Jurnal TICOM: Technology of Information and Communication
Volume 13, Nomor 1, September 2024 Hal: 38-44 E-ISSN: 2962-7982

Topic Modeling Tugas Akhir Mahasiswa Menggunakan Metode Latent Dirichlet Allocation Dengan Gibbs Sampling

Zulfikar Rosadi[1*], Achmad Solichin[2]

[1,2] Teknik Informatika, Fakultas Teknologi Informasi, Universitas Budi Luhur, Jakarta, Indonesia
Jl. Ciledug Raya, RT.10/RW.2, Petukangan Utara, Kec. Pesanggrahan, Jakarta Selatan, 12260
Email: [1*]2011501273@student.budiluhur.ac.id, [2]achmad.solichin@budiluhur.ac.id
(* : corresponding author)

Abstrak*—Penyusunan tugas akhir adalah kewajiban bagi seluruh mahasiswa Universitas Budi Luhur di semester akhir untuk mendapatkan gelar sarjana. Salah satu langkah penting dalam persiapan ini adalah memilih topik penelitian yang tepat dan relevan. Untuk mendapatkan topik yang tepat, mahasiswa biasanya membaca laporan tugas akhir dari angkatan sebelumnya, baik di perpustakaan fisik maupun melalui situs *web* repositori kampus. Akan tetapi, situs ini belum memiliki fitur pengelompokan topik, sehingga mahasiswa harus membaca laporan satu per satu untuk menemukan topik yang sesuai dengan minat mereka. Penelitian ini mengusulkan penggunaan metode pemodelan topik, *Latent Dirichlet Allocation* (LDA) dengan *Gibbs Sampling*, untuk mengidentifikasi tren topik dalam laporan tugas akhir secara otomatis. LDA dengan *Gibbs Sampling* dipilih karena efektif dalam menemukan pola topik utama dalam teks yang tidak terstruktur. Hasil penelitian menunjukkan bahwa LDA dapat mengidentifikasi topik tersembunyi dengan nilai *coherence* 0,56 pada iterasi ke-6 dari 10 iterasi yang dijalankan. Topik yang ditemukan meliputi: *Web Service*, *Internet of Things*, Sistem Pakar, Sentimen Analisis, *Data Mining*, dan Kriptografi. Penelitian ini juga menunjukkan bahwa vektorisasi *Bag of Words* efektif dalam LDA, memberikan distribusi topik yang akurat dan membantu mahasiswa dalam menentukan topik penelitian yang relevan dan menarik. Dengan demikian, penggunaan LDA dapat menjadi solusi untuk mempermudah mahasiswa dalam memilih topik tugas akhir yang sesuai dengan minat dan kebutuhan akademis mereka.

Kata Kunci**— Pemodelan Topik, *Latent Dirichlet Allocation*, Nilai *Coherence***

Abstract—***The preparation of a final project is an obligation for all Budi Luhur University students in their final semester to obtain a bachelor's degree. One of the important steps in this preparation is choosing the right and relevant research topic. To get the right topic, students usually read final project reports from previous batches, either in the physical library or through the campus repository website. However, these sites do not have a topic grouping feature, so students have to read the reports one by one to find topics that match their interests. This research proposes the use of a topic modeling method, Latent Dirichlet Allocation (LDA) with Gibbs Sampling, to automatically identify topic trends in final project reports. LDA with Gibbs Sampling was chosen because it is effective in finding main topic patterns in unstructured text. The results show that LDA can identify hidden topics with a coherence value of 0,56 at the 6th iteration out of 10 iterations. The topics found include 'Web Service', 'Internet of Things', 'Expert System', 'Sentiment Analysis', 'Data Mining', and 'Cryptography'. This research also shows that Bag of Words vectorization is effective in LDA, providing accurate topic distribution and assisting students in determining relevant and interesting research topics. Thus, the use of LDA can be a solution to make it easier for students to choose a final project topic that suits their interests and academic needs.***

Keywords— ***Topic Modeling, Latent Dirichlet Allocation, Coherence Score***

I. PENDAHULUAN

Penyusunan tugas akhir adalah kewajiban bagi mahasiswa Universitas Budi Luhur di semester akhir untuk mendapatkan gelar sarjana. Salah satu persiapan penting adalah menentukan topik penelitian. Mahasiswa biasanya membaca laporan tugas akhir mahasiswa angkatan sebelumnya pada perpustakaan fisik atau situs *web* repositori perpustakaan kampus, yaitu https://lib.budiluhur.ac.id/. Akan tetapi, situs ini belum memiliki fitur pengelompokan topik, sehingga mahasiswa harus membaca daftar laporan satu per satu agar mendapatkan topik penelitian yang diinginkan. Penulis melihat ini sebagai suatu cara yang kurang efisien dalam mencari tren topik pada laporan-laporan sebelumnya, sehingga penulis ingin menerapkan metode *topic modeling* (pemodelan topik) yang dapat mendistribusikan tren topik pada berbagai laporan. Berdasarkan riset yang penulis lakukan, terdapat berbagai jenis algoritma dalam membuat model untuk pemodelan topik, salah satunya yang cukup terkenal adalah *Latent Dirichlet Allocation* (LDA).

Penelitian terkait LDA di Indonesia dalam konteks akademik masih terbatas, dengan beberapa contoh penerapan di sektor lain seperti media sosial, artikel berita, dan ulasan produk. Misalnya, penelitian [1] menggunakan LDA pada data *Covid-19* dari Wikipedia, menghasilkan topik kesehatan sebagai pembahasan terbanyak. Penelitian [2] menerapkan LDA pada portal berita *Detikcom*, mengidentifikasi topik konflik, krisis, kemanusiaan, dan korupsi. Penelitian [3] menganalisis sentimen ulasan produk video game di *Steam* dengan menerapkan algoritma *Naïve Bayes* dan LDA, menghasilkan istilah yang sering muncul dari topik-topik yang dominan seperti *story*, *character*, *music*, dan *art*. Meskipun begitu, penulis menemukan sedikitnya beberapa penelitian terkait penerapan metode LDA dalam konteks akademik.

Misalnya seperti penelitian [4] yang menggunakan LDA untuk mencari tren topik penelitian mahasiswa Fakultas Informatika IT Telkom Purwokerto, menghasilkan nilai *coherence* sebesar 0,44 dengan 8 topik terbaik. Lalu penelitian [5] untuk membuat sistem rekomendasi topik tugas akhir berdasarkan transkrip akademik menggunakan LDA dan *Gibbs Sampling*.

Berdasarkan beberapa penelitian terkait penerapan LDA yang sudah disebutkan, penulis memutuskan untuk menggunakan LDA dengan *Gibbs Sampling* pada penelitian ini untuk mendistribusikan topik pada tugas akhir mahasiswa. Akan tetapi, berbeda dengan penelitian sebelumnya yang menggunakan vektorisasi *TF-IDF* ((*term frequency–inverse document frequency*), penelitian ini menggunakan vektorisasi *Bag of Words*, hal ini didasari pada penelitian [6] yang menyatakan bahwa LDA adalah metode pemodelan topik berbasis *Bag of Words*. Penggunaan vektorisasi *Bag of Words* didasari oleh kemampuan metode ini dalam mengabaikan urutan kata, yang sesuai dengan karakteristik LDA. Perbedaan lainnya adalah distribusi kata-topik yang didapatkan pada penelitian sebelumnya tidak dialokasikan kembali pada setiap dokumen pada abstrak tugas akhir yang digunakan, penelitian kali ini mengalokasikan topik yang sudah dilabeli ke setiap dokumen abstrak tugas akhir berdasarkan probabilitas topik tertinggi pada distribusi topik-dokumen. Sehingga dengan ini, mahasiswa dapat mencari topik tugas akhir berdasarkan distribusi topik yang ditampilkan. Hal ini diharapkan dapat meningkatkan efisiensi dalam mencari topik-topik utama dalam laporan tugas akhir. Dengan demikian, penelitian ini tidak hanya memberikan kontribusi pada peningkatan efisiensi dalam penentuan topik penelitian mahasiswa, tetapi juga menawarkan pendekatan baru untuk penerapan LDA dalam konteks akademik.

Terakhir, dengan identifikasi tren topik yang lebih jelas, mahasiswa dapat lebih mudah menemukan referensi yang relevan dan *up-to-date* untuk penelitian mereka. Ini akan membantu dalam penyusunan tugas akhir yang lebih terarah dan berbobot, serta meningkatkan kualitas penelitian di lingkungan akademik Universitas Budi Luhur. Penelitian ini diharapkan dapat menjadi landasan untuk pengembangan lebih lanjut dalam penggunaan teknik pemodelan topik di berbagai institusi pendidikan di Indonesia.

Pada penelitian terdahulu dengan topik Modelling Skripsi Menggunakan Metode Latent yang mempunyai permasalahan Program Studi Sastra Inggris UINSA menghadapi kesulitan dalam mengidentifikasi topik penelitian skripsi secara efisien. Metode *topic modeling* menawarkan solusi untuk identifikasi kluster topik dengan cepat. Penelitian ini menggunakan Metode *Latent Dirichlet Allocation.* Hasil Penelitian yaitu Percobaan sebanyak 5 uji iterasi dengan iterasi berbeda yakni: 100, 500, 1000, dan 5000. Setiap uji iterasi dimasukkan jumlah topik yang berbeda yaitu: 2, 3, 4, dan 5. Hasil *cluster* topik terbaik didapat pada jumlah topik 3 [7].

Penelitian terdahulu lainnya dengan topik Deteksi Topik Tentang Tokoh Publik Politik Menggunakan Latent Dirichlet Allocation (LDA), permasalahan yang dihadapi yaitu Bagaimana cara mengekstraksi *tweet* tentang tokoh politik (Jokowi, Ahok, Anies, Sandiaga, Habib Rizieq) untuk memudahkan pengguna untuk mengetahui suatutopik apakah yang sedang dibicarakan tokoh publik politik. Metode yang digunakan adalahan *Agglomerative Hierarchical Clustering* dan *Latent Dirichlet Allocation* dan Hasil dari penelitian ini yaitu Algoritma A*gglomerative Hierarchical Clustering* dapat digunakan untuk klastering tugas akhir berdasarkan pengujian *silhouette coefficient* dan metode LDA berhasil mendeteksi topik-topik yang relevan berdasarkan *tweet* tentang tokoh publik politik [8].

Penelitian terdahulu lain yang membahas topik Analisis Metoda Latent Dirichlet Allocation untuk Klasifikasi Dokumen Laporan Tugas Akhir Berdasarkan Pemodelan Topik, Metode yang digunakan *Latent Dirichlet Allocation.* Hasil Penelitiannya yaitu Probabilitas kata dalam LDA dipengaruhi oleh jumlah topik dan dokumen, LDA sensitif terhadap kata umum sehingga mengurangi presisi, LDA dapat mengelompokkan dokumen tanpa label [9].

Penelitian lain yang membahsah topik Pemodelan Topik dengan LDA untuk Temu Kembali Informasi dalam Rekomendasi Tugas Akhir dengan permasalahan Perlu adanya suatu sistem yang membantu mahasiswa menentukan topik tugas akhir dengan cepat sesuai kemampuan pada transkrip akademik. Penelitian ini mengusulkan sistem rekomendasi topik tugas akhir berdasarkan kompetensi dalam transkrip akademik. Metode yang digunakan *K-Means* dan *Latent Dirichlet Allocation*. Hasil Penelitian nya adalah Sistem merekomendasikan topik tugas akhir dengan akurasi tinggi, menghasilkan kemiripan 0,44 dan kesepakatan 92%, berdasarkan kecocokan topik rekomendasi dengan transkrip akademik mahasiswa [5].

Penelitian lain dengan topik Klasterisasi Cerita Berbahasa Bali dengan permasalahan Cerita-cerita berbahasa Bali memiliki topik yang beragam namun memuat nilai kearifan lokal yang perlu untuk dilestarikan. Jika cerita-cerita tersebut dapat dikelompokkan berdasarkan topik, akan sangat memudahkan bagi para pembaca dalam memilih bacaan yang diinginkan. Meode yang digunakan *Latent Dirichlet Allocation.* Hasil Penelitian nya adalahan Akurasi tertinggi yang diperoleh dari klasterisasi cerita adalah 62% ketika jumlah kata yang digunakan sebagai representasi dokumen adalah 3000 kata. akurasi klasterisasi sangat dipengaruhi oleh ukuran kesamaan yang digunakan dan jumlah kata sebagai representasi dokumen [10].

II. METODE PENELITIAN

Subjek penelitian ini adalah data abstrak tugas akhir mahasiswa program studi Teknik Informatika dari tahun 2021 sampai 2023 yang diambil dari situs *web* repositori perpustakaan Universitas Budi Luhur. Sedangkan objek penelitian ini adalah penerapan *topic modeling* dari data tersebut menggunakan metode *Latent Dirichlet Allocation* dengan *Gibbs Sampling*.

Pada penelitian ini, penulis menggunakan metode penelitian yang meliputi beberapa proses, yaitu: pengumpulan data penelitian, *preprocessing*, *data extraction*, evaluasi model dengan *coherence score*, dan pemodelan menggunakan LDA *Gibbs Sampling*. Pada tahap awal, penulis melakukan pengumpulan data dengan mengumpulkan sejumlah data abstrak mahasiswa pada *web* repositori perpustakaan. Data yang terkumpul kemudian diproses pada tahap *preprocessing* dengan tujuan meningkatkan kualitas data dan mengubah bentuk data ke dalam format yang lebih mudah dipahami. Tahap *preprocessing* yang digunakan meliputi: *remove punctuation*, *case folding*, *stopword removal*, dan *stemming*. Setelah data melalui tahap *preprocessing*, penulis melakukan tahap *data extraction* dengan tujuan membentuk beberapa format data sehingga sesuai dengan kebutuhan data masukan pada model LDA. Tahap ini terdiri dari: perhitungan jumlah kemunculan kata dengan vektorisasi *Bag of Words*, pembuatan data *vocabulary* dari data kemunculan kata, dan pembuatan data *corpus* dari data *vocabulary*. Selanjutnya, sebelum

dilakukan pemodelan LDA, penulis melakukan evaluasi model dengan *coherence score* untuk menemukan jumlah topik yang optimal dari sejumlah iterasi yang diberikan. Terakhir, penulis melakukan pemodelan LDA *Gibbs Sampling* dengan menggunakan data *corpus* dan jumlah topik yang didapat dari tahap evaluasi model dengan *coherence score*.

A. Data Penelitian

Penulis melakukan pengumpulan data secara manual dengan mengunduh 217 berkas abstrak tugas akhir mahasiswa program studi Teknik Informatika yang terdapat pada situs *web* repositori perpustakaan Universitas Budi Luhur dengan pilihan pencarian sebagai 'Skripsi', kategori program studi diisi dengan 'Teknik Informatika', dan tahun diisi dengan '2021', '2022', dan '2023'. Berkas abstrak yang sudah terkumpul kemudian disimpan ke dalam *file* dengan ekstensi *.xlsx*. Pada sampel data abstrak yang ditunjukkan Tabel 2, ditunjukkan tahun ajaran untuk tiap teks abstrak yang didapatkan.

TABEL 1.
DATA MENTAH ABSTRAK TUGAS AKHIR

Tahun Ajaran	Teks
2021	PT. Primajaya Multisindo adalah perusahaan dibidang penjualan berbagai produk IT seperti laptop, *PC Desktop*, AIO, printer dan *accessories* komputer dengan sistem penjualan bersifat *offline* maupun *online*.
	Di era modern sekarang ini semua kebutuhan sudah bisa didapatkan dengan cara yang mudah dan cepat karena diakibatan oleh teknologi yang dari tahun ketahun makin maju.
2022	Penggajian merupakan proses penting yang melibatkan pengolahan informasi gaji pegawai. Data gaji pegawai mengandung informasi sensitif seperti jumlah gaji, tunjangan, dan bonus yang harus dijaga kerahasiaannya.
	Perkembangan Teknologi yang signifikan memiliki banyak keuntungan terhadap perusahaan-perusahaan baru untuk berkembang dengan cepat.
2023	Kriptografi adalah bidang ilmu untuk menjaga keamanan pesan (*message*). Kriptografi telah banyak diimplementasikan di banyak hal. *Smart card*, Anjungan Tunai Mandiri (ATM), *pay* TV, *Mobile Phone*, dan komputer.
	Sistem yang ada saat ini di SMAS Daya Utama Bekasi masih bersifat manual. Sehingga sering kali ditemukan masalah seperti pencatatan awal peminjaman buku, hingga pengembalian buku. *RESTful API* bisa menjadi sarana agar data dapat dengan mudah diakses dengan bentuk data *JSON*.

B. *Preprocessing*

Setelah data terkumpul, penulis melakukan tahap *preprocessing* dengan tujuan meningkatkan kualitas data sehingga menghasilkan kinerja yang lebih baik [11]. Tahap-tahap dalam *preprocessing* yang penulis lakukan pada penelitian ini adalah sebagai berikut:

a. *Remove Punctuation*

Remove Punctuation merupakan teknik penghilangan tanda baca yang digunakan dalam sebuah teks untuk membedakan antara kalimat dan bagian penyusunnya dan untuk memperjelas maknanya [12].

b. *Case Folding*

Case Folding adalah tahap proses yang mengubah kata menjadi bentuk seragam. Tujuan dari *case folding* adalah mengubah semua kata menjadi huruf kecil agar teks yang diproses berada dalam format yang konsisten [13].

c. *Stopword Removal*

Proses stopword removal bertujuan untuk menghilangkan kata-kata penghubung dan kata-kata yang tidak relevan dalam *dataset* [1]. Pada tahap ini, penulis menghilangkan kata-kata yang dianggap tidak berpengaruh terhadap kalimat, kata-kata yang umum dan sering muncul dalam teks seperti kata depan, kata ganti, dan kata penghubung [14].

d. *Stemming*

Stemming adalah proses penghapusan imbuhan kata untuk mengubah setiap kata ke dalam bentuk dasar. Pada proses ini, *penulis* mengubah kata menjadi kata dasar (*stem*) dengan cara menghilangkan imbuhan kata berupa awalan maupun akhiran [15].

C. *Data Extraction*

Pada tahap ini, penulis menghitung *word frequency* atau kemunculan setiap kata pada seluruh dokumen abstrak menggunakan vektorisasi *Bag of Words*. Setelah perhitungan selesai, penulis membuat *vocabulary* atau kosakata dari kemunculan kata dengan *frequency threshold* atau minimum jumlah kemunculan. Berdasarkan dari jumlah data abstrak yang dikumpulkan, penulis menetapkan angka *frequency threshold* sebesar 10. Artinya, hanya kata yang memiliki minimal jumlah kemunculan 10 kali pada seluruh dokumen yang dapat dimasukkan ke dalam kosakata. Hal ini penulis lakukan guna menghilangkan kata-kata yang tidak memberikan informasi yang berguna karena memiliki jumlah kemunculan yang rendah. Setelah kosakata terbentuk, penulis membuat *corpus* dari kumpulan dokumen abstrak dengan pemfilteran setiap kata yang digunakan berdasarkan kosakata yang sudah penulis buat.

D. *Topic Modeling*

Topic Modeling atau Pemodelan Topik adalah model statistik dan model *unsupervised* yang mampu menemukan topik-topik tersembunyi di dalam korpus dokumen yang besar. Pada intinya, algoritma pemodelan topik mengklasifikasikan kelompok dokumen ke dalam tema-tema yang koheren tanpa campur tangan pengguna. Pemodelan topik dianggap sebagai masalah pengelompokan yang terdiri dari campuran acak topik dalam sebuah dokumen, dan setiap kata dianggap termasuk dalam setidaknya satu topik dari topik-topik tersebut [16]. Hasil dari pemodelan topik adalah sekumpulan topik yang berisi gugus-gugus kata yang telah ter-*clustering* berdasarkan pola dokumen [15]. *Latent Dirichlet Allocation* (LDA) adalah algoritma pemodelan topik yang populer dan digunakan untuk mengekstraksi topik utama dari dokumen. LDA bekerja tanpa pengawasan manusia (*unsupervised*) dan bergantung pada pendekatan *Bag of Words*. Pada LDA, pemilihan jumlah topik yang tepat sangat penting untuk menghasilkan topik yang koheren, karena topik yang terlalu sedikit cenderung menghasilkan kata-kata umum dan tidak spesifik [16].

E. LDA dengan Gibbs *Sampling*

Proses penerapan LDA dengan *Gibbs Sampling* yang dilakukan penulis meliputi: perhitungan variabel yang diperlukan, inisialisasi topik secara acak, dan kemudian menjalankan sejumlah iterasi yang diinginkan di mana pada

setiap iterasi, penulis mengambil sebuah sampel topik untuk setiap kata dalam *corpus*. Setelah iterasi dilakukan, hasil perhitungan dapat digunakan untuk menghitung distribusi laten θ_d dan φ_k. Variabel-variabel yang diperlukan adalah:

n_{d_ik} = jumlah dokumen *d* pada topik *k*
$n_{k,w}$ = jumlah kata *w* yang diberikan ke topik *k*
n_k = jumlah total dari setiap kata yang diberikan ke topik *k*
n_d = jumlah total seluruh dokumen *d* dalam semua topik

Karena prosedur pengambilan sampel *Gibbs* melibatkan pengambilan sampel dari distribusi yang dikondisikan pada semua variabel, penulis menghapus penugasan topik saat ini sebelum membangun distribusi dari persamaan (1).

$$p(z_i = k|w_i, d_i, \alpha, \beta) \propto \frac{n_{d_ik} + \alpha}{\sum_{k'=1}^{K} n_{d_ik'} + K\alpha} \cdot \frac{n_{kw_i} + \beta}{\sum_{w=1}^{W} n_{kw} + W\beta} \quad (1)$$

Dimana:

$z_i = k$ = pemberian topik *k* untuk kata ke-*i*
w_i = kata ke-*i*
d_i = dokumen ke-*i*
α = *hyperparameter* untuk prior *dirichlet* pada distribusi topik-dokumen
β = *hyperparameter* untuk prior *dirichlet* pada distribusi kata-topik
n_{d_ik} = jumlah dokumen *d* pada topik *k*
$\sum_{k'=1}^{K} n_{d_ik'}$ = total seluruh dokumen *d* dalam semua topik *k*
n_{kw_i} = jumlah kata *w* yang muncul pada topik *k*
$\sum_{W=1}^{W} n_{kw}$ = jumlah kata pada topik *k*
K = jumlah topik
W = panjang dari kamus kata

Penghapusan dilakukan dengan mengurangi jumlah yang terkait dengan penugasan saat ini karena urutan penugasan topik tidak mempengaruhi hasil akhir. Setelah itu, penulis menghitung probabilitas dari setiap topik menggunakan persamaan (1). Distribusi ini kemudian diambil sampelnya oleh penulis dan topik yang dipilih diatur dalam suatu himpunan dan jumlah yang sesuai kemudian ditambah.

Dengan kata lain, penugasan topik untuk sebuah kata dihapus sementara dengan mengurangi hitungan terkait pada distribusi topik-dokumen dan topik-kata, sehingga penulis menghitung ulang probabilitas kondisional tanpa pengaruh kata tersebut. Selanjutnya, probabilitas penugasan ulang topik dihitung dengan mempertimbangkan distribusi kata dalam topik dan distribusi topik dalam dokumen, masing-masing dibobot dengan *hyperparameter* α dan β. Setelah probabilitas dihitung, kata tersebut diberi penugasan topik baru berdasarkan sampel dari distribusi probabilitas yang telah dihitung, dan hitungan terkait ditambahkan kembali untuk melanjutkan ke iterasi berikutnya.

F. Hasil Evaluasi Model dengan Coherence Score

Evaluasi model adalah bagian penting dari setiap proses pengembangan sistem karena memungkinkan untuk menilai, menganalisis, dan memahami seberapa akurat atau kesesuaian hasil yang telah dicapai oleh sistem yang dirancang. Evaluasi model pada LDA tidak menggunakan pengujian akurasi serta konsep 'benar' dan 'salah' tidak sejelas dalam model *supervised learning*, hal ini karena LDA adalah metode *unsupervised learning* yang tidak memiliki label atau *ground truth* yang dapat digunakan untuk mengukur akurasi. LDA juga tidak berusaha untuk mengklasifikasikan data berdasarkan label yang sudah ada, sehingga pengukuran seperti akurasi yang bergantung pada perbandingan prediksi dengan label tidak relevan untuk jenis model ini. Untuk menilai kualitas topik yang dihasilkan oleh LDA, penulis melakukan pengujian evaluasi untuk menemukan jumlah topik yang optimal pada data *corpus* dengan mengukur nilai *coherence* pada setiap iterasi jumlah topik yang diuji dan dijalankan oleh model LDA.

Coherence Score adalah penilaian kualitas sebuah topik dengan mengukur seberapa mirip makna kata-kata dalam topik tersebut. Semakin tinggi nilai *coherence*, semakin baik pemahaman manusia terhadap topik tersebut, yang pada akhirnya meningkatkan kualitas topik tersebut. Ada empat tahapan dalam menentukan nilai *coherence*, yaitu *segmentation* yang merupakan proses pengelompokkan data menjadi pasangan kata, *probability estimation* yang menghitung kemungkinan kemunculan kata atau pasangan kata tersebut, *confirmation measure* yang menunjukkan seberapa kuat satu kumpulan kata mendukung kumpulan lainnya, dan *aggregation* yang digunakan untuk melihat skor koherensi keseluruhan. Dalam menentukan jumlah topik, perlu diperhatikan nilai *coherence* terbaik. Tahapan untuk menentukan nilai *coherence* ditunjukkan pada persamaan (2).

$$C = S \times M \times P \times \Sigma \quad (2)$$

Dimana:
C = *Coherence Score*
S = *Segmentation*
M = *Confirmation Measure*
P = *Probability Estimation*
Σ = *Aggregation*

Jumlah iterasi yang penulis berikan adalah sebanyak 10 iterasi. Tiap iterasi dijalankan pada pemodelan LDA untuk didapatkan 20 kata tertinggi. Nilai *coherence* kemudian dihitung dengan menggunakan kata tertinggi tersebut. Hasil perhitungan ini dapat dilihat pada Tabel 3.

TABEL 2.
NILAI COHERENCE

Num Topic	*Coherence Score*
1	0,32
2	0,36
3	0,38
4	0,51
5	0,44
6	0,56
7	0,52
8	0,55
9	0,54
10	0,50

Dapat dilihat pada Tabel 3 bahwa nilai *coherence* tertinggi terdapat pada iterasi topik ke-6, yaitu senilai 0,56. Dapat penulis simpulkan berdasarkan nilai-nilai yang ada bahwa jumlah 6 topik menjadi distribusi topik yang mempunyai

interpretasi atau makna topik yang paling baik untuk dipahami oleh manusia. Oleh karena itu, jumlah topik pada iterasi ke-6 akan penulis gunakan kembali dalam proses pemodelan dan untuk mendapatkan distribusi topik-dokumen dan distribusi kata-topik dari 6 jumlah topik yang digunakan.

III. HASIL DAN PEMBAHASAN

Sebelum dilakukan tahap pemodelan, penulis mencari terlebih dahulu jumlah topik yang optimal dengan menghitung nilai *coherence* pada setiap iterasi topik yang dijalankan. Iterasi topik yang memiliki nilai *coherence* tertinggi kemudian digunakan oleh penulis pada model LDA sebagai tahap akhir dari penerapan algoritma.

A. Algoritma LDA *Gibbs Sampling*

Algoritma ini menunjukkan proses pemodelan LDA *Gibbs Sampling* dan mencakup semua langkah-langkahnya. Berikut ini adalah penjabarannya yang ditunjukkan pada Gambar 1.

```
1   mulai
2   lakukan proses inisialisasi jumlah dokumen
3   lakukan proses inisialisasi topik awal (Z)
4   lakukan proses hitung jumlah kata dengan topik-k pada dokumen-d (n_dk)
5   lakukan proses hitung jumlah total kata-w dengan topik-k (n_kw)
6   lakukan proses hitung jumlah total semua kata dengan topik-k (n_k)
7   lakukan proses hitung jumlah total Semua topik dengan dokumen-d (n_d)
8   lakukan proses inisialisasi array dengan panjang num_topic (topic_list)
9   for _ in range(num_iter)
10      for item in corpus
11          for item in len(doc)
12              lakukan proses inisialisasi kata dan topik saat ini
13              lakukan proses kurangi 1 jumlah kemunculan topik saat ini pada n_dk,n_kw,n_k,n_d
14              lakukan proses hitung probabilitas tiap topik saat ini dengan rumus LDA
15              lakukan proses ambil sampel topik baru dari hasil probabilitas
16              lakukan proses tambahkan 1 jumlah kemunculan topik baru pada n_dk,n_kw,n_k,n_d
17          endfor
18      endfor
19  endfor
20  return n_dk,n_kw,n_k,n_d
21  selesai
```

Gambar 1. Algoritma LDA *Gibbs Sampling*

B. Algoritma Proses Pencarian Jumlah Topik Optimal

Algoritma ini menunjukkan proses pencarian jumlah topik optimal dan mencakup semua langkah-langkahnya. Berikut ini adalah penjabarannya yang ditunjukkan pada Gambar 2.

```
1   mulai
2   select tabel database word_corpus, word_vocabulary
3   lakukan proses inisialisasi panjang kosakata, kamus kata atau dictionary
4   lakukan proses inisialisasi jumlah topik
5   for item in topics_range
6       lakukan predefined proses LDA Gibbs Sampling
7       lakukan proses probabilitas distribusi kata pada topik
8       for item in num_topic
9           lakukan proses ambil 20 kata teratas dari distribusi kata pada topik
10          lakukan proses masukkan 20 kata teratas ke dalam variabel array top_words
11      endfor
12      if top_words != None
13          lakukan proses hitung nilai coherence
14          insert ke tabel database coherence_score
15      else
16          insert ke tabel database coherence_score = 0.0
17      endif
18  endfor
19  select tabel database coherence_score
20  lakukan proses ambil nilai coherence_score tertinggi
21  selesai
```

Gambar 2. Algoritma Proses Pencarian Jumlah Topik Optimal

C. Algoritma Proses Pelabelan Topik

Algoritma ini menunjukkan proses pelabelan topik dan mencakup semua langkah-langkahnya. Berikut ini adalah penjabarannya yang ditunjukkan pada Gambar 3.

```
1   mulai
2   input nama label pada semua form topik
3   lakukan proses ekstrak semua input form
4   update tabel database lda_topic dengan label
5   select tabel database lda_topic
6   for item in lda_topic
7       lakukan proses ekstrak dan inisialisasi variabel nama topik sebelum (topic_idx) dan sesudah diberikan pelabelan (label)
8       select tabel database lda_document
9       count = 0
10      for item in docs
11          lakukan proses ekstrak dan Inisialisasi variabel nama topik pada distribusi topik - dokumen (topic_keys)
12          if topic_keys[0] == topic_idx
13              count += 1
14          endif
15      endfor
16      update and insert tabel database lda_topic_count dengan label dan count
17  endfor
18  selesai
```

Gambar 3. Algoritma Proses Pelabelan Topik

D. Hasil Distribusi dengan LDA Gibbs Sampling

Pemodelan topik menggunakan LDA dengan *Gibbs Sampling* menghasilkan 217 kelompok distribusi topik-dokumen dan 6 kelompok distribusi kata-topik. Masing-masing kelompok mengandung *term* beserta bobotnya yang ditunjukkan pada Tabel 4 dan Tabel 5. Setiap *term* memiliki bobot yang merepresentasikan probabilitas *term* tersebut pada setiap kelompok. Pada distribusi topik-dokumen, ditunjukkan sampel dari keseluruhan dokumen yang ada dengan sejumlah topik pada kolom 'Topik' sebagai *term* yang mengandung bobot dari tiap kelompok dokumen. Sedangkan pada distribusi kata-topik, ditunjukkan keseluruhan jumlah topik dengan sejumlah kata pada kolom 'Kata' sebagai *term* yang mengandung bobot dari tiap kelompok topik.

TABEL 3.
DISTRIBUSI TOPIK-DOKUMEN

Dokumen	Topik	Bobot
Dokumen 1	topic_6	0,96
	topic_1	0,03
Dokumen 2	topic_5	1
Dokumen 8	topic_4	0,96
	topic_2	0,02
	topic_3	0,01
Dokumen 161	topic_4	1

TABEL 4.
DISTRIBUSI KATA-TOPIK

Topik	Kata	Bobot
Topik 1	web	0,059
	aplikasi	0,041
	sistem	0,038
	service	0,031
	api	0,028
Topik 2	sensor	0,041
	sistem	0,029
	air	0,022
	deteksi	0,019
	alat	0,018
Topik 3	sakit	0,018
	sistem	0,060
	pakar	0,044
	metode	0,032
	teliti	0,022
Topik 4	sentimen	0,032
	data	0,028
	indonesia	0,026
	hasil	0,024
	masyarakat	0,023
Topik 5	data	0,023
	jual	0,026
	hasil	0,025
	metode	0,025
	nilai	0,024
Topik 6	data	0,075
	aman	0,058
	file	0,049
	enkripsi	0,034
	kriptografi	0,024

Distribusi yang dihasilkan menunjukkan bahwa setiap dokumen memiliki dominasi topik tertentu dengan bobot yang signifikan. Pada Tabel 4, dokumen 1 mengandung topik 6 dengan bobot sebesar 0,96 dan topik 1 dengan bobot yang lebih kecil, yaitu sebesar 0,03. Hal ini menunjukkan bahwa dalam dokumen 1, lebih banyak kata yang di asosiasikan dengan topik 6 daripada dengan topik 1. Pada Tabel 5, bobot tertinggi dari semua kata yang dimiliki pada topik 6 adalah 'data' dengan bobot sebesar 0,075. Ini menunjukkan bahwa kata 'data' menjadi kata dengan jumlah kata terbanyak yang berasosiasi dengan topik 6 daripada kata lainnya dalam seluruh dokumen. Oleh karena itu, jika melihat kumpulan kata yang terdapat pada topik 6 yang mengandung kata 'data', 'aman', 'file', 'enkripsi', dan 'kriptografi', hal ini mengindikasikan bahwa dokumen 1 kemungkinan besar membahas isu-isu terkait keamanan data dan kriptografi.

E. Hasil Pelabelan

Setelah didapatkan distribusi topik-dokumen dan distribusi kata-topik, diperlukan pelabelan nama topik pada distribusi kata-topik dengan mengidentifikasi kata-kata yang memiliki probabilitas tinggi dalam setiap topik yang dihasilkan oleh model. Setiap topik kemudian diberi label yang mewakili tema atau kategori umum dari kata-kata tersebut. Misalnya, jika sebuah topik didominasi oleh kata-kata seperti 'data', 'aman', 'file' dan 'enkripsi', maka topik tersebut dapat diberi label mengenai Kriptografi. Pelabelan ini bertujuan untuk memudahkan interpretasi topik pada distribusi topik-dokumen, sehingga mempermudah dalam mengetahui topik apa saja yang terdapat pada setiap dokumen. Topik yang diambil pada distribusi topik-dokumen hanyalah topik dengan bobot tertinggi agar tiap topik menjadi lebih informatif dalam menggambarkan isi dokumen.

Proses ini melibatkan pemahaman konteks dan makna kata dalam setiap topik untuk memastikan label yang diberikan tepat dan informatif. Untuk memastikan akurasi dan relevansi label yang diberikan, penulis melibatkan seorang pakar dalam bidang tersebut. Pakar yang dipilih untuk penelitian ini adalah seorang Kepala Program Studi Teknik Informatika, yang memiliki pemahaman mendalam tentang terminologi dan konteks dalam data peneltian yang digunakan, yaitu data abstrak tugas akhir mahasiswa program studi Teknik Informatika. Tabel 6 ini menunjukkan distribusi kata-topik yang sudah diberikan label oleh seorang pakar beserta jumlah dokumen yang memuat topik tersebut. Label topik yang diberikan meliputi: topik '*Web Service*' dengan jumlah 29 dokumen, topik '*Internet of Things*' dengan jumlah 51 dokumen, topik 'Sistem Pakar' dengan jumlah 20 dokumen, topik 'Sentimen Analisis' dengan jumlah 40 dokumen, topik '*Data Mining*' dengan jumlah 27 dokumen, dan topik 'Kriptografi' dengan jumlah 50 dokumen.

TABEL 5.
HASIL PELABELAN DISTRIBUSI KATA-TOPIK

Topik	Kata	Jumlah Dokumen
Web Service	web, aplikasi, sistem, service, api	29
Internet of Things	sensor, sistem, air, deteksi, alat	51
Sistem Pakar	sakit, sistem, pakar, metode, teliti	20
Sentimen Analisis	sentimen, data, indonesia, hasil, masyarakat	40

Data Mining	data, jual, hasil, metode, nilai	27
Kriptografi	data, aman, file, enkripsi, kriptografi	50

Berdasarkan hasil pelabelan, terlihat bahwa topik *Internet of Things* menjadi topik yang paling banyak muncul, dengan 51 dokumen yang memuatnya. Topik ini dikaitkan dengan kata-kata seperti 'sensor', 'sistem', 'air', 'deteksi', dan 'alat', yang menunjukkan bahwa penelitian-penelitian ini berfokus pada teknologi deteksi atau monitoring air**,** seperti sistem yang menggunakan sensor untuk mendeteksi atau memonitor kualitas air, aliran air, atau kebocoran. Sebaliknya, topik Sistem Pakar menjadi topik yang paling sedikit muncul, hanya berjumlah 20 dokumen yang memuatnya. Kata-kata seperti 'sakit', 'sistem', 'pakar', 'metode', dan 'teliti' mendominasi topik ini, yang menunjukkan bahwa penelitian-penelitian ini berkaitan dengan sistem pakar di bidang kesehatan, suatu sistem yang menggunakan metode tertentu untuk mendiagnosis atau menganalisis kondisi kesehatan berdasarkan pengetahuan dari pakar atau dokter.

IV. KESIMPULAN

Dari proses uji coba dan analisis yang penulis lakukan, hasil penelitian menunjukkan bahwa pengaplikasian pemodelan topik dengan metode *Latent Dirichlet Allocation* dengan *Gibbs Sampling* dapat diterapkan pada 217 data abstrak tugas akhir mahasiswa Universitas Budi Luhur program studi Teknik Informatika. Proses penelitian mencakup: pengumpulan data, *preprocessing*, *data extraction*, penerapan algoritma, evaluasi model, dan pelabelan topik. Dari hasil implementasi metode, penulis menemukan bahwa metode LDA mampu mengidentifikasi topik-topik yang tersembunyi dalam kumpulan dokumen abstrak, dimana hasil ini ditunjukkan oleh nilai *coherence* senilai 0,56 pada iterasi topik ke-6 dari 10 iterasi yang dijalankan.

Hasil dari pemodelan LDA menggunakan nilai *coherence* tertinggi yang penulis dapatkan pada iterasi topik ke-6 adalah distribusi topik-dokumen dan distribusi kata-topik dengan topik yang sudah diberikan label oleh pakar. Topik ke-1 dikategorikan sebagai topik mengenai *Web Service*, topik ke-2 dikategorikan sebagai topik *Internet of Things*, topik ke-3 dikategorikan sebagai topik Sistem Pakar, topik ke-4 dikategorikan sebagai topik Sentimen Analisis, topik ke-5 dikategorikan sebagai topik *Data Mining*, dan topik ke-6 dikategorikan sebagai topik 'Kriptografi'.

REFERENSI

[1] W. A. N. Sari and H. D. Purnomo, "Topic Modeling Using the Latent Dirichlet Allocation Method on Wikipedia Pandemic Covid-19 Data in Indonesia," *J. Tek. Inform.*, vol. 3, no. 5, pp. 1223–1230, 2022.

[2] Y. Matira, Junaidi, and I. Setiawan, "Pemodelan Topik pada Judul Berita Online Detikcom Menggunakan Latent Dirichlet Allocation," *Estimasi J. Stat. Its Appl.*, vol. 4, no. 1, pp. 53–63, 2023.

[3] M. Y. Febrianta, S. Widiyanesti, and S. R. Ramadhan, "Analisis Ulasan Indie Video Game Lokal pada Steam Menggunakan Analisis Sentimen dan Pemodelan Topik Berbasis Latent Dirichlet Allocation," *J. Animat. Games Stud.*, vol. 7, no. 2, pp. 117–144, 2021.

[4] N. Novarian, S. Khomsah, and A. B. Arifa, "Topic Modeling Tugas Akhir Mahasiswa Fakultas Informatika Institut Teknologi Telkom Purwokerto Menggunakan Metode Latent Dirichlet Allocation Nathanael," *LEDGER J. Inform. Inf. Technol.*, vol. 2, no. 1, pp. 14–27, 2023.

[5] D. Purwitasari, A. Muflichah, N. A. Hasanah, and A. Z. Arifin, "Pemodelan Topik dengan LDA untuk Temu Kembali Informasi dalam Rekomendasi Tugas Akhir," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 3, pp. 421–428, 2021.

[6] W. Qiu and Z. Pan, "Polarimetric Synthetic Aperture Radar Ship Potential Area Extraction Based on Neighborhood Semantic Differences of the Latent Dirichlet Allocation Bag-of-Words Topic Model," *Remote Sens.*, vol. 15, no. 23, pp. 1–26, 2023.

[7] A. I. Alfanzar, Khalid, and I. S. Rozas, "Topic Modelling Skripsi Menggunakan Metode Latent," *JSiI (Jurnal Sist. Informasi)*, vol. 7, no. 1, pp. 7–13, 2020.

[8] F. N. Hikmah, S. Basuki, and Y. Azhar, "Deteksi Topik Tentang Tokoh Publik Politik Menggunakan Latent Dirichlet Allocation (LDA)," *J. Repos.*, vol. 2, no. 4, pp. 415–426, 2020.

[9] U. T. Setijohatmo, S. Rachmat, T. Susilawati, Y. Rahman, and K. Kunci, "Analisis Metoda Latent Dirichlet Allocation untuk Klasifikasi Dokumen Laporan Tugas Akhir Berdasarkan Pemodelan Topik," *Pros. 11th Ind. Res. Work. Natl. Semin.*, vol. 11, no. 1, pp. 402–408, 2020.

[10] N. A. Sanjaya ER, "Implementasi Latent Dirichlet Allocation (LDA) untuk Klasterisasi Cerita Berbahasa Bali," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 1, pp. 127–134, 2021.

[11] E. P. Laksono, A. Basuki, and F. A. Bachtiar, "Optimasi Nilai K pada Algoritma KNN untuk Klasifikasi Spam dan Ham Email," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 2, pp. 377–383, 2020.

[12] M. Lestandy, A. Abdurrahim, and L. Syafa'ah, "Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Recurrent," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 10, pp. 802–808, 2021.

[13] M. U. Albab, Y. Karuniawati, and M. N. Fawaiq, "Optimization of the Stemming Technique on Text preprocessing President 3 Periods Topic," *J. Transform.*, vol. 20, no. 2, pp. 1–10, 2023.

[14] S. S. Tandiapa and G. C. Rorimpandey, "Analisis Sentimen Ulasan Pengguna Pada Aplikasi Threads Dengan Metode Lexicon Based dan Naive Bayes Classifier," *J. Cahaya Mandalika*, vol. 3, no. 1, pp. 339–353, 2024.

[15] S. Roiqoh, B. Zaman, and K. Kartono, "Analisis Sentimen Berbasis Aspek Ulasan Aplikasi Mobile JKN dengan Lexicon Based dan Naïve Bayes," *J. Media Inform. Budidarma*, vol. 7, no. 3, pp. 1582–1592, 2023.

[16] M. B. Mutanga and A. Abayomi, "Tweeting on COVID-19 pandemic in South Africa: LDA-based topic modelling approach," *African J. Sci. Technol. Innov. Dev.*, vol. 14, no. 1, pp. 163–172, 2020.