Pemodelan Topik Berita pada Portal Berita Online Berbahasa Indonesia Menggunakan Latent Dirichlet Allocation (LDA)

Muhammad Andika Nugraha dan Lulu Chaerani Munggaran

Manajemen Sistem Informasi, Program Pasca Sarjana
Universitas Gunadarma
Jl. Margonda Raya No. 100, Depok 16424, Jawa Barat
E-mail: m.andika.nugraha@gmail.com, lulu@staff.gunadarma.ac.id

Abstrak

Media *online* merupakan sarana untuk mendapatkan informasi melalui Internet. Salah satu informasi yang dapat diperoleh adalah berita *online*. Pada tahun 2019 Badan Pusat Statistik melakukan survei tentang pemanfaatan Internet oleh pengguna berumur lima tahun keatas dimana hasil survei menunjukan bahwa 69.9% pengguna Internet memanfaatkan internet untuk mencari informasi atau berita. Berita yang beredar secara *online* di internet terus bertambah seiring berjalanya waktu. Data yang didapat dari laman resmi Kementerian Komunikasi dan Informatika menyatakan bahwa di Indonesia terdapat kurang dari 100 portal berita *online* yang terverifikasi oleh Badan Pers dari total seluruh lebih dari 43.000 portal. Dengan banyaknya data berita yang beredar menimbulkan bermacam tantangan yang harus dihadapi pada era digital sekarang. Salah satu tantangan yang dihadapi adalah melakukan pembuatan pemodelan topik untuk mengkategorikan teks berita berdasarkan topik yang dibahas. Pemodelan topik dalam bentuk teks berita dapat dilakukan dengan memodelkan topik menggunakan *Latent Dirichlet Allocation* (LDA). Proses-proses yang dilakukan untuk membuat suatu pemodelan topik menggunakan LDA diantaranya *input data*, *preprocessing*, *feature extraction*, pembuatan model LDA, dan pengukuran koheren. Setelah dilakukan serangkaian percobaan dan membandingkan hasil dari beragam jumlah pembentukan topik didapatkan nilai koheren terbaik sebesar 0.67 dengan pembentukan topik sebanyak lima topik.

Kata Kunci: Data, Latent Dirichlet Allocation, Media Online, Pemodelan Topik.

Pendahuluan

Kemajuan teknologi memungkinkan seseorang untuk mendapatkan berita dan informasi dengan mudah dan cepat melalui media *online*. Media *online* merupakan satu diantara banyak sarana penyebaran informasi melalui Internet. Pemanfaatan media *online* adalah untuk mencari informasi atau berita melalui laman portal media *online* atau aplikasi seluler yang terhubung dengan Internet.

Di Indonesia terdapat banyak portal media *online*. Data yang dimuat pada laman resmi Menteri Komunikasi dan Informatika (Menkominfo) mengatakan bahwa terdapat kurang dari 100 portal media *online* yang terverifikasi oleh Badan Pers dari total seluruh sebanyak 43.000 [1]. Dengan jumlah portal media yang banyak maka artikel berita yang beredar juga semakin banyak. Data tersebut terus bertambah seiring dengan berjalanya waktu sehingga menyebabkan semakin menumpuknya data yang tersedia.

Data dengan jumlah besar menjadi tantangan tersendiri untuk dapat diolah menjadi bentuk yang lebih bermanfaat. Salah satu bentuk pemanfaatan data-data berjumlah besar tersebut adalah dengan melakukan ekstraksi topik dari data teks berita dengan pemodelan topik *Latent Dirichlet Allocation* (LDA) agar data-data tersebut dapat dikategorikan berdasarkan topik pembahasan di dalamnya.

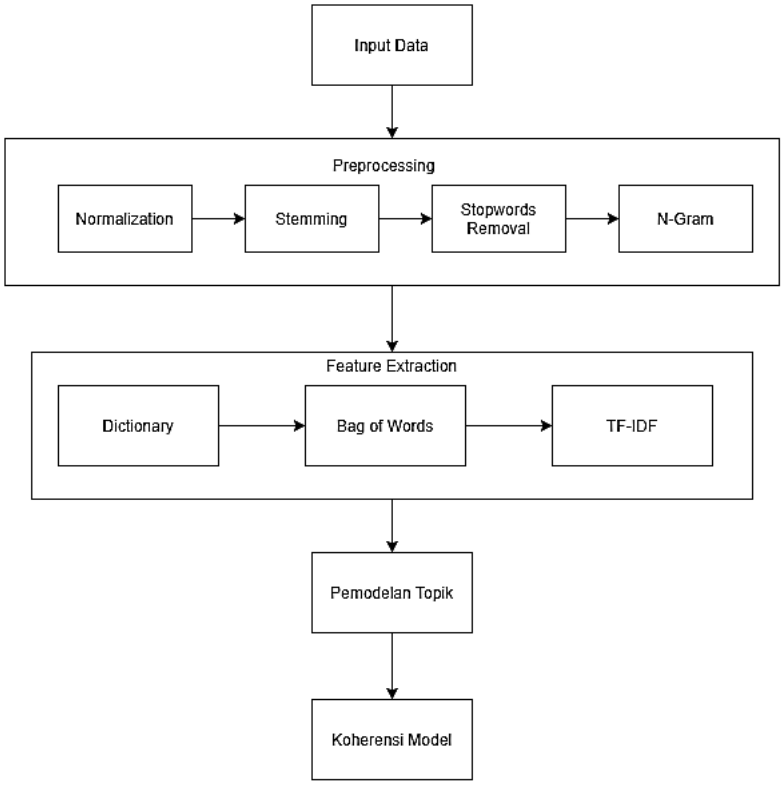
Beberapa penelitian terkait dengan pemodelan topik yang pernah dilakukan diantaranya adalah penerapan LDA untuk klasterisasi cerita berbahasa Bali [2], LDA sebagai pemodelan topik untuk menganalisa relasi antara *hot news* dan *stock market* [3], pemodelan topik untuk menemukan topik pembangunan di Indonesia melalui berita *online* [4], Eksplorasi NMF dan LDA pada artikel berita swedia [5], Pemodelan topik dengan metode *aggregate topic classifier* (ATC) untuk diklasifikasi menggunakan *similarity – based and confidence based topic* classifi*ers* (STC, CTC) pada data *clinical report* [6], Pemodelan topik dengan LDA untuk klasifikasi per-

DOI : http://dx.doi.org/10.32409/jikstik.20.2.2719

tanyaan [7], Pemodelan topik LDA dan klasifikasi Naive Bayes untuk mendeteksi *cyber bullying* di twitter [8], Kombinasi LSTM dan LDA untuk memprediksi kematian berdasarkan data pasien pada unit perawatan kritis [9], Klasifikasi jenis *settlement* menggunakan LDA dan LSTM [10], Pemodelan topik LDA dan LSTM model untuk mempelajari opini [11], pemodelan topik LDA untuk mengklasifikasi data medis [12]. Pada penelitian ini dilakukan pemodelan topik dengan LDA bertujuan untuk membuat model topik yang dapat mengelompokan berita berdasarkan topik bahasan dengan data berasal dari portal berita berbahasa Indonesia.

Metode Penelitian

Dalam melakukan pemodelan topik diperlukan tahapan-tahapan proses yang harus dilakukan. Tahapan proses pada penelitian ini terdapat pada Gambar 1.



Gambar 1: Tahapan Proses Penelitian

Berdasarkan Gambar 1 Tahapan penelitian ini terdiri dari *input data*, *preprocessing*, *feature extraction*, pemodelan topik menggunakan LDA, dan pengukuran koherensi model.

Input Data

Dataset yang digunakan sebagai *input* adalah data berupa teks artikel pada portal berita detikcom dengan kanal berita detikNews. Dikutip dari laman resmi detikcom Kanal detikNews merupakan kumpulan berita tentang berita terbaru setiap harinya dengan bahasan peristiwa, kecelakaan, kriminal, hukum, berita unik, Politik, dan liputan khusus di Indonesia dan Internasional. Data artikel yang digunakan merupakan data yang diterbitkan pada tanggal 1 Oktober 2019 hingga 31 maret 2020.

Preprocessing

Preprocessing merupakan tahap awal untuk melakukan pemodelan topik menggunakan LDA. Tujuan *preprocessing* adalah standarisasi terhadap teks berita dimana pada artikel-artikel yang tersedia terdapat beragam gaya tulisan masing-masing jurnalis. Beberapa proses yang dilakukan pada tahap *preprocessing* adalah *normalization*, *stemming*, *stopwords removal* dan pembentukan frasa dengan N-gram.

Normalization dalam data berbentuk teks merupakan proses transformasi suatu teks menjadi suatu bentuk teks yang dapat diproses oleh model. Beberapa *normalization* yang dilakukan pada penelitian ini adalah *tokenization* yang akan mengubah kalimat menjadi satuan token atau kata-kata, penghilangan tanda baca, penghilangan angka, dan penghilangan karater spesial.

Stemming merupakan proses reduksi suatu kata menjadi bentuk dasar. Tujuanya adalah mengurangi noise dan dimensi pada data. Bahasa Indonesia memiliki karakteristik tersendiri pada kata-katanya sehingga pada proses *stemming* dibutuhkan *library* khusus yang menangani proses *stemming*. Pada penelitian ini menggunakan *libary* sastrawi untuk melakukan *stemming*. Contoh hasil *stemming* adalah perubahan kata "menulis" menjadi bentuk dasarnya "tulis".

Stopwords removal merupakan penghapusan kata-kata yang termasuk ke dalam *stopwords*. *Stopwords* sendiri merupakan kumpulan kata yang sering muncul tetapi memiliki sedikit informasi yang terkandung di dalamnya. Kata-kata yang masuk dalam daftar *stopwords* berbeda-beda pada setiap bahasanya sehingga dibutuhkan librari khusus yang dapat mengandung *stopwords* dalam bahasa Indonesia. *Library* yang digunakan dalam proses *syopwords removal* pada penelitian ini adalah NLTK. Contoh dari kata yang masuk dalam kategori *stopwords* adalah "yang", "jika" "untuk" dan masih banyak lagi.

N-gram merupakan distribusi frekuensi yang akan menggabungkan dua kata menjadi satu kesatuan atau disebut frasa berdasarkan frekuensi kemunculan kedua kata tersebut secara berurutan. Contoh apabila pada suatu datasets terdapat banyak teks yang selalu memunculkan kata "universitas" dan kata "gunadarma" maka kedua kata tersebut akan menjadi "universitas_gunadarma".

Feature Extraction

Feature extraction merupakan proses transformasi suatu format tekstual dari format yang tidak terstruktur menjadi terstruktur agar data dapat diproses pada model. Proses yang dilakukan pada tahap *feature extra*ction adalah pembentukan *dictionary*, mengubah dokumen menjadi $bag - of - words$, dan *term frequency* pada dokumen. *Dictionary* adalah pembentukan token atau kumpulan kata yang di *mapping* dengan suatu id integer. Tujuan dari pembuatan *dictionary* adalah agar model dapat mengenali kata-kata dalam bentuk integer. $Bag - of - words$ adalah perhitungan kemunculan token yang terdapat pada dictionary pada suatu data. TF-IDF atau *Term frequency–Inverse document* adalah proses perhitungan secara statistik yang mencerminkan pentingnya sebuah kata atau token pada suatu corpus. Untuk menghitung TF-IDF dapat menggunakan persamaan (3) yang merupakan turunan dari persamaan (1) dan (2).

$$TF_{ij} = \frac{f_{ij}}{max_k f_{kj}} \quad (1)$$

f_{ij} : frekuensi term i pada dokumen j.

$$IDF_i = log_2\left(\frac{N+1}{n_i+1}\right) + 1 \quad (2)$$

$max_k f_{kj}$: frekuensi term umum atau term dengan frekuensi tertinggi pada dokumen j.

N: jumlah dokumen yang digunakan.

n_i : jumlah dokumen yang memuat term$i$.

Dari persamaan (1) dan (2) maka didapat persamaan (3) untuk menghitung TF-IDF.

$$TFIDF_{ij} = TF_{ij} \cdot IDF_i \quad (3)$$

Pemodelan Topik

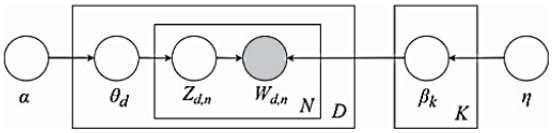
Pemodelan topik merupakan suatu pendekatan untuk menganalisis teks dengan jumlah besar yang tidak terstruktur [13]. Pemodelan topik menggunakan pendekatan secara statistik untuk menemukan topik yang bersifat abstrak pada suatu teks. Cara kerjanya adalah dengan mengelompokan teks berdasarkan kemiripanya. Pemodelan topik merupakan pendekatan *unsupervised leaning* yang memproses korpus dokumen sehingga menghasilkan topik.

Salah satu metode yang digunakan untuk melakukan pemodelan topik adalah *Latent Dirichlet Allocation* (LDA) Menurut Jelodar LDA adalah model probabilistik generatif dari sebuah korpus [14]. LDA merupakan salah satu metode *unsupervised learning* probabilistik paling stabil dalam menemukan latern struktur [15]. LDA pertama kali ditemukan oleh Nlei, N dan Jordan pada tahun 2003 [16].

Langkah-langkah pemodelan topik menggunakan LDA pada setiap dokumen adalah sebagai berikut:

1. Asumsikan terdapat k topik pada semua dokumen.
2. Distribusikan topik k ke seluruh dokumen.
3. Untuk setiap kata w dalam dokumen, asumsikan topiknya salah tetapi setiap kata lain diberi topik yang benar.
4. Secara probabilistik tetapkan kata w pada suatu topik berdasarkan dua hal yaitu:
   (a) Topik apa yang terdapat pada dokumen.
   (b) Berapa kali kata w telah diberi topik tertentu pada setiap dokumen.
5. Melakukan proses untuk setiap dokumen.

Gambar 2 merupakan arsitektur pemodelan topik LDA.



Gambar 2: Arsitektur LDA

Diberikan sebuah dokumen dalam bentuk (w_1, w_1, ..., w_n) dengan jumlah k topik yang diminta, model LDA akan memperkirakan parameter θ dan β. Dimana θ merupakan distribusi topik yang tersembunyi pada setiap dokumen dan β adalah probabilitas setiap kata yang diberi topik. Pada gambar 2, *node* merupakan variabel acak sedangkan *edge* adalah hubungan antar *node*. Variabel w diarsir dengan warna abu-abu karena w adalah satu-satunya variabel yang dapat diamati dalam sistem sedangkan variabel lainya bersifat laten (tersembunyi).

Parameter α adalah data $set-level\ Dirichlet\ prior$ yang diartikan sebagai jumlah observasi sebelumnya dari topik yang dijadikan sampel dalam dokumen sebelum mengamati kata-kata dari dokumen sedangkan parameterη adalah $set-level\ data\ Dirichlet\ prior$ yang dapat diartikan sebagai jumlah observasi sebelumnya dari kata-kata yang diambil sampelnya dari suatu topik sebelum kata apa pun dari kumpulan data tersebut diamati [17]. Berdasarkan gambar 2 variabel-variabel di gambarkan dengan simbol dalam suatu *node* yang terhubung dengan *edge*. Berikut adalah keterangan untuk setiap variabelnya:

α: distribusi topik per dokumen

η: distribusi kata per topik

θ_{d} : sebaran topik tersembunyi dokumen ke-d berdasarkan distribusi multinomial dengan parameter α

β_{k} : variabel level topik dari distribusi probabilitas kata-kata pada topik$k$

$\mathrm{Z_{d,n}}$: topik yang dihasilkan oleh distribusi multinomial dengan parameter θ

$\mathrm{W_{d,n}}$: sampel kata dari distribusi multinomial dengan parameter β dan Z

Variabel $\mathrm{Z_{d,n}}$ dan $\mathrm{W_{d,n}}$ merupakan variabel tingkat kata yang diambil sampelnya sekali untuk setiap kata pada setiap dokumen (n $\in$ {1, 2, ..., N}). Proses pengelompokan metode LDA menyiratkan distribusi gabungan atas variabel laten dan acak yang diamati (W, Z, β, θ) didefinisikan pada persamaan (4).

$$p(W,Z,\beta,\theta|\alpha,\eta) = \prod_{k=1}^{K} p(\beta_k|\eta) \times \prod_{d=1}^{D} p(\theta_d|\alpha)\left(\prod_{n=1}^{N} p(Z_{d,n}|\theta_d)p(W_{d,n}|Z_{d,n},\beta_{d,k})\right) \quad (4)$$

Dalam menggunakan LDA, masalah inferensial utama yang harus dipecahkan adalah menghitung distribusi posterior dari variabel acak tersembunyi yang diberikan kata-kata diamati dalam dokumen. Distribusi posterior ini didefinisikan dalam persamaan (5).

$$p(W,Z,\beta,\theta|\alpha,\eta) = \frac{p(W,Z,\beta,\theta|\alpha,\eta)}{p(W|\alpha,\eta)} \quad (5)$$

$$p(W|\alpha,\eta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n \epsilon Z} p(z_n|\theta)p(w_n|z_n,\beta)p(\beta|\eta)\right)d\theta \quad (6)$$

Pengukuran Model

Pengukuran model adalah tahap penilaian kualitas model topik yang telah dibuat. Untuk mengetahui kualitas model topik yang terbentuk perlu dilakukan suatu pengukuran terhadap model tersebut. Pengukuran yang dapat dilakukan untuk mengetahui kualitas model LDA adalah dengan mengukur nilai koheren antar dokumen pada topiknya.

Salah satu jenis pengukuran koheren pada pemodelan topik adalah pengukuran cv. Pengukuran cv didasarkan pada *sliding windows*, yang merupakan satu set segmentasi kata-kata teratas dan ukuran konfirmasi tidak langsung yang menggunakan *Normalized Pointwise Mutual Information* (NPMI) dan *cosine similarity* [18]. Pada pengukuran koheren dengan metode cv terbagi menjadi empat bagian. Bagian pertama adalah segmentasi data menjadi term yang berpasangan. Bagian kedua adalah menghitung probabilitas term atau term yang berpasangan. Bagian ketiga menghitung ukuran konfirmasi keterkaitan antar satu term dengan term lainya. Bagian terakhir adalah agregasi langkah konfirmasi individu menjadi koherensi keseluruhan.

Hasil dan Pembahasan

Setelah serangkaian proses penelitian mulai dari *input* data, *preprocessing*, *feature extraction*, pemodelan topik dengan LDA, dan pengukuran koheren maka didapatkan hasil-hasil sebagai berikut.

- Data input merupakan kumpulan berita pada portal berita detikcom kanal detikNews yang di terbit pada 1 Oktober 2019 hingga 31 maret 2020 dengan jumlah sebanyak 68.537 artikel berita. Data diolah dengan melakukan proses *preprocessing* sehingga data yang tersedia dapat diproses oleh model.
- Proses *preprocessing* yang dilakukan diantaranya adalah *normalization*, *stemming*, *stopwords removal*, dan pembentukan N-Gram dengan N sebanyak 2 (Bi-Gram) sehingga menghasilkan datasets yang dapat pada proses *feature extraction*. Beberapa contoh hasil dari tahap preprocessing adalah "tumbuh ekonomi kreatif Indonesia dorong

usaha ...", "kondisi jakarta kondusif wilayah duduk massa a..." dan "operasi koridor transjakarta blok m kota buka ...".

- Proses *feature extraction* dilakukan agar datasets dapat diolah pada pemodelan topik. Proses *feature extraction* yang dilakukan adalah pembuatan *dictionary* yang menghasilkan token unik sebanyak 32.057 token. Token-token tersebut dilakukan proses *mapping* pada tahap $bag - of - words$ sehingga dokumen direpresentasikan dalam bentuk id integer. Tahap terakhir pada feature extraction adalah menghitung *term frequency* menggunakan TF-IDF dan menghasilkan dokumen yang dalam bentuk term frequency terhadap kumpulan dokumen atau korpus.
- Hasil dari *feature extraction* diproses pada pemodelan topik menggunakan LDA. Proses LDA dilakukan dengan membuat 19 percobaan dengan target topik sebanyak dua hingga dua puluh topik.
- Seluruh hasil percobaan dihitung dengan menilai nilai koheren dengan menggunakan pendekatan cv. Hasil dari perhitungan koheren dapat dilihat pada Tabel 1 yang merupakan nilai koheren berdasarkan banyak topik pada setiap pembentukan model topik dengan LDA.

Tabel 1: Nilai Koheren Pemodelan Topik

Jumlah Topik	Nilai Koheren
2	0,478398
3	0,583241
4	0,654224
5	0,674946
6	0,556033
7	0,473711
8	0,584828
9	0,579336
10	0,551277
11	0,50422
12	0,406168
13	0,466745
14	0,516764
15	0,494738
16	0,428583
17	0,477565
18	0,491564
19	0,522031
20	0,462459

Tabel 1 dan Gambar 3 merupakan hasil koheren dari *input* data berupa artikel bertita sebanyak 68.537 artikel yang telah dilakuukan *preprocessing* terhadap model LDA dengan jumlah target topik dua hingga dua puluh topik yang terbentuk. Berdasarkan Tabel 1 dan Gambar 3 dapat disimpulkan bahwa nilai koheren terbaik adalah sebesar 0.67 dengan jumlah topik yang dibentuk oleh pemodelan LDA sebanyak 5 buah topik berbeda. Setiap pemodelan topik memiliki keyword yang merepresentasikan topik dan nilai kontribusi terhadap topiknya. Tabel 2 merupakan topik yang terbentuk dan nilai kontribusi setiap keyword pada pemodelan topik yang terbentuk dengan 5 topik berbeda.

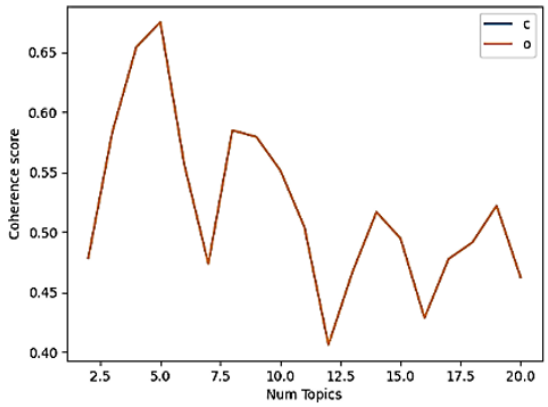
Tabel 2: Top keyword LDA pada Bentuk 5 Topik

Topik	Keyword	Bobot
Topik_0	jokowi	0.005
	partai	0.004
	presiden	0.004
	dki	0.004
	dpr	0.004
	anies	0.003
	ketua	0.003
	menteri	0.003
	pan	0.003
	agama	0.003
Topik_1	kpk	0.014
	dakwa	0.006
	sidang	0.006
	jaksa	0.006
	sangka	0.005
	hukum	0.005
	uang	0.004
	hakim	0.004
	adil	0.004
	pidana	0.004
Topik_2	korban	0.009
	sangka	0.004
	mobil	0.004
	tangkap	0.004
	kendara	0.004
	video	0.003
	aksi	0.003
	polres	0.003
	aman	0.003
	motor	0.003
Topic_3	banjir	0.008
	desa	0.005
	air	0.005
	camat	0.004
	bakar	0.004
	kabupaten	0.004
	korban	0.004
	sungai	0.003
	makam	0.003
	lokasi	0.003
Topic_4	corona	0.016
	virus	0.012
	pasien	0.011
	covid	0.009
	positif	0.007
	sehat	0.006
	sakit	0.004
	sebar	0.004
	pdp	0.004
	isolasi	0.003

Tabel 2 menuunjukan sepuluh token yang paling berpengaruh pada setiap topik yang terbentuk. Setiap token memiliki bobot yang menunjukan relevansi antar artikel dalam suatu topik. Semakin besar bobot token yang dimiliki maka se-

makin berpengaruh token tersebut dalam menentukan topik suatu artikel. Tabel 3 merupakan persentase banyaknya data yang tersebar pada masing-masing topik dan persentasi token yang dimiliki pada masing-masing topik.

Gambar 3 merupakan nilai koheren model LDA yang disajikan dalam bentuk grafik.



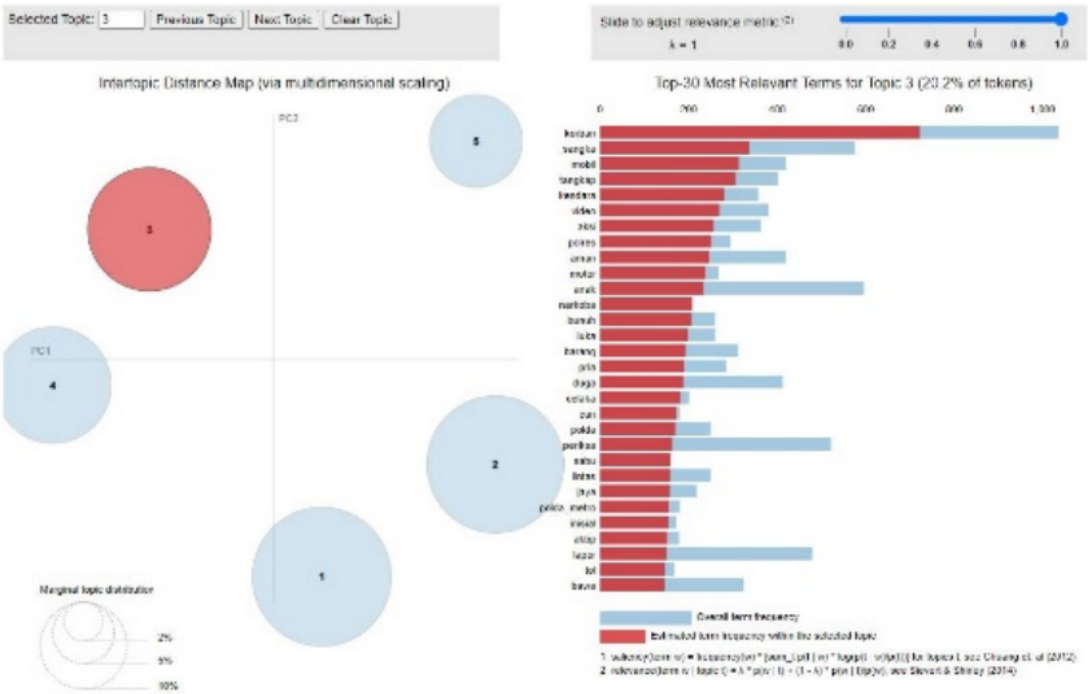
Gambar 3: Grafik Nilai Koheren

Tabel 3 menunjukan persebaran artikel adalah persentase terhadap total artikel, sedangkan persebaran token adalah persentase terhadap jumlah *dictionary* yang terbentuk. Persentase yang dihasilkan adalah hasil permodelan topik dengan target sebanyak lima topik dari input artikel sebanyak 68.537 artikel. Setiap topik yang terbentuk memiliki jarak kedekatan dengan topik lainya kedekatan tersebut menggambarkan kesamaan atau similiaritas antar topik yang dapat dapat di visualisasikan dengan bantuan peta jarak.

Gambar 4 merupakan visualisasi jarak antar topik yang menunjukan jarak kedekatan atau similiaritas satu topik dengan topik lainya dan token yang terdapat pada topik-topik yang terbentuk. Lingkaran 3 (topic_2) berdekatan dengan lingkaran 4 (topic_3) menunjukan bahwa kedua topik memiliki kedekatan atau similiaritas yang dekat berdasarkan model yang dibentuk. Lingkaran 3 (topic_2) menunjukan bahwa pada topik memiliki token teratas berupa "korban", "sangka", "mobil", dan lainya. Token yang diarsir dengan warna merah menunjukan banyaknya kemunculan token pada topik yang dipilih sedangkan arsir berwarna biru menunjukan banyaknya token yang dipilih pada topik-topik lainya.

Tabel 3: Persentase Persebaran Data dan Token pada Topik

Nama Topik	Total Artikel	Persebaran Artikel	Persebaran Token
Topik_0	17556	25.62%	24.9%
Topik_1	6668	9.73%	11.4%
Topik_2	14326	20.90%	20.2%
Topik_3	11677	17.04%	17.9%
Topik_4	18310	26.72%	25.6%



Gambar 4: Peta Jarak Antar Topik dan Distribusi Token

Penutup

Penelitian ini berhasil membuat pemodelan topik menggunakan *Latent Dirichlet Allocation* (LDA) pada datasets berita berbahasa Indonesia. Hasil terbaik didapat dengan membentuk lima buah topik dari total sebanyak 68.537 artikel dengan nilai koheren sebesar 0.67 menggunakan pengukuran cv. Model juga berhasil mengelompokan artikel pada setiap topik dengan persentase penyebaran artikel sebesar 25.62% artikel pada topik_0, 9.73% artikel pada topik_1, 20.90% artikel pada topik_2, 17.04% artikel pada topic_3, dan 26.72% artikel pada topic_4. Untuk penelitian selanjutnya dapat dilakukan dengan beberapa penambahan diantaranya adalah memperbanyak artikel agar semakin banyak jumlah token dan topik yang terbentuk, membuat frase dengan tiga kata guna menambah variasi token dengan model Tri-Gram.

Daftar Pustaka

[1] Anonym, "Menkominfo: Baru 100 Portal Berita Online Terverifikasi," Komunikasi dan Informatika, 2018, diakses daring pada: https:// kominfo.go.id/content/detail/12345/ menkominfo-baru-100-portal-berita-online-terverifikasi /0/berita_satker (accessed Mar. 01, 2021).

[2] N. A. Sanjaya ER, "Implementasi Latent Dirichlet Allocation (LDA) untuk Klasterisasi Cerita Berbahasa Bali", Jurnal Teknologi Informasi dan Ilmu Komputer, doi: 10.25126/jtiik.0813556, 2021 .

[3] C. Liu, "Analysis of relationship between hot news and stock market - Based on LDA model and event study", Journal Physic Conferences Series, vol. 1616, no. 1, doi: 10.1088/1742-6596/1616/1/012048, 2020.

[4] A. F. Hidayatullah and M. Rifqi, "Indonesia Infrastructure Development Topic Discovery on Online News with Latent Dirichlet Allocation Indonesia Infrastructure Development Topic Discovery on Online News with Latent Dirichlet Allocation", IOP Conference Series: Materials Science and Engineering, vol. 1077-012012, doi: 10.1088/1757-899X/1077/1/012012, 2021.

[5] J. Blad, and K. Svensson, "Exploring NMF and LDA Topic Models of Swedish News Articles News Articles", Thesis, ISSN 1650-8319, Uppsala University, December, 2020.

[6] E. S. Kayi, K. Yadav, and H.-A. Choi, "Topic Modeling for Classification of Clinical Reports", 51st Annual Meeting of the Association for Computational Linguistics, pp:67-73, DOI:10.13140/2.1.1740.6720, 2013.

[7] S. Momtazi, "Unsupervised Latent Dirichlet Allocation for supervised question classification", Inf. Process. Manag., vol. 54, no. 3, pp. 380–393, doi: 10.1016/j.ipm.2018.01.001, 2018.

[8] K. Nalini and L. Jaba Sheela, "Classification using Latent Dirichlet Allocation with Naive Bayes Classifier to detect Cyber Bullying in Twitter", Indian J. Sci. Technol., vol. 9, no. 28, pp. 3–7, doi: 10.17485/ijst/2016/v9i28/93825, 2016, .

[9] Y. Jo, L. Lee, and S. Palaskar, "Combining LSTM and Latent Topic Modeling for Mortality Prediction", [Online]. Available: http://arxiv.org/abs/1709.02842, 2017.

[10] R. Huang, H. Taubenböck, L. Mou, and X. X. Zhu, "Classification of settlement types from tweets using LDA and LSTM", Int. Geosci. Remote Sens. Symp., vol. 2018-July, pp. 6408–6411, doi: 10.1109/IGARSS.2018.8519240, 2018 .

[11] Q. Ye et al., "Using LDA and LSTM Models to Study People ' s Opinions and Critical Groups Towards Congestion Pricing in New York City through 2007 to 2019", Business, Computer Science, pp. 1–12, 2020.

[12] M. Selvi, K. Thangaramya, M. S. Saranya, K. Kulothungan, S. Ganapathy, and A. Kannan, "Classification of medical dataset along with topic modeling using LDA", vol. 511. Springer Singapore, 2019.

[13] B. Sohrabi, I. Raeesi Vanani Mohsen Baranizade Shineh, and M. Baranizade Shineh, "Topic Modeling and Classification of Cyberspace Papers Using Text Mining", J. Cybersp. Stud., vol. 2, no. 1, pp. 103–125, doi: 10.22059/jcss.2017.239847.1009, 2018.

[14] H. Jelodar et al., "Latent Dirichlet Allocation (LDA) and Topic modeling: Models, applications, a survey", arXiv, pp. 15169–15211, 2017.

[15] C. Karmakar, C. O. Dumitru, G. Schwarz, and M. Datcu, "Feature-Free Explainable Data Mining in SAR Images Using Latent Dirichlet Allocation," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 14, pp. 676–689, doi: 10.1109/JSTARS.2020.3039012, 2021 .

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation", J. Mach. Learn. Res., doi: 10.1016/b978-0-12-411519-4.00006-9, 2003.

[17] A. Bagheri, A. Sammani, P. G. M. van der Heijden, F. W. Asselbergs, and D. L. Oberski, "ETM: Enrichment by topic modeling for automated clinical sentence classification to detect patients' disease history", J. Intell. Inf. Syst., vol. 55, no. 2, pp. 329–349, 2020, doi: 10.1007/s10844-020-00605-w.

[18] S. Syed and M. Spruit, "Full-Text or abstract? Examining topic coherence scores using latent dirichlet allocation", Proc. - 2017 Int. Conf. Data Sci. Adv. Anal. DSAA 2017, vol. 2018-Janua, pp. 165–174, doi: 10.1109/DSAA.2017.61, 2017.