ELINVO (Electronics, Informatics, and Vocational Education), November 2019; 4(2):154-161
ISSN 2580-6424 (printed), ISSN 2477-2399 (online,) DOI: 10.21831/elinvo.v4i2. 28254

Topic Modeling Penelitian Dosen JPTEI UNY pada *Google Scholar* Menggunakan *Latent Dirichlet Allocation*

Akhsin Nurlayli[1], **Moch. Ari Nasichuddin**[2]
[1]Jurusan Pendidikan Teknik Elektronika dan Informatika, Universitas Negeri Yogyakarta;
[2]PT. Atmatech Global Informatika
E-mail: akhsinnurlayli@uny.ac.id

ABSTRACT

The mapping of research *topic*s for lecturers is necessary to determine the research tendencies in a department or study program. This study aims to implement *topic modeling* in the publication titles of the Department of Electronics and Informatics Education Engineering of Universitas Negeri Yogyakarta (JPTEI UNY) lecturers taken from *Google Scholar*. The method used for *topic modeling* is the Latent Dirichlet Allocation (LDA). LDA is a generative probabilistic model for finding the semantic structure of a corpus collection based on the hierarchical bayesian analysis. After the *topic modeling* process, the results showed that JPTEI UNY lecturers tend to have four research clusters consisting of vocational education, system development, learning media, and vocational learning systems.

Keywords*: clustering*, LDA, *research, topic modeling*

ABSTRAK

Pemetaan topik penelitian dosen diperlukan untuk mengetahui kecenderungan penelitian dosen pada suatu jurusan atau program studi. Penelitian ini bertujuan untuk mengimplementasikan *topic modeling* pada judul publikasi dari dosen Jurusan Pendidikan Teknik Elektronika dan Informatika (JPTEI UNY) yang diambil dari *Google Scholar*. Metode yang digunakan adalah *Latent Dirichlet Allocation* (LDA). LDA adalah model probabilistik generatif untuk mencari struktur semantik dari kumpulan korpus yang berdasarkan *hierarchical bayesian analysis*. Setelah dilakukan *topic modeling*, data hasil menunjukkan bahwa dosen JPTEI UNY cenderung memiliki judul penelitian tentang pendidikan vokasi, pengembangan sistem, media pembelajaran, dan sistem pembelajaran di SMK.

Kata kunci: *clustering*, LDA, penelitian, *topic modeling*

PENDAHULUAN

Penelitan menjadi entitas yang penting dalam dunia akademik. Kemajuan lini penelitian dapat dilihat dari segi jumlah penelitian dan arah topik penelitian yang dibahas. Untuk melihat arah penelitian perlu membaca satu persatu publikasi ilmiah dari seluruh dosen di sebuah instansi, kegiatan tersebut tentu akan memerlukan waktu dan tenaga. Oleh karena itu diperlukan suatu langkah yang bertujuan mengotomatisasi pembacaan arah topik penelitan dosen di suatu instansi akademik.

Langkah otomatisasi bisa dilakukan dengan pendekatan *topic modeling*. *topic modeling* merupakan suatu pendekatan untuk menganalisis kumpulan dokumen berbentuk teks dan mengelompokkan menjadi beberapa topik. Pendekatan tersebut masuk dalam pendekatan *Clustering* dalam studi *Machine Learning*.

Metode yang digunakan untuk melakukan pendekatan *topic modeling* bermacam-macam, yakni *Latent Semantic Analysis* (LSA), *Probabilistic Latent Sematic Analysis* (pLSA), *Latent Dirichlet Allocation* (LDA), dan lain-lain [1]. Contoh beberapa penelitian sebelumnya yang melakukan *topic modeling*, yakni

penerapan *topic modeling* untuk menggambarkan topik mengenai kondisi cuaca dan iklim di Pulau Jawa berdasarkan data dari twitter BMKG (Badan Meteorologi, Klimatologi, dan Geofisika) [2], penerapan *topic modeling* untuk menghasilkan topik atau klaster secara semantik agar mampu merangkum keluhan konsumen pada Biro Perlindungan Keuangan Konsumen di Amerika [3], penerapan *topic modeling* menggunakan LDA untuk mendeteksi keadaan darurat perkotaan di Virginia menggunakan data yang dikumpulkan dari Twitter [4], penerapan *topic modeling* untuk analisis topik dari domain penelitian di University of Kentucky [5], penerapan *topic modeling* untuk klasifikasi ketertarikan penelitian berdasarkan data abstrak pada publikasi [6], penerapan *topic modeling* menggunakan LDA untuk menautkan *user-generated content* dan data *e-commerce* [7], pemodelan penyebaran informasi multi-topik di jejaring sosial menggunakan LDA dan proses Hawkes [8], penerapan *topic modeling* menggunakan LDA dalam pelabelan posting blog pada Wikipedia [9], penerapan topic modeling pada klasifikasi teks *dataless multi-label* [10]. Adapun penelitian terhadap perkembangan *topic modeling* menunjukkan bahwa LDA merupakan metode yang telah membuat dampak besar di bidang Natural Language Processing (NLP) dan *machine learning*, sehingga dengan cepat menjadi salah satu teknik *topic modeling* yang paling populer dalam *machine learning*. [1], [11].

Berdasarkan hasil analisis beberapa penelitian pada *topic modeling*, kami menggunakan metode LDA yang merupakan *state-of-the-art* [1], [6], [11] dalam pemodelan topik penelitian dosen berdasarkan data judul publikasi masing-masing dosen yang diambil dari *Google Scholar*. Setelah data diproses menggunakan LDA, hasilnya menunjukkan apa saja topik penelitian yang dibahas, kata-kata apa saja yang digunakan untuk penelitian, dan apa saja topik yang sering dibahas oleh dosen Jurusan Pendidikan Teknik Elektronika dan Informatika (JPTEI UNY).

METODE

Data Retrieval

Dataset diambil dari akun masing-masing dosen Jurusan Pendidikan Teknik Elektronika dan Informatika Universitas Negeri Yogyakarta (JPTEI UNY) pada *Google Scholar*. *Google Scholar* merupakan *repository* yang berisi publikasi dari hasil penelitian-penelitian yang telah dilakukan. Pengambilan data dilakukan menggunakan teknik *scrapping* yang dibantu oleh ekstensi Google Chrome bernama *Web Scrapper*. Hasil *scrapping* disimpan dalam format .csv dan data yang didapatkan adalah 909 judul publikasi.

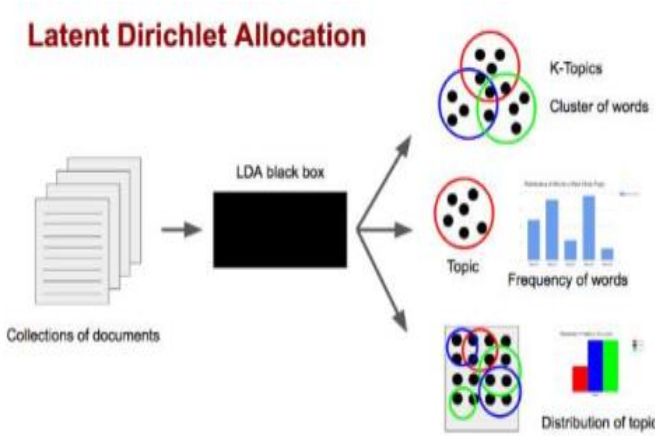
Data Pre-Processing

Data Pre-Processing dibutuhkan untuk menghasilkan data *clustering* yang tepat. Berikut adalah *data pre-processing* yang dilakukan: (1) Mengubah kalimat menjadi kata; (2) Menghilangkan beberapa kata dalam Bahasa Indonesia dan Bahasa Inggris yang tidak memiliki arti. Kata dalam Bahasa Inggris misalnya: *using, of, the, in, on, as, and, based,* dan lain-lain. Kata dalam Bahasa Indonesia misalnya: berbasis, kasus, dan, pada, dan lain-lain. (3) Mengubah susunan suatu kalimat menjadi dalam bentuk bigram, misalnya: "Inovasi Media Pembelajaran Sain Teknologi di SMP Berbasis Mikrokontroler", ketika diubah menjadi bigram: [inovasi, media], [pembelajaran, sain], [teknologi, di], [SMP, berbasis], [mikrokontroller].

Topic Modeling

Topic Modeling atau pemodelan topik adalah sebuh metode *unsupervised machine learning* yang menerapkan pengelompokan untuk menemukan variable laten dari data teks yang besar. Metode yang paling populer untuk pemodelan topik adalah *Latent Dirichlet Allocation (LDA)* yang diperkenalkan oleh Blei dan Jordan, dijelaskan sebagai model probabilistik generatif untuk mencari struktur semantik dari kumpulan korpus yang berdasarkan *hierarchical bayesian analysis* [8]. LDA merupakan kumpulan dokumen sebagai topik campuran yang berisi kata-kata dengan

probabilitas tertentu. Adapaun alur kerja dari LDA dapat dilihat pada Gambar 1.



Gambar 1. Alur Kerja LDA [2]

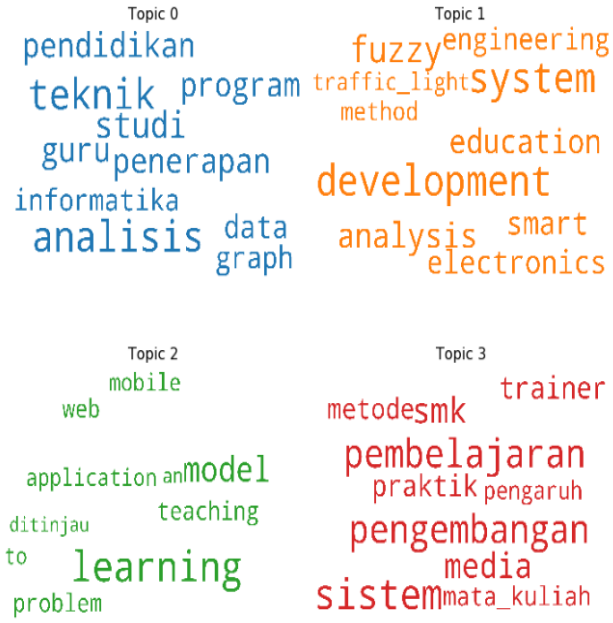
Prosedur cara kerja LDA adalah sebagai berikut: (1) Inisialisasi beberapa parameter, termasuk jumlah dokumen, topik, dan iterasi. Dalam LDA, parameter yang paling penting adalah jumlah topik k. (2) Menetapkan kata untuk topik tertentu secara acak sesuai dengan distribusi dirichlect. (3) Mengulangi masing-masing alur proses untuk semua kata dalam korpus.

Parameter yang digunakan ketika proses perhitungan LDA sebagai berikut: (1) *Random state*: 100; (2) *Update Every*: 1; (3) *Chunk Size*: 10; (4) *Passes*: 10; (5) *Alpha*: *Symmetric;* (6) *Iterations*: 100; (7) *Per Word Topics*: *True*

Dalam menentukan jumlah topik alat ukur yang dipergunakan adalah *coherence value*. Pada penelitian kali dari percobaan menggunakan 1-9 topik, nilai *coherence value* yang didapat sama, yaitu 0,4076. Dari nilai tersebut maka penelitian ini menentukan banyaknya topik secara acak, yaitu 4.

HASIL DAN PEMBAHASAN

Penelitian ini mengelompokkan topik penelitian berdasarkan data publikasi dosen JPTEI pada *Google Scholar* ke dalam empat klaster topik. Visualisasi kata yang muncul pada empat klaster data tersebut dapat dilihat pada Gambar 2.



Gambar 2.Visualisasi Kata

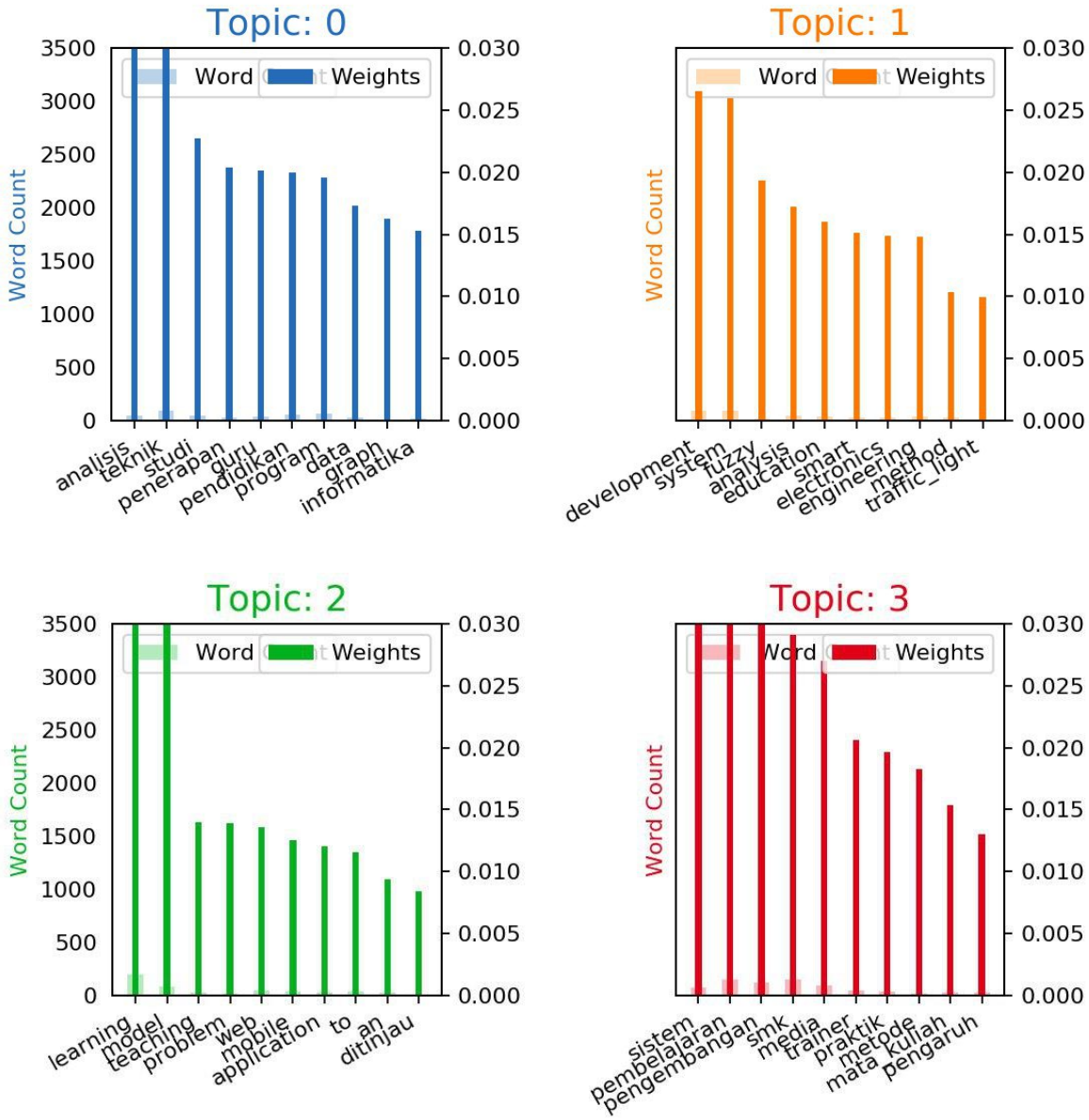
Gambar 2. Menunjukkan bahwa pada Klaster ke-1 (*Topic* 0) berisi kata tentang Pendidikan, tenik, informatika, analisis, graph, dan sebagainya. Dari kata yang muncul pada Klaster ke-1 dapat interpretasikan berbicara tentang Pendidikan Vokasi.

Pada Klaster ke-2 (*Topic* 1) mengandung kata electronics, development, education, system, method, dan sebagainya. Dari kata yang muncul pada Klaster ke-2 dapat interpretasikan berbicara tentang Pengembangan Sistem.

Sedangkan pada Klaster ke-3 (*Topic* 2) adalah mobile, web, model, teaching, problem, dan lain-lain. Dari kata yang muncul pada Klaster ke-3 dapat interpretasikan berbicara tentang Media Pembelajaran.

Dan pada Klaster ke-4 (*Topic* 3) mengandung kata SMK, trainer, pembelajaran, praktik, dan lain sebagainya. Dari kata yang muncul pada Klaster ke-4 dapat interpretasikan berbicara tentang Sistem Pembelajaran di SMK.

Interpretasi di atas diperkuat dengan adanya visualisasi jumlah masing-masing kata yang muncul pada masing setiap topik (Gambar 3).



Gambar 3. Visualisasi Jumlah Kata

Gambar 3. menunjukkan kata yang sering muncul pada Klaster ke-1 (*Topic* 0) adalah Analisis, Teknik, Studi, Penerapan, Guru, Pendidikan, dan Program. Klaster ke-2 (*Topic* 1) adalah *Development*, System, *Fuzzy, Analysis, Education*, dan *Smart*. Klaster ke-3 (*Topic* 2) adalah *Learning*, *Model, Teaching, Problem*, dan Web. Klaster ke-4 (*Topic* 3) adalah Sistem, Pembelajaran, Pengembangan, SMK, Media, dan *Trainer*.

Jika kita lihat secara keseluruhan pada Gambar 3, kata yang sering dipakai untuk judul adalah Analisis, Teknik, *Development, System, Learning*, Model, Sistem, Pembelajaran, Pengembangan, SMK, Media, dan *Trainer*.

Lebih detail lagi beberapa contoh judul penelitian dapat dilihat pada setiap klaster atau topik. Judul-judul penelitian pada keempat topik dapat dilihat pada Tabel 1, 2, 3, dan 4. Judul penelitian yang masuk dalam Klaster ke-1 (*Topic* 0) dapat dilihat pada Tabel 1.

Tabel 1. Judul Penelitian pada Klaster ke-1 (*Topic* 0)

Klaster ke-1 (*Topic* 0)
Analisis penerapan sistem informasi akademik (siakad) 2013 menggunakan model *end-user computing satisfaction* (eucs) di program studi pendidikan teknik informatika
Penerapan *problem based learning* untuk *higher order thinking skills* pada siswa pendidikan teknik informatika
Speech recognizing for presentation tool navigation using back propagation artificial neural network
Studi penelusuran (*tracer study*) terhadap alumni program studi pendidikan teknik informatika jurusan pendidikan teknik elektronika fakultas teknik universitas negeri yogyakarta
Studi Penelusuran Alumni Teknik Elektronika D3 sebagai Upaya Peningkatan Mutu Penyelenggaraan Program Studi

Judul penelitian yang masuk dalam Klaster ke-2 (*Topic* 1) dapat dilihat pada Tabel 2.

Tabel 2. Judul Penelitian pada Klaster ke-2 (*Topic* 1)

Klaster ke-2 (*Topic* 1)
Smart Traffic Light based on IoT and mBaaS using High Priority Vehicles Method
E-module Development for the Subject of Measuring Instruments and Measurement in Electronics Engineering Education
Aplikasi Platform Komputasi Software-Defined Radio (SDR) untuk *Digital Spectrum Analizer*
Android-based applications on teaching skills based on TPACK analysis
Smart System for Lung Disease Early Detection

Judul penelitian yang masuk dalam Klaster ke-3 (*Topic* 2) dapat dilihat pada Tabel 3.

Tabel 3. Judul Penelitian pada Klaster ke-3 (*Topic* 2)

Klaster ke-3 (*Topic* 2)
The development of mobile gamification learning application for web programming learning
Multi-connection phone-based mobile internet to support e-learning and ict literacy for rural community
Mobile internetberbasis a multiconally mobile phone for supporting e-learning
Design of interaction model for interactive e–book
Collaborative learning model with computer supported learning approach

Judul penelitian yang masuk dalam Klaster ke-4 (*Topic* 3) dapat dilihat pada Tabel 4.

Tabel 4. Judul Penelitian pada Klaster ke-4 (*Topic* 3)

Klaster ke-4 (*Topic* 3)
Trainer PID Controller sebagai Media Pembelajaran Praktik Sistem Kendali
Pengembangan *Trainer Internet Of Things* Sebagai Media Pembelajaran Pada Mata Kuliah *Internet Of Things*
Pengembangan Trainer Signal Conditioning
Pengembangan Media Pembelajaran *Trainer Audio Power Amplifier* OCL Dilengkapi VU *Meter* Dan Protektor *Speaker* Untuk Mata Pelajaran Perekayasaan Sistem Audio Di SMK Negeri 1 Magelang 1
Peningkatan Kualitas Pembelajaran Praktik Rangkaian Listrik Melalui Penerapan *Lesson Study*

Dari segi penulis pada setiap topik, dapat dilihat pada Tabel 5, 6, 7, dan 8. Sepuluh penulis dengan jumlah judul penelitian terbanyak pada Klaster ke-1 (*Topic* 0) dapat dilihat pada Tabel 5.

Tabel 5. Nama Penulis pada Klaster ke-1 (*Topic* 0)

Nama	Jumlah Penelitian
Putu Sudira	28
Soenarto	28
Herman Dwi Surjono	17
Fatchul Arifin	13
Adi Dewanto	10
Pramudi Utomo	9
Handaru Jati	8
Dessy Irmawati	7
Sri Waluyanti	5
Eko Marpanaji	5

Sepuluh penulis dengan jumlah judul penelitian terbanyak pada Klaster ke-2 (*Topic* 1) dapat dilihat pada Tabel 6.

Tabel 6. Nama Penulis pada Klaster ke-2 (*Topic* 1)

Nama	Jumlah Penelitian
Adi Dewanto	21
Soenarto	21
Handaru Jati	21
Totok Sukardiyono	20
Sri Waluyanti	17
Suprapto	15
Nurkhamid	15
Herman Dwi Surjono	13
Eko Marpanaji	12
Dessy Irmawati	12

Sepuluh penulis dengan jumlah judul penelitian terbanyak pada Klaster ke-3 (*Topic* 2) dapat dilihat pada Tabel 7.

Tabel 7. Nama Penulis pada Klaster ke-2 (*Topic* 1)

Nama	Jumlah Penelitian
Herman Dwi Surjono	44
Soenarto	33
Handaru Jati	26
Priyanto	15
Fatchul Arifin	11
Nurkhamid	9
Totok Sukardiyono	8
Sri Waluyanti	7
Pipit Utami	5
Suprapto	5

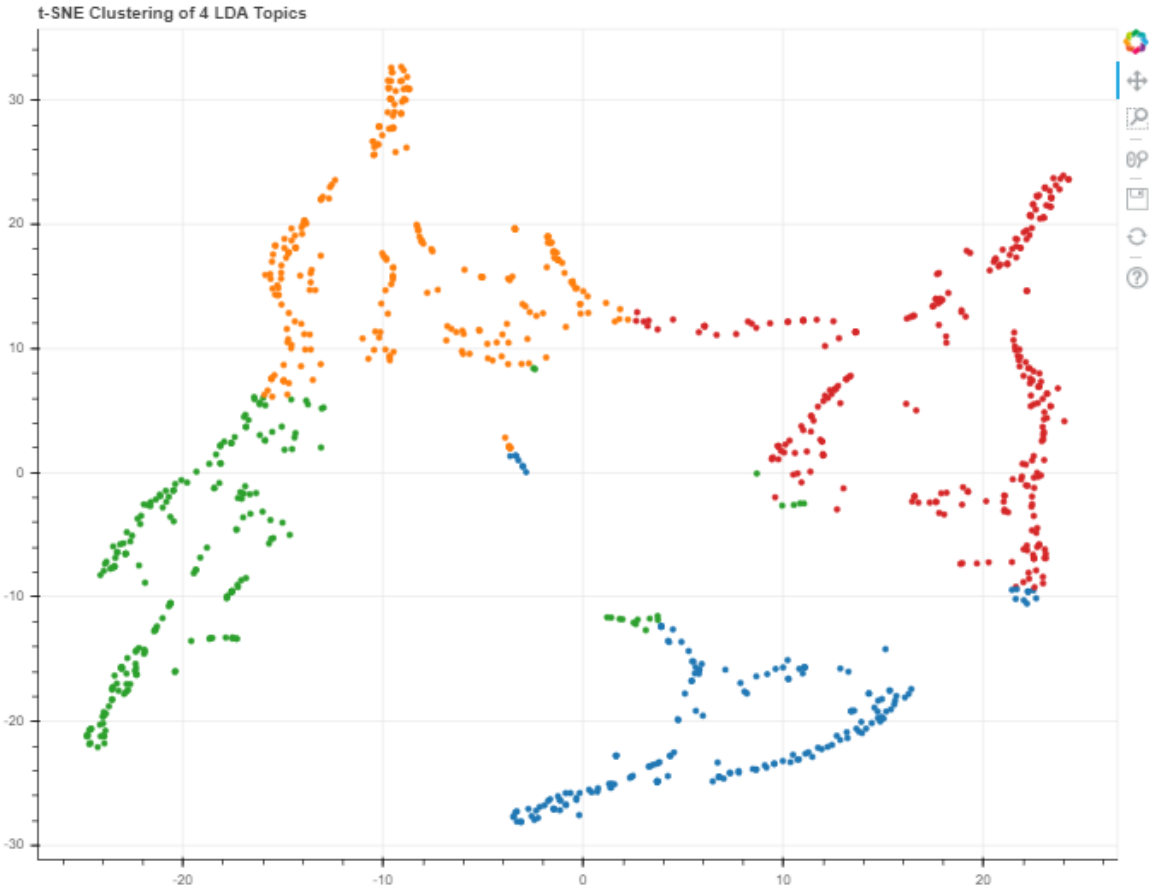
Sepuluh penulis dengan jumlah judul penelitian terbanyak pada Klaster ke-4 (*Topic* 3) dapat dilihat pada Tabel 8.

Tabel 8. Nama Penulis pada Klaster ke-4 (*Topic* 3)

Nama	Jumlah Penelitian
Herman Dwi Surjono	44
Soenarto	34
Sri Waluyanti	18
Putu Sudira	17
Fatchul Arifin	17
Bekti Wulandari	13
Handaru Jati	12
Eko Marpanaji	11
Umi Rochayati	11
Djoko Santoso	10

Dari Tabel 5, 6, 7, dan 8 menunjukkan bahwa Herman Dwi Surjono menjadi penulis terbanyak di Klaster ke-3 (*Topic* 2) dan Klaster ke-4 (*Topic* 3). Adi Dewanto, Soenarto, dan Handaru Jati menjadi penulis terbanyak di Klaster ke-2 (*Topic* 1). Putu Sudira dan Soenarto menjadi penulis terbanyak di Klaster ke-1 (*Topic* 0). Khusus untuk Soenarto, beliau memiliki jumlah publikasi cukup banyak di masing-masing topik. 28 judul penelitian pada Klaster ke-1 (*Topic* 0), 21 judul penelitian pada Klaster ke-2 (*Topic* 1), 33 judul penelitian pada Klaster ke-3 (*Topic* 2), 34 judul penelitian pada Klaster ke-4 (*Topic* 3).

Berdasarkan hasil pengelompokan Klaster 1, 2, 3, dan 4 ditemukan adanya keterkaitan antar klaster atau *topic*. Pada Gambar 4. menunjukkan visualisasi keterkaitan antar klaster atau *topic*.



Gambar 4. Keterkaitan antar Klaster atau *Topic*

Dari segi keterkaitan antar klaster atau *topic* (Gambar 4.), Klaster ke-4 (*Topic* 3) yang berwarna merah, Klaster ke-2 (*Topic* 1) yang berwarna oranye, dan Klaster ke-3 (*Topic* 2) yang berwarna hijau saling memiliki keterkaitan. Hal ini karena Klaster ke-2 (*Topic* 1) dan Klaster ke-3 (*Topic* 2) memiliki banyak kata berbahasa Inggris. Klaster ke-1 (*Topic* 0) yang berwarna biru meskipun menjadi klaster atau *topic* terbesar, ternyata cenderung minim keterkaitan dengan klaster yang lainnya.

SIMPULAN

Pada penelitian ini telah menerapkan pemodelan topik menggunakan LDA (*Latent Dirichlet Allocation*) pada data judul penelitian dosen JPTEI UNY yang telah terpublikasi dan ada pada *Google Scholar*. LDA telah menjadi algoritme yang baik dalam pemodelan topik dan telah dibuktikan pada beberapa penelitian terkait pemodelan topik.

Berdasarkan hasil penelitian kami, hasil yang paling optimum adalah pengelompokan data judul penelitian menjadi empat klaster atau empat topik. Sehingga dapat disimpulkan bahwa topik penelitian dosen JPTEI UNY adalah tentang pendidikan vokasi (Klaster ke-1/*Topic 0*), pengembangan system (Klaster ke-2/*Topic 1*), media pembelajaran (Klaster ke-3/*Topic 2*), dan sistem pembelajaran di SMK (Klaster ke-4/*Topic 3*). Jumlah penelitian dosen JPTEI UNY paling banyak adalah pada Klaster ke-1 dan paling sedikit adalah pada Klaster ke-4.

Adapun kelemahan dalam penelitian ini adalah belum adanya pendekatan khusus untuk teks tidak berbahasa Indonesia pada proses *pre-processing*. Sehingga untuk penelitian selanjutnya, diperlukan optimalisasi pada *data pre-processing.*

REFERENSI

[1] R. Alghamdi and K. Alfalqi, "A Survey of Topic Modeling in Text Mining," *Int. J. Adv. Comput. Sci. Appl.*, vol. 6, no. 1, pp. 147–153, 2015.

[2] A. F. Hidayatullah, S. K. Aditya, Karimah, and S. T. Gardini, "Topic modeling of weather and climate condition on twitter using latent dirichlet allocation (LDA)," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 482, no. 1, 2019.

[3] K. Bastani, H. Namavari, and J. Shaffer, "Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints," *Expert Syst. Appl.*, vol. 127, pp. 256–271, 2019.

[4] Y. Wang and J. E. Taylor, "DUET: Data-Driven Approach Based on Latent Dirichlet Allocation Topic Modeling," *J. Comput. Civ. Eng.*, vol. 33, no. 3, 2019.

[5] S. Joo, I. Choi, and N. Choi, "Topic analysis of the research domain in knowledge organization: A latent dirichlet allocation approach," *Knowl. Organ.*, vol. 45, no. 2, pp. 170–183, 2018.

[6] S. W. Kim and J. M. Gil, "Research paper classification systems based on TF-IDF and LDA schemes," *Human-centric Comput. Inf. Sci.*, vol. 9, no. 1, 2019.

[7] S. Zoghbi, I. Vulić, and M. F. Moens, "Latent Dirichlet allocation for linking user-generated content and e-commerce data," *Inf. Sci. (Ny).*, vol. 367–368, pp. 573–599, 2016.

[8] J. C. L. Pinto and T. Chahed, "Modeling multi-topic information diffusion in social networks using latent dirichlet allocation and hawkes processes," *Proc. - 10th Int. Conf. Signal-Image Technol. Internet-Based Syst. SITIS 2014*, pp. 339–346, 2015.

[9] D. Yokomoto *et al.*, "LDA-based topic modeling in labeling blog posts with wikipedia entries," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7234 LNCS, pp. 114–124, 2012.

[10] D. Zha and C. Li, "Multi-label dataless text classification with topic modeling," *Knowl. Inf. Syst.*, vol. 61, no. 1, pp. 137–160, 2019.

[11] H. Jelodar *et al.*, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimed. Tools Appl.*, vol. 78, no. 11, pp. 15169–15211, 2019.