

**TUGAS AKHIR – EF234801**

**PERBANDINGAN KINERJA WORD EMBEDDING DAN PEMODELAN TOPIK DALAM IDENTIFIKASI TOPIK TUGAS AKHIR BERDASARKAN JUDUL DAN ABSTRAK PENELITIAN**

**DAVID FISCHER SIMANJUNTAK**

NRP 5025201123

Dosen Pembimbing I

**Prof. Dr. Eng. Nanik Suciati, S.Kom., M.Kom.**

NIP 197104281994122001

Dosen Pembimbing II

**Hadziq Fabroyir, S.Kom., Ph.D.**

NIP 198602272019031006

**Program Studi S1 Teknik Informatika**

Departemen Teknik Informatika

Fakultas Teknologi Elektro dan Informatika Cerdas

Institut Teknologi Sepuluh Nopember

Surabaya

1

2024



**TUGAS AKHIR – EF234801**

**PERBANDINGAN KINERJA WORD EMBEDDING DAN PEMODELAN TOPIK DALAM IDENTIFIKASI TOPIK TUGAS AKHIR BERDASARKAN JUDUL DAN ABSTRAK PENELITIAN**

**DAVID FISCHER SIMANJUNTAK**

NRP 5025201123

Dosen Pembimbing I

**Prof. Dr. Eng. Nanik Suciati, S.Kom., M.Kom.**

NIP 197104281994122001

Dosen Pembimbing II

**Hadziq Fabroyir, S.Kom., Ph.D.**

NIP 198602272019031006

**Program Studi S1 Teknik Informatika**

Departemen Teknik Informatika

Fakultas Teknologi Elektro dan Informatika Cerdas

Institut Teknologi Sepuluh Nopember

Surabaya

2024



**FINAL PROJECT – IF184802**

**COMPARISON OF WORD EMBEDDING AND TOPIC MODELING PERFORMANCE IN IDENTIFYING THESIS TOPICS BASED ON RESEARCH TITLES AND ABSTRACS**

**DAVID FISCHER SIMANJUNTAK**

NRP 5025201123

Advisor I

**Prof. Dr. Eng. Nanik Suciati, S.Kom., M.Kom.**

NIP 197104281994122001

Advisor II

**Hadziq Fabroyir, S.Kom., Ph.D.**

NIP 198602272019031006

**Study Program Bachelor of Informatics**

Department of Informatics

Faculty of Intelligent Electrical and Informatics Technology

Institut Teknologi Sepuluh Nopember

Surabaya

2024

# LEMBAR PENGESAHAN

**PERBANDINGAN KINERJA WORD EMBEDDING DAN PEMODELAN TOPIK DALAM IDENTIFIKASI TOPIK TUGAS AKHIR BERDASARKAN JUDUL DAN ABSTRAK PENELITIAN**

**TUGAS AKHIR**

Diajukan untuk memenuhi salah satu syarat

Memperoleh gelar Sarjana Komputer pada

Program Studi S-1 Teknik Informatika

Departemen Teknik Informatika

Fakultas Teknologi Elektro dan Informatika Cerdas

Institut Teknologi Sepuluh Nopember

Oleh : **DAVID FISCHER SIMANJUNTAK**

NRP. 5025201123

Disetujui oleh Tim Penguji Tugas Akhir:

|  |  |  |
| --- | --- | --- |
| 1. | Prof. Dr. Eng. Nanik Suciati, S.Kom., M.Kom. | Pembimbing |
| 2. | Hadziq Fabroyir, S.Kom., Ph.D. | Ko-pembimbing |
| 3. | Nama dan gelar penguji | Penguji |
| 4. | Nama dan gelar penguji | Penguji |
| 5. | Nama dan gelar penguji | Penguji |

**SURABAYA**

**Bulan, Tahun**

**APPROVAL SHEET**

**COMPARISON OF WORD EMBEDDING AND TOPIC MODELING PERFORMANCE IN IDENTIFYING THESIS TOPICS BASED ON RESEARCH TITLES AND ABSTRACTS**

**FINAL PROJECT**

Submitted to fulfill one of the requirements

for obtaining a degree Computer Science at

Undergraduate Study Program of Informatics Engineering

Department of Informatics Engineering

Faculty of Intelligent Electrical and Informatics Technology

Institut Teknologi Sepuluh Nopember

By: **DAVID FISCHER SIMANJUNTAK**

NRP. 5025201123

Approved by Final Project Examiner Team:

|  |  |  |
| --- | --- | --- |
| 1. | Prof. Dr. Eng. Nanik Suciati, S.Kom., M.Kom. | Advisor |
| 2. | Hadziq Fabroyir, S.Kom., Ph.D. | Co-Advisor |
| 3. | Name of Examiner and academic title | Examiner |
| 4. | Name of Examiner and academic title | Examiner |
| 5. | Name of Examiner and academic title | Examiner |

**SURABAYA**

**Month, Year**

# PERNYATAAN ORISINALITAS

Yang bertanda tangan di bawah ini:

Nama mahasiswa / NRP : David Fischer Simanjuntak / 5025201123

Departemen : Teknik Informatika

Dosen Pembimbing I / NIP : Prof. Dr. Eng. Nanik Suciati, S.Kom., M.Kom. / 197104281994122001

Dosen Pembimbing II / NIP : Hadziq Fabroyir, S.Kom., Ph.D. / 198602272019031006

Dengan ini menyatakan bahwa Tugas Akhir dengan judul “Perbandingan Kinerja Word Embedding dan Pemodelan Topik Dalam Identifikasi Topik Tugas Akhir Berdasarkan Judul dan Abstrak Penelitian” adalah hasil karya sendiri, bersifat orisinal, dan ditulis dengan mengikuti kaidah penulisan ilmiah.

Bilamana di kemudian hari dietemukan ketidaksesuaian dengan pernyataan ini, maka saya bersedia menerima saksi sesuai dengan ketentuan yang berlaku di Institut Teknologi Sepuluh Nopember.

|  |  |
| --- | --- |
|  | Surabaya,  Mahasiswa  David Fischer Simanjuntak  NRP. 5025201123 |
| Mengetahui |  |
| Dosen Pembimbing I | Dosen Pembimbing II |
|  |  |
| Prof. Dr. Eng. Nanik Suciati, S.Kom., M.Kom. | Hadziq Fabroyir, S.Kom., Ph.D. |
| NIP. 197104281994122001 | NIP. 198602272019031006 |

**STATEMENT OF ORIGINALITY**

The undersigned below:

Name of student / NRP : David Fischer Simanjuntak / 5025201123

Department : Informatics

Advisor I / NIP : Prof. Dr. Eng. Nanik Suciati, S.Kom., M.Kom. / 197104281994122001

Advisor II / NIP : Hadziq Fabroyir, S.Kom., Ph.D. / 198602272019031006

Hereby declare that Final Project with the title of “Comparison of Word Embedding and Topic Modeling Performance in Identifying Thesis Topics Based on Research Titles and Abstracts” is the result of my own work, is original, and is written by following the rules of scientific writing.

If in the future there is a discrepancy with this statement, then I am willing to accept sanctions

in accordance with the provisions that apply at Institut Teknologi Sepuluh Nopember.

|  |  |
| --- | --- |
|  | Surabaya,  Student  David Fischer Simanjuntak  NRP. 5025201123 |
| Acknowledge |  |
| Advisor I | Advisor II |
|  |  |
| Prof. Dr. Eng. Nanik Suciati, S.Kom., M.Kom. | Hadziq Fabroyir, S.Kom., Ph.D. |
| NIP. 197104281994122001 | NIP. 198602272019031006 |

**PERBANDINGAN KINERJA WORD EMBEDDING DAN PEMODELAN TOPIK DALAM IDENTIFIKASI TOPIK TUGAS AKHIR BERDASARKAN JUDUL DAN ABSTRAK PENELITIAN**

**Nama Mahasiswa / NRP : David Fischer Simanjuntak / 5025201123**

**Departemen : Teknik Informatika FTEIC - ITS**

**Dosen Pembimbing I : Prof. Dr. Eng. Nanik Suciati, S.Kom., M.Kom.**

**Dosen Pembimbing II : Hadziq Fabroyir, S.Kom., Ph.D**

# ABSTRAK

**Abstrak**

Identifikasi Topik Tugas Akhir memegang peranan penting dalam membantu mahasiswa menemukan proyek penelitian yang sesuai dengan minat dan keterampilan mahasiswa. Penelitian ini membandingkan kinerja dari tiga metode *Word Embedding*, yaitu Word2Vec, GloVe, dan FastText dengan tiga metode pemodelan topik, yaitu *Latent Semantic Analysis* (LSA), *Non-negative Matrix Factorization* (NMF), dan *Latent Dirichlet Allocation* (LDA) untuk mengidentifikasi topik Tugas Akhir. Implementasi *Word Embedding* melibatkan representasi numerik kata-kata menggunakan model Word2Vec, GloVe, dan FastText, sementara pemodelan topik menggunakan LSA, NMF, dan LDA. Untuk mengevaluasi kinerja metode-metode tersebut menggunakan metrik *Coherence Score*, *Perplexity*, dan *Word Similarity*. Data yang digunakan mencakup judul dan abstrak dari 301 Tugas Akhir mahasiswa Departemen Teknik Informatika Universitas Surabaya dari tahun 2016-2021. Hasil penelitian ini memberikan pemahaman mengenai keunggulan dan kelemahan masing-masing metode dalam konteks identifikasi topik Tugas Akhir.

**Kata kunci: Identifikasi Topik, *Word Embedding*, Pemodelan Topik, *Coherence Score*, *Perplexity*, *Word Similarity*.**

**COMPARISON OF WORD EMBEDDING AND TOPIC MODELING PERFORMANCE IN IDENTIFYING THESIS TOPICS BASED ON RESEARCH TITLES AND ABSTRACTS**

**Student Name / NRP : David Fischer Simanjuntak / 5025201123**

**Department : Teknik Informatika FTEIC - ITS**

**Advisor I : Prof. Dr. Eng. Nanik Suciati, S.Kom., M.Kom.**

**Advisor II : Hadziq Fabroyir, S.Kom., Ph.D**

**Abstract**

The identification of Final Project topics plays a crucial role in assisting students in finding research projects that align with their interests and skills. This study compares the performance of three Word Embedding methods, namely Word2Vec, GloVe, and FastText, with three topic modeling methods, namely Latent Semantic Analysis (LSA), Non-negative Matrix Factorization (NMF), and Latent Dirichlet Allocation (LDA), to identify Final Project topics. The implementation of Word Embedding involves numerical representations of word using Word2Vec, GloVe, and FastText models, while Topic Modeling is carried out through LSA, NMF, and LDA. Performance evaluation is conducted using metrics such as Coherence Score, Perplexity, and Word Similarity. The data used includes titles and abstracts from 301 Final Projects of students from the Departement of Informatics Engineering, University of Surabaya, spanning from 2016 to 2021. The results of this research provide depth insights into the strengths and weaknesses of each method in the context of Final Project topic identification.

**Keywords: Topic Identification, Word Embedding, Topic Modeling, Coherence Score, Perplexity, Word Similarity.**

# KATA PENGANTAR

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

# DAFTAR ISI

[LEMBAR PENGESAHAN v](#_Toc187033442)

[PERNYATAAN ORISINALITAS ix](#_Toc187033443)

[ABSTRAK xiii](#_Toc187033444)

[KATA PENGANTAR xvii](#_Toc187033445)

[DAFTAR ISI xix](#_Toc187033446)

[DAFTAR GAMBAR/GRAFIK/DIAGRAM xxi](#_Toc187033447)

[DAFTAR TABEL xxii](#_Toc187033448)

[DAFTAR SIMBOL (jika ada) xxiii](#_Toc187033449)

[DAFTAR SINGKATAN xxiv](#_Toc187033450)

[BAB I PENDAHULUAN 1](#_Toc187033451)

[1.1 Latar belakang 1](#_Toc187033452)

[1.2 Rumusan Permasalahan 1](#_Toc187033453)

[1.3 Batasan Masalah 2](#_Toc187033454)

[1.4 Tujuan 2](#_Toc187033455)

[1.5 Manfaat 2](#_Toc187033456)

[BAB II TINJAUAN PUSTAKA 3](#_Toc187033457)

[2.1 Penelitian Terkait 3](#_Toc187033458)

[2.1.1 Using Topic Modeling Methods for Short-Text Data: A Comperative Analysis 4](#_Toc187033459)

[2.1.2 Perbandingan Kinerja Word Embedding Word2Vec, GloVe, dan FastText pada Klasifikasi Teks 4](#_Toc187033460)

[2.1.3 Performance Comparison of Topic Models for Text Classification 4](#_Toc187033461)

[2.1.4 Improving Topic Modeling with Pretrained Word Embeddings 5](#_Toc187033462)

[2.1.5 A Comparative Study of Word Embeddings and Topic Models on Text Clustering 5](#_Toc187033463)

[2.2 Dasar Teori 5](#_Toc187033464)

[2.2.1 Identifikasi Topik 5](#_Toc187033465)

[2.2.2 Natural Language Processing 5](#_Toc187033466)

[2.2.3 Word Embedding 6](#_Toc187033467)

[2.2.4 Pemodelan Topik 10](#_Toc187033468)

[2.2.5 Metode Evaluasi 13](#_Toc187033469)

[BAB III METODOLOGI 16](#_Toc187033470)

[3.1 Metode yang dirancang 16](#_Toc187033471)

[3.2 Peralatan pendukung 16](#_Toc187033472)

[3.3 Implementasi 16](#_Toc187033473)

[BAB IV HASIL DAN PEMBAHASAN 17](#_Toc187033474)

[4.1 Hasil eksperimen 17](#_Toc187033475)

[4.2 Pembahasan/Diskusi 17](#_Toc187033476)

[BAB V KESIMPULAN DAN SARAN 18](#_Toc187033477)

[5.1 Kesimpulan 18](#_Toc187033478)

[5.2 Saran (jika dianggap perlu) 18](#_Toc187033479)

[DAFTAR PUSTAKA 19](#_Toc187033480)

[LAMPIRAN-LAMPIRAN ATAU APPENDIKS (jika ada) 22](#_Toc187033481)

[BIODATA PENULIS 23](#_Toc187033482)

# DAFTAR GAMBAR/GRAFIK/DIAGRAM

[**Gambar 2. 1** Rangkaian Proses Word Embedding (Liu dkk., 2017) 4](#_Toc187021690)

[**Gambar 2. 2** Model CBOW pada Word2Vec (Mikolov et al., 2013) 5](#_Toc187021691)

[**Gambar 2. 3** Model Skip-Grams pada Word2Vec (Mikolov et al., 2013) 6](#_Toc187021692)

[**Gambar 2. 4** Alur Model LDA (Blei et al., 2003) 9](#_Toc187021693)

# DAFTAR TABEL

[**Tabel 2. 1** Perbandingan Penelitian Terkait 2](#_Toc187021910)

# DAFTAR SIMBOL (jika ada)

# DAFTAR SINGKATAN

CBOW : Continuous Bag-of-Words

LDA : Latent Dirichlet Allocation

LSA : Latent Semantic Analysis

LSI : Latent Semantic Indexing

NLP : Natural Language Processing

NMF : Non-negative Matrix Factorization

PCA : Principal Component Analysis

RP : Random Projection

SVD : Singular Value Decomposition

TA : Tugas Akhir

TF-IDF : Term Frequency-Inverse Document Frequency

# BAB I PENDAHULUAN

## Latar belakang

Tugas Akhir (TA) merupakan karya ilmiah yang dibuat oleh mahasiswa sarjana yang melibatkan penelitian lapangan dan/atau penelitian literatur (Sari, 2020). Topik penelitian dalam Tugas Akhir menjadi elemen penting karena menentukan arah dan cakupan penelitian yang akan dilakukan oleh mahasiswa. Dalam proses pemilihan topik, mahasiswa sering merujuk pada judul dan abstrak penelitian sebelumnya untuk menemukan topik yang relevan dengan minat dan kemampuan mereka. Namun, banyaknya judul dan abstrak penelitian yang tersedia dapat menjadi tantangan bagi mahsiswa dalam menentukan topik yang sesuai. Oleh karena itu, diperlukan pendekatan yang mampu mengidentifikasi dan mengelompokkan topik penelitian berdasarkan kemiripan judul dan abstrak. Salah satu pendekatan yang dapat digunakan adalah kombinasi metode *Word Embedding* dan Pemodelan Topik, yang memungkinkan analisis yang lebih mendalam terhadap data teks untuk menghasilkan topik-topik utama dari kumpulan data judul dan abstrak penelitian. Langkah awal dalam mengelompokkan topik adalah identifikasi topik, yang merupakan proses untuk menemukan tema utama dari teks yang tidak terstruktur dengan cara menggabungkan informasi yang tersebar diberbagai dokumen (Daniel Veles et al., 2020). Salah satu pendekatan yang umum digunakan untuk identifikasi topik adalah *Word Embedding*. Metode ini mampu menangkap makna semantik kata-kata dalam berbagai konteks, sehingga sangat efektif untuk tugas analisis teks (Pouriyeh et al., 2018). Selain itu, metode lain yang juga populer adalah Pemodelan Topik. Pemodelan topik dapat digunakan untuk menemukan struktur laten dalam data teks berukuran besar dan mengelompokkan dokumen berdasarkan distribusi topik (Yakunin et al., 2020).

*Word Embedding* merupakan teknik representasi numerik yang bertujuan memahami hubungan semantik antar kata. Metode ini banyak digunakan dalam tugas *Natural Language Processing* (NLP), seperti klasifikasi teks, analisis sentimen, dan pembuatan teks otomatis (Lebret et al., 2016). Terdapat tiga metode *Word Embedding* yang umum digunakan, yaitu Word2Vec, GloVe, dan FastText, yang masing-masing memiliki pendekatan unik dalam menangkap hubungan antar kata. Di sisi lain, Pemodelan Topik adalah metode yang dirancang untuk mengidentifikasi topik laten dalam dokumen dengan menganalisis distribusi kata dalam korpus teks (Rani dan Kumar, 2021). Tiga metode utama yang sering digunakan dalam Pemodelan Topik adalah *Latent Dirichlet Allocation* (LDA), *Non-negative Matrix Factorization* (NMF), dan *Latent Semantic Analysis* (LSA).

Penelitian ini bertujuan untuk membandingkan kinerja dari *Word Embedding* dan Pemodelan Topik dalam mengidentifikasi topik dari judul dan abstrak Tugas Akhir. Untuk evaluasi, digunakan tiga metrik evaluasi, seperti *Coherence Score*, *Perplexity*, dan *Word Similarity*. Hasil dari penelitian ini diharapkan dapat memberikan wawasan yang mendalam mengenai keunggulan dan kelemahan masing-masing metode dalam konteks identifikasi topik Tugas Akhir mahasiswa.

## Rumusan Permasalahan

Rumusan masalah yang akan dibahas dalam tugas akhir ini sebagai berikut:

1. Bagaimana membangun model identifikasi topik Tugas Akhir menggunakan Word Embedding berdasarkan data judul TA?
2. Bagaimana membangun model identifikasi topik Tugas Akhir menggunakan Word Embedding berdasarkan data asbtrak TA?
3. Bagaimana membangun model identifikasi topik Tugas Akhir menggunakan Pemodelan Topik berdasarkan data judul TA?
4. Bagaimana membangun model identifikasi topik Tugas Akhir menggunakan Pemodelan Topik berdasarkan data abstrak TA?
5. Bagaimana mengevaluasi hasil identifikasi topik Tugas Akhir?

## Batasan Masalah

Permasalah yang dibahas pada Tugas Akhir ini memiliki beberapa batasan, diantaranya sebagai berikut:

1. Dataset yang digunakan adalah data judul dan abstrak dari 301 Tugas Akhir mahasiswa Departemen Teknik Informatika Universitas Surabaya dari tahun 2016-2021.
2. Metode Word Embedding yang digunakan adalah Word2Vec, GloVe, dan FastText.
3. Metode Pemodelan Topik yang digunakan adalah LDA, NMF, dan LSA.
4. Metode evaluasi yang digunakan adalah Coherence Score, Perplexity, dan Word Similarity.

## Tujuan

Tujuan dari Tugas Akhir ini adalah membandingkan kinerja Word Embedding dan Pemodelan Topik dalam identifikasi topik Tugas Akhir berdasarkan data judul dan abstrak penelitian.

## Manfaat

Tugas Akhir ini dapat bermanfaat dalam memberikan pemahaman mengenai Word Embedding dan Pemodelan Topik. Tugas Akhir ini bisa dimanfaatkan juga untuk pembuatan sistem rekomendasi topik TA mahasiswa untuk penelitian berikutnya.

# BAB II TINJAUAN PUSTAKA

## Penelitian Terkait

Pada bagian ini akan dipaparkan beberapa penelitian-penelitian sebelumnya yang akan digunakan sebagai referensi dan pertimbangan dalam pengerjaan Tugas Akhir. Tabel 2.1 berisi beberapa penelitian yang berkaitan dengan *Word Embedding* dan Pemodelan Topik.

**Tabel 2. 1** Perbandingan Penelitian Terkait

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Nama Peneliti** | **Judul** | **Metode** | **Dataset** | **Hasil** |
| Rania Albalawi,  Tet Hin Yeap,  Morad Benyoucef (2020) | *Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis* | LSA, LDA, NMF, RP, PCA | 20 *Newsgroup*, *Facebook* | LDA dan NMF lebih efektif pada teks pendek dibandingkan LSA, RP, dan PCA. Hasilnya lebih bermakna dan berkinerja baik untuk analisis teks pendek. |
| Arliyanti Nurdin,  Bernadus Anggo Seno Aji,  Anugrayani Bustamin,  Zaenal Abidin  (2020) | Perbandingan Kinerja *Word Embedding* *Word2Vec*, *GloVe*, dan *FastText* pada Klasifikasi Teks | *Word2Vec, GloVe, FastText* | *Reuters Newswire*, 20 *Newsgroup* | *FasText* unggul dalam menangkap semantik kata, terutama untuk kata yang tidak ada di korpus, karena teknik n-gram yang digunakan. |
| Arora, P.,  Li, J.,  Liang, L (2017) | *Performance Comparison of Topic Models for Text Classification* | LDA, NMF, Doc2Vec | 20 *Newsgroup*, *DBLP dataset* | LDA unggul dalam menghasilkan distribusi topik yang stabil, sementara NMF lebih efisien pada teks pendek dengan dimensi rendah. |
| Dong, Cai.,  Liu, Z.,  Sun, M.  (2022) | *Improving Topic Modeling with Pretrained Word Embeddings* | Word2Vec, GloVe, LDA | 20 *Newsgroup, Wikipedia* | Pretrained *Word Embedding* seperti Word2Vec dan GloVe meningkatkan koherensi topik dan stabilitas distribusi topik dalam model LDA. |
| Chen, J.,  Qiu, W.,  Zhang, Y.  (2019) | *A Comparative Study of Word Embeddings and Topic Models on Text Clustering* | FastText, GloVe, NMF | *Reuters Newswire, StackExhange* | FastText unggul dalam variasi semantik tinggi, sementara NMF lebih efisien dalam pola laten data berstruktur sederhana. |

### 2.1.1 Using Topic Modeling Methods for Short-Text Data: A Comperative Analysis

Penelitian (Rania Albalawi et al., 2020) bertujuan untuk mengevaluasi dan membandingkan performa dari lima metode dalam *topic modeling*, yaitu LSA, LDA, NMF, RP dan PCA pada data teks pendek. Penelitian ini berfokus kepada kemampuan metode dalam menghasilkan topik yang bermakna dan relevan untuk data informasi yang terbatas. Dataset yang digunakan pada penelitian ini terdiri dari 20 *Newsgroup* data koleksi berita online dan data percakapan pendek dari jejaring sosial Facebook. Hasil dari penelitian menunjukkan LDA dan NMF lebih aktif pada teks pendek dan menghasilkan topik yang lebih bermakna dan relevan dibandingkan metode lainnya.

### 2.1.2 Perbandingan Kinerja Word Embedding Word2Vec, GloVe, dan FastText pada Klasifikasi Teks

Penelitian (Nurdin et al., 2020) membandingkan kinerja tiga metode *Word Embedding*, yaitu *Word2Vec*, *GloVe*, dan *FastText* dalam klasifikasi teks. Penelitian ini memberikan pendekatan komperatif dengan memerhatikan akurasi model, generalisasi model, dan efisiensi dari tiap metode yang digunakan. Dataset yang digunakan dalam penelitian ini terdiri dari *Reuters Newswire*, yang mencakup lebih dari 18.000 dokumen berita dari berbagai topik. Evaluasi dalam penelitian ini dilakukan dengan mengubah kata menjadi vektor fitur yang akan digunakan untuk klasifikasi teks. Performa dari ketiga metode akan dibandingkan berdasarkan akurasi hasil klasifikasi, efisiensi waktu pelatihan, serta kemampuan metode dalam menangkap makna kata yang tidak terdapat dalam korpus. Hasil yang didapatkan menunjukkan *FastText* unggul dalam menangkap makna semantik untuk kata-kata yang jarang muncul, *Word2Vec* memberikan hasil penelitian yang lebih cepat, serta *GloVe* menghasilkan vektor yang stabil terutama untuk dataset yang besar.

### 2.1.3 Performance Comparison of Topic Models for Text Classification

Penelitian (Arora et al., 2017) membandingkan kinerja metode dari tiga *topic modeling*, yaitu LDA, NMF, dan Doc2Vec. Penelitian ini menggunakan dataset dari 20 *Newsgroup* dan *DBLP*, yang mencakup koleksi dokumen teks dengan beragam topik. Evaluasi dilakukan berdasarkan stabilitas distribusi topik, kecepatan pemrosesan, serta efektivitas dalam menghasilkan representasi topik yang relevan. Hasil penelitian menunjukkan bahwa LDA unggul dalam menghasilkan distribusi topik yang stbail dan bermakna, menjadikannya pilihan yang cocok untuk analisis teks berukuran besar. Sementara itu, NMF lebih efisien dalam analisis teks pendek, terutama untuk data dengan dimensi rendah. Doc2Vec menawarkan pendekatan berbasis *embedding* yang fleksibel, tetapi memiliki kelemahan dalam stabilitas distribusi topik.

### 2.1.4 Improving Topic Modeling with Pretrained Word Embeddings

Penelitian (Dong et al., 2022) mengeksplorasi penggunaan *Word Embedding* pretrained, seperti Word2Ve dan GloVe , untuk meningkatkan kinerja model *Latent Dirichlet Allocation* (LDA). Dataset yang digunakan adalah 20 *Newsgroup* dan *Wikipedia*, yang mencakup teks dari berbagai domain. Penelitian ini menambahkan representasi semantik kata dari *Word Embedding* ke dalam proses pelatihan LDA untuk menghasilkan distribusi topik yang lebih koheren. Hasil penelitian menunjukkan bahwa pendekatan ini tidak hanya meningkatkan koherensi topik, tetapi juga menghasilkan stabilitas distribusi topik yang lebih baik dibandingkan dengan LDA standar. Hal ini menunjukkan pentingnya menggabungkan representasi semantik dalam model *generatif* untuk analisis teks.

### 2.1.5 A Comparative Study of Word Embeddings and Topic Models on Text Clustering

Penelitian (Chen et al., 2019) membandingkan kinerja *Word Embedding* dan *Topic Modeling* pada tugas pengelompokan teks. Metode yang digunakan mencakup FastText, GloVe, dan *Non-negative Matrix Factorization* (NMF). Dataset yang digunakan adalah *Reuters Newswire* dan *StackExchange*, yang mewakili teks dengan variasi semantik tinggi dan struktur sederhana. Penelitian ini mengevaluasi kemampuan metode untuk menghasilkan representasi topik yang bermakna dan melakukan pengelompokan dokumen secara efektif. Hasilnya menunjukkan bahwa FastText unggul dalam menangkap variasi semantik tinggi karena pendekatan berbasis *subword*, sementara NMF lebih efisien dalam mengidentifikasi pola laten pada data yang memiliki struktur sederhana. GloVe menawarkan hasil yang lebih stabil, tetapi kurang optimal pada teks dengan variasi semantik kompleks.

## Dasar Teori

Pada bab ini akan dijelaskan berbagai konsep yang akan digunakan pada pengerjaan Tugas Akhir ini. Berikut adalah beberapa penjelasan dari konsep yang digunakan.

### 2.2.1 Identifikasi Topik

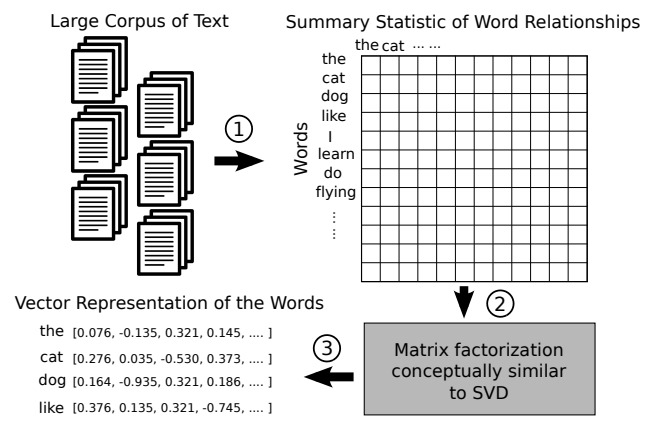
Identifikasi topik adalah proses menentukan topik utama dari suatu data teks dengan menentukan tema atau topik utama dari suatu dokumen. Proses ini membantu dalam memahami struktur informasi pada dokumen, mengelompokkan data teks berdasarkan tema, dan menggali wawasan dari data yang besar dan tidak terstruktur. Identifikasi topik banyak digunakan dalam berbagai bidang, seperti klasifikasi artikel ilmiah, ekstraksi pengetahuan, dan penemuan topik terbaru (De Nooijer et al., 2000). Kemajuan bidang *Natural Language Processing* (NLP), cabang kecerdasan buatan yang berfokus kepada pemrosesan bahasa manusia, memberikan kontribusi signifikan terhadap pengembangan metode identifikasi topik.

### 2.2.2 Natural Language Processing

*Natural language processing* atau NLP merupakan cabang dari kecerdasan buatan yang berfokus kepada interaksi komputer dan bahasa manusia. NLP memungkinkan komputer untuk memahami, menafsirkan, dan menghasilkan teks dalam bahasa manusia secara otomatis. Dalam beberapa dekade terakhir, NLP telah menjadi dasar dari berbagai aplikasi penerjemah mesin, *chatbot,* analisis sentimen, dan sistem pencarian informasi (Manning, Schütze, & Raghavan, 2008). Pendekatan dalam NLP sering memanfaatkan teknik pembelajaran mesin dan *deep learning* untuk menganalisis data teks dengan lebih efisien. Representasi data teks dalam format numerik menjadi kunci dalam pengolahan teks. Dua pendekatan utama yang digunakan dalam NLP untuk analisis teks adalah *word embedding* dan *topic modeling. Word embedding* menyediakan representasi kata dalam bentuk numerik untuk mencerminkan hubungan semantik dan konteks antar kata, sementara *topic modeling* menggunakan informasi tersebut untuk menemukan pola topik dan tema utama dalam teks.

### 2.2.3 Word Embedding

*Word embedding* adalah teknik dalam *Natural Language Processing* (NLP) yang digunakan untuk merepresentasikan kata dalam bentuk vektor numerik. Representasi ini menangkap makna semantik dan hubungan antar kata, di mana kata-kata dengan makna serupa direpresentasikan oleh vektor yang berdekatan. Teknik ini menjadi dasar bagi berbagai aplikasi NLP, seperti analisis sentimen, penerjemahan mesin, dan klasifikasi teks. Penelitian Liu, S. (Liu dkk., 2017) mengilustrasikan alur proses pembentukan *Word Embedding,* seperti yang ditunjukkan pada Gambar 2.1.



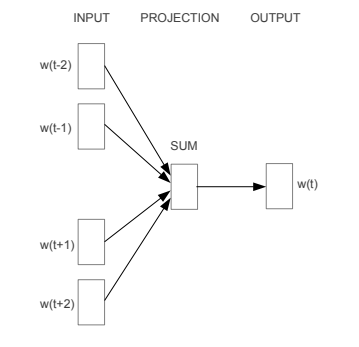
**Gambar 2. 1** Rangkaian Proses Word Embedding (Liu dkk., 2017)

Proses *Word Embedding* dimulai dengan pengumpulan korpus teks besar, yang dirangkum dalam sebuah matriks hubungan antar kata (*co-occurrence matrix*) *M*. Matriks ini berukuran x , dimana adalah jumlah kata unik dalam korpus. Setiap elemen *M* (,) mencatat probabilitas kemunculan bersama antara kata dan dalam konteks yang sama. Pada tahap berikutnya, matriks *M* difaktorkan menggunakan metode seperti *Singular Value Decomposition* (SVD) untuk menghasilkan representasi vektor berdimensi lebih rendah bagi setiap kata. Faktorisasi bertujuan untuk meringkas informasi hubungan antar kata ke dalam bentuk yang lebih efisien dan bermakna. Hasil akhirnya berupa vektor numerik yang merepresentasikan kata-kata dengan makna serupa secara berdekatan dalam ruang vektor. Representasi ini memungkinkan analisis semantik dan sintaksis pada tingkat yang lebih tinggi, tetapi juga membuka jalan bagi pengembangan berbagai metode *Word Embedding* yang lebih efektif. Sejumlah metode telah dirancang untuk menghasilkan representasi kata yang optimal, seperti *Word2Vec*, *GloVe*, dan *FastText*, yang masing-masing memiliki pendekatan unik dalam menghasilkan vektor numerik.

#### 2.2.3.1 Word2Vec

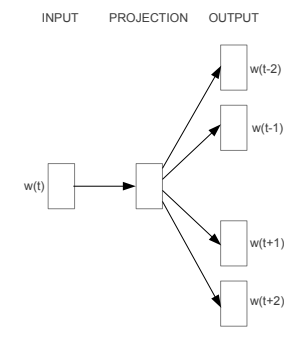
*Word2Vec* adalah salah satu metode populer dalam NLP yang digunakan untuk merepresentasikan kata-kata dalam bentuk vektor numerik berdimensi rendah. Metode ini dirancang oleh Mikolov et al. (2013) untuk mempelajari hubungan semantik dan sintaksis antar kata dari sebuah corpus teks besar. Word2Vec mengubah kata-kata menjadi vektor numerik dalam ruang berdimensi rendah (biasanya 100 hingga 300 dimensi), yang dimana kata-kata dengan makna serupa akan memiliki vektor yang berdekatan. Word2Vec bekerja dengan dua model utama, yaitu *Continuous Bag-of-Words* (CBOW) dan *Skip-Grams*.

*Continuous Bag-of-Words* (CBOW) merupakan salah satu metode dalam *Word2Vec* yang bertujuan untuk memprediksi kata target berdasarkan kata-kata konteks di sekitarnya. CBOW menggunakan kata-kata di sekitar target dalam jendela tertentu (*window size*), seperti dua kata sebelum dan dua kata setelah kata target, sebagai input. Setiap kata konteks diubah menjadi representasi vektor, lalu vektor-vektor ini dirata-rata untuk menghasilkan satu vektor input yang mewakili konteks. Model akan memprediksi kata target yang paling mungkin muncul berdasarkan distribusi probabilitas kosakata yang dihasilkan dari vektor konteks tersebut. CBOW lebih cepat dilatih dibandingkan *Skip-Grams*, sehingga cocok untuk kumpulan data besar karena hanya memprediksi satu kata taget dalam setiap iterasi pelatihan. Gambar 2.2 menunjukkan arsitektur CBOW pada *Word2Vec* pada penelitian Mikolov et al. (2013).



**Gambar 2. 2** Model CBOW pada Word2Vec (Mikolov et al., 2013)

*Skip-Grams* merupakan salah satu metode utama dalam *Word2Vec* yang bertujuan untuk memprediksi kata-kata konteks berdasarkan kata target. Berbeda dengan CBOW, yang memprediksi kata target dari konteksnya, *Skip-Grams* menggunakan kata target sebagai input untuk menghasilkan kata-kata konteks relevan disekitarnya sebagai output. Pada metode ini, kata target terlebih dahulu diubah menjadi representasi vektor melalui lapisan embedding. Vektor ini digunakan untuk memprediksi kata-kata konteks melalui distribusi probabilitas di kosakata. Model dilatih untuk memaksimalkan probabilitas pasangan kata taget-konteks yang sebenarnya (positif) dan meminimalkan pasangan kata target-konteks yang tidak relevan (negatif), seringkali menggunakan teknik seperti *negative sampling* atau *hierarchical softmax* untuk efisiensi pelatihan. Keunggulan *Skip-Grams* terletak pada kemampuannya untuk menangkap hubungan semantik yang lebih mendalam antar kata-kata. Model ini juga unggul dalam memberikan representasi yang lebih bermakna untuk kata-kata yang jarang muncul dalam teks. Namun, metode ini memiliki kecepatan pelatihan yang lebih lambat dibandingkan CBOW karena memprediksi banyak kata konteks untuk setiap kata target. Gambar 2.3 menunjukkan arsitektur *Skip-Grams* pada *Word2Vec* pada penelitian Mikolov et al. (2013).



**Gambar 2. 3** Model Skip-Grams pada Word2Vec (Mikolov et al., 2013)

#### 2.2.3.2 GloVe

*Global Vectors for Word Representation* (GloVe) adalah metode *Word Embedding* yang memetakan kata-kata menjadi vektor numerik berdimensi rendah untuk menangkap hubungan semantik dan sintaksis antar kata. Metode ini menggunakan matriks *co-occurrence* global untuk mempelajari hubungan antar kata dan merepresentasikannya secara bermakna. GloVe dirancang untuk efisiensi tinggi dan menghasilkan representasi kata yang kaya informasi dalam analisis teks berskala besar (Pennington et al., 2014).

Proses pembuatan vektor kata pada GloVe dimulai dengan membangun matriks *co-occurrence* , berukuran x , dimana adalah jumlah kata unik dalam korpus. Setiap elemen mencatat jumlah kemunculan bersama antara kata dan kata dalam konteks yang sama. Hubungan semantik antar kata dievaluasi dengan menggunakan proporsi *co-occurrence*, yang dirumuskan pada persamaan 2.1 berikut:

|  |  |
| --- | --- |
|  | 2.1 |

dengan sebagai probabilitas kata yang muncul dalam konteks kata dan adalah probabilitas kata yang muncul dalam konteks .

Tahap berikutnya adalah penggunaan fungsi objektif untuk meminimalkan kesalahan antara nilai logaritma *co-occurrence* dengan kombinasi linear vektor kata , vektor kata , bias skalar , dan bias skalar yang dituliskan seperti pada persamaan 2.2. Fungsi ini memastikan bahwa kata-kata dengan hubungan semantik serupa memiliki representasi vektor yang berdekatan.

|  |  |
| --- | --- |
|  | 2.2 |

GloVe memiliki beberapa keunggulan, seperti efisiensi waktu pelatihan pada korpus teks skala besar, kemampuan menangkap hubungan semantik global antar kata dengan memanfaatkan matriks *co-occurrence*, serta menghasilkan representasi kata yang kaya dengan informasi semantik dan sintaksis. Namun, GloVe juga memiliki beberapa kelemahan, seperti kebutuhan memori yang besar untuk menyimpan matriks *co-occurrence*, representasi vektor yang bersifat statis sehingga satu kata hanya memiliki satu vektor, serta kurangnya perhatian terhadap urutan kata atau hubungan sintaksis yang lebih spesifik dalam teks.

#### 2.2.3.3 FastText

FastText adalah metode *Word Embedding* yang memperluas pendekatannya dengan memasukkan informasi *subword* dalam proses pelatihan. Dengan pendekatan ini, FastText dapat menangkap struktur internal kata seperti awalan, akhiran, atau bagian-bagian kata lainnya. Selain itu, FastText juga lebih mampu menangkap hubungan semantik antar kata karena representasi vektor yang dihasilkan menjadi lebih kaya (Bojanowski et al., 2016). FastText menggunakan pendekatan *Skip-Grams* sebagai dasar pembentukan modelnya. Model ini bertujuan untuk mempelajari distribusi probabilitas antara kata target dan kata konteks didalam korpus pelatihan. Persamaan yang digunakan dalam model ini dapat dirumuskan pada persamaan 2.3.

|  |  |
| --- | --- |
|  | 2.3 |

Distribusi probabilitas ini digambarkan melalui fungsi *sigmoid* (), yang berperan untuk memetakan nilai antara 0 dan 1, sehingga dapat merepresentasikan probabilitas antara kata target () dan kata konteks (). Representasi kata target dan konteks masing-masing dilambangkan sebagai dan , yang merupakan vektor representasi dari kata-kata tersebut. Pada persamaan tersebut, *D* melambangkan korpus pelatihan, yang terdiri atas pasangan kata target dan kata konteks. Sedangkan merujuk kepada kumpulan kata negatif yang dipilih secara acak dari korpus menggunakan pendekatan *negative sampling* (Goldberg & Levy, 2014). Dengan menggunakan fungsi sigmoid ini, model dapat mengevaluasi hubungan semantik antar kata melalui optimalisasi probabilitas bahwa kata target muncul bersama dengan kata konteks yang relevan di dalam korpus pelatihan.

FastText dapat memperluas modelnya dengan pendekatan berbasis *subword*. Dalam pendekatan ini, kata target () tidak hanya direpresentasikan sebagai satu vektor tunggal, tetapi sebagai kombinasi dari vektor *n-grams* (*subword*) yang menyusun kata tersebut. Dengan kata lain, representasi kata target adalah hasil penjumlahan dari semua vektor *subword* yang membentuk kata tersebut. Persamaan yang digunakan pada persamaan 2.4.

|  |  |
| --- | --- |
|  | 2.4 |

Dalam persamaan tersebut, merepresentasikan himpunan *n-grams* atau *subword* yang membentuk kata , sedangkan melambangkan vektor representasi dari setiap *subword* . Sebagai contoh kata “running” dapat dipecah menjadi beberapa *n-grams*, seperti “run”,”unn”,”nni”,”nin”, dan ”ing”. Setiap *n-grams* memiliki vektor tersendiri, yang kemudian dikombinasikan untuk membentuk vektor keseluruhan kata “running”.

Melalui pendekatan *subword* ini, FastText memiliki keunggulan yang signifikan, seperti kemampuan untuk menangkap makna kata yang tidak ada dalam korpus pelatihan (*out-of-vocabulary*) dengan memanfaatkan representasi *n-grams*. Selain itu, pendeketan ini menghasilkan vektor yang lebih akurat dengan menangkap struktur internal kata, sekaligus efisiensi komputasi dengan menggunakan *negative sampling* yang memungkinkan pelatihan pada korpus besar dengan lebih cepat. Namun, FastText juga memiliki keterbatasan, seperti kebutuhan memori tambahan untuk menyimpan *n-grams* dan representasi *subword*, serta sifat vektor yang bersifat statis yang menyebabkan kata dengan makna berbeda dalam konteks yang berbeda direpresentasikan dengan satu vektor yang sama.

### 2.2.4 Pemodelan Topik

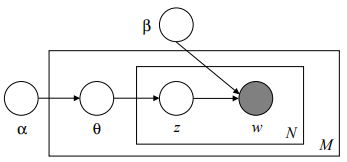
Pemodelan topik adalah pendekatan Natural Language Processing(NLP) yang bertujuan untuk mengekstrak topik, mengelompokkan topik, dan mengidentifikasi topik dalam sekumpulan dokumen dan teks. Metode ini membantu dalam menganalisis dan memahami struktur dan tema utama yang terkandung didalam teks (Blei et al., 2003). Untuk mencapai tujuan, pemodelan topik melibatkan serangkaian langkah sistematis dalam pengerjaan nya. Proses diawali dengan pra-pemrosesan data, seperti pembersihan teks melalui tokenisasi, penghapusan stopwords, dan stemming atau lemmatization. Setelah itu, teks akan direpresentasikan dalam bentuk matriks kata-dokumen. Model topic modeling, seperti Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), atau Latent Semantic Analysis (LSA), kemudian dipilih dan dilatih untuk menghasilkan distribusi topik dalam dokumen serta kata-kata dalam setiap topik. Hasil pelatihan model akan dievaluasi menggunakan metrik, seperti coherence score untuk menilai relevansi topik atau perplexity untuk mengukur kualitas probabilistik model. Tahap akhir melibatkan interpretasi dan visualisasi hasil, misalnya melalui word cloud atau diagram distribusi topik, untuk menyajikan pola tematik yang ditemukan.

Pemodelan topik memiliki aplikasi yang luas, seperti memberikan rekomendasi konten yang relevan berdasarkan referensi pengguna, mengidentifikasi topik utama dari media sosial atau berita, serta mengelompokkan dokumen berdasarkan kemiripan topik untuk mempermudah pengelolaan informasi. Selain itu, pendekatan ini memiliki manfaat lain dalam mengungkap wawasan tersembunyi dari data teks yang besar dan tidak terstruktur dan membantu dalam pengambilan keputusan dalam bidang ekonomi, kesehatan, dan pendidikan.

#### 2.2.4.1 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) adalah metode pemodelan topik berbasis probabilistik yang memodelkan dokumen sebagai campuran distribusi topik, dan topik sebagai campuran distribusi kata (Blei et al., 2003). Dengan pendekatan ini, LDA dapat menemukan struktur laten dari dokumen yang tidak terstruktur tanpa memerlukan label data yang disebut *unsupervised learning*. Alur generatif model pada LDA dapat dilihat pada Gambar 2.4.

LDA menggunakan pendekatan generatif yang diawali dengan parameter prior , yaitu distribusi *Dirichlet* yang mengatur distribusi topik dalam dokumen (). Nilai menentukan distribusi topik di setiap dokumen, apakah lebih seragam atau berpusat pada beberapa topik saja. Untuk setiap dokumen , distribusi topik diambil dari distribusi *Dirichlet* yang ditentukan oleh . Selanjutnya, untuk setiap kata dalam dokumen , sebuah topik laten dipilih dari distribusi . Setelah topik ditentukan, kata yang diamati dipilih berdasarkan distribusi kata-kata yang terkait dengan topik, yang diatur oleh prior . Parameter adalah distribusi *Dirichlet* yang mengontrol bagaimana kata-kata terdistribusi dalam setiap topik. Proses ini dilakukan secara *iteratif* untuk setiap kata dalam dokumen () dan diulang untuk seluruh dokumen dalam korpus ().



**Gambar 2. 4** Alur Model LDA (Blei et al., 2003)

Penelitian oleh Wahid, Ahmad (2021) menggunakan LDA untuk analisis tren dalam media sosial dan berita. Penelitian ini menunjukkan bahwa LDA efektif dalam mengidentifikasi topik-topik dominan yang berkembang dalam teks yang tidak terstruktur. Zhang, Mei (2023) dalam penelitiannya menggunakan LDA dalam klasifikasi dokumen hukum. Hasilnya menunjukkan bahwa LDA mampu membentuk klaster topik dan menghasilkan visualisasi distribusi topik yang membantu pengambilan keputusan hukum.

#### 2.2.4.2 Non-Negative Matrix Factorization (NMF)

*Non-negative Matrix Factorization* adalah teknik dekomposisi matriks yang bertujuan untuk menemukan representasi laten dari data dengan memfaktorkan matriks menjadi dua matriks yang lebih kecil, yaitu matriks basis () dan matriks koefisien (), dengan semua elemen bernilai non-negatif (Lee & Seung., 1999). Dalam konteks pemodelan topik, NMF digunakan untuk mengekstrak topik laten dari dokumen teks dengan merepresentasikan setiap dokumen sebagai kombinasi linier dari topik-topik tersebut. Hubungan ini dapat dituliskan seperti pada persamaan 2.5 berikut.

|  |  |
| --- | --- |
|  | 2.5 |

merupakan matriks dokumen kata-kata ( x ), yang merepresentasikan frekuensi kata dalam dokumen. adalah matriks basis ( x ) yang menunjukkan distribusi dokumen terhadap topik. H adalah matriks koefisien ( x ) yang menunjukkan distribusi kata terhadap topik, sedangkan merupakan jumlah topik laten.

Proses NMF dimulai dengan tahapan representasi data, dimana dokumen teks dikonversi menjadi sebuah matriks . Dalam matriks ini, setiap baris merepresentasikan sebuah dokumen, setiap kolom merepresentasikan kata, dan elemen menunjukkan frekuensi kemunculan kata dalam dokumen . Selanjutnya, dilakukan tahap faktorisasi matriks, yaitu memecah matriks menjadi dua matriks yang lebih kecil, yaitu matriks dan . Matriks memiliki dimensi x , dimana adalah jumlah dokumen dan adalah jumlah topik latern, sedangkan matriks memiliki dimensi x , dimana adalah jumlah kata dalam korpus. Proses faktorisasi ini dilakukan menggunakan algoritma optimasi tertentu, seperti multiplicative update rules yang diusulkan oleh Lee & Seung (2001) atau metode alternating least squares (ALS). Nilai-nilai dalam dan dipertahankan non-negatif, yang memudahkan interpretasi karena distribusi probabilitas tidak memiliki nilai negatif. Setelah faktorisasi, tahap optimasi dilakukan untuk meminimalkan kesalahan rekonstruksi antara matriks asli dengan hasil perkalian matriks dan . Fungsi yang diminimalkan adalah kuadrat kesalah rekonstruksi, yang dinyatakan pada persamaan 2.6. Proses ini dilakukan secara iteratif untuk memastikan bahwa dan memberikan hasil rekonstruksi yang mendekati matriks dengan akurasi tinggi. Pada tahap terakhir, hasil faktorisasi diinterpretasikan untuk mengidentifikasi distribusi topik dan kata-kata dalam dokumen. Matriks memberikan distribusi dokumen, sedangkan matriks menunjukkan distribusi kata untuk setiap topik. Elemen-elemen dalam matriks dan digunakan untuk menentukan pentingnya suatu topik dalam dokumen tertentu serta kata-kata yang paling relevan untuk setiap topik.

|  |  |
| --- | --- |
|  | 2.6 |

NMF memiliki beberapa kelebihan, seperti interpretasi yang intuitif karena sifat non-negatif nya, efesiensi dalam analisis data besar, serta fleksibilitas untuk berbagai jenis data (Cheng et al., 2022). Namun, metode ini juga memiliki beberapa keterbatasan, seperti ketergantungan pada inisialisasi dan , risiko terjebak pada solusi lokal, serta sulitnya emnentukan jumlah topik () yang optimal (Kim et al., 2019).

#### 2.2.4.3 Latent Semantic Analysis (LSA)

*Latent Semantic Analysis* (LSA), yang juga dikenal sebagai *Latent Semantic Indexing* (LSI), adalah metode berbasis dekomposisi matriks yang bertujuan untuk menemukan hubungan laten antara kata-kata dalam dokumen dan topik-topik yang mendasarinya. Metode ini diperkenalkan oleh Deerwester et al. (1990) untuk mengatasi masalah sinonim (kata berbeda dengan makna serupa) dan polisemik (kata dengan banyak makna) dalam teks. LSA memanfaatkan *Singular Value Decomposition* (SVD) untuk mereduksi dimensi dari matriks kata-dokumen, sehingga hanya informasi semantik yang relevan yang dipertahankan. Pendekatan ini memungkinkan analisis hubungan semantik antar dokumen sekaligus mengurangi *noise* dari data teks yang besar dan tidak terstruktur (Berry et al., 1995).

Proses Utama dalam LSA melibatkan tiga tahap utama yang saling berkaitan. Tahap pertama adalah representasi data, di mana dokumen teks direpresentasikan dalam bentuk matriks istilah-dokumen dengan dimensi x . Dalam matriks ini merepresentasikan jumlah kata unik (*terms*) yang muncul dalam sebuah korpus, sedangkan adalah jumlah dokumen dalam korpus tersebut. Setiap elemen menunjukkan frekuensi atau bobot kemunculan kata dalam dokumen . Pembobotan elemen-elemen pada matriks ini sering kali menggunakan metode seperti *Term Frequency-Inverse Document Frequency* (TF-IDF) untuk memberikan bobot yang lebih proporsional berdasarkan pentingnya kata dalam suatu dokumen dan dalam keseluruhan korpus (Manning et al., 2008). Tahap kedua adalah dekomposisi matriks dengan metode *Singular Value Decomposition* (SVD). Dalam proses ini, matriks difaktorkan menjadi tiga matriks komponen, yaitu matriks , , dan , yang dituliskan pada persamaan 2.7. Matriks berdimensi x memuat distribusi kata terhadap topik laten, matriks berdimensi x berisi nilai *singular* yang mencerminkan relevansi dimensi laten, dan matriks menunjukkan distribusi dokumen terhadap topik laten. Tahap terakhir adalah reduksi dimensi, dimana hanya dimensi laten yang signifikan dipertahankan, menghasilkan representasi teks yang lebih sederhana dan informatif. Hasil akhir dari proses ini adalah matriks yang memuat distribusi kata tehradap dimensi laten, dan matriks yang memuat informasi distribusi dokumen terhadap dimensi laten. Elemen-elemen pada matriks dapat digunakan untuk menentukan pentingnya kata dalam sebuah topik atau relevansi suatu dokumen terhadap dimensi laten tertentu (Berry et al., 1995).

|  |  |
| --- | --- |
|  | 2.7 |

LSA menjadi populer karena mampu mengatasi masalah sinonim dan polisemik dalam analisis teks. Selain itu, LSA cukup stabil secara komputasi dan efektif dalam mengidentifikasi hubungan semantik dalam data teks. Namun, metode ini juga memiliki keterbatasan, terutama asumsi distribusi Gaussian untuk data yang dapat menurunkan performa pada data yang sangat kompleks atau non-linier (Hoffman, 1999).

### 2.2.5 Metode Evaluasi

Metode evaluasi adalah pendekatan yang digunakan untuk menilai kinerja sistem dalam mencapai tujuan atau memenuhi standar tertentu. Dalam penelitian ini, evaluasi digunakan untuk mengukur performa *Word Embedding* dan pemodelan topik dalam menghasilkan topik yang koheren, memprediksi data teks, dan menangkap hubungan semantik antar kata. Tiga metode evaluasi yang digunakan adalah *Coherence Score*, *Perplexity*, dan *Word Similarity*.

#### 2.2.5.1 Coherence Score

*Coherence Score* adalah metrik evaluasi yang digunakan untuk menilai sejauh mana kata-kata dalam suatu topik saling berhubungan secara semantik. Metrik ini memastikan bahwa kata-kata dalam suatu topik tidak hanya muncul bersama tetapi juga memiliki hubungan semantik yang bermakna, sehingga meningkatkan kualitas interpretasi topik (Röder et al., 2015). *Coherence Score* dihitung menggunakan rumus matematis yang mempertimbangkan hubungan antar kata dalam topik yang sama. Persamaan matematis *Coherence Score* dapat dituliskan seperti pada persamaan 2.8. Pada persamaan ini, merepresentasikan jumlah kata dalam topik, sedangkan dan adalah kata-kata dalam topik tersebut. Nilai menunjukkan tingkat hubungan antar kata yang dapat dihitung menggunakan metode seperti *cosine similarity* atau probabilitas *co-occurrence*. Pendekatan populer dalam menghitung *Coherence Score* meliputi metode dan *UMass*. Pendekatan menggabungkan informasi probabilistik dan statistik untuk menghitung koherensi antar kata, sedangkan *UMass* menggunakan log-probabilitas kata-kata dalam topik berdasarkan frekuensi kata dalam korpus. *Coherence Score* memiliki aplikasi luas dalam pemodelan topik, seperti LDA, NMF, dan LSA. Metrik ini membantu mengevaluasi model topik untuk memastikan bahwa topik-topik yang dihasilkan memiliki makna secara semantik.

|  |  |
| --- | --- |
|  | 2.8 |

#### 2.2.5.2 Perplexity

*Perplexity* adalah metrik evaluasi yang digunakan untuk menilai seberapa baik model probabilistik, seperti LDA, dalam memprediksi kata-kata dalam dataset uji. *Perplexity* mencerminkan tingkat ketidakpastian model dalam menghasilkan prediksi. Nilai *Perplexity* yang lebih rendah menunjukkan bahwa model mampu menangkap struktur data teks dengan lebih baik, sedangkan nilai yang tinggi mengindikasikan bahwa model kurang optimal dalam memahami distribusi kata-kata dalam dokumen (Blei et al., 2003). Perhitungan *Perplexity* dirumuskan pada persaman 2.9 berikut.

|  |  |
| --- | --- |
|  | 2.9 |

Pada persamaan ini, adalah jumlah dokumen dalam korpus uji dan merepresentasikan jumlah kata dalam dokumen . Probabilitas dokumen dihitung berdasarkan distribusi kata terhadap topik  dan distribusi topik terhadap dokumen , yang dirumuskan pada persamaan 2.10.

|  |  |
| --- | --- |
|  | 2.10 |

Proses perhitungan dimulai dengan menghitung probabilitas dokumen berdasarkan distribusi kata terhadap topik dan distribusi topik terhadap dokumen. Langkah ini diikuti dengan penerapan logaritma pada probabilitas dokumen untuk memperhalus nilai kecil, yang membuatnya lebih stabil untuk perhitungan skala besar. Terakhir, logaritma probabilitas dinormalisasi dengan rata-rata jumlah kata dalam dokumen untuk mendapatkan nilai *Perplexity*.

*Perplexity* sering digunakan untuk mengevaluasi kemampuan model dalam memprediksi distribusi kata dalam dokumen. Namun, nilai *Perplexity* yang rendah tidak selalu mencerminkan topik yang lebih bermakna secara semantik. Oleh karena itu, *Perlplexity* sering digunakan bersama metrik evaluasi lain, seperti *Coherence Score,* untuk memberikan evaluasi yang lebih komprehensif terhadap kualitas model topik (Wallach et al., 2009).

#### 2.2.5.2 Word Similarity

*Word Similarity* adalah metrik evaluasi yang digunakan untuk menilai sejauh mana representasi vektor kata mencerminkan hubungan semantik antar kata. Metriks ini sangat penting dalam mengevaluasi kualitas *Word Embedding*, karena membantu menilai kemampuan model dalam menangkap makna kata secara kontekstual (Mikolov et al., 2013). *Word Similarity* biasanya dihitung menggunakan *Cosine Similarity*, yang mengukur sudut kosinus antara dua vektor. Rumus *Cosine Similarity* dapat dilihat pada persamaan 2.11.

|  |  |
| --- | --- |
|  | 2.11 |

dan adalah vektor kata yang akan dibandingkan, menunjukkan operasi *dot product* antara kedua vektor, dan serta adalah norma atau panjang dari masing-masing vektor. Nilai *Cosine Similarity* berkisar antara -1 hingga 1, yang dimana nilai mendekati 1 menunjukkan hubungan semantik yang kuat, sementara nilai yang mendekati -1 menunjukkan perbedaan hubungan semantik yang besar.

*Word Similarity* memiliki aplikasi luas dalam NLP, seperti pencarian informasi, sistem tanya jawab, dan analisis sentimen. Namun, metrik ini memiliki keterbatasan dalam menangkap hubungan sintaksis antar kata dan cenderung gagal dalam mengidentifikasi hubungan kompleks dalam kata (Pennington ett al., 2014). Oleh karena itu, *Word Similarity* sering dikombinasikan dengan metrik evaluasi lain untuk memberikan gambaran yang lebih komprehensif mengenai kualitas representasi kata.

# BAB III METODOLOGI

## Metode yang dirancang

(termasuk di dalamnya arsitektur sistem)

## Peralatan pendukung

## Implementasi

(termasuk pseudocode dan desain user interface)

# BAB IV HASIL DAN PEMBAHASAN

## Hasil eksperimen

## Pembahasan/Diskusi

(analisis, sintesis, dan evaluasi)

# BAB V KESIMPULAN DAN SARAN

## Kesimpulan

Berupa hasil penelitian/perancangan yang menjawab permasalahan atau yang berupa konsep, program, dan karya rancangan

## Saran (jika dianggap perlu)

berisi hal-hal yang masih dapat dikerjakan dengan lebih baik dan dapat dikembangkan lebih lanjut, atau berisi masalahmasalah yang dialami pada saat proses pengerjaan tugas/proyek akhir.

# DAFTAR PUSTAKA

Af'idah, D. I., Dairoh, Handayani, S. F., & Pratiwi, R. W. (2021). Pengaruh parameter Word2Vec terhadap performa deep learning pada klasifikasi sentimen. Jurnal Informatika: Jurnal Pengembangan IT (JPIT), 6(3), 156–162. ISSN: 2477-5126, e-ISSN: 2548-9356.

Albalawi, R., Yeap, T., & Benyoucef, M. (2020). Using topic modeling methods for short-text data: A comparative analysis. Frontiers in Artificial Intelligence, 3, 42.

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). **A Neural Probabilistic Language Model**. Journal of Machine Learning Research, 3, 1137–1155.

Berry, M. W., Dumais, S. T., & O’Brien, G. W. (1995). Using Linear Algebra for Intelligent Information Retrieval. SIAM Review, 37(4), 573–595.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3(4), 993-1022.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. arXiv preprint arXiv:1607.04606.

Cichocki, A., Zadeh, R., Proakis, A., & Kumar, N. (2015). Nonnegative matrix factorization techniques. In Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(1), 203–221.

Cheng, X., Li, Y., & Zhang, Z. (2022). Non-negative matrix factorization for text clustering: A comparative study. Journal of Machine Learning Research.

De Nooijer, W., Borst, L., & van den Berg, E. (2000). Identifying topics in scientific literature. Proceedings of the 23rd International Conference on Knowledge Discovery and Data Mining (pp. 257-266). ACM

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 41(6), 391–407.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2001). Placing Search in Context: The Concept Revisited. In Proceedings of the 10th International Conference on World Wide Web (WWW 2001).

Goldberg, Y., & Levy, O. (2014). **Word2Vec Explained: Deriving Mikolov et al.’s Negative-Sampling Word-Embedding Method**. arXiv preprint arXiv:1402.3722.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences, 101(suppl 1), 5228-5235.

Hill, F., Reichart, R., & Korhonen, A. (2015). SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation. Computational Linguistics, 41(4), 665–695.

Kim, J., Park, H., & Eldén, L. (2019). Optimization approaches for non-negative matrix factorization. SIAM Journal on Matrix Analysis and Applications.

Lebret, R. (2016). Word embeddings for natural language processing. [Thesis]. EPFL

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature.

Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. Advances in Neural Information Processing Systems.

Liu, S., Bremer, P.-T., Thiagarajan, J. J., Srikumar, V., Wang, B., Livnat, Y., & Pascucci, V. (2017). Visual exploration of semantic relationships in neural word embeddings. IEEE Transactions on Visualization and Computer Graphics, 24(1), 553–562.

Manning, C. D., Schütze, H., & Raghavan, P. (2008). Introduction to Information Retrieval. Cambridge University Press

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems (NIPS).

Newman, D. (2006). Latent Dirichlet allocation and beyond. The Journal of Machine Learning Research, 7, 1289-1341.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, 100-108.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543.

Pouriyeh, S., Allahyari, M., Kochut, K., Cheng, G., & Arabnia, H. (2018, CCNIS). Combining word embedding and knowledge-based topic modeling for entity summarization. In Proceedings of the 2018 International Conference on Computing, Networking and Information Systems (pp. 1-6)

Rani, S., & Kumar, M. (2021). Topic modeling and its applications in materials science and engineering. Materials Today: Proceedings, 45, 5591-5596.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM).

Sari, M., & Asmendri. (2020). Penelitian Kepustakaan (Library Research) Dalam Penelitian Pendidikan IPA. NATURAL SCIENCES, 6(1), 41-53. DOI: 10.24265/nsp.2020.6.1.41

Seth, N. (2021). Part 2: Topic Modeling and Latent Dirichlet Allocation (LDA) using Gensim and Sklearn. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2021/06/part-2-topic-modeling-and-latent-dirichlet-allocation-lda-using-gensim-and-sklearn>

Umair, M., Qureshi, M. A., Shoaib, M., & Ahmad, F. (2020). A Survey on Latent Semantic Analysis for Natural Language Processing. arXiv preprint arXiv:2009.07872.

Vélez, D., Villarino, G., Rodriguez, J. T., & Gómez, D. (2020). A Method to Generate Soft Reference Data for Topic Identification. arXiv preprint arXiv:2005.07864.

Wahid, Ahmad. (2021). Using LDA for Trend Analysis in Social Media and News. Journal of Data Analytics.

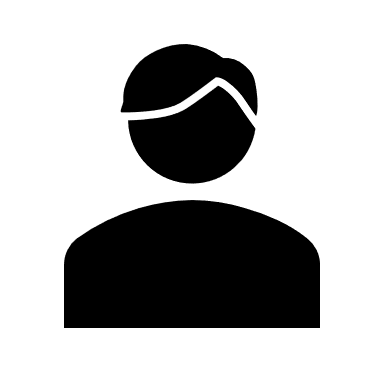
Wallach, H. M., Mimno, D., & McCallum, A. (2009). Rethinking LDA: Why priors matter. Advances in Neural Information Processing Systems, 22, 1973–1981.

Yakunin, K., Ionescu, G. M., Murzakhmetov, S., & Mussabayev, R. (2020). Propaganda identification using topic modelling. Social Psychology, 51(1), 1-15.

Zhang, Mei. (2023). Topic Modeling for Legal Document Classification. Law & Data Science Journal.

# LAMPIRAN-LAMPIRAN ATAU APPENDIKS (jika ada)

# BIODATA PENULIS

Penulis dilahirkan di Madiun, 29 Januari 1985, merupakan anak pertama dari 4 bersaudara. Penulis telah menempuh pendidikan formal yaitu di TK ABA 18 Madiun, SDN Beteng 1 Madiun, SMPN 2 Madiun dan SMAN 2 Madiun. Setelah lulus dari SMAN tahun 2020, Penulis mengikuti SBMPTN dan diterima di Departemen Teknik Mesin FTIRS - ITS pada tahun 2020 dan terdaftar dengan NRP 02112040000130.

Di Departemen Teknik Mesin Penulis sempat aktif di beberapa kegiatan Seminar yang diselenggarakan oleh Departemen, Himpunan Mahasiswa Teknik Mesin (HMM) dan aktif sebagai Asisten Praktikum Mesin Konversi Enersi maupun Grader mata kuliah Termodinamika.