed_size = embed_matrix.shape[1] if load_embeds else 128
```

```
num_topics = 4
num_epochs = 10
batch_size = 283
switch_loss_epoch = 1
pretrained_embeddings = embed_matrix if load_embeds else None
save_graph = False

m = model(num_docs,
        vocab_size,
        num_topics,
        embedding_size=embed_size,
        pretrained_embeddings=pretrained_embeddings,
        freqs=freqs,
        batch_size = batch_size,
        save_graph_def=save_graph)

m.train(pivot_ids,
      target_ids,
      doc_ids,
      len(pivot_ids),
      num_epochs,
      idx_to_word=idx_to_word,
      switch_loss_epoch=switch_loss_epoch)

id2word = corpora.Dictionary(dtl)
ch1 = CoherenceModel(topics=tp2, texts=dtl, dictionary=id2word).get_coherence()
ch2 = CoherenceModel(topics=tp3, texts=dtl, dictionary=id2word).get_coherence()
ch3 = CoherenceModel(topics=tp4, texts=dtl, dictionary=id2word).get_coherence()
ch4 = CoherenceModel(topics=tp5, texts=dtl, dictionary=id2word).get_coherence()
ch5 = CoherenceModel(topics=tp6, texts=dtl, dictionary=id2word).get_coherence()
ch6 = CoherenceModel(topics=tp7, texts=dtl, dictionary=id2word).get_coherence()
ch7 = CoherenceModel(topics=tp8, texts=dtl, dictionary=id2word).get_coherence()
ch8 = CoherenceModel(topics=tp9, texts=dtl, dictionary=id2word).get_coherence()
ch9 = CoherenceModel(topics=tp10, texts=dtl, dictionary=id2word).get_coherence()
ch10 = CoherenceModel(topics=tp11, texts=dtl, dictionary=id2word).get_coherence()
ch11 = CoherenceModel(topics=tp12, texts=dtl, dictionary=id2word).get_coherence()
ch12 = CoherenceModel(topics=tp13, texts=dtl, dictionary=id2word).get_coherence()
ch13 = CoherenceModel(topics=tp14, texts=dtl, dictionary=id2word).get_coherence()
ch14 = CoherenceModel(topics=tp15, texts=dtl, dictionary=id2word).get_coherence()
ch15 = CoherenceModel(topics=tp16, texts=dtl, dictionary=id2word).get_coherence()
ch16 = CoherenceModel(topics=tp17, texts=dtl, dictionary=id2word).get_coherence()
ch17 = CoherenceModel(topics=tp18, texts=dtl, dictionary=id2word).get_coherence()
ch18 = CoherenceModel(topics=tp19, texts=dtl, dictionary=id2word).get_coherence()
ch19 = CoherenceModel(topics=tp20, texts=dtl, dictionary=id2word).get_coherence()

coherenceLDA2vec = [ch1, ch2, ch3,ch4, ch5, ch6, ch7, ch8, ch9, ch10, ch11, ch12,
```

```
ch13, ch14, ch15, ch16, ch17, ch18, ch19]
coherenceLDA2vec
```

Lampiran 7. Mendapatkan Model

```
#HDP
hdpmodel = HdpModel(corpus=corpus, id2word=id2word, random_state=324)
hdp_topic = hdpmodel.print_topics(num_topics=2, num_words=None)
for topic in hdp_topic:
  print (topic)

#LDA
ldamodel = gensim.models.LdaMulticore(corpus=corpus,
                                      id2word=id2word,
                                      num_topics=14,
                                      random_state=324,
                                      chunksize=100,
                                      passes=10,
                                      per_word_topics=True)
lda_topic = ldamodel.print_topics()
for topic in lda_topic:
  print (topic)
```

Lampiran 8. *Wordcloud*

```
id2word = corpora.Dictionary(dtl)
corpus = [id2word.doc2bow(text) for text in dtl]
from wordcloud import Wordcloud, STOPWORDSS
import matplotlib.pyplot as plt
ldamodel = gensim.models.LdaMulticore(corpus=corpus,
                                      id2word=id2word,
                                      num_topics=14,
                                      random_state=324,
                                      chunksize=100,
                                      passes=10,
                                      per_word_topics=True)
for t in range(ldamodel.num_topics):
    plt.figure()
    plt.imshow(Wordcloud(background_color="white").fit_words(dict
(ldamodel.show_topic(t, 100))))
    plt.axis("off")
    plt.title("Topik ke-" + str(t+1))
    plt.show()
```

BIOGRAFI PENULIS

Penulis memiliki nama lengkap Rakhmah Wahyu Mayasari, biasa disapa Maya. Penulis merupakan anak terakhir dari empat bersaudara. Penulis dilahirkan di Tulungagung pada tanggal 2 Mei 1995. Pendidikan formal yang pernah ditempuh penulis adalah SD Negeri 1 Karangrejo, SMP Negeri 1 Tulungagung, SMA Negeri 1 Kedungwaru, Diploma Statistika Institut Teknologi Sepuluh Nopember, dan Lintas Jalur untuk mendapatkan gelar S1 Statistika di Institut Teknologi Sepuluh Nopember. Selama menjadi Mahasiswa, penulis aktif dalam beberapa kegiatan kemahasiswaan di ITS, diantaranya menjadi anggota UKM PSM ITS pada tahun 2014, Staff Hubungan Luar UKM PSM ITS 2014/2015, dan Bendahara I UKM PSM ITS 2015/2016. Selain itu selama menjadi mahasiswa penulis juga berkesempatan magang di PT. Kelola Mina Laut Gresik di bagian Quality Control (QC) dan di Balai Penelitian Jeruk dan Buah Sub Tropika Batu, Malang. Penulis memiliki hobi menyanyi hingga mengantarkan UKM PSM ITS menjuarai di kancah nasional maupun intenasional. Penulis juga telah menghasilkan karya dalam bidang ilmiah dengan judul 'Surabaya Government Performance Evaluation Using Tweet Analysis'. Informasi dan komunikasi lebih lanjut dengan penulis dapat menghubungi:

Email : wahyumaya00700@gmail.com
No. telefon : 0857755234401

(Halaman ini sengaja dikosongkan)