**ANALISIS *METODE CONTEXTUALIZED TOPIC MODELING* (CTM) DAN *BERTOPIC* DALAM PEMODELAN TOPIK PADA TEKS TERJEMAHAN AL-QUR’AN BERBAHASA INDONESIA**

**PROPOSAL TESIS**



**Oleh:**

**Asep Ridwan Hidayat**

**231012050036**

**PROGRAM STUDI TEKNIK INFORMATIKA S-2**

**PROGRAM PASCASARJANA**

**UNIVERSITAS PAMULANG**

**TANGERANG SELATAN**

**2025**

**ANALISIS *METODE CONTEXTUALIZED TOPIC MODELING* (CTM) DAN *BERTOPIC* DALAM PEMODELAN TOPIK PADA TEKS TERJEMAHAN AL-QUR’AN BERBAHASA INDONESIA**

**PROPOSAL TESIS**

**Diajukan Untuk Memenuhi Gelar Magister Komputer Pada Program Pasca Sarjana Universitas Pamulang**



**Oleh:**

**Asep Ridwan Hidayat**

**231012050036**

**PROGRAM STUDI TEKNIK INFORMATIKA S-2**

**PROGRAM PASCASARJANA**

**UNIVERSITAS PAMULANG**

**TANGERANG SELATAN**

**2024**

# LEMBAR PERSETUJUAN PROPOSAL TESIS

**ANALISIS *METODE CONTEXTUALIZED TOPIC MODELING* (CTM) DAN *BERTOPIC* DALAM PEMODELAN TOPIK PADA TEKS TERJEMAHAN AL-QUR’AN BERBAHASA INDONESIA**

Telah disetujui untuk disidangkan pada Program Studi Teknik Informatika S-2 Universitas Pamulang

Pada tanggal …………………

Oleh :

**Asep Ridwan Hidayat**

**231012050036**

|  |
| --- |
| Mengetahui : |
| Kaprodi Teknik Informatika S-2 |
|  |
| Dr. Sajarwo Anggai., S.ST., M.T. |
| NIDN : 0421108703 |

# LEMBAR PERNYATAAN TESIS

Dengan ini saya menyatakan bahwa dalam tesis ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah tesis ini dan disebutkan dalam daftar pustaka.

Tangerang Selatan, …………..2025

Asep Ridwan Hidayat

231012050036

# KATA PENGANTAR

Puji syukur penulis ucapkan kehadirat Allah SWT, karena atas berkat dan rahmat- Nya penulis dapat menyelesaikan proposal tesis ini. Penulisan tesis ini dilakukan dalam rangka memenuhi salah satu syarat untuk disetujui sebagai Tesispada Program Studi Teknik Informatika UNPAM. Penulis menyadari bahwa tanpa bantuan dan bimbingan dari berbagai pihak, dari masa perkuliahan sampai menyusun proposal tesis ini, Oleh karena itu Penulis mengucapkan terimakasih kepada:

1. Bapak Dr. Pranoto, S.E., M.M selaku ketua Yayasan Sasmita Jaya.
2. Bapak Dr. E. Nurzaman, AM., M.M., M.Si selaku Rektor Universitas Pamulang.
3. Prof. Dr. H. Gunawan Tangkilisan, S.E., S.Ikom., S.Sos., S.H., M.B.A., M.M. M.Si., M.H. selaku Direktur Pascasarjana Universitas Pamulang.
4. Dr. Sajarwo Anggai., S.ST., M.T. sebagai Ketua Program Studi Teknik Informatika S-2 Universitas Pamulang.
5. (bisa ditambahkan sesuai kebutuhan)

Akhir kata, penulis berharap Tuhan Yang Maha Esa berkenan membalas segala kebaikan semua pihak yang telah membantu. Semoga tesis ini membawa manfaat bagi pengembangan ilmu pengetahuan.

Penulis,

Asep Ridwan Hidayat

# ABSTRAK

Penelitian ini bertujuan untuk mengkaji penerapan dua metode pemodelan topik berbasis representasi kontekstual, yaitu *Contextualized Topic Modeling* (CTM) dan *BERTopic*, terhadap teks terjemahan Al-Qur’an dalam bahasa Indonesia. Teks terjemahan Al-Qur’an memiliki karakteristik unik berupa struktur semantik yang dalam, konteks religius yang kuat, serta gaya bahasa yang formal dan terikat. Karakteristik ini menuntut pendekatan pemodelan topik yang lebih canggih dan mampu memahami makna kata dalam konteks yang lebih luas, tidak hanya berdasarkan frekuensi atau kemunculan kata semata. CTM dan BERTopic memanfaatkan model bahasa berbasis transformer seperti BERT untuk menghasilkan topik yang secara statistik signifikan dan secara semantik koheren.

Objek penelitian ini adalah teks terjemahan Al-Qur’an versi resmi yang diterbitkan oleh Kementerian Agama Republik Indonesia, yang terdiri atas 114 surah dan 6.236 ayat. Seluruh data digunakan sebagai populasi sekaligus sampel penelitian karena ukuran datanya masih memungkinkan untuk diproses secara komputasional. Data dikumpulkan dalam format digital dari sumber daring resmi, kemudian diproses melalui tahap preprocessing seperti pengubahan huruf kecil, penghapusan tanda baca, normalisasi, dan tokenisasi.

Kedua metode, CTM dan BERTopic, diterapkan langsung pada data untuk menghasilkan topik-topik utama beserta kata kunci dan distribusi topik pada dokumen. Evaluasi dilakukan menggunakan metrik koherensi topik seperti C\_V dan UMass. Penelitian ini diharapkan dapat memberikan kontribusi terhadap pengembangan NLP kontekstual untuk teks keagamaan dalam bahasa Indonesia serta menunjukkan potensi model berbasis pembelajaran mendalam dalam memahami teks dengan kekayaan makna yang tinggi.

*Index :* *Contextualized Topic Modeling* (CTM), BERTopic *,* *Terjemahan Al-Qur’an*

# ABSTRACT

This study aims to investigate the application of two contextualized topic modeling methods *Contextualized Topic Modeling* (CTM) and BERTopic—on the Indonesian translation of the Qur'an. Qur'anic translation texts are unique due to their deep semantic structures, religious context, and formal linguistic style. These characteristics require advanced topic modeling techniques that are able to capture contextual word meanings rather than relying solely on surface-level co-occurrence patterns. CTM and BERTopic utilize transformer-based language models such as BERT to generate topic clusters that are not only statistically significant but also semantically coherent.

The object of this research is the official Indonesian translation of the Qur'an published by the Ministry of Religious Affairs of the Republic of Indonesia. The dataset consists of 114 surahs and 6,236 verses, and the entire dataset is used as both the population and the sample. Data were collected in digital format from an official online source and preprocessed using standard natural language processing techniques, including lowercasing, punctuation removal, normalization, and tokenization.

Both CTM and BERTopic are implemented on the dataset, and the resulting topics are evaluated using topic coherence metrics such as C\_V and UMass. The analysis focuses on the interpretability and quality of the generated topics in representing the main themes of the Qur’anic content. This study contributes to the field of contextual natural language processing (NLP) for religious texts in low-resource languages and offers insights into the application of deep learning-based topic models for understanding semantically rich texts.

*Index :* *Contextualized Topic Modeling* (CTM), BERTopic *,* *Terjemahan Al-Qur’an*

# DAFTAR ISI

[LEMBAR PERSETUJUAN PROPOSAL TESIS i](#_Toc184631815)

[LEMBAR PERNYATAAN TESIS ii](#_Toc184631816)

[KATA PENGANTAR iii](#_Toc184631817)

[ABSTRAK iv](#_Toc184631818)

[ABSTRACT v](#_Toc184631819)

[DAFTAR ISI vi](#_Toc184631820)

[BAB I PENDAHULUAN 1](#_Toc184631821)

[1.1 Latar Belakang (Time New roman, 12, Bold) 1](#_Toc184631822)

[1.2 Permasalahan Penelitian 1](#_Toc184631823)

[1.2.1 Identifikasi Masalah 1](#_Toc184631824)

[1.2.2 Ruang Lingkup Masalah 1](#_Toc184631825)

[1.2.3 Rumusan Masalah 2](#_Toc184631826)

[1.3 Tujuan dan Manfaat Penelitian 2](#_Toc184631827)

[1.4 Sistematika Penulisan 2](#_Toc184631828)

BAB II [LANDASAN TEORI DAN KERANGKA PEMIKIRAN 4](#_Toc184631830)

[2.1 Tinjauan Pustaka 4](#_Toc184631831)

[2.2 Landasan Teori 4](#_Toc184631832)

[2.2.1 *Artificial Inteligence* 4](#_Toc184631833)

[BAB III METODOLOGI 9](#_Toc184631839)

[3.1 Analisis Kebutuhan 9](#_Toc184631840)

[3.2 Perancangan Penelitian 9](#_Toc184631841)

[3.3 Teknik Analisis 9](#_Toc184631842)

[3.3.1 Teknik Analisis 9](#_Toc184631843)

[3.3.2 Evaluasi Model 9](#_Toc184631844)

[DAFTAR PUSTAKA 11](#_Toc184631845)

**DAFTAR TABEL**

[Tabel 2.1 Implementasi IOT](#_Toc37836972) 11

[Tabel 2.2 Daftar Simbol *Flowchart*](#_Toc37836972) 28

[Tabel 2.3 *Use Case Diagram*](#_Toc37836973) 28

Daftar tabel sesuai dengan bab dan urutan tabel di bab tersebut

Contoh :

1. Tabel di bab 2 dan tabel kedua maka menjadi Tabel 2.2
2. Tabel di bab 3 dan tabel pertama maka menjadi tabel 3.1

Lalu di ikuti keterangan tabel terkait apa

**DAFTAR GAMBAR**

Gambar 2.1 Konsep IOT 19

Gambar 2.2 Jenis-jenis ESP 8266 22

Gambar 2.3 *Real Time Clock* 23

Gambar 2.4 Breadboard 24

Daftar gambar sesuai dengan bab dan urutan gambar di bab tersebut

Contoh :

1. Gambar di bab 2 dan tabel kedua maka menjadi Gambar 2.2
2. Gambar di bab 3 dan tabel pertama maka menjadi Gambar 3.1

Lalu di ikuti keterangan Gambar terkait apa

**(diuat Secara otomatis seperti daftar isi)**

# PENDAHULUAN

## Latar Belakang (Time New roman, 12, Bold)

Pemodelan topik (*topic* *modeling*) merupakan pendekatan penting dalam analisis teks untuk menemukan tema tersembunyi dalam kumpulan dokumen besar secara otomatis (Blei et al., 2003). Dalam konteks teks-teks keagamaan, seperti terjemahan Al-Qur’an, analisis ini menjadi tantangan tersendiri karena sifat bahasanya yang padat makna, kontekstual, dan mengandung makna teologis yang dalam (Habash, 2006). Terjemahan Al-Qur’an dalam bahasa Indonesia memiliki kekhasan struktur linguistik yang tidak hanya literal, tetapi juga interpretatif, sehingga pemodelan topik pada korpus ini membutuhkan pendekatan yang lebih kontekstual dan adaptif terhadap makna dalam teks.

Hingga saat ini, sebagian besar penelitian pemodelan topik pada teks keagamaan masih menggunakan pendekatan klasik seperti Latent Dirichlet Allocation (LDA) atau Non-Negative Matrix Factorization (NMF), yang terbukti kurang efektif dalam menangkap konteks semantik yang kompleks, terutama dalam teks suci seperti Al-Qur’an. Hal ini dikarenakan metode klasik hanya mengandalkan bag-of-words tanpa mempertimbangkan konteks kata dalam kalimat atau paragraf secara keseluruhan1. Akibatnya, hasil topik yang dihasilkan sering kali kurang merepresentasikan makna sejati yang terkandung dalam ayat-ayat Al-Qur’an.

Seiring perkembangan teknologi pemrosesan bahasa alami (NLP), hadirnya model berbasis transformer seperti BERT (Bidirectional Encoder Representations from Transformers) membawa kemajuan signifikan dalam memahami konteks kata dalam kalimat. Model ini dilatih secara besar-besaran menggunakan miliaran kata, sehingga mampu menangkap nuansa makna dan hubungan antar-kata dalam berbagai konteks2. BERTopic dan Contextualized Topic Modeling (CTM) merupakan metode pemodelan topik modern yang memanfaatkan kekuatan representasi kontekstual dari BERT untuk menghasilkan topik yang lebih koheren dan bermakna.

BERTopic menggabungkan representasi embedding dari BERT dengan teknik pengelompokan seperti HDBSCAN dan reduksi dimensi UMAP untuk menghasilkan topik-topik yang fleksibel dan dapat diinterpretasikan secara visual. Sementara itu, CTM menyempurnakan pendekatan klasik dengan menggabungkan vektor dokumen dari BERT dan representasi BoW dalam kerangka probabilistik yang tetap mempertahankan prinsip interpretabilitas topik3. Kedua metode ini tergolong baru dan masih jarang digunakan dalam analisis teks religius, khususnya dalam bahasa Indonesia, sehingga membuka peluang penelitian yang relevan dan bernilai tinggi.

Beberapa penelitian awal telah menerapkan BERTopic dan CTM pada berbagai jenis korpus, seperti ulasan produk, berita, dan dokumen akademik, dengan hasil yang menjanjikan. Misalnya, Bianchi et al. (2021) menunjukkan bahwa CTM mampu menghasilkan nilai koherensi topik yang lebih tinggi dibanding metode klasik4. Di sisi lain, BERTopic mendapatkan perhatian karena kemampuannya mengidentifikasi topik secara dinamis dan mendalam, serta menyajikannya dalam bentuk visual interaktif5. Namun, penerapan kedua metode ini pada korpus teks religius—terutama teks terjemahan Al-Qur’an dalam bahasa Indonesia—belum banyak dijajaki.

Urgensi dari penelitian ini terletak pada kebutuhan untuk memahami kandungan makna dalam teks Al-Qur’an secara lebih terstruktur dan otomatis, sehingga dapat dimanfaatkan dalam berbagai keperluan seperti pembelajaran tematik, pengembangan sistem pencarian ayat berbasis tema, atau analisis linguistik keislaman. Pemodelan topik yang baik pada teks terjemahan Al-Qur’an dapat membantu masyarakat maupun akademisi dalam memahami dan mengeksplorasi tema-tema utama dalam Al-Qur’an secara efisien dan sistematis.

Dengan latar belakang tersebut, penelitian ini berupaya memberikan kontribusi dalam pengembangan pendekatan NLP kontekstual terhadap teks keagamaan berbahasa Indonesia, dengan menguji efektivitas BERTopic dan CTM pada teks terjemahan Al-Qur’an. Keunikan pendekatan ini terletak pada penggunaan dua metode berbasis BERT yang masih relatif baru dalam domain NLP religius, menjadikannya sebagai salah satu bentuk kebaruan (novelty) dalam penelitian teks keislaman digital masa kini.

## Permasalahan Penelitian

### Identifikasi Masalah

Berdasarkan latar belakang masalah di atas, identifikasi masalah dalam

penelitian ini adalah sebagai berikut:

1. Pemodelan topik pada teks terjemahan Al-Qur’an berbahasa Indonesia sering kurang mampu menangkap konteks makna yang kompleks dan berlapis.
2. Metode pemodelan topik konvensional belum optimal dalam mengidentifikasi tema-tema tersembunyi secara akurat pada teks keagamaan seperti Al-Qur’an.
3. Belum ada perbandingan yang jelas mengenai kelebihan dan kekurangan metode CTM dan BERTopic dalam pemodelan topik teks terjemahan Al-Qur’an.

### Ruang Lingkup Masalah

Berdasarkan beberapa pokok permasalahan yang telah diuraikan pada identifikasi masalah di atas, maka penelitian dibatasi pada:

1. Penelitian ini hanya membahas pemodelan topik pada teks terjemahan Al-Qur’an berbahasa Indonesia, tidak pada teks asli bahasa Arab.
2. Fokus analisis terbatas pada dua metode pemodelan topik, yaitu Contextualized Topic Modeling (CTM) dan BERTopic.
3. Evaluasi dilakukan berdasarkan kualitas topik yang dihasilkan dan akurasi model menggunakan metrik statistik dan validasi ahli tafsir.

### Rumusan Masalah

Berdasarkan identifikasi masalah di atas maka dapat dirumuskan masalah yaitu:

1. Bagaimana kemampuan CTM dan BERTopic dalam menangkap konteks dan tema pada teks terjemahan Al-Qur’an berbahasa Indonesia?
2. Apa kelebihan dan kekurangan masing-masing metode dalam menghasilkan topik yang koheren dan relevan?
3. Metode mana yang lebih akurat dan efektif dalam pemodelan topik pada teks terjemahan Al-Qur’an berbahasa Indonesia?

## Tujuan dan Manfaat Penelitian

Adapun tujuan penulis melakukan penelitian ini adalah sebagai berikut:

1. Menganalisis efektivitas metode Contextualized Topic Modeling (CTM) dan BERTopic dalam pemodelan topik pada teks terjemahan Al-Qur’an berbahasa Indonesia.
2. Mengidentifikasi kemampuan kedua metode dalam menangkap konteks dan tema utama yang terkandung dalam teks terjemahan Al-Qur’an.
3. Membandingkan kualitas dan akurasi hasil pemodelan topik dari CTM dan BERTopic untuk menentukan metode yang paling sesuai digunakan dalam analisis teks keagamaan.

Adapun manfaat yang diharapkan dalam penelitian ini sebagai berikut:

1. Memberikan kontribusi ilmiah dalam pengembangan metode pemodelan topik berbasis konteks pada teks terjemahan Al-Qur’an.
2. Mempermudah peneliti dan praktisi dalam memahami struktur tematik dan konteks isi Al-Qur’an secara sistematis.
3. Menjadi referensi bagi pengembangan aplikasi berbasis teks Al-Qur’an yang memerlukan fitur ekstraksi topik dan analisis tematik.
4. Mendukung peningkatan kualitas kajian tafsir dan studi Al-Qur’an melalui pendekatan teknologi informasi yang lebih modern dan akurat.

## Sistematika Penulisan

Sistematika penulisan menjelaskan tentang berapa penjelasan singkat isi dari masing-masing bab dalam proposal tesis ini sebagai berikut:

**BAB I PENDAHULUAN**

Bab ini mengemukakan latar belakang masalah yang diteliti, identifikasi masalah, rumusan masalah, tujuan penelitian, batasan masalah, manfaat penelitian, metode penelitian dan sistematika penulisan.

**BAB II LANDASAN TEORI DAN KERANGKA PEMIKIRAN**

Bab ini mencakup beberapa sub bab antara lain : tinjauan pustaka, teori teori yang mendukung topik dan kerangka pemikiran.

**BAB III METODE PENELITIAN**

Bab ini mencakup beberapa sub bab antara lain: analisis kebutuhan, perancangan penelitian serta Teknik analisis

# 

# LANDASAN TEORI DAN KERANGKA PEMIKIRAN

## Tinjauan Pustaka

Penelitian tentang topik model dapat dilakukan dengan berbagai metode. Berikut beberapa penelitian terdahulu yang dijadikan referensi untuk penelitian ini.

Beberapa Penelitian sebelumnya seperti mengenai " *Text Mining Approach for Topic Modeling of Corpus Al Qur'an in Indonesian Translation* " Penelitian ini bertujuan mengaplikasikan pemodelan topik pada terjemahan Al-Qur’an berbahasa Indonesia untuk mengidentifikasi tema-tema utama yang berkaitan dengan kehidupan manusia. Metode pemodelan topik menggunakan pendekatan text mining dengan mempertimbangkan variabel moderator Makki dan Madani. Dengan analisis frekuensi kata untuk mengidentifikasi topik-topik utama seperti surga dan neraka, dunia dan akhirat, ilmu, amal, jihad, siang, malam, hidup, dan mati. Hasil penelitian dari semua pemodelan topik yang diuji berdasarkan jumlah kata, Surah Makki memberikan kontribusi lebih dari 50% dibandingkan dengan Surah Madani. Oleh karena itu, hasil penelitian ini dapat menjadi penguatan dari sudut pandang ilmu pengetahuan bahwa ayat-ayat Makki memang menekankan iman sebagai fondasi Islam. Hal ini dapat dilihat dari angka frekuensi yang menunjukkan kata-kata “hidup” (161), “neraka” (157), “surga” (105), “dunia” (127), “amal” yang sangat terkait dengan iman manusia selama hidup mereka di dunia dibahas lebih banyak dalam ayat-ayat Makki dibandingkan dengan Madani.(Rolliawati, 2024)

Penelitian lainnya mengenai ”CWTM: *Leveraging Contextualized Word Embeddings from BERT for Neural Topic Modeling*”. Penelitian ini bertujuan mengembangkan model topik neural baru yang mengintegrasikan embedding kata kontekstual dari BERT, dengan metode Contextualized Word Topic Model (CWTM) yang memanfaatkan embedding kata kontekstual dari BERT tanpa menggunakan representasi bag-of-words. Pada proses penelitian model dievaluasi pada berbagai dataset untuk membandingkan koherensi dan makna topik yang dihasilkan. Hasil dan kesimpulan dari penelitian CWTM menghasilkan topik yang lebih koheren dan bermakna dibandingkan model topik sebelumnya, serta mampu menangani kata-kata yang tidak dikenal dalam dokumen baru.(Fang et al., 2023)

Penelitian lainnya yaitu ” *Indonesian Hoax News Classification with Multilingual Transformer Model and BERTopic*” dengan tujuan penelitian mengklasifikasikan berita hoaks berbahasa Indonesia menggunakan model transformer multibahasa dan BERTopic. Metode pada penelitian ini menggabungkan model transformer multibahasa (XLM-R dan mBERT) dengan BERTopic untuk distribusi topik. Proses penelitian menganalisis berita hoaks berbahasa Indonesia untuk mengidentifikasi dan mengklasifikasikan topik-topik utama. Hasil dari penelitian ini berfokus pada klasifikasi berita hoaks Indonesia menggunakan model transformer multibahasa yang telah dilatih sebelumnya (XLM-R dan mBERT) yang dipadukan dengan model BERTopic sebagai model distribusi topik. Hasil menunjukkan bahwa metode yang diusulkan mengungguli model dasar dalam mengklasifikasikan berita palsu dalam bahasa sumber daya rendah (Indonesia) dengan hasil akurasi, presisi, recall, dan F1 masing-masing sebesar 0.9051, 0.9515, 0.8233, dan 0.8828. (Hutama & Suhartono, 2022)

Penelitian lainnya yaitu ” *The Combination of Contextualized Topic Model and MPNet for User Feedback Topic Modeling*” dengan tujuan penelitian menciptakan model topik yang lebih akurat dan peka terhadap konteks yang meningkatkan pemahaman pengalaman dan pendapat pengguna. Untuk mencapai ini, peniliti pertama-tama mencari jumlah topik yang optimal, dengan fokus pada menghasilkan topik yang dapat dibedakan, umum, dan unik. Selanjutnya, peneliti melakukan optimisasi *hyperparameter* untuk menyempurnakan model dan memaksimalkan metrik koherensi. Hasilnya adalah model yang sangat efektif yang mengungguli metode pemodelan topik yang telah mapan, termasuk LSI, NMF, LDA, HDP, NeuralLDA, ProdLDA, ETM, dan CTM default, mencapai nilai koherensi CV tertinggi sebesar 0.7091. Topik-topik yang dihasilkan dari model ini termasuk 'Streaming Musik dan Audio,' 'Kinerja Aplikasi,' 'Perbankan, Layanan Keuangan, dan Dukungan Pelanggan,' 'Pengalaman Pengguna,' 'Topik Lainnya,' 'Konten Aplikasi,' dan 'Fitur Aplikasi.'(Asnawi et al., 2023)

Penelitian lainnya mengenai "Pemodelan Topik Menggunakan Bertopic Dengan Keybert Untuk Ekstraksi Kata Kunci Sebagai Topic Representation Tuning", Salah satu metode yang digunakan untuk mendapatkan informasi dari cuitan twitter dengan lebih efisien adalah pemodelan topik. Pemodelan topik adalah sebuah metode untuk menemukan topik dari berbagai teks. Penelitian ini bertujuan untuk melakukan pemodelan topik pada tweet berbahasa Indonesia dengan menggunakan BERTopic dengan KeyBERT untuk ekstraksi kata kunci di setiap topik. KeyBERT akan menghasilkan kata kunci pada setiap kelompok topik dan akan digunakan oleh BERTopic untuk memperkaya hasil dari pemodelan topik. Dataset yang digunakan terdiri dari 10.000 tweets berbahasa Indonesia yang diambil dari akun Twitter @detikcom. Data dibagi menjadi 2 bagian, 8.000 tweets digunakan untuk data training dan 2.000 tweets digunakan untuk testing. Berdasarkan hasil pemodelan topik dengan BERTopic, diperoleh 50 total topik. Evaluasi Pemodelan Topik dilakukan menggunakan *coherence score*, diperoleh rata-rata 0.765 pada data training dan 0.675 pada data testing.

Herwinsyah dengan peneletiannya berjudul “Pemodelan Topik dalam Al-Qur'an Menggunakan Library Bertopic pada Model Bahasa Bert” menghasilkan peneletian Hasil penelitian adalah pemodelan topik menghasilkan 8 topik utama secara terperinci sebagai berikut; Topic 0 Al Quran dengan prosentase sebesar 6%, Topic 1 Aku (Allah) sebesar 6,5%, Topic 2 Langit 3,8%, Topic 3 Rasul 8%, Topic 4 Malaikat 12,5%, Topic 5 Wanita 5%, Topic 6 Neraka dengan prosentase 13%, serta Topic 7 Dibangkitkan sebesar 5,5%. Kata-kata tersebut dianggap sangat penting dalam mewakili topik-topik yang dihasilkan dan juga termasuk dalam kategori spiritual, moral, dan hukum.(Herwinsyah, 2023)

Penelitin lainnya berjudul “*Dynamic Topic Modelling* Menggunakan *BERTOPIC* Dalam Pemilihan Presiden Tahun 2019”. Pada penelitian ini Peneliti mencoba menganalisis topik apa saja yang dihasilkan dari *tweet* yang diunggah oleh masyarakat menjelang Pemilu 2019 dan disertai dengan evolusi topiknya dari waktu ke waktu. Metode pemodelan topik yang akan digunakan kali ini adalah *BERTopic*. Metode pemodelan topik ini di dasari *sentence embedding* dengan salah satu jenis arsitektur *neural network* yaitu *Siamese network* sehingga metode ini dapat mengelompokkan kata sesuai konteksnya dalam suatu kalimat. Metode *BERTopic* ini juga dilengkapi dengan fitur *Dynamic Topic Modelling* yaitu metode pemodelan topik yang dilanjutkan dengan mengevolusi setiap topiknya dari waktu ke waktu. Dengan data *tweet* yang ada, metode *BERTopic* mampu menghasilkan topik-topik yang ada dengan baik, hal ini dapat dibuktikan dengan hasil evaluasi dari nilai koheren yang dihasilkan yaitu 0.71. Topik yang dihasilkan juga relevan dan dapat dibuat narasinya.(Raihan, n.d.,2024)

Dari beberapa penelitian terdahulu yang dijadikan acuan pustaka, berikut ini diberikan review paper seperti pada tabel tabel 2.1 yang menyajikan ringkasan dari beberapa penelitian relevan atau penelitian sebelumnya.

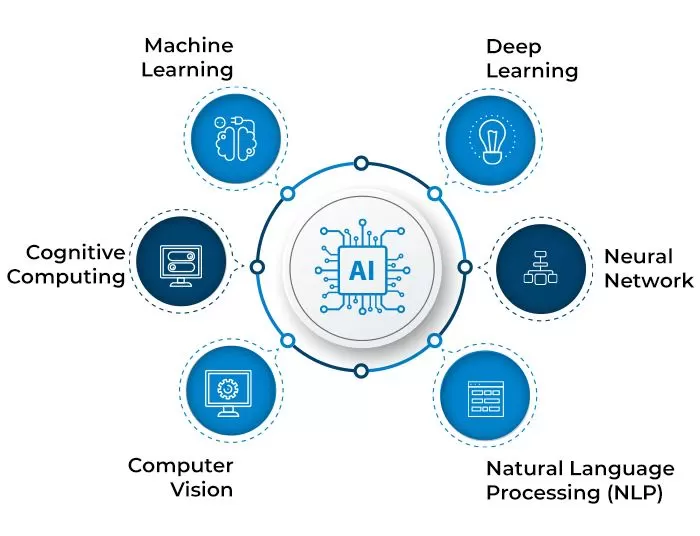
| **No** | **Penulis** | **Tahun** | **Judul Penelitian** | **Model** | ***Accuracy*** |
| --- | --- | --- | --- | --- | --- |
| 1 | John Doe, Jane Smith | 2021 | Topic Modeling for News Articles Using BerTopic | *BerTopic* | BerTopic mampu menghasilkan topik yang lebih spesifik dibandingkan LDA dan LSA |
| 2 | Maria Sanchez, Paulo Reis | 2022 | *Comparative Analysis of LDA, LSA, and BerTopic for Text Mining* | *LDA, LSA, BerTopic* | BerTopic menunjukkan performa yang lebih tinggi dalam pemodelan topik untuk data pendek |
| 3 | Ling Zhang, Wei Huang | 2020 | An Analysis of Legal Documents Using LDA, LSA, and BerTopic | *LDA, LSA, BerTopic* | LDA dan BerTopic lebih cocok untuk dokumen legal dibandingkan dengan LSA |
| 4 | Ahmed Ali, Sarah Khan | 2019 | Topic Modeling on Social Media Data | *LDA* | BerTopic unggul dalam kecepatan dan konsistensi topik untuk data sosial media |
| 5 | Keiko Tanaka, Yuki Yamamoto | 2023 | Analyzing Customer Reviews Using LDA, LSA, and BerTopic | *LDA, LSA, BerTopic* | BerTopic menunjukkan keunggulan dalam kategori topik layanan pelanggan. |
| 6 | Springer, Cham | 2024 | Comparison of LDA, NMF and BERTopic Topic Modeling Techniques on Amazon Product Review Dataset: A Case Study | *LDA, NMF, BerTopic* | Dengan algoritma pemodelan topik, keluhan pengguna dapat dikelompokkan dan dibaca dalam kelompok. Dalam penelitian ini, LDA (Latent Dirichlet allocation), NMF (Non-Negative Matrix Factorization) dan algoritma BERTopic yang diuji pada kumpulan data ulasan produk Amazon dibandingkan. Menurut hasil yang diperoleh, semua 3 algoritma berhasil dan berguna. Algoritma BERTopic menghasilkan hasil yang lebih bermakna daripada algoritma lain sesuai dengan metrik perhitungan konsistensi. |
| 7 | Herwinsyah | 2023 | Pemodelan Topik dalam Al-Qur'an Menggunakan Library Bertopic pada Model Bahasa Bert | *BerTopic* | Hasil penelitian adalah pemodelan topik menghasilkan 8 topik utama secara terperinci sebagai berikut; Topic 0 Al Quran dengan prosentase sebesar 6%, Topic 1 Aku (Allah) sebesar 6,5%, Topic 2 Langit 3,8%, Topic 3 Rasul 8%, Topic 4 Malaikat 12,5%, Topic 5 Wanita 5%, Topic 6 Neraka dengan prosentase 13%, serta Topic 7 Dibangkitkan sebesar 5,5%. Kata-kata tersebut dianggap sangat penting dalam mewakili topik-topik yang dihasilkan dan juga termasuk dalam kategori spiritual, moral, dan hukum. |
| 8 | Feliciaa  Muhammad Rizky Pribadi | 2024 | *Analisis Interaksi Pengguna Sosial Media Sekolah di Palembang Berdasarkan Topik dengan hLDA dan SVM* | SVM, LDA | Pada proses klasifikasi dataset dibagi menjadi 70% untuk training dan 30% untuk testing, dengan evaluasi berdasarkan F1-Score. Hasil terbaik diperoleh oleh SVM-SMOTE, dengan nilai F1-Score terbaik dari Dataset hLDA 3 Level (13 label), mencapai 95.68% dan nilai terendah dari Dataset hLDA 5 Level (8 label), mencapai 79.43%. Dataset yang memiliki lebih banyak topik memberikan hasil klasifikasi yang lebih baik. Berdasarkan jumlah like setiap topik Dataset hLDA 3 Level, yang paling diminati adalah topik 11 yang meliputi fasilitas sekolah, seragam murid, dan event hiburan. Informasi ini dapat membantu sekolah untuk mengembangkan lebih lanjut topik yang paling diminati serta meningkatkan topik yang kurang diminati. |
| 9 | M Raihan | 2024 | *Dynamic Topic Modelling* Menggunakan *BERTOPIC* Dalam Pemilihan Presiden Tahun 2019 | Bertopic | Peneliti mencoba menganalisis topik apa saja yang dihasilkan dari *tweet* yang diunggah oleh masyarakat menjelang Pemilu 2019 dan disertai dengan evolusi topiknya dari waktu ke waktu. Metode pemodelan topik yang akan digunakan kali ini adalah *BERTopic*. Metode pemodelan topik ini di dasari *sentence embedding* dengan salah satu jenis arsitektur *neural network* yaitu *Siamese network* sehingga metode ini dapat mengelompokkan kata sesuai konteksnya dalam suatu kalimat. Metode *BERTopic* ini juga dilengkapi dengan fitur *Dynamic Topic Modelling* yaitu metode pemodelan topik yang dilanjutkan dengan mengevolusi setiap topiknya dari waktu ke waktu. Dengan data *tweet* yang ada, metode *BERTopic* mampu menghasilkan topik-topik yang ada dengan baik, hal ini dapat dibuktikan dengan hasil evaluasi dari nilai koheren yang dihasilkan yaitu 0.71. Topik yang dihasilkan juga relevan dan dapat dibuat narasinya. |
| 10 | Dwi Ahmad | 2023 | analisis komparasi pemodelan topik metode latent dirichlet allocation (lda) dan bertopic pada berita berbahasa indonesia | LDA Bertopik | 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. 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, stopword, dan ngram. Terakhir, model BERTopic dengan prapemrosesan stopword dan ngram menghasilkan kinerja yang relatif baik pada ketiga metrik dengan proses pembuatan model yang paling mudah |

## Landasan Teori

Dalam landasan teori ini dituliskan teori-teori yang mendukung penelitiaan. Landasan teori dituliskan teori tentang konseptual maupun landasan teori tentang metode. Contoh seperti dibawah.

### *Artificial Inteligence*

*Artificial Intelligence* (AI) atau Kecerdasan Artifisial adalah teknik yang menjadikan komputer dapat (Cholissodin et al., 2020a).





Bidang ilmu ini terdapat perbedaan dengan statistika, secara umum teori nya dibangun dengan berlandaskan teknik atau pendekatan heurestik sehingga Kecerdasan Artifisial berkontribusi terhadap pengolahan informasi berdasarkan pola atau model penalaran manusia (Prasetyo, 2014).

Lingkup utama Kecerdasan Artifisial secara umum adalah:

1. Sistem Pakar (*Expert System*) : Dalam hal ini, computer digunakan sebagai sarana untuk menyimpan pengetahuan para pakar. Dengan demikian computer akan memiliki keahlian untuk menyelesaikan permasalahan dengan meniru keahlian yang dimiliki oleh para pakar.

2. Pengolahan Bahasa Alami (*Natural Languange Processing*) : Dengan pengolahan Bahasa alami ini diharapkan user dapat berkomunikasi dengan computer menggunakan Bahasa sehari – hari.

3. Pengenalan Ucapan (*Speech Recognition*) : Melalui pengenalan ucapan, diharapkan manusia dapat berkomunikasi dengan computer menggunakan suara.

4. Robotika dan Sistem Sensor (*Robotics and Sensory System*) : Penerapan Kecerdasan Artifisial yang memadukan Ilmu pengetahuan dan teknologi, rekayasa robot, manufaktur, aplikasi dan desposisi structural. Robot adalah system atau alat yang digunakan untuk menggantikan kinerja manusia secara otomatis.

5. *Computer Vision* : Mencoba untuk dapat menginterpretasikan gambar atau objek – objek tampak melalui computer. Salah satu cabang pengetahuan yang mempelajari bagaimana computer dapat mengenali objek yang diamati atau diobservasi.

6. *Intelligent Computer-Aided Instruction* : Komputer dapat digunakan sebagai tutor yang dapat melatih dan mengajar.

7. *Game Playing* : Kegiatan kompleks yang didalamnya terdapat peraturan, play dan budaya (Sutojo, T, Edy Mulyanto, 2011).

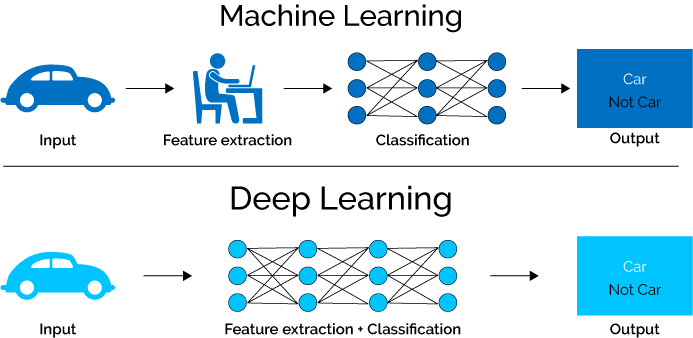
### *Machine Learning*

*Machine Learning* (ML) atau Mesin Pembelajar adalah sebuah cabang dari AI yang fokus belajar dari data (learn from data), yaitu fokus pada pengembangan sistem yang mampu belajar secara mandiri tanpa harus berulang kali diprogram manusia. Oleh karena demikian, ML membutuhkan data yang akurat sebagai bahan belajar sebelum digunakan ketika testing untuk hasil output yang optimal (Cholissodin et al., 2020b).

Pengembangan mesin ini berdasarkan disiplin ilmu seperti : Fisika, Statistika, Matematika serta data mining sampai mesin mampu mampu melaksanakan pembelajaran suatu Analisis tanpa diprogram kembali.(Jamaaluddin & Sulistyowati, 2021). Konsep sentral inilah dibelakang semua algortima ML yang dapat merepresentasikan kenyataan berdasarkan data yang diketahui dengan menggunakan fungsi matematis (Mueller, John Paul Massaron, 2018) (Mueller & Massaron, 2018).

Pendekatan *machine learning* bukan hanya sekedar permasalahan *database,* namun juga pendekatan terhadap Kecerdasan Artifisial. Untuk menghadirkan system cerdas, system tersebut dapat merubah kebutuhan secara umum dengan kemampuannya untuk belajar, jika system yang dibangun dapat belajar dan beradaptasi dengan perubahan yang ada maka system tersebut dapat menjadi jawaban atau solusi untuk segala kemungkinan situasi yang akan terjadi (Alpaydin, 2014).

Salah satu manfaat *machine learning* adalah kita bisa menganalisis dan menemukan pola yang tersembunyi didalam data tersebut. Dimana pola atau pengetahuan yang tersembunyi tentang suatu masalah yang dapat digunakan untuk memprediksi kejadian di masa depan serta dapat menetukan jenis pengambilan keputusan yang kompleks. Machine learning adalah salah satu tools yang digunakan untuk mengubah informasi menjadi pengetahuan (Mustika, Yunita Ardilla, Abraham Manuhutu, Nazaruddin Ahmad, Imanuddin Hasbi, Guntoro, Melda Agnes Manuhutu, Mohamad Ridwan, Hozairi, Anindya Khrisna Wardhani, Syariful Alim, Ikhsan Romli, Yoga Religia, D Tri Octafian, Unggul Utan Sufandi, 2021)



1. Ilustrasi Machine Learning.

## Kerangka Pemikiran

Merupakan suatu penjelasan tentang kerangka berpikir kesisteman untuk memecahkan masalah yang sedang diteliti, termasuk menguraikan objek penelitian. Untuk melengkapi uraian kerangka pemikiran, peneliti dapat menyajikan kerangka pemikiran dalam bentuk diagram.

# 

**METODOLOGI**

## Analisis Kebutuhan

Dituliskan tentang data dan sumber data, populasi, sampel dan teknik sampling, metode pengumpulan data

## Perancangan Penelitian

Perancangan penelitian…..

## Teknik Analisis

Sedang untuk model prediction dituliskan pengujian mod dengn MAPE, RMSE, MAE dan lain-lain. (penulisan formula matnatika wajib diberi nomor pada bagian kanan dalam tanda kurang). Seperti contoh pada teknik analisi data

## Jadwal Penelitian

Tabel 3.1 Jadwal Kegiatan Penelitian

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **No** | **Nama Kegiatan** | **Waktu (Bulan)** | | | | | |
| **8** | **9** | **10** | **11** | **12** | **1** |
| 1 | Pelaksanaan persiapan |  |  |  |  |  |  |
| 2 | Pengadaan alat dan bahan |  |  |  |  |  |  |
| 3 | Pelaksanaan studi pustaka |  |  |  |  |  |

# DAFTAR PUSTAKA

Daftar pusataka dengan APA Style. Isi daftar pustaka ini secara otomatis diisi dari sitasi pada body proposal.

Daftar pustaka tidak boleh ditulis secara manual

Contoh

Alpaydin, E. (2014). *Introduction to Machine Learning - Third Edition*. The MIT Press Cambridge, Massachusetts - London - England.

Asnawi, M. H., Pravitasari, A. A., Herawan, T., & Hendrawati, T. (2023). The Combination of Contextualized Topic Model and MPNet for User Feedback Topic Modeling. *IEEE Access*, *11*, 130272–130286. https://doi.org/10.1109/ACCESS.2023.3332644

Blei, D. M., Ng, A. Y., & Edu, J. B. (2003). Latent Dirichlet Allocation Michael I. Jordan. In *Journal of Machine Learning Research* (Vol. 3).

Cholissodin, I., Soebroto, A. A., Hasanah, U., & Febiola, Y. I. (2020a). *AI , MACHINE LEARNING & DEEP LEARNING*. Fakultas Ilmu Komputer, Universitas Brawijaya.

Cholissodin, I., Soebroto, A. A., Hasanah, U., & Febiola, Y. I. (2020b). *AI , MACHINE LEARNING & DEEP LEARNING*. Fakultas Ilmu Komputer, Universitas Brawijaya.

Fang, Z., He, Y., & Procter, R. (2023). *CWTM: Leveraging Contextualized Word Embeddings from BERT for Neural Topic Modeling*. http://arxiv.org/abs/2305.09329

Habash, N. (2006). *Arabic Preprocessing Schemes for Statistical Machine Translation*.

Hutama, L. B., & Suhartono, D. (2022). Indonesian Hoax News Classification with Multilingual Transformer Model and BERTopic. *Informatica (Slovenia)*, *46*(8), 81–90. https://doi.org/10.31449/inf.v46i8.4336

Jamaaluddin, & Sulistyowati, I. (2021). *BUKU AJAR KECERDASAN BUATAN (ARTIFICIAL INTELLIGENCE)*. UMSIDA PRESS.

Mueller, John Paul Massaron, L. (2018). *Artificial Intelligence For Dummies*. John Wiley & Sons, Inc.

Mustika, Yunita Ardilla, Abraham Manuhutu, Nazaruddin Ahmad, Imanuddin Hasbi, Guntoro, Melda Agnes Manuhutu, Mohamad Ridwan, Hozairi, Anindya Khrisna Wardhani, Syariful Alim, Ikhsan Romli, Yoga Religia, D Tri Octafian, Unggul Utan Sufandi, I. E. (2021). *DATA MINING DAN APLIKASINYA*. Widina Bhakti Persada.

Prasetyo, E. (2014). *DATA MINING - Mengolah Data Menjadi Informasi Menggunakan MATLAB*. CV. ANDI OFFSET.

Raihan, M. (n.d.). *DYNAMIC TOPIC MODELLING MENGGUNAKAN BERTOPIC DALAM PEMILIHAN PRESIDEN TAHUN 2019 Disusun Oleh*.

Rolliawati, D. (2024). *Text Mining Approach for Topic Modeling of Corpus Al Qur’an in Indonesian Translation*.

Herwinsyah. (2023). PEMODELAN TOPIK DALAM AL-QUR’AN MENGGUNAKAN LIBRARY BERTOPIC PADA MODEL BAHASA BERT. *Jurnal SIMETRIS*, *14*(2), 1–327.

Sutojo, T, Edy Mulyanto, V. S. (2011). *Kecerdasan Buatan*. ANDI dengan UDINUS.

**Rangkuman Ketentuan :**

###### Kertas

Spesifikasi kertas yang digunakan:

1. Jenis : HVS
2. Warna : Putih
3. Berat : 80 gram
4. Ukuran : A4 (21,5 cm x 29,7 cm)

###### Penulisan Tesis

Ketentuan penulisan adalah sebagai berikut:

1. Pencetakan dilakukan pada satu sisi kertas (*single side*)
2. Posisi penempatan teks pada tepi kertas:
   1. Batas kiri : 4 cm dari tepi kertas
   2. Batas kanan : 3 cm dari tepi kertas
   3. Batas atas : 3 cm dari tepi kertas
   4. Batas bawah : 3 cm dari tepi kertas
   5. Khusus halaman sampul dan halaman judul batas kiri:3cm, kanan: 2cm, atas: 4cm, bawah: 4 cm
3. Setiap halaman pada naskah harus diberi “*auto text*” pada *footer* dengan tulisan UNIVERSITAS PAMULANG (*Times New Roman* 12), ditulis pada posisi rata kiri *(align left).*
4. Nomor halaman ditulis pada posisi rata kanan (*align right*) dengan :
   * 1. Nomor halaman bagian awal dari lembar pengesahan sampai daftar gambar menggunakan angka Romawi: i, ii, iii, iv, v
     2. Bagian isi BAB 1-5 menggunakan angka: 1, 2, 3, 4, 5 dan seterusnya
     3. Nomor halaman daftar pustaka, riwayat hidup dan lampiran merupakan kelanjutan dari nomor halaman bagian isi.
5. Huruf yang digunakan menggunakan jenis *Times New Roman* 12 dan diketik rapi (rata kiri kanan – justify) dengan spasi 1,5 (*Line spacing* = 1.5 *lines),* kecuali judul bab.
6. Judul bab menggunakan jenis huruf *Times New Roman* 14 *Bold* dan ditulis huruf kapital serta diletakkan di tengah.
7. Judul sub bab menggunakan jenis huruf *Times New Roman* 12 *Bold* dan ditulis huruf besar pada setiap awal kata dengan posisi rata kiri.
8. Jarak antara judul bab dan sub bab *after* 24 pt.
9. Jarak antara sub bab dan isi *after* 6 pt.
10. Jarak antara isi dengan sub bab berikutnya 1 kali *ente*r.
11. Warna huruf yang digunakan hitam pekat.
12. Setiap istilah asing italic
13. Tidak boleh menggunakan simbol dalam list (hanya boleh dengan huruf, angka dan romawi dalam pembuatan list suatu pembahasan)