PEMODELAN TOPIK MENGGUNAKAN *PRE-TRAINED LANGUAGE MODEL* ROBERTA DAN *VARIATIONAL AUTOENCODER*

Diajukan Sebagai Syarat Untuk Menyelesaikan
Pendidikan Program Strata-1 Pada
Jurusan Teknik Informatika

Oleh:

Fadhil Zahran Muwafa
NIM: 09021282025077

Jurusan Teknik Informatika

FAKULTAS ILMU KOMPUTER UNIVERSITAS SRIWIJAYA

2024

LEMBAR PENGESAHAN SKRIPSI

PEMODELAN TOPIK MENGGUNAKAN *PRE-TRAINED LANGUAGE MODEL* ROBERTA DAN *VARIATIONAL AUTOENCODER*

Oleh:

Fadhil Zahran Muwafa
NIM: 09021282025077

Inderalaya, 01 April 2024

Pembimbing I,

Novi Yusliani, M.T.
NIP. 198211082012122001

Pembimbing II,

M. Naufal Rachmatullah, S.Kom., M.T.
NIP. 19921201202231008

Mengetahui,
Ketua Jurusan Teknik Informatika

Dr. M. Fachrurrozi, S.Si., M.T.
NIP. 198005222008121002

TANDA LULUS UJIAN KOMPREHENSIF

Pada hari Jum'at tanggal 22 Maret 2024 telah dilaksanakan ujian komprehensif skripsi oleh Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Nama : Fadhil Zahran Muwafa

NIM : 09021282025077

Judul : Pemodelan Topik Menggunakan *Pre-trained Language Model* RoBERTa dan *Variational Autoencoder*

Dan dinyatakan **LULUS.**

1. Ketua Penguji

   Kanda Januar Miraswan, M.T. .....
   NIP. 199001092019031012

2. Penguji

   Desty Rodiah, M.T. .....
   NIP. 198912212020122011

3. Pembimbing I

   Novi Yusliani, M.T. .....
   NIP. 198211082012122001

4. Pembimbing II

   M. Naufal Rachmatullah, S.Kom., M.T. .....
   NIP. 19921201202231008

Mengetahui,
Ketua Jurusan Teknik Informatika

Dr. M. Fachrurrozi, S.Si., M.T.
NIP. 198005222008121002

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini:

Nama	: Fadhil Zahran Muwafa
NIM	: 09021282025077
Program Studi	: Teknik Informatika
Judul Skripsi	: Pemodelan Topik Menggunakan *Pre-trained Language Model* RoBERTa dan *Variational Autoencoder*

Hasil Pengecekan *Sofiware iThenticate/Turnitin:* 12%

Menyatakan bahwa laporan proyek saya merupakan hasil karya sendiri dan bukan hasil penjiplakan/plagiat. Apabila ditemukan unsur penjiplakan/plagiat dalam laporan proyek ini maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya sesuai dengan ketentuan yang berlaku.

Demikian pernyataan ini saya buat dengan sebenarnya dan tidak ada paksaan dari pihak mana pun.

Palembang, 22 Maret 2024

Fadhil Zahran Muwafa
NIM. 09021282025077

MOTTO DAN PERSEMBAHAN

Motto:

"Kunci dalam menggapai sebuah kesuksesan adalah berusaha, berdoa, dan bertawakal kepada Allah SWT."

- *Fadhil ZM*

Kupersembahkan Karya Tulis ini Kepada:

- Allah SWT
- Orang Tua
- Keluarga Besar
- Fakultas Ilmu Komputer
- Universitas Sriwijaya

ABSTRACT

The rapid and widespread flow of information highlights the importance of efficient text data management, making it even more important to organize and classify information from text data as more news is published online all the time. Topic modeling is useful in clustering news texts from the ever-growing sea of online news based on the topic of each text data. One method of topic modeling is to use Variational Autoencoder combined with a trained language model, RoBERTa. This research aims to create a topic modeling system using the Pre-trained Language Model RoBERTa and Variational Autoencoder. The dataset used consists of 5000 news data with 10 different topics taken from cnnindonesia, kompas, and detik.com. Topic modeling evaluation is done using coherence score cv, homogeneity score, and v-measure. With a coherence score cv of 77.3%, homogeneity score of 6.5%, and v-measure of 7.1%.

Keywords: Topic Modeling, Variational Autoencoder, Pre-trained Language Model, RoBERTa, K-Means, Coherence Score cv, Homogeneity Score, V-Measure

ABSTRAK

Arus informasi yang cepat dan luas menyoroti pentingnya pengelolaan data teks yang efisien, sehingga semakin pentingnya mengorganisir dan mengelompokkan informasi dari data teks karena banyaknya berita yang dipublikasikan secara *online* setiap saat. Pemodelan topik berguna dalam mengelompokkan teks berita dari lautan berita online yang terus berkembang berdasarkan topik setiap data teks. Salah satu metode pemodelan topik adalah dengan menggunakan *Variational Autoencoder* dikombinasikan dengan model bahasa yang telah dilatih yaitu RoBERTa. Penelitian ini bertujuan untuk membuat sistem pemodelan topik menggunakan *Pre-trained Language Model* RoBERTa dan *Variational Autoencoder*. Dataset yang digunakan terdiri dari 5000 data berita dengan 10 topik berbeda diambil dari cnnindonesia, kompas, dan detik.com. Evaluasi pemodelan topik dilakukan menggunakan *coherence score cv, homogeneity score,* dan *v-measure*. Dengan nilai *coherence score cv* sebesar 77.3%, *homogeneity score* sebesar 6.5%, dan *v-measure* sebesar 7.1%.

Kata Kunci: Pemodelan Topik, *Variational Autoencoder, Pre-trained Language Model,* RoBERTa, *Coherence Score cv, Homogeneity Score, V-Measure*

KATA PENGANTAR

Puji syukur kepada Allah SWT atas rahmat dan nikmat Nya yang lebih diberikan kepada penulis sehingga dapat menyelesaikan skripsi ini dengan baik. Skripsi ini disusun sebagai salah satu persyaratan untuk menyelesaikan pendidikan program Strata-1 di Fakultas Ilmu Komputer Universitas Sriwijaya. Dalam proses penyelesaian skripsi ini, penulis mendapatkan bantuan, bimbingan, dan dukungan dari berbagai pihak, baik secara langsung maupun tidak langsung. Oleh karena itu, penulis ingin mengucapkan terima kasih kepada:

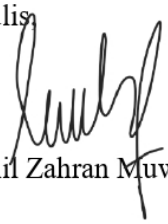
1. Allah SWT atas rahmat dan nikmat-Nya penulis dapat menyelesaikan skripsi ini dengan baik.
2. Kedua orang tua dan keluarga besar yang telah mendoakan, memberi semangat, memotivasi, dan nasihat untuk menyelesaikan skripsi ini.
3. Bapak Dr. M. Fachrurrozi, S.Si., M.T. selaku Ketua Jurusan Teknik Informatika Universitas Sriwijaya.
4. Bapak Osvari Arsalan, S.Kom., M.T. selaku Dosen Pembimbing Akademik yang telah memberikan banyak sekali bantuan dan arahan kepada penulis selama perkuliahan.
5. Ibu Novi Yusliani, M.T. selaku Dosen Pembimbing I dan Bapak Muhammad Naufal Rachmatullah, M. T. selaku Dosen Pembimbing II yang telah membimbing serta memberikan arahan kepada penulis selama proses pengerjaan skripsi.
6. Seluruh dosen program studi serta admin Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.
7. Saudara Alif Toriq Alkausar, M. Farhan Ghifari, M. Bintang Khadafi, Sheva Satrian, Anwaripasha Akbar, Alfaris Oktavian, dan Bayu Daru Pangestu sebagai teman yang selalu memotivasi penulis untuk semangat mengerjakan skripsi.

8. Seluruh Staf Administrasi dan Pegawai Fakultas Ilmu Komputer yang telah membantu dalam urusan administrasi tugas akhir penulis.
9. Seluruh teman-teman yang telah memberikan saran, motivasi, dan semangat kepada penulis.
10. Pihak-pihak lain yang tidak dapat penulis sebutkan satu persatu.

Penulis menyadari bahwa dalam penyusunan skripsi ini masih banyak sekali kekurangan dikarenakan kurangnya pengalaman dan pengetahuan penulis. Oleh karena itu penulis mengharapkan saran dan kritik yang membangun guna kemajuan penelitian selanjutnya. Semoga tugas akhir ini dapat bermanfaat. Terima kasih.

Palembang, 22 Maret 2024

Penulis,



Fadhil Zahran Muwafa

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I

PENDAHULUAN

1.1 Pendahuluan

Bab pendahuluan akan membahas latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan masalah penelitian, dan sistematika penulisan. Secara keseluruhan, skripsi ini menguraikan tentang bagaimana membangun sebuah sistem pemodelan topik dengan menggunakan pra-pelatihan model bahasa (*Pre-Trained Language Model*) dalam kasus ini yaitu RoBERTa dan menggunakan metode *Variational Autoencoder* yang berbasis dengan jaringan *neural*. Sistem ini dapat digunakan dalam pemodelan topik yang berguna untuk mendapatkan topik yang semantik terhadap keseluruhan konten yang dianalisis.

1.2 Latar Belakang

Arus informasi yang mengalir pada zaman sekarang ini dapat dikatakan sangat cepat dengan didukung oleh perkembangan teknologi informasi serta komunikasi sehingga membuat dunia saat ini tenggelam dalam lautan data yang terus-menerus bertambah seiring berjalannya waktu. Khususnya mulai dari media sosial, portal berita, forum, dan platform lainnya membanjiri pengguna dengan teks dalam jumlah yang belum pernah terjadi sebelumnya. Data teks tersebut, meskipun kaya akan informasi, seringkali tidak terstruktur dan memiliki keberagaman dalam topik, serta bervariasi dalam kualitas, namun jika dikelola dengan benar data teks dapat menjadi sumber wawasan yang sangat berharga. Oleh karena itu, menghadapi volume dan kecepatan data teks yang kian meningkat, ada kebutuhan mendesak

untuk metode yang dapat mengorganisir, mengelompokkan, dan mengekstrak wawasan dari data teks besar ini secara otomatis dan efisien. Dalam konteks ini, pemodelan topik muncul sebagai salah satu solusi yang menjanjikan.

Pemodelan topik menyediakan algoritma teknik dari berbagai perspektif untuk menemukan semantik tersembunyi dalam koleksi dokumen dan mengelompokkan tema-tema tersebut sebagai topik (Kherwa & Bansal, 2018). Topik modeling akan mengelompokkan korpus berbentuk kumpulan kata menjadi topik-topik yang dapat menggambarkan korpus tersebut. Pemodelan topik adalah proses mengidentifikasi dan mengelompokkan kata-kata dari satu set dokumen ke dalam topik yang berbeda, di mana setiap topik adalah kumpulan kata-kata yang sering muncul bersama-sama pada suatu *cluster* topik (Meng et al., 2022). Pemodelan topik merupakan sebuah *unsupervised model* yang secara statistik bertujuan untuk menemukan potensi topik dari koleksi besar dokumen untuk dapat digunakan dalam tugas-tugas lanjutan dalam bidang *natural language processing*, seperti pengelompokan teks, analisis sentimen, dan sebagainya (Cheng et al., 2023). Nurlayli dan Nasichuddin (2019) mengemukakan bahwa *topic modeling* berperan dalam menguraikan dan mengorganisir informasi yang terkandung dalam teks yang besar sehingga memungkinkan pemahaman dan analisis yang lebih efisien.

Proses pemodelan topik yang mencakup komponen *neural* (*Neural Topic Modeling*) salah satunya yaitu *Variational Autoencoder* (VAE). Berbeda dengan model topik konvensional, model topik *neural* dapat langsung menerapkan propagasi gradien yang meningkatkan fleksibilitas dan skalabilitas. Model ini

menggunakan teknik seperti *Variational Autoencoder* (VAE) yang mampu menghasilkan topik yang beragam serta koheren (Wu et al., 2023). VAE juga mampu menghasilkan topik-topik berkualitas dan kinerja yang unggul dibandingkan dengan model topik konvensional lain dengan melakukan inferensi *variational* melalui jaringan saraf untuk mengambil sampel vektor topik potensial dari dokumen (Cheng et al., 2023). VAE adalah jaringan saraf yang terdiri dari dua bagian utama, yaitu encoder dan decoder (Song et al., 2019).

Di dalam proses *neural topic modeling* dengan menggunakan konsep metode VAE, penggunaan *pre-trained language model* (PLM) memiliki peranan yang sangat penting. Salah satu jenis *Pre-trained Language Model* adalah RoBERTa yang merupakan versi lebih canggih dan lebih baik sebagai hasil dari pengembangan PLM sebelumnya, yaitu BERT sehingga memberikan hasil yang lebih baik lagi dalam tugas-tugas pemrosesan bahasa alami (Liu et al., 2019). RoBERTA memiliki beberapa keunggulan yang terletak pada beberapa modifikasi kunci yang dilakukan selama proses pra-pelatihan. Mulai dari model yang dilatih untuk jangka waktu yang lebih lama dan meningkatkan jumlah langkah pra-pelatihan, hal ini menghasilkan peningkatan signifikan dalam kinerja tugas *downstream* dalam pemrosesan bahasa alami. Disamping itu juga, RoBERTa menggunakan ukuran batch yang lebih besar selama pelatihannya, sehingga secara dinamis mengubah pola *masking* yang diterapkan pada data pelatihan (Liu et al., 2019). Strategi optimasi ini telah terbukti efektif dalam meningkatkan kinerja model BERT dan menjadikan RoBERTa sebagai standar baru dalam beberapa

tugas GLUE. Sehingga penelitian ini akan memanfaatkan penggunaan *Pre-trained Language Model* RoBERTa sebagai fondasi untuk mengekstraksi fitur-fitur linguistik yang mendalam dan kontekstual dari data teks yang akan diolah lebih lanjut oleh *Variational Autoencoder* dalam proses pemodelan topik.

1.3 Rumusan Masalah

Berdasarkan Penjelasan pada latar belakang sebelumnya maka rumusan masalah dari penelitian ini yaitu:

1. Bagaimana menerapkan metode *Variational Autoencoder* dengan menggunakan *Pre-Trained Language Model* RoBERTa untuk menghasilkan representasi topik yang relevan?
2. Bagaimana nilai *Coherence Score, Homegeneity Score,* dan *V-Measure* dari sistem pemodelan topik dengan menggunakan *Variational Autoencoder* dan *Pre-Trained Language Model* RoBERTa?

1.4 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah sebagai berikut:

1. Menghasilkan sebuah sistem yang dapat melakukan representasi topik dengan menggunakan *Pre-Trained Language Model* RoBERTa dan *Variational Autoencoder.*
2. Mengetahui nilai *Coherence Score, Homegeneity Score,* dan *V-Measure* dari sistem pemodelan topik dengan menggunakan *Pre-Trained Language Model* RoBERTa dan *Variational Autoencoder.*

1.5 Manfaat Penelitian

Manfaat penelitian yang diperoleh yaitu:

1. Diharapkan dapat menjadi referensi untuk penelitian atau pengembangan selanjutnya.
2. Sistem dapat digunakan untuk representasi topik.

1.6 Batasan Masalah

Agar permasalahan tidak menyimpang dari batasan yang telah ditetapkan, maka adapun batasan dari penelitian ini adalah:

1. Data yang digunakan adalah data judul berita berbahasa Indonesia.
2. Data judul berita yang digunakan berjumlah 5.000 data.
3. Data judul berita terdiri dari 10 kategori berita yaitu edukasi, ekonomi & bisnis, hiburan, kesehatan, *lifestyle*, makanan, olahraga, otomotif, politik, dan teknologi.

1.7 Sistematika Penulisan

Sistematika penulisan yang digunakan pada penelitian ini mengikuti standar operasional penulisan tugas akhir Fakultas Ilmu Komputer Universitas Sriwijaya, yaitu:

BAB I. PENDAHULUAN

Bab ini membahas tentang latar belakang, perumusan masalah, tujuan, dan manfaat dari penelitian, batasan masalah, metodologi, penelitian, dan

sistematika penulisan pada penyusunan penelitian ini. Pokok-pokok fikiran ini akan menjadi landasan pada bab selanjutnya.

BAB II. KAJIAN LITERATUR

Bab ini menjelaskan mengenai landasan teori yang digunakan dalam menunjang penelitian. Dalam bab ini dimuat mengenai literatur dan penelitian terkait sebelumnya yang berkaitan dengan penelitian ini, seperti penjelasan mengenai *Pre-Trained Language Model* RoBERTa, metode *Variational Autoencoder,* serta penjelasan lain yang terkait.

BAB III. METODOLOGI PENELITIAN

Bab ini membahas mengenai langkah-langkah dalam penelitian yang akan dijalani mulai dari proses pengumpulan data, perancangan dari sistem yang dibuat dan rincian dari setiap tahapan dalam melakukan penelitian sesuai dengan kerangka kerja yang telah disusun.

BAB IV. PENGEMBANGAN PERANGKAT LUNAK

Bab ini akan membahas mengenai perancangan perangkat lunak mulai dari analisis kebutuhan perangkat lunak hingga pengujian pada perangkat lunak guna mengevaluasi pengembangan perangkat lunak.

BAB V. HASIL DAN ANALISIS PENELITIAN

Bab ini memaparkan hasil penelitian berdasarkan langkah dan metode yang telah direncanakan sebelumnya. Analisis tersebut diberikan sebagai dasar kesimpulan yang akan diambil dari penelitian ini.

BAB VI. KESIMPULAN DAN SARAN

Bab ini memaparkan kesimpulan dari penelitian yang dilakukan berdasarkan uraian pada bab-bab sebelumnya dan memuat saran yang diharapkan dapat membuat sistem lebih baik lagi kedepannya.

1.8 Kesimpulan

Bab ini telah menjelaskan terkait rencana penelitian yang akan dilakukan mulai dari latar belakang penelitian, rumusan masalah, tujuan dan manfaat yang diperoleh dari penelitian yang dilakukan, batasan masalah serta sistematika dari penulisan yang akan dibuat.

DAFTAR PUSTAKA

Adiya, M. H., & Desnelita, Y. (2019). *Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan Pada RSUD Pekanbaru*. *05*(01).

Bahrainian, S. A., Jaggi, M., & Eickhoff, C. (2021). Self-Supervised Neural Topic Modeling. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3341–3350. https://doi.org/10.18653/v1/2021.findings-emnlp.284

Cheng, H., Liu, S., Sun, W., & Sun, Q. (2023). A Neural Topic Modeling Study Integrating SBERT and Data Augmentation. *Applied Sciences*, *13*(7), 4595. https://doi.org/10.3390/app13074595

Davidson, T. R., Falorsi, L., De Cao, N., Kipf, T., & Tomczak, J. M. (2022). *Hyperspherical Variational Auto-Encoders* (arXiv:1804.00891). arXiv. http://arxiv.org/abs/1804.00891

Diyasa, F. G. (n.d.). *Jurusan Teknik Informatika FAKULTAS ILMU KOMPUTER UNIVERSITAS SRIWIJAYA 2023*.

Kherwa, P., & Bansal, P. (2018). Topic Modeling: A Comprehensive Review. *ICST Transactions on Scalable Information Systems*, *0*(0), 159623. https://doi.org/10.4108/eai.13-7-2018.159623

KuÅ, U. (n.d.). *Performance of Multi-Clustering Recommender System after Selection of Clusters based on V-Measures*.

Kumar, A., Esmaili, N., & Piccardi, M. (2021). A REINFORCEd Variational Autoencoder Topic Model. In T. Mantoro, M. Lee, M. A. Ayu, K. W. Wong, & A. N. Hidayanto (Eds.), *Neural Information Processing* (Vol. 1516, pp. 360–

369). Springer International Publishing. https://doi.org/10.1007/978-3-030-92307-5_42

Liang, D., Krishnan, R. G., Hoffman, M. D., & Jebara, T. (2018). *Variational Autoencoders for Collaborative Filtering* (arXiv:1802.05814). arXiv. http://arxiv.org/abs/1802.05814

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. http://arxiv.org/abs/1907.11692

Meng, Y., Zhang, Y., Huang, J., Zhang, Y., & Han, J. (2022). Topic Discovery via Latent Space Clustering of Pretrained Language Model Representations. *Proceedings of the ACM Web Conference 2022*, 3143–3152. https://doi.org/10.1145/3485447.3512034

Nan, F., Ding, R., Nallapati, R., & Xiang, B. (2019). Topic Modeling with Wasserstein Autoencoders. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6345–6381. https://doi.org/10.18653/v1/P19-1640

Rungta, K., Chau, G., Dewangan, A., Wagner, M., & Huang, J.-L. (n.d.). *Sentence Generation and Classification with Variational Autoencoder and BERT*.

Sato-Ilic, M. (2018). Homogeneous Cluster Analysis. *Procedia Computer Science*, *140*, 269–275. https://doi.org/10.1016/j.procs.2018.10.320

Song, T., Sun, J., Chen, B., Peng, W., & Song, J. (2019). Latent Space Expanded Variational Autoencoder for Sentence Generation. *IEEE Access*, *7*, 144618–144627. https://doi.org/10.1109/ACCESS.2019.2944630

Srivastava, A., & Sutton, C. (2017). *AUTOENCODING VARIATIONAL INFERENCE FOR TOPIC MODELS*.

Wu, X., Dong, X., Nguyen, T., & Luu, A. T. (n.d.). *Effective Neural Topic Modeling with Embedding Clustering Regularization*.

Zhang, L., Hu, X., Wang, B., Zhou, D., Zhang, Q.-W., & Cao, Y. (2022). Pre-training and Fine-tuning Neural Topic Model: A Simple yet Effective Approach to Incorporating External Knowledge. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5980–5989. https://doi.org/10.18653/v1/2022.acl-long.413