PEMODELAN TOPIK PADA TWEET BAHASA INDONESIA MENGGUNAKAN BERTOPIC

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

Oleh :

Fahmi Guntara Diyasa
NIM : 09021181823002

Jurusan Teknik Informatika
FAKULTAS ILMU KOMPUTER UNIVERSITAS SRIWIJAYA
2023

LEMBAR PENGESAHAN SKRIPSI

PEMODELAN TOPIK PADA TWEET BAHASA INDONESIA MENGGUNAKAN BERTOPIC

Oleh:

Fahmi Guntara Diyasa
NIM: 09021181823002

Indralaya, Agusutus 2023

Mengetahui,
Ketua Jurusan Teknik Informatika

Alvi Syahrini Utami, M.Kom
NIP. 197812222006042003

Pembimbing,

Dr. Abdiansah, S.Kom., M.Cs.
NIP. 198410012009121005

TANDA LULUS UJIAN KOMPREHENSIF

Pada hari senin tanggal 31 juli 2023 telah dilaksanakan ujian komprehensif skripsi oleh Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Nama : Fahmi Guntara Diyasa
NIM : 09021181823002
Judul : Pemodelan Topik Pada Tweet Bahasa Indonesia Menggunakan BERTopic

Dan dinyatakan LULUS

1. Ketua Penguji

   Yunita, M.Cs.
   NIP. 198306062015042002 .....

2. Penguji

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

3. Pembimbing

   Dr. Abdiansah, S.Kom., M.Cs.
   NIP. 198410012009121005 .....

Mengetahui
Ketua Jurusan Teknik Informatika

Alvi Syahrini Utami, M.Kom
NIP. 197812222006042003

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini:

Nama : Fahmi Guntara Diyasa

NIM : 09021181823002

Program Studi : Teknik Informatika

Judul Skripsi : Pemodelan Topik Pada Tweet Bahasa Indonesia Menggunakan BERTopic

Hasil pengecekan software iThenticate/Turnitin: 14%

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

Demikian penyataan ini saya buat dengan sebenarnya dan tidak ada paksaan dari pihak manapun.

Indralaya, 14 Agustus 2023

Fahmi Guntara Diyasa

NIM. 09021181823002

MOTTO DAN PERSEMBAHAN

Motto:

Keep it simple

Kupersembahkan Karya Tulis ini kepada:

- Allah SWT
- Kedua orang tua, saudara dan teman saya
- Fakultas Ilmu Komputer
- Universitas Sriwijaya

ABSTRACT

The increasing use of information and communication technology in recent years has had a significant impact on the trend of communicating and expressing aspirations through social media, especially the Twitter platform. Every tweet on Twitter carries information about a particular topic that can be identified through the Topic Modeling method. Topic Modeling is a tool used to uncover hidden topics in a group of documents. This research aims to perform topic modeling on Indonesian tweets using BERTopic. The Topic Modeling process using BERTopic includes steps such as document embedding, dimension reduction with UMAP, document clustering using HDBSCAN, and representing topics using c-TF-IDF. The dataset used consists of 10,000 Indonesian tweets taken from the Twitter account @detikcom. From the 10,000 tweets, 119 main topics and 1 outlier topic were found. Topic Modeling Evaluation is done using coherence score cv, with the average coherence score cv of 0.685, the highest coherence score cv of 0.995, and the lowest coherence score cv of 0.119..

***Keywords**: Topic Modeling, BERTopic, Cohrence Score cv, Tweets*

ABSTRAK

Peningkatan penggunaan teknologi informasi dan komunikasi dalam beberapa tahun terakhir telah membawa dampak signifikan terhadap tren berkomunikasi dan penyampaian aspirasi melalui media sosial, khususnya platform Twitter. Setiap *tweet* di Twitter membawa informasi mengenai topik tertentu yang dapat diidentifikasi melalui metode Pemodelan Topik. Pemodelan Topik merupakan alat yang digunakan untuk mengungkap topik tersembunyi dalam sekelompok dokumen. Penelitian ini bertujuan untuk melakukan pemodelan topik pada *tweet* berbahasa Indonesia dengan menggunakan *BERTopic*. Proses Pemodelan Topik menggunakan *BERTopic* meliputi langkah-langkah seperti penyematan dokumen, pengurangan dimensi dengan UMAP, pengelompokan dokumen menggunakan HDBSCAN, dan merepresentasikan topik menggunakan c-TF-IDF. Dataset yang digunakan terdiri dari 10.000 *tweet* berbahasa Indonesia yang diambil dari akun Twitter *@detikcom*. Dari 10.000 *tweet* tersebut, ditemukan 119 topik utama dan 1 topik *outlier*. Evaluasi Pemodelan Topik dilakukan menggunakan *coherence score cv*, dengan hasil rata-rata *coherence score cv* sebesar 0,685, *coherence score cv* tertinggi sebesar 0,995, dan *coherence score cv* terendah sebesar 0,119.

Kata Kunci: Pemodelan Topik, *BERTopic*, *Cohrence Score cv*, *Tweet*

KATA PENGANTAR

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

1. Allah SWT atas rahmat dan nikmat-Nya penulis dapat menyelesaikan skripsi ini dengan baik.
2. Kedua orang tua, saudara dan teman yang telah mendoakan, memberi semangat, motivasi, dan nasihat untuk menyelesaikan skripsi ini.
3. Ibu Alvi Syahrini Utami, M.Kom. selaku Ketua Jurusan Teknik Informatika Universitas Sriwijaya.
4. Bapak Dr. Abdiansah, S.Kom., M.Cs. selaku Dosen Pembimbing yang telah membimbing, memberikan motivasi serta arahan kepada penulis dalam proses pengerjaan skripsi.
5. Ibu Novi Yusliani, M.T. selaku Dosen Penguji Tugas Akhir yang telah memberikan ilmu, nasihat serta saran yang membangun.
6. Bapak Rifkie Primartha, M.T. selaku Pembimbing Akademik selama di Universitas Sriwijaya.
7. Seluruh dosen, staf dan pegawai Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.
8. Pihak-pihak lain yang tidak dapat penulis sebutkan satu-persatu.

Penulis menyadari bahwa dalam penyusuna 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.

Indralaya, 14 Agustus 2023

Fahmi Guntara Diyasa

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I

PENDAHULUAN

1.1 Pendahuluan

Pada Bab ini membahas latar belakang masalah, rumusan masalah, tujuan dan manfaat penelitian serta batasan masalah. Bab ini memberikan penjelasan umum mengenai keseluruhan penelitian.

1.2 Latar Belakang

Meningkatnya penggunaan teknologi informasi dan komunikasi dalam beberapa tahun terakhir telah memunculkan satu tren di kalangan masyarakat untuk berkomunikasi ataupun menyampaikan aspirasi melalui media sosial. Salah satu media sosial yang banyak digunakan adalah twitter (Chilmi, 2021). Setiap *tweet* pada twitter memiliki topik tertentu, dan untuk mengetahui topik utama dari kumpulan *tweet* tersebut dapat menggunakan *topic modeling* (Patmawati dan Yusuf, 2021).

Pemodelan topik merupakan teknik yang digunakan dalam pendekatan *text mining* dan *text analysis* dalam menemukan data teks yang tersembunyi dan hubungan antar teks yang saling berkaitan dari suatu korpus (Jelodar et al., 2018). Pemodelan topik telah dibahas pada banyak literatur menggunakan metode-metode yang bervariasi seperti Top2Vec (Angelov, 2020), *Non-negative Matrix Factorization* (NMF) (Carbonetto et al., 2022) dan *Latent Dirichlet*

Allocation (LDA) (Blei et al., 2013). LDA merupakan metode pemodelan topik yang paling populer dan banyak digunakan (Angelov, 2020).

LDA sangat cocok untuk tugas pemodelan topik umum menggunakan berbagai data. Akan tetapi LDA tidak mampu memodelkan hubungan data yang lebih maju dan berkinerja buruk ketika dokumen tidak cukup panjang (Vayansky and Kumar, 2020). NMF adalah pendekatan tanpa pengawasan untuk mengurangi dimensi matriks nonnegatif (Lee and Seung, 1999), dan telah banyak digunakan untuk menemukan hubungan yang mendasari antara teks dan mengidentifikasi topik *laten* (Arora et al, 2012).

Meskipun demikian LDA dan NMF membutuhkan upaya yang cukup besar untuk penyetelan *hyperparameter* guna menciptakan topik yang bermakna (Abuzayed and Al-Khalifa, 2021). Top2Vec dapat digunakan ketika banyak bahasa muncul dalam sebuah korpus (Hendry et al., 2021). Top2Vec juga memungkinkan untuk tidak menggunakan *preprocessing* pada data asli karena telah menggunakan teknik *embeddings* (Egger and Yu, 2022). Akan tetapi Top2Vec tidak mampu bekerja baik dengan data yang kecil misalnya kurang dari 1000 dokumen (Abuzayed and Al-Khalifa, 2021).

BERTopic adalah teknik pemodelan topik yang memanfaatkan penyematan BERT dan TF-IDF berbasis kelas untuk menghasilkan representasi topik yang koheren, juga menggunakan teknik *Uniform Manifold Approximation and Projection* (UMAP) untuk menurunkan dimensi penyematan sebelum pengelompokan dokumen-dokumen (Grootendorst, 2022). *BERTopic* bekerja secara luar biasa dengan penyematan yang telah dilatih sebelumnya dan juga

karena pemisahan antara pengelompokan dokumen dan penggunaan *c-TF-IDF* untuk mengekstraksi representasi topik (Franco and Moreno, 2022).

Karena penggunanan prosedur *c-TF-IDF*, *BERTopic* dapat mendukung beberapa variasi pemodelan topik, seperti pemodelan topik terpadu, pemodelan topik dinamis, atau pemodelan topik berbasis kelas (Abuzayed and Al-Khalifa, 2021). Kelebihan utamanya terletak pada kenyataan bahwa algoritma ini bekerja dengan baik pada sebagian besar aspek domain pemodelan topik, sedangkan yang lain biasanya unggul dalam satu aspek (Egger and Yu, 2022). Berdasarkan referensi penelitian yang dilakukan sebelumnya, metode *BERTopic* akan digunakan dalam penelitian pemodelan topik pada *tweet* bahasa indonesia.

1.3 Rumusan Masalah

Berdasarkan permasalahan yang telah dijelaskan pada latar belakang maka rumusan masalah dari penelitian ini adalah

1. Bagaimana mengembangkan pemodelan topik pada *tweet* bahasa Indonesia menggunakan *BERTopic*?
2. Bagaimana kinerja topik yang dihasilkan dari pemodelan topik pada *tweet* bahasa Indonesia menggunakan *BERTopic*?

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah :

1. Menghasilkan pemodelan topik pada *tweet* bahasa Indonesia menggunakan *BERTopic*.

2. Mengetahui hasil pemodelan topik pada *tweet* bahasa Indonesia menggunakan *BERTopic*.

1.5 Manfaat Penelitian

Manfaat yang diperoleh dari penelitian ini adalah :

1. Membantu pembaca untuk mengetahui topik *laten* pada kumpulan *tweet* berbahasa Indonesia.
2. Hasil penelitian dapat dijadikan sebagai rujukan penelitian terkait.

1.6 Batasan Masalah

Batasan masalah dari penelitian ini adalah :

1. Data yang digunakan adalah data *tweet* berbahasa Indonesia.
2. Data *tweet* yang digunakan berjumlah 10.000 *tweet*.

1.7 Sistematika Penulisan

Sistematika penulisan tugas akhir mengikuti standar penulisan tugas akhir Fakultas Ilmu Komputer Universitas Sriwijaya yaitu sebagai berikut:

BAB I. PENDAHULUAN

Pada bab ini membahas latar belakang, rumusan masalah, tujuan dan manfaat penelitian, batasan masalah serta sistematika penelitian yang akan dijadikan sebagai pokok pikiran penelitian ini.

BAB II. KAJIAN LITERATUR

Pada bab ini membahas landasan teori yang digunakan dalam penelitian, seperti definisi pemodelan topik dan model *BERTopic*, serta beberapa literatur yang relevan dengan penelitian ini.

BAB III. METODOLOGI PENELITIAN

Pada bab ini membahas proses yang akan dilaksanakan selama penelitian, Seperti pengumpulan data, analisis data dan perancangan perangkat lunak. Setiap tahap akan dijelaskan berdasarkan kerangka kerja yang dibuat.

BAB IV. PENGEMBANGAN PERANGKAT LUNAK

Pada bab ini membahas analisis dan rancangan perangkat lunak yang akan dikembangkan. Diawali dari analisis kebutuhan, perancangan dan konstruksi perangkat lunak, dan diakhiri dengan evaluasi untuk memastikan sistem yang dikembangkan sudah sesuai dengan rancangan dan kebutuhan penelitian.

BAB V. HASIL DAN ANALISIS PENELITIAN

Pada bab ini menyajikan hasil pengujian berdasarkan langkah-langkah yang telah direncanakan. Analisis diberikan sebagai dasar kesimpulan yang akan diambil dari penelitian ini.

BAB VI. KESIMPULAN DAN SARAN

Pada bab ini membahas kesimpulan yang diambil berdasarkan uraian dalam bab sebelumnya serta saran yang diberikan berdasarkan penelitian yang telah dilakukan.

1.8 Kesimpulan

Pada bab ini telah dijelaskan latar belakang, rumusan masalah, tujuan dan manfaat penelitian, batasan masalah serta sistematika penelitian yang akan dijadikan sebagai pokok pikiran penelitian ini.

DAFTAR PUSTAKA

Abuzayed, A., & Al-Khalifa, H. (2021). BERT for Arabic topic modeling: an experimental study on BERTopic technique. Procedia Computer Science, 189, 191-194.

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001, January). On the surprising behavior of distance metrics in high dimensional space. In International conference on database theory (pp. 420-434). Springer, Berlin, Heidelberg.

Allaoui, M., Kherfi, M. L., & Cheriet, A. (2020, June). Considerably improving clustering algorithms using UMAP dimensionality reduction technique: a comparative study. In International Conference on Image and Signal Processing (pp. 317-325). Springer, Cham.

Al-khairi, Y. U., Wibisono, Y., & Putro, B. L. (2017). Deteksi topik fashion pada twitter dengan latent dirichlet allocation.

Angelov, D. (2020). Top2vec: Distributed representations of topics. arXiv preprint arXiv:2008.09470.

Arora, S., Ge, R., & Moitra, A. (2012, October). Learning topic models--going beyond SVD. In 2012 IEEE 53rd annual symposium on foundations of computer science (pp. 1-10). IEEE.

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999, January). When is "nearest neighbor" meaningful?. In International conference on database theory (pp. 217-235). Springer, Berlin, Heidelberg.

Blei, D. M. (2013). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Carbonetto, P., Sarkar, A., Wang, Z., & Stephens, M. (2021). Non-negative matrix factorization algorithms greatly improve topic model fits. arXiv preprint arXiv:2105.13440.

Chilmi, M. L. C. (2021). Latent dirichlet allocation lda untuk mengetahui topik pembicaraan warganet twitter tentang omnibus law (Bachelor's thesis, Fakultas Sains Dan Teknologi UIN Syarif Hidayatullah Jakarta).

Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. Frontiers in Sociology, 7.

Gornik, D. (2004). IBM Rational Unified Process: Best practices for software development teams. Rational Software White Paper TP026B, Rev, 11(01).

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.

Hendry, D., Darari, F., Nurfadillah, R., Khanna, G., Sun, M., Condylis, P. C., & Taufik, N. (2021, October). Topic modeling for customer service chats. In 2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS) (pp. 1-6). IEEE.

Joachims, T. (1996). A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. Carnegie-mellon univ pittsburgh pa dept of computer science.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications, 78(11), 15169-15211.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788-791.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

McInnes, L., & Healy, J. (2017, November). Accelerated hierarchical density based clustering. In 2017 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 33-42). IEEE.

Meeks, E., & Weingart, S. B. (2012). The digital humanities contribution to topic modeling. Journal of Digital Humanities, 2(1), 1-6.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010, June). Automatic evaluation of topic coherence. In Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics (pp. 100-108).

Newman, D., Noh, Y., Talley, E., Karimi, S., & Baldwin, T. (2010, June). Evaluating topic models for digital libraries. In Proceedings of the 10th annual joint conference on Digital libraries (pp. 215-224).

Pandove, D., Goel, S., & Rani, R. (2018). Systematic review of clustering high-dimensional and large datasets. ACM Transactions on Knowledge Discovery from Data (TKDD), 12(2), 1-68.

Patmawati, P., & Yusuf, M. (2021). Analisis Topik Modelling Terhadap Penggunaan Sosial Media Twitter oleh Pejabat Negara. Building of Informatics, Technology and Science (BITS), 3(3), 122-129.

Pradha, S., Halgamuge, M. N., & Vinh, N. T. Q. (2019, October). Effective text data preprocessing technique for sentiment analysis in social media data. In 2019 11th international conference on knowledge and systems engineering (KSE) (pp. 1-8). IEEE.

Putra, I. M. K. B. (2017). Analisis topik informasi publik media sosial di surabaya menggunakan pemodelan latent dirichlet allocation (LDA) (Doctoral dissertation, Institut Teknologi Sepuluh Nopember).

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

Sánchez-Franco, M. J., & Rey-Moreno, M. (2022). Do travelers' reviews depend on the destination? An analysis in coastal and urban peer-to-peer lodgings. Psychology & Marketing, 39(2), 441-459.

Steinbach, M., Ertöz, L., & Kumar, V. (2004). The challenges of clustering high dimensional data. In New directions in statistical physics (pp. 273-309). Springer, Berlin, Heidelberg.

Thakur, N., Reimers, N., Daxenberger, J., & Gurevych, I. (2020). Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. arXiv preprint arXiv:2010.08240.

Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. Information Systems, 94, 101582.