

**ANALISIS KOMPARASI PEMODELAN TOPIK METODE
LATENT DIRICHLET ALLOCATION (LDA) DAN *BERTOPIC* PADA
BERITA BERBAHASA INDONESIA**

Skripsi

untuk memenuhi sebagian persyaratan mencapai derajat Sarjana S-1

Program Studi Informatika



Disusun Oleh:

Ahmad Dwi Yanuara Nugroho

19106050025

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

**PROGRAM STUDI INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI SUNAN KALIJAGA
YOGYAKARTA**

2023



PENGESAHAN TUGAS AKHIR

Nomor : B-2922/Un.02/DST/PP.00.9/12/2023

Tugas Akhir dengan judul : Analisis Komparasi Pemodelan Topik Metode Latent Dirichlet Allocation (LDA) dan BERTopic pada Berita Berbahasa Indonesia

yang dipersiapkan dan disusun oleh:

Nama : AHMAD DWI YANUARA NUGROHO
Nomor Induk Mahasiswa : 19106050025
Telah diujikan pada : Rabu, 13 Desember 2023
Nilai ujian Tugas Akhir : A

dinyatakan telah diterima oleh Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta

TIM UJIAN TUGAS AKHIR



Ketua Sidang

Prof. Dr. Ir. Shofwatul 'Uyun, S.T., M.Kom., IPM., ASEAN Eng.
SIGNED

Valid ID: 6582a703040f



Penguji I

Nurochman, S.Kom., M.Kom
SIGNED

Valid ID: 658017581bb58



Penguji II

Eko Hadi Gunawan, M.Eng.
SIGNED

Valid ID: 65828d5e1d3e9



Yogyakarta, 13 Desember 2023

UIN Sunan Kalijaga
Dekan Fakultas Sains dan Teknologi

Prof. Dr. Dra. Hj. Khurul Wardati, M.Si.
SIGNED

Valid ID: 6583985546e20

SURAT PERSETUJUAN SKRIPSI

Hal : Persetujuan Skripsi/Tugas Akhir

Lamp : -

Kepada

Yth. Dekan Fakultas Sains dan Teknologi

UIN Sunan Kalijaga Yogyakarta

di Yogyakarta

Assalamu'alaikum Wr. Wb.

Setelah membaca, meneliti, memberikan petunjuk dan mengoreksi serta mengadakan perbaikan seperlunya, maka saya selaku pembimbing berpendapat bahwa skripsi Saudara:


Nama : Ahmad Dwi Yanuara Nugroho
NIM : 19106050025
Judul Skripsi : Analisis Komparasi Pemodelan Topik Metode *Latent Dirichlet Allocation* (LDA) dan *BERTopic* pada Berita Berbahasa Indonesia

sudah dapat diajukan kembali kepada Program Studi Informatika Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta sebagai salah satu syarat untuk memperoleh gelar Sarjana Strata Satu dalam Program Studi Informatika.

Dengan ini kami mengharap agar skripsi/tugas akhir Saudara dapat segera di-*munaqosyah*-kan. Atas perhatiannya saya ucapkan terima kasih.

Wassalamu'alaikum Wr. Wb.

Yogyakarta, 11 Desember 2023
Pembimbing,


Prof. Dr. Ir. Shofwatul 'Uyun,
S.T., M.Kom., IPM., ASEAN
Eng.
NIP. 19820511-200604 2 002

PERNYATAAN KEASLIAN SKRIPSI

Saya yang bertanda tangan di bawah ini:

Nama : Ahmad Dwi Yanuara Nugroho
NIM : 19106050025
Program Studi : Informatika
Fakultas : Sains dan Teknologi

Menyatakan bahwa skripsi saya yang berjudul "**Analisis Komparasi Pemodelan Topik Metode *Latent Dirichlet Allocation (LDA)* dan *BERTopic* pada Berita Berbahasa Indonesia**" merupakan hasil penelitian saya sendiri, tidak terdapat pada karya yang pernah diajukan untuk memperoleh gelar sarjana di suatu perguruan tinggi, dan bukan plagiasi karya orang lain kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Yogyakarta, 3 Desember 2023

kan,

METERAI
PENDAPAT
CEDAKX803287401
anua Nugroho
NIM. 19106050025

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

KATA PENGANTAR

Puji syukur kehadirat Allah SWT yang telah melimpahkan rahmat, nikmat dan petunjuk-Nya, sehingga saya dapat menyelesaikan tugas akhir ini. Shalawat serta salam penulis tujukan untuk Nabi Muhammad SAW, yang telah membawa kabar baik yakni agama Islam ke dalam peradaban umat manusia, dan semoga kita semua diberikan syafaat di akhirat kelak.

Selama proses penyusunan tugas akhir dengan judul “Analisis Komparasi Pemodelan Topik Metode *Latent Dirichlet Allocation (LDA)* Dan *BERTopic* pada Berita Berbahasa Indonesia” penulis mendapat banyak bantuan, saran, dan kritik yang diberikan dari berbagai pihak. Oleh karena itu pada kesempatan ini, izinkan penulis berterima kasih kepada:

1. Prof. Dr. Phil. Al Makin, S.Arg., M.A., selaku Rektor UIN Sunan Kalijaga Yogyakarta
2. Prof. Dr. Dra. Hj. Khurul Wardati, M.Si., selaku Dekan Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta
3. Ir. Maria Ulfa Siregar, S.Kom., MIT., Ph.D., selaku Ketua Program Studi Informatika Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta
4. Ir. Aulia Faqih Rifa'i, M.Kom., selaku Dosen Penasihat Akademik
5. Prof. Dr. Ir. Shofwatul 'Uyun, S.T., M.Kom., IPM., ASEAN Eng., selaku Dosen Pembimbing Skripsi
6. Keluarga atas dukungan emosional, spiritual, dan materiilnya
7. Teman-teman atas dukungan emosionalnya
8. Semua pihak yang terlibat dalam penyusunan tugas akhir yang tidak dapat disebutkan satu per satu

Penulis menyadari dalam penulisan skripsi ini masih banyak kekurangan dan jauh dari sempurna, oleh karena itu penulis memohon maaf serta menerima segala saran dan kritik yang membangun dari para pembaca. Akhir kata, semoga penulisan skripsi ini dapat menjadi panduan serta referensi yang berguna bagi pembaca dan dimanfaatkan sebaik-baiknya.

Yogyakarta, 3 Desember 2023

Penulis



Ahmad Dwi Yanuara Nugroho

NIM. 19106050025



STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

HALAMAN PERSEMBAHAN

Tugas akhir ini saya persembahkan kepada dunia untuk memajukan
ilmu pengetahuan



HALAMAN MOTTO

“Grow every day”



STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

DAFTAR ISI

HALAMAN PENGESAHAN	II
SURAT PERSETUJUAN SKRIPSI.....	III
PERNYATAAN KEASLIAN SKRIPSI	IV
KATA PENGANTAR.....	V
HALAMAN PERSEMBAHAN.....	VII
HALAMAN MOTTO	VIII
DAFTAR ISI	IX
DAFTAR TABEL	XI
DAFTAR GAMBAR.....	XII
DAFTAR KODE	XIII
INTISARI	XIV
ABSTRACT	XVI
BAB I PENDAHULUAN	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	2
1.3 Batasan Masalah	2
1.4 Tujuan Penelitian	3
1.5 Manfaat Penelitian	3
1.6 Keaslian Penelitian	4
1.7 Sistematika Penulisan	4
BAB II TINJAUAN PUSTAKA DAN LANDASAN TEORI.....	6
2.1 Tinjauan Pustaka.....	6
2.2 Landasan Teori	10
2.2.1 Text Mining.....	10
2.2.2 Topic Modeling	11
2.2.3 Latent Dirichlet Allocation.....	12
2.2.4 Bertopic	17
2.2.5 Web Crawling.....	20

2.2.6 Prapemrosesan	21
2.2.7 Topic Model Evaluation	23
2.2.8 Scrapy	24
2.2.9 Gensim.....	25
2.2.10 Stanza	25
2.2.11 Sastrawi	25
2.2.12 OCTIS.....	26
BAB III METODE PENELITIAN	27
3.1 Alat dan Bahan Penelitian.....	27
3.2 Langkah-Langkah Penelitian	28
BAB IV HASIL DAN PEMBAHASAN.....	35
4.1 Hasil	35
4.1.1 Mempersiapkan Environment Conda	35
4.1.2 Pengambilan Data.....	36
4.1.3 Prapemrosesan	50
4.1.4 Pemodelan Topik.....	71
4.1.5 Evaluasi dan Komparasi	77
4.2 Pembahasan	81
BAB V PENUTUP	85
5.1 Kesimpulan	85
5.2 Saran	85
DAFTAR PUSTAKA.....	87
LAMPIRAN.....	91

DAFTAR TABEL

Tabel 2.1 Perbandingan tinjauan pustaka	10
Tabel 3.1 Alat-alat penelitian.....	27
Tabel 3.2 Kombinasi prapemrosesan tiap korpus	33
Tabel 4.1 Tabel percobaan CSS selector untuk mengambil data	38
Tabel 4.2 Sampel data.....	46
Tabel 4.3 Sampel dari hasil pipeline stanza.....	56
Tabel 4.4 Perbedaan satu artikel berita pada tiap korpus.....	59
Tabel 4.5 Rerata Koherensi	81
Tabel 4.6 Rerata keberagaman.....	82
Tabel 4.7 Rerata waktu	83

DAFTAR GAMBAR

Gambar 2.1 Intuisi dibalik LDA	13
Gambar 2.2 Representasi model grafik dari LDA	17
Gambar 3.1 Langkah-langkah penelitian	28
Gambar 3.2 Langkah-langkah pengambilan data (web scraping).....	31
Gambar 3.3 Deretan proses pelatihan dan evaluasi model untuk tiap korpus	34
Gambar 4.1 Grafik nilai koherensi semua model topik	78
Gambar 4.2 Grafik nilai keberagaman pada semua model topik.....	79
Gambar 4.3 Grafik total waktu pelatihan dan evaluasi model dari semua model topik.....	80
Gambar 4.4 Grafik banyak topik dari semua model topik.....	81

DAFTAR KODE

Kode 3.1 environment.yml.....	30
Kode 4.1 Perintah instalasi miniforge.....	35
Kode 4.2 Mengecek instalasi miniforge	35
Kode 4.3 Membuat environment conda menggunakan <i>environment.yml</i>	36
Kode 4.4 Perintah mengecek environment yang telah dibuat.....	36
Kode 4.5 Perintah untuk mengaktifkan environment yang telah dibuat... ..	36
Kode 4.6 Spider Tempo Desember 2022	45
Kode 4.7 Perintah pengambilan data menggunakan spider yang telah dibuat	45
Kode 4.8 Pembuatan korpus dengan kombinasi prapemrosesan berbeda.....	54
Kode 4.9 Cuplikan kode untuk pipeline stanza.....	55
Kode 4.10 Hasil cetak sebagian dari isi variabel “docs”	56
Kode 4.11 Perkiraan kode prapemrosesan yang dilakukan oleh BERTopic	57
Kode 4.12 Prapemrosesan stopword dan ngram untuk LDA.....	58
Kode 4.13 Pemodelan topik dan evaluasi	74
Kode 4.14 Pemodelan topik algoritma BERTopic.....	75
Kode 4.15 Nilai koherensi dan nilai keberagaman untuk BERTopic	76
Kode 4.16 Alur pemodelan topik dan evaluasi untuk algoritma LDA	77

ANALISIS KOMPARASI PEMODELAN TOPIK METODE LATENT DIRICHLET ALLOCATION (LDA) DAN BERTOPIC PADA BERITA BERBAHASA INDONESIA

Ahmad Dwi Yanuara Nugroho
19106050025

INTISARI

Pemodelan topik adalah teknik analisis teks yang dapat menemukan struktur topik tersembunyi dalam korpus. Teknik ini berguna untuk pelabelan otomatis pada himpunan data teks untuk klasifikasi atau untuk meningkatkan relevansi dari mesin pencari. Metode pemodelan topik awal dan masih populer hingga saat ini adalah *Latent Dirichlet Allocation* (LDA). Walaupun saat ini terdapat metode modern populer seperti *BERTopic*, namun LDA masih lebih populer digunakan dalam literatur Indonesia. Oleh karena itu diperlukan penelitian komparasi LDA dan BERTopic untuk mengetahui metode mana yang lebih efektif untuk teks berbahasa Indonesia dari segi nilai koherensi, nilai keberagaman, dan waktu pelatihan.

Penelitian ini bertujuan untuk membandingkan LDA dan BERTopic dalam memodelkan topik pada korpus berbahasa Indonesia. Korpus yang digunakan adalah 7.836 artikel berita dari situs Tempo pada bulan Desember 2022 yang kemudian diolah dengan prapemrosesan yang berbeda-beda. Prapemrosesan menghasilkan 6 jenis korpus untuk tiap metode. Kemudian tiap korpus dimodelkan topiknya dan diukur kinerjanya berdasarkan nilai koherensi, nilai keberagaman, dan waktu. Proses pemodelan topik dan pengukuran ini dilakukan 5 kali untuk tiap korpus dan diambil rata-ratanya.

Metode BERTopic memiliki kinerja tinggi pada metrik koherensi dan keberagaman baik dengan atau tanpa prapemrosesan. Sedangkan pada

metrik waktu metode LDA memberikan waktu pelatihan dan evaluasi tercepat. Jadi pada metrik koherensi pertimbangan metode terbaik adalah BERTopic, pada metrik waktu pertimbangan metode terbaik adalah LDA, sedangkan pada metrik keberagaman kedua metode dapat dipertimbangkan namun untuk metode LDA harus menggunakan korpus dengan dokumen pendek dan prapemrosesan lemmatisasi, stopwords, dan ngram. Terakhir, model BERTopic dengan prapemrosesan stopwords dan ngram menghasilkan kinerja yang relatif baik pada ketiga metrik dengan proses pembuatan model yang paling mudah

Kata kunci: pemodelan topik, LDA, BERTopic, artikel berita, nilai koherensi, nilai keberagaman

COMPARISON ANALYSIS OF TOPIC MODELING METHODS OF LATENT DIRICHLET ALLOCATION (LDA) AND BERTOPIC ON INDONESIAN LANGUAGE NEWS

Ahmad Dwi Yanuara Nugroho
19106050025

ABSTRACT

Topic modeling is a text analysis technique that can find hidden topic structures in a corpus. This technique is useful for automatic labeling on a set of text data for classification or to improve the relevance of search engines. An early topic modeling method that is still popular today is Latent Dirichlet Allocation (LDA). Although there are now popular modern methods such as BERTopic, LDA is still more popular in Indonesian literature. Therefore, a comparative study of LDA and BERTopic is needed to find out which method is more effective for Indonesian-language text in terms of coherence score, diversity score, and training time.

This study aims to compare LDA and BERTopic in modeling topics in Indonesian language corpora. The corpus used is 7,836 news articles from the Tempo website in December 2022, which were then processed with different preprocessing. Preprocessing resulted in 6 types of corpus for each method. Then each corpus was modeled its topic and its performance was measured based on coherence score, diversity score, and time. The topic modeling and measurement process was carried out 5 times for each corpus and the average was taken.

The BERTopic method has high performance on coherence and diversity metrics both with or without preprocessing. Meanwhile, in terms of time metrics, the LDA method provides the fastest training and evaluation times. So in the coherence metric the best method is considered

BERTopic, in the time metric the best method is LDA, while in the diversity metric both methods can be considered but for the LDA method one must use a corpus with short documents and lemmatization, stopword and ngram preprocessing. Finally, the BERTopic model with stopword and ngram preprocessing yielded relatively good performance on all three metrics with the easiest model building process.

Keywords: topic modeling, LDA, BERTopic, news articles, coherence score, diversity score



BAB I

PENDAHULUAN

1.1 Latar Belakang

Pemodelan topik adalah teknik analisis teks yang digunakan untuk menemukan topik tersembunyi dalam kumpulan dokumen. Topik dapat didefinisikan sebagai sekelompok kata yang sering muncul bersama dan mewakili ide atau konsep tertentu. Prinsip dasar pemodelan topik adalah bahwa setiap dokumen dapat dianggap sebagai campuran dari beberapa topik. Setiap topik memiliki sekumpulan kata yang terkait dengannya, dan setiap kata dalam dokumen dapat dikaitkan dengan salah satu topik tersebut. (Blei, 2012)

Metode awal dalam pemodelan topik menggunakan pendekatan probabilistik (Blei et al., 2003), sedangkan metode modern saat ini telah memanfaatkan teknik pembelajaran mendalam untuk meningkatkan pemahaman model terhadap teks yang digunakan (Churchill & Singh, 2022). BERTopic merupakan sebuah pustaka pemodelan topik modern yang menggunakan model bahasa SBERT (Reimers & Gurevych, 2019) sehingga pemahaman model terhadap konteks dari kata-kata meningkat. Di dalam artikelnya BERTopic membandingkan kinerja algoritmanya dengan LDA berdasarkan nilai koherensi dan nilai keberagaman. Hasil dari evaluasi tersebut adalah BERTopic memiliki nilai yang lebih besar dibandingkan LDA. (Grootendorst, 2022)

Semua korpus yang digunakan dalam artikel BERTopic telah melalui prapemrosesan dan berbahasa inggris. Jadi tidak ada pemodelan topik yang

dilakukan menggunakan korpus asli atau belum diproses sama sekali. Karena hal ini maka tidak ada data empiris mengenai bagaimana perbedaan kinerja dari BERTopic pada korpus dengan prapemrosesan dan bahasa yang berbeda. Apalagi mengingat kemampuan dari SBERT yang dapat memproses berbagai bahasa namun kemampuan ini tidak terpakai. Oleh karena itu peneliti berminat untuk melakukan penelitian yang mengisi kekurangan dari penelitian BERTopic pada kinerja pemodelan topik dengan prapemrosesan yang berbeda dan menggunakan korpus berbahasa Indonesia.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah yang akan diselesaikan dalam penelitian ini adalah:

1. Bagaimana kinerja LDA dan BERTopic berdasarkan nilai koherensi, nilai keberagaman, dan waktunya?

1.3 Batasan Masalah

Agar penyusunan dan pembahasan penelitian lebih terarah, diperlukan batasan masalah dalam penelitian ini. Berikut adalah batasan masalah yang digunakan di penelitian ini:

1. Pemodelan topik LDA dilakukan menggunakan pustaka Gensim
2. Pemodelan topik BERTopic dilakukan menggunakan pustaka BERTopic
3. Korpus yang akan digunakan adalah judul dan isi dari artikel berita Tempo pada bulan Desember 2022 berjumlah 7.836 artikel

4. Penelitian dilakukan menggunakan bahasa pemrograman Python versi 3.10

1.4 Tujuan Penelitian

Tujuan penelitian ini adalah membandingkan pemodelan topik dari BERTopic dan LDA pada Gensim dengan korpus berbahasa Indonesia. Kemudian tiap hasil model topik akan dievaluasi kualitasnya berdasarkan nilai koherensi topik, nilai keberagaman topik, dan lama waktu proses pelatihan.

1.5 Manfaat Penelitian

Dengan tercapainya tujuan penelitian ini, diharapkan akan memiliki manfaat penelitian sebagai berikut:

1. Manfaat Akademis
 - a. Hasil dari komparasi metode akan memberikan pemahaman yang lebih baik mengenai kedua metode sehingga peneliti selanjutnya dapat memilih metode yang terbaik untuk penelitiannya.
 - b. Komparasi metode akan menggunakan panjang dokumen dan kombinasi prapemrosesan yang berbeda sehingga peneliti selanjutnya dapat menyesuaikan panjang dokumen dan kombinasi prapemrosesan sesuai kebutuhan.
 - c. Penelitian ini dapat dijadikan rujukan dan pertimbangan untuk penelitian selanjutnya.
2. Manfaat Umum

- a. Penelitian ini dapat digunakan untuk meningkatkan kesadaran mengenai metode untuk mengetahui topik-topik yang ada pada banyak dokumen teks baik berita atau teks dari sosial media.
- b. Penelitian ini dapat digunakan sebagai media pembelajaran pemodelan topik.

1.6 Keaslian Penelitian

Penelitian analisis komparasi pemodelan topik Latent Dirichlet Allocation (LDA) dan BERTopic pada berita berbahasa Indonesia belum pernah dilakukan sebelumnya. Selain itu, penelitian untuk mengetahui pengaruh kombinasi prapemrosesan dan panjang teks berbahasa Indonesia pada pemodelan topik juga belum pernah dilakukan sebelumnya.

1.7 Sistematika Penulisan

Penelitian tugas akhir ini disusun secara sistematis dalam 5 bab, dimulai dari BAB I hingga BAB V. Berikut penjelasan pada masing-masing bab:

BAB I PENDAHULUAN

Bab ini berisi tentang penjelasan latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, keaslian penelitian, dan sistematika penulisan penelitian.

BAB II TINJAUAN PUSTAKA DAN LANDASAN TEORI

Bab ini berisi penjelasan tentang tinjauan pustaka dan landasan teori yang berkaitan dengan penelitian ini.

BAB III METODE PENELITIAN

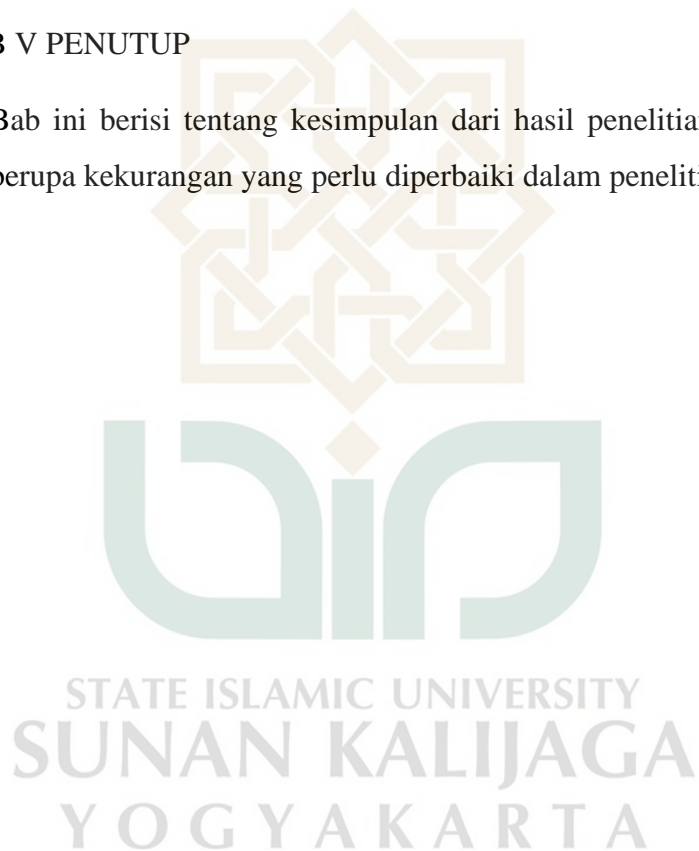
Bab ini berisi penjelasan tentang tahap-tahap atau metode yang digunakan untuk mencapai tujuan dan memperoleh hasil penelitian.

BAB IV HASIL DAN PEMBAHASAN

Bab ini berisi tentang hasil penerapan metode yang digunakan dan analisis hasil penelitian

BAB V PENUTUP

Bab ini berisi tentang kesimpulan dari hasil penelitian dan saran berupa kekurangan yang perlu diperbaiki dalam penelitian ini.



BAB V

PENUTUP

5.1 Kesimpulan

Berdasarkan penelitian yang telah dilakukan dapat disimpulkan analisis komparasi pemodelan topik algoritma LDA menggunakan pustaka gensim dan algoritma BERTopic menggunakan pustaka dengan nama yang sama pada teks berita berbahasa Indonesia berhasil dilakukan. Metode BERTopic memiliki kinerja tinggi pada metrik koherensi dan keberagaman baik tanpa atau dengan prapemrosesan. Sedangkan pada metrik waktu metode LDA memberikan waktu pelatihan dan evaluasi tercepat. Jadi pada metrik koherensi pertimbangan metode terbaik adalah BERTopic, pada metrik waktu pertimbangan metode terbaik adalah LDA, sedangkan pada metrik keberagaman kedua metode dapat dipertimbangkan namun untuk metode LDA harus menggunakan korpus dengan dokumen pendek dan prapemrosesan lemmatisasi, stopwords, dan ngram. Terakhir, model BERTopic dengan prapemrosesan stopwords dan ngram menghasilkan kinerja yang relatif baik pada ketiga metrik dengan proses pembuatan model yang paling mudah.

5.2 Saran

Pada penelitian ini masih terdapat kekurangan. Maka dari itu penulis menyarankan beberapa hal untuk penelitian selanjutnya, di antaranya:

1. Penelitian selanjutnya dapat melakukan analisis pada BERTopic dengan memanfaatkan GPU. Penelitian ini tidak menggunakan GPU karena Gensim tidak memiliki dukungan GPU.
2. Penelitian selanjutnya dapat menggunakan melakukan *guided topic modeling* atau pemodelan topik terbimbing sehingga peneliti membantu model topik dalam penentuan topik dengan memberikan daftar topik yang dipastikan ada pada korpus.
3. Penelitian selanjutnya dapat membangun korpus sendiri. Korpus yang dibangun memiliki banyak topik dan dokumen yang ditentukan. Jadi korpus dapat digunakan sebagai data latih yang sudah memiliki struktur topik

DAFTAR PUSTAKA

- Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the Surprising Behavior of Distance Metrics in High Dimensional Space. In J. Van Den Bussche & V. Vianu (Eds.), *Database Theory—ICDT 2001* (Vol. 1973, pp. 420–434). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-44503-X_27
- Allaoui, M., Kherfi, M. L., & Cheriet, A. (2020). Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study. In A. El Moataz, D. Mammass, A. Mansouri, & F. Nouboud (Eds.), *Image and Signal Processing* (Vol. 12119, pp. 317–325). Springer International Publishing. https://doi.org/10.1007/978-3-030-51935-3_34
- Balakrishnan, V., & Ethel, L.-Y. (2014). Stemming and Lemmatization: A Comparison of Retrieval Performances. *Lecture Notes on Software Engineering*, 2(3), 262–267. <https://doi.org/10.7763/LNSE.2014.V2.134>
- Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When Is “Nearest Neighbor” Meaningful? In C. Beeri & P. Buneman (Eds.), *Database Theory—ICDT’99* (Vol. 1540, pp. 217–235). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-49257-7_15
- Blei, D. M. (2012). Probabilistic Topic Models. *Communications of the ACM*, 55(4), 77–84. <https://doi.org/10.1145/2133806.2133826>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3, 993–1022.
- Churchill, R., & Singh, L. (2022). The Evolution of Topic Modeling. *ACM Computing Surveys*, 54(10s), 1–35. <https://doi.org/10.1145/3507900>
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407. [https://doi.org/10.1002/\(SICI\)1097-4571\(199009\)41:6<391::AID-ASII>3.0.CO;2-9](https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASII>3.0.CO;2-9)
- Denny, M. J., & Spirling, A. (2018). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Political Analysis*, 26(2), 168–189. <https://doi.org/10.1017/pan.2017.44>
- Dieng, A. B., Ruiz, F. J., & Blei, D. M. (2020). Topic Modeling in Embedding Spaces. *Transactions of the Association for*

- Computational Linguistics*, 8, 439–453.
https://doi.org/10.1162/tacl_a_00325
- Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*.
<https://doi.org/10.48550/ARXIV.2203.05794>
- Habibi, M., Priadana, A., & Rifqi Ma'arif, M. (2021). Sentiment Analysis and Topic Modeling of Indonesian Public Conversation about COVID-19 Epidemics on Twitter. *IJID (International Journal on Informatics for Development)*, 10(1), 23–30.
<https://doi.org/10.14421/ijid.2021.2400>
- har07. (2018). *PySastrawi* (1.2) [Python].
<https://github.com/har07/PySastrawi>
- Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2022). Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations. *Organizational Research Methods*, 25(1), 114–146. <https://doi.org/10.1177/1094428120971683>
- Humam, A. I. (2022). *PEMODELAN TOPIK DOKUMEN SKRIPSI MAHASISWA SI TEKNIK INFORMATIKA UIN SUNAN KALIJAGA YOGYAKARTA MENGGUNAKAN ANALISIS LATENT DIRICHLET ALLOCATION* [Thesis (Skripsi), UIN SUNAN KALIJAGA YOGYAKARTA]. <https://digilib.uin-suka.ac.id/id/eprint/50970/>
- Joachims, T. (1997). *A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization*. 97, 143–151.
- Lau, J. H., Newman, D., & Baldwin, T. (2014). Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 530–539.
<https://doi.org/10.3115/v1/E14-1056>
- Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.
- McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11), 205.
<https://doi.org/10.21105/joss.00205>
- McInnes, L., Healy, J., & Melville, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*.
<https://doi.org/10.48550/ARXIV.1802.03426>
- Olston, C., & Najork, M. (2010). Web Crawling. *Foundations and Trends® in Information Retrieval*, 4(3), 175–246.
<https://doi.org/10.1561/1500000017>

- Pandove, D., Goel, S., & Rani, R. (2018). Systematic Review of Clustering High-Dimensional and Large Datasets. *ACM Transactions on Knowledge Discovery from Data*, 12(2), 1–68. <https://doi.org/10.1145/3132088>
- pankaj1707k. (2023, February 3). *Scrapy/docs/intro/overview.rst at 2.11 · scrapy/scrapy*. GitHub. <https://github.com/scrapy/scrapy/blob/2.11/docs/intro/overview.rst>
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages*. <https://doi.org/10.48550/ARXIV.2003.07082>
- R Wahyudi, M. D., Fatwanto, A., Kiftiyani, U., & Galih Wonoseto, M. (2021). Topic Modeling of Online Media News Titles during COVID-19 Emergency Response in Indonesia Using the Latent Dirichlet Allocation (LDA) Algorithm. *Telematika*, 14(2), 101–111. <https://doi.org/10.35671/telematika.v14i2.1225>
- Rachmawati, A. (2022). *Analisis Topic Modeling Pada Jurnal Menggunakan Metode Latent Dirichlet Allocation (LDA)* [Undergraduate Thesis, UIN Sunan Kalijaga Yogyakarta]. <https://digilib.uin-suka.ac.id/id/eprint/54439/>
- Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. <https://doi.org/10.48550/ARXIV.1908.10084>
- Reimers, N., & Gurevych, I. (2020). *Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation*. <https://doi.org/10.48550/ARXIV.2004.09813>
- Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408. <https://doi.org/10.1145/2684822.2685324>
- Santoso, K. R. A. P., Husna, A., Putri, N. W., & Rakhmawati, N. A. (2022). Analisis Topik Tagar Covidindonesia pada Instagram Menggunakan Latent Dirichlet Allocation. *JISKA (Jurnal Informatika Sunan Kalijaga)*, 7(1), 1–9. <https://doi.org/10.14421/jiska.2022.7.1.1-9>
- Shahnaz, F., Berry, M. W., Pauca, V. P., & Plemmons, R. J. (2006). Document clustering using nonnegative matrix factorization. *Information Processing & Management*, 42(2), 373–386. <https://doi.org/10.1016/j.ipm.2004.11.005>
- Steinbach, M., Ertöz, L., & Kumar, V. (2004). The Challenges of Clustering High Dimensional Data. In L. T. Wille (Ed.), *New Directions in Statistical Physics* (pp. 273–309). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-08968-2_16

- Terragni, S., Fersini, E., Galuzzi, B. G., Tropeano, P., & Candelieri, A. (2021). OCTIS: Comparing and Optimizing Topic models is Simple! *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 263–270. <https://doi.org/10.18653/v1/2021.eacl-demos.31>
- Thakur, N., Reimers, N., Daxenberger, J., & Gurevych, I. (2020). *Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks*. <https://doi.org/10.48550/ARXIV.2010.08240>
- VijayGaikwad, S., Chaugule, A., & Patil, P. (2014). Text Mining Methods and Techniques. *International Journal of Computer Applications*, 85(17), 42–45. <https://doi.org/10.5120/14937-3507>

