

Analisis Trend Topik Penelitian Tesis Pada Program Studi Magister Ilmu Komputer Universitas Budi Luhur Menggunakan Metode *Latent Dirichlet Allocation* (LDA)

Luhur Bayuaji¹, Arief Wahyudi²

^{1,2}Program Studi Magister Ilmu Komputer, Universitas Budi Luhur, Indonesia

Article Info

Article history:

Received Nov 29, 2023

Revised Apr 17, 2024

Accepted Apr 29, 2024

Keywords:

Trend

Tesis

Latent Dirichlet Allocation

ABSTRACT

Every year, thousands of studies are conducted by researchers from various institutions and places, focusing on various fields and topics. This also applies to thesis research conducted by students of the Master of Computer Science Program at the Faculty of Information Technology, Universitas Budi Luhur. Given the significant amount of research over time at Budi Luhur University's Master of Computer Science Program, it has become increasingly difficult to effectively understand research trends and focus. The purpose of this study is to identify trending thesis research topics in the Master of Computer Science Program at Universitas Budi Luhur. The data used in this research includes thesis research titles conducted from 2016 to 2021. The method used in this research is *Latent Dirichlet Allocation* (LDA). The results of the study produced the best pass value at 28 and the best number of topics was 5 topics. LDA modeling produces 5 research topics that are trending in the period 2016 to 2021, namely sentiment analysis, data analysis, prediction analysis, decision support systems and machine learning.

Copyright © 2024 Universitas Indraprasta PGRI.
All rights reserved.

Corresponding Author:

Arief Wahyudi,

Program Studi Magister Ilmu Komputer,

Universitas Budi Luhur,

Jl. Ciledug Raya, Rt.10/Rw 02, Petukangan Utara, Kec. Pesanggrahan, Jakarta Selatan.

Email: 2111600249@student.budiluhur.ac.id

1. PENDAHULUAN

Data mining atau penambangan data adalah sebuah proses pencarian secara otomatis informasi yang berguna dari sekumpulan data yang berukuran besar [1]. Tidak semua pekerjaan penemuan informasi dianggap sebagai *data mining*. Contohnya *query* atau pencarian informasi di internet dengan mengetikkan kata kunci tertentu di mesin pencari. *Text mining* adalah proses intensif pengetahuan di mana seperangkat alat analisis digunakan untuk berinteraksi dengan kumpulan dokumen dari waktu ke waktu [2]. *Text mining* bertujuan untuk mendapatkan informasi bermanfaat dari sumber data melalui pengenalan dan eksplorasi pola yang menarik. Meskipun memiliki tujuan yang sama namun terdapat perbedaan antara *data mining* dan *text mining* yaitu pada bentuk data yang diolah. Jika pada *data mining* data yang akan diproses merupakan data yang sudah terstruktur sedangkan pada *text mining* data yang akan dianalisa merupakan data yang tidak terstruktur sehingga harus diubah terlebih dahulu menjadi data terstruktur [3].

Pemodelan topik merupakan metode yang *text mining* yang cukup efektif untuk menemukan data teks tersembunyi dan menemukan hubungan antar teks (Jelodar dkk). Metode yang paling banyak digunakan dalam pemodelan topik adalah *Latent dirichlet Allocation* (LDA) [4].

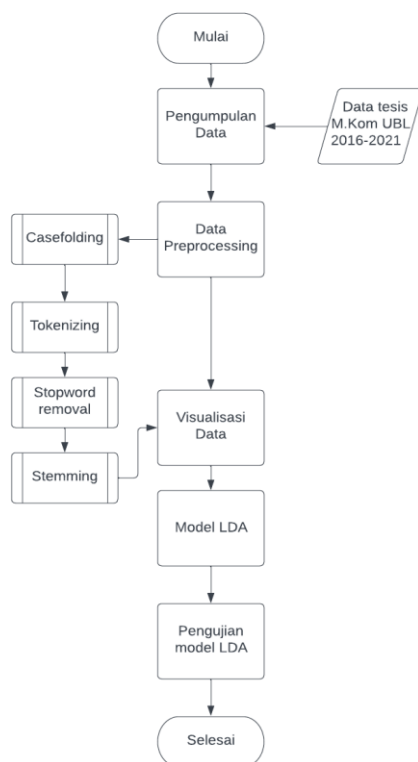
Penggunaan LDA dalam pemodelan topik telah dilakukan untuk *clustering* teks dikombinasikan dengan TFxICF berbasis frasa [5], analisis informasi publik media sosial [6], pemodelan topik jurnal [7], penentuan topik konsultasi pada portal *telemedicine* [8], analisis topik tagar pada media sosial [9].

Analisis trend topik secara umum telah banyak dilakukan terutama untuk topik-topik di media sosial. Namun belum banyak penelitian untuk data yang tidak bersumber dari media sosial, salah satunya yaitu data judul penelitian tesis. Analisis trend topik penelitian dapat memberikan manfaat untuk memetakan bidang-bidang penelitian apa saja yang telah dilakukan dan bidang apa saja yang masih membutuhkan penelitian lebih lanjut.

Penelitian ini bertujuan untuk menganalisis topik-topik yang diangkat menjadi penelitian tesis di Fakultas Ilmu Komputer Universitas Budi Luhur selama kurun waktu 2016 sampai dengan 2021. Adapun metode yang akan digunakan yaitu *Latent Dirichlet Allocation* (LDA).

2. METODE

Metode yang digunakan dalam penelitian ini dapat dilihat pada Gambar 1 berikut.



Gambar 1. Langkah-langkah penelitian

Langkah-langkah dalam penelitian ini diawali dengan pengumpulan data, data preprocessing, visualisasi data, pemodelan menggunakan LDA dan pengujian model LDA.

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Langkah pertama yang dilakukan dalam penelitian ini adalah pengumpulan data. Pengumpulan data dilakukan dengan mengajukan permohonan data kepada lembaga yang berwenang mengelola pengarsipan penelitian tesis di lingkungan Universitas Budi Luhur yaitu Perpustakaan Universitas Budi Luhur. Dari pengajuan tersebut diperoleh data judul dan abstrak tesis mahasiswa Magister Ilmu Komputer Universitas Budi Luhur pada periode 2016 sampai dengan 2021 sebanyak 237 judul yang kemudian diambil sebanyak 211 judul dikarenakan terdapat file yang tidak lengkap yaitu data yang hanya memuat abstrak saja tanpa judul.

Cuplikan data yang digunakan dapat dilihat pada gambar berikut.

	NIM	JUDUL
1	1611601475	Analisis dan Perancangan Data Warehouse Sebagai Alat Integrasi Data Untuk Mendukung Strategi Pengambilan Keputusan: Studi Kasus Pada PT. Herlina Indah (Enesis Group)
2	1611601962	Knowledge Management System Menggunakan Metode Kano Dan SECI Pada Unit Kemahasiswaan STMIK Widuri Jakarta
3	1711600088	Perbandingan Metode MPE dan Waspas Dalam Menentukan Guru Teladan: Studi Kasus di PKBM Surya Harapan Kota Tangerang
4	1711600393	Modifikasi Framework Togaf dan Penyelarasan Multidimensi Business Architecture Untuk Peningkatan Efektifitas Enterprise Architecture dan Analisis Beban Kerja Pegawai
5	1711600823	Peningkatan Akurasi Prediksi Pengadaan Bahan Baku Produksi Dengan Menggunakan Metode Neural Network: Studi Kasus Pada PT. International Paint Indonesia
6	1711600963	Peramalan Persediaan Sparepart Menggunakan Metode Double Exponential Smoothing Pada PT. Mayora Indah Tbk
7	1711601169	Knowledge Management System Dengan Model SECI, Studi Kasus: Politeknik LP3I Kampus Pasar Minggu
8	1711601276	Penentuan Kenaikan Jabatan Dengan Metode Analytical Hierarchy Process dan Profile Matching, Studi Kasus: Kementerian Agraria dan Tata Ruang/Badan Pertahanan Nasional
9	1711601284	Perancangan Knowledge Management System Pada Bagian Packing PT.Surya Toto Indonesia Tbk
10	1711601292	Pengoptimalan Performa Database Pada Proses Transformasi Data Pada SQL Server

Gambar 2. Cuplikan data.

3.2 Pre-processing Data

Untuk melakukan analisis trend topik penelitian, data yang telah dikumpulkan tidak dapat langsung digunakan dalam pemodelan topik. Data akan melalui tahap *preprocessing* terlebih dahulu. Adapun langkah-langkah dalam *preprocessing* data diantaranya: *case folding*, *tokenizing*, dan *stopword removal*.

A. Case folding

Proses ini akan mengubah teks dengan huruf besar menjadi huruf kecil. Gambar berikut menunjukkan ilustrasi dari proses *case folding*.

Teks sebelum case folding :
 Analisis dan Perancangan Data Warehouse Sebagai Alat Integrasi Data untuk Mendukung Strategi Pengambilan Keputusan: Studi Kasus Pada PT. Herlina Indah (Enesis Group)

Teks setelah case folding:
 analisis dan perancangan data warehouse sebagai alat integrasi data untuk mendukung strategi pengambilan keputusan: studi kasus pada pt. herlina indah (enesis group)

Gambar 3. Ilustrasi case folding

Pada gambar 3 terlihat teks sebelum case folding masih mengandung huruf besar sedangkan setelah mengalami proses *case folding*, teks menjadi huruf kecil.

B. Tokenizing

Pada proses ini, teks akan dibagi kedalam *token* atau kata dengan menggunakan spasi sebagai pemisah sehingga setiap kata dapat berdiri sendiri. Proses *tokenizing* juga akan menghilangkan tanda baca yang terdapat pada teks. Gambar berikut menunjukkan ilustrasi dari proses *tokenizing*.

Teks judul :
 Analisis dan Perancangan Data Warehouse Sebagai Alat Integrasi Data untuk Mendukung Strategi Pengambilan Keputusan: Studi Kasus Pada PT. Herlina Indah (Enesis Group)

Teks judul setelah tokenizing:
 {'analisis', 'dan', 'perancangan', 'data', 'warehouse', 'sebagai', 'alat', 'integrasi', 'data', 'untuk', 'mendukung', 'strategi', 'pengambilan', 'keputusan', 'studi', 'kasus', 'pada', 'pt', 'herlina', 'indah', 'enesis', 'group'}

Gambar 4. Ilustrasi proses tokenizing

C. Stopword Removal

Proses ini akan menghilangkan kata-kata non-topikal yang tidak dianggap penting seperti, “ini”, “adalah”, “atau”, “yang” dan lainnya. Proses *stopword removal* dilakukan dengan memanfaatkan *library*

stopwords yang sudah tersedia pada modul NLTK *Python*. Gambar {} menunjukkan ilustrasi proses *stopword removal*.

Teks judul hasil tokenizing:

{'analisis', 'dan', 'perancangan', 'data', 'warehouse', 'sebagai', 'alat', 'integrasi', 'data', 'untuk', 'mendukung', 'strategi', 'pengambilan', 'keputusan', 'studi', 'kasus', 'pada', 'pt', 'herlina', 'indah', 'enesis', 'group'}

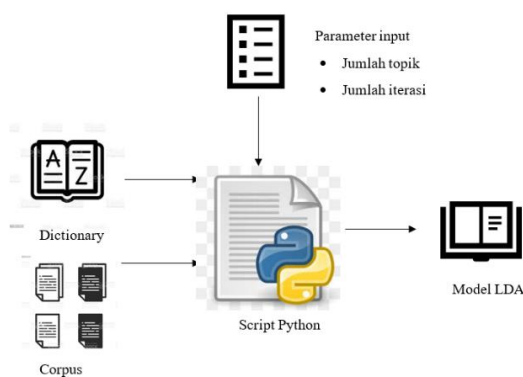
Teks judul setelah *stopword removal*:

['analisis', 'perancangan', 'data', 'warehouse', 'alat', 'integrasi', 'data', 'mendukung', 'strategi', 'pengambilan', 'keputusan', 'studi', 'herlina', 'indah', 'enesis', 'group']

Gambar 5. Ilustrasi proses *stopword removal*

3.3 Pemodelan *Latent Dirichlet Allocation*

Gambar berikut menunjukkan alur pemodelan LDA.



Gambar 6. Alur pemodelan LDA

Pada pemodelan topik LDA, *dictionary* (kumpulan kata-kata unik) dan *corpus* (kumpulan dokumen) digunakan secara berhubungan. *Dictionary* digunakan untuk membangun representasi numerik dari dokumen-dokumen dalam *corpus*. Representasi inilah yang digunakan untuk menemukan topik tersembunyi dari dokumen.

Cuplikan *dictionary* dari data judul tesis yang telah mengalami *preprocessing* dapat dilihat pada gambar berikut.

Index	Word
0	jakarta
1	kano
2	kemahasiswaan
3	knowledge_management
4	metode
5	saci
6	stmik
7	system
8	unit
9	widuri
10	adaptive
11	aural
12	berbasis
13	kineshtetic
14	klasifikasi
15	learning

Gambar 7. Cuplikan dictionary

Adapun *corpus* yang terbentuk dari data judul tesis dapat dilihat pada tabel berikut.

Tabel 1. cuplikan *Corpus*

Corpus: [(0, 1), (1, 1), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1), (7, 1), (8, 1), (9, 1)]

Tabel 1 memperlihatkan cuplikan *corpus* yang terbentuk dari data teks judul penelitian tesis. *Corpus* tersebut mengandung pasangan (indeks_kata, frekuensi_kata) untuk setiap dokumen dalam *corpus*. Indeks_kata merujuk pada kata dan indeks yang terdapat pada *dictionary*. Pada contoh *corpus* di atas memiliki pengertian dokumen tersebut memiliki 10 kata berbeda mulai dari kata dengan indeks 1 sampai indeks 9.

(0, 1) = indeks kata 0, frekuensi kemunculan 1

(1, 1) = indeks kata 1, frekuensi kemunculan 1

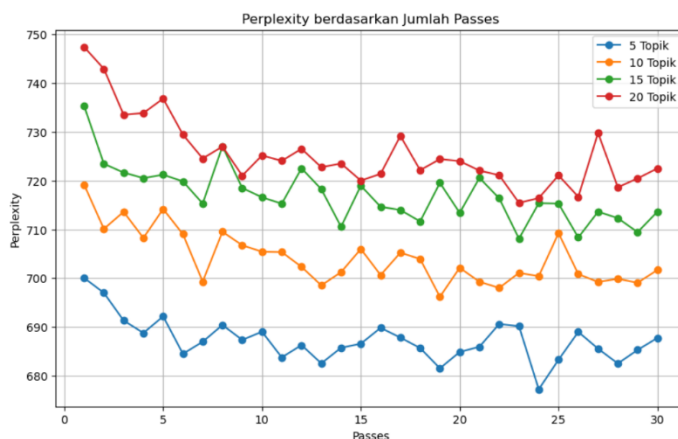
(2, 1) = indeks kata 2, frekuensi kemunculan 1

Demikian seterusnya untuk *corpus* lainnya.

Untuk menentukan jumlah iterasi dilakukan dengan melakukan analisis *perplexity*. Nilai *perplexity* dihitung berdasarkan probabilitas kemunculan kata-kata dalam data pengujian oleh model yang sudah dilatih. Semakin baik model dapat memprediksi kata-kata dalam data pengujian, semakin rendah nilai *perplexity* yang dihasilkan. *Perplexity* yang lebih rendah menunjukkan kualitas model yang lebih baik. Pada penelitian ini, perhitungan nilai *perplexity* dilakukan dengan memanfaatkan fungsi yang sudah tersedia pada *library python* yaitu *log_perplexity()*.

Analisis nilai *perplexity* dilakukan dengan cara menjalankan perhitungan *perplexity* dengan nilai parameter jumlah topik yang berbeda-beda yaitu 5, 10, 15 dan 20. Hasil perhitungan tersebut kemudian dicatat dan divisualisasikan untuk kemudian di analisis. Jika terjadi penurunan nilai *perplexity* yang signifikan kemudian stabil, hal itu dapat mengindikasikan bahwa model sudah *konvergen*.

Gambar 5 menunjukkan visualisasi hasil perhitungan nilai *perplexity*.



Gambar 8. Visualisasi perhitungan nilai *perplexity*

Pada gambar 5 nilai *perplexity* cenderung stabil untuk setiap parameter jumlah topik yaitu pada nilai *passes* 28 oleh karena itu pada penelitian ini nilai iterasi atau *passes* yang digunakan adalah 28.

Parameter lainnya yang menjadi input dalam model LDA adalah jumlah topik. Tidak ada aturan praktis dalam menentukan jumlah topik yang ada dalam suatu koleksi dokumen. Salah satu pendekatan yang dapat digunakan dalam menentukan jumlah topik adalah skor *coherence* dengan jumlah topik berbeda. Semakin tinggi nilai skor *coherence* mengindikasikan akurasi model yang lebih baik. Pada penelitian ini, penghitungan skor *coherence* dilakukan dengan memanfaatkan fungsi *get_coherence()* yang terdapat pada modul *CoherenceModel*. Nilai skor *coherence* yang dihasilkan kemudian di analisis *trend* nilainya. Jumlah topik yang dipilih adalah jumlah topik dengan nilai skor *coherence* tertinggi.

Selanjutnya dilakukan interpretasi terhadap tiap-tiap topik dan menghasilkan kesimpulan bahwa topik-topik yang dihasilkan pada pemodelan LDA dengan jumlah topik 5 adalah yang paling mudah diinterpretasikan. Hal ini sejalan dengan hasil perhitungan skor *coherence* yang menunjukkan angka terbaik pada jumlah topik 5.

Tabel berikut menunjukkan topik-topik hasil pemodelan LDA untuk jumlah topik 5.

Tabel 3. hasil pemodelan LDA jumlah topik 5

Model LDA Topik 1: ('0.040*metode" + 0.022*naive_bayes" + 0.018*analisis" + 0.017*algoritma" + 0.014*studi" + 0.014*perbandingan" + 0.013*mahasiswa" + 0.011*prediksi" + 0.010*sentimen" + 0.009*machine"')
Model LDA Topik 2: ('0.018*studi" + 0.018*algoritma" + 0.016*mean" + 0.015*metode" + 0.015*data" + 0.014*segmentasi" + 0.013*fp_growth" + 0.011*model" + 0.011*analisis" + 0.011*penjualan"')
Model LDA Topik 3: ('0.025*metode" + 0.018*prediksi" + 0.014*technique" + 0.012*model" + 0.010*analisis" + 0.009*pemilihan" + 0.009*algoritma" + 0.009*studi" + 0.009*term" + 0.009*short"')
Model LDA Topik 4: ('0.044*sistem" + 0.043*metode" + 0.032*keputusan" + 0.028*pendukung" + 0.027*analytical" + 0.027*hierarchy" + 0.025*process" + 0.021*studi" + 0.019*simple_additive" + 0.017*weighting_saw"')
Model LDA Topik 5: ('0.028*metode" + 0.021*learning" + 0.020*sistem" + 0.011*berbasis" + 0.010*informasi" + 0.010*jaringan" + 0.010*peramalan" + 0.009*system" + 0.008*recognition" + 0.008*karawang"')

Pada tabel di atas, kata-kata yang dicetak tebal adalah kata-kata yang berkontribusi pada tiap-tiap topik sehingga dapat dilakukan interpretasi pembahasan dari tiap-tiap topik. Kata-kata yang berkontribusi pada topik 1 adalah *naive_bayes*, *perbandingan*, *mahasiswa*, *sentimen*, *machine*. Kata-kata tersebut memberikan gambaran topik 1 adalah tentang analisis sentimen dan *machine learning*. Kata-kata yang berkontribusi pada topik 2 adalah *mean*, *data*, *segmentasi*, *fp_growth*, *penjualan*. Kata-kata tersebut memberikan gambaran topik 2 adalah tentang analisis data dan *machine learning*. Kata-kata yang berkontribusi pada topik 3 adalah *technique*, *pemilihan*, *term*, *short*. Kata-kata tersebut memberikan gambaran topik 3 adalah tentang analisis teks. Kata-kata yang berkontribusi pada topik 4 adalah *keputusan*, *pendukung*, *analytical*, *hierarchy*, *process*, *simple_additive*, *weighting_saw*. Kata-kata tersebut memberikan gambaran topik 4 adalah tentang sistem pendukung keputusan. Kata-kata yang berkontribusi pada topik 5 adalah *learning*, *berbasis*, *informasi*, *jaringan*, *peramalan*, *system*, *recognition*, *karawang*. Kata-kata tersebut memberikan gambaran topik 5 adalah tentang *machine learning*.

Hasil pemodelan LDA dapat divisualisasikan melalui *wordcloud*. Gambar 10 menunjukkan *wordcloud* topik-topik yang muncul dari hasil pemodelan LDA dengan jumlah topik 5.



Gambar 10. *wordcloud* topik hasil pemodelan LDA

Gambar di atas memperlihatkan kata-kata yang berkontribusi pada masing-masing topik, semakin besar ukuran dari teks semakin besar kontribusi kata tersebut pada sebuah topik.

4. PENUTUP

Analisis trend topik penelitian tesis dilakukan dengan mengimplementasikan pemodelan topik menggunakan metode *Latent Dirichlet Allocation* (LDA). Data yang digunakan yaitu data judul tesis Program Studi Magister Ilmu Komputer Universitas Budi Luhur dengan jumlah data 237 kemudian dipilih yang sesuai dan didapatkan sejumlah 211. Data tersebut kemudian dipersiapkan melalui tahap *pre-processing* untuk mempermudah dalam pemodelan topik yang meliputi proses *case folding*, *tokenizing*, dan *stopword removal*. Tahap selanjutnya yaitu perhitungan nilai *perplexity* dan skor *coherence* kemudian dilakukan pemodelan LDA.

Dari hasil pemodelan LDA didapatkan topik-topik yang menjadi trend dalam penelitian tesis Program Magister Ilmu Komputer Universitas Budi Luhur pada periode 2016 sampai dengan 2021 yaitu analisis sentimen, *machine learning*, analisis teks, analisis prediksi dan sistem pendukung keputusan

DAFTAR PUSTAKA

- [1] P.-N. Tan, M. Steinbach, A. Karpatne, dan V. Kumar, *Introduction to Data Mining*, 2nd ed. United Kingdom : Pearson , 2019.
- [2] O. Y. Findawati, M. M. Muhammad, A. Rosid, S. Kom, dan M. Kom, *Buku Ajar Text Mining*, vol. cetakan I. Sidoarjo : UMSIDA Press , 2020.
- [3] M. R. Faisal, D. Kartini, A. R. Arrahimi, dan T. H. Saragih, *Belajar Data Science: Text Mining Untuk Pemula I*, I. Banjarbaru : Scripta Cendekia , 2022.
- [4] D. M. Blei, A. Y. Ng, dan J. B. Edu, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, hlm. 993–1022, 2003. Diakses: 11 Februari 2023. [Daring]. Tersedia pada: <https://dl.acm.org/doi/10.5555/944919.944937>
- [5] L. H. Suadaa dan A. Purwarianti, "Combination of Latent Dirichlet Allocation (LDA) and Term Frequency-Inverse Cluster Frequency (TFxICF) in Indonesian text clustering with labeling," dalam *2016 4th International Conference on Information and Communication Technology, ICoICT 2016*, Institute of Electrical and Electronics Engineers Inc., Sep 2016. doi: 10.1109/ICoICT.2016.7571885
- [6] I. M. K. B. Putra dan R. P. Kusumawardani, "Analisis Topik Informasi Publik Media Sosial di Surabaya Menggunakan Pemodelan Latent Dirichlet Allocation (LDA)," *Jurnal Teknik ITS* , vol. 6, no. 2, hlm. 311–316, 2017.
- [7] Y. Sahlia, N. Isnaini Febriarini, dan P. Dwi Oktavianti, "Pemodelan Topik Penelitian Bidang Keperawatan Indonesia pada Repository Jurnal Sinta Menggunakan Metode Topic Modelling LDA (Latent Dirichlet Allocation)," dalam *Prosiding-Seminar Teknologi Informasi dan Ilmu Komputer* , hlm. 90–102, 2020.
- [8] S. Bahri, D. Wintana, R. Wajhillah, dan S. Suhada, "Penerapan Metode Topic Modelling untuk Penentuan Topik Konsultasi pada Portal Telemedicine Menggunakan LDA (Latent Dirichlet Allocation)," *Kumpulan Jurnal Ilmu Komputer* , vol. 09, no. 03, hlm. 402–410, 2022.
- [9] K. R. Adjie, P. Santoso, A. Husna, N. W. Putri, dan A. Rakhmawati, "Analisis Topik Tagar Covidindonesia pada Instagram Menggunakan Latent Dirichlet Allocation," *Jurnal Informatika Sunan Kalijaga*, vol. 7, no. 1, hlm. 1–9, Jan 2022.