IDENTIFIKASI TOPIK ARTIKEL BERITA MENGGUNAKAN *TOPIC MODELLING* DENGAN *LATENT DIRICHLET ALLOCATION*

[1]Vira Faradhiba Rusdhi, [2]Ilmiyati Sari
[1, 2]Fakultas Teknologi Industri Universitas Gunadarma
Jl. Margonda Raya No. 100, Depok 16424, Jawa Barat
[1] virafaradhibarusdhi@gmail.com, [2]ilmiyati@staff.gunadarma.ac.id

Abstrak

Portal berita memberikan informasi yang sangat beragam, namun judul berita tidak dapat dijadikan acuan utama dalam penentuan topik suatu berita secara keseluruhan karena judul berita bersifat hipebola untuk menarik pembaca. Oleh karena itu, penelitian ini mengusulkan sistem identifikasi topik artikel berita menggunakan topic modelling dengan algoritma Latent Dirichlet Allocation (LDA). Tahapan penelitian diawali dengan pengambilan data secara otomatis dari situs web detik.com dan tempo.co dengan proses web scrapping, kemudian dilakukan preprocessing terhadap data. Ada 4 tahap preprocessing yaitu tokenization, case folding, stopword removal, dan stemming. Tahap terakhir adalah topic modelling dengan algoritma LDA. Topic modelling merupakan model statistik untuk menentukan inti atau topik pada kumpulan dokumen. Identifikasi topik dengan algoritma LDA didasarkan pada probabilitas kemunculan kata dalam kumpulan dokumen. Penelitian ini menghasilkan topik yang paling sering muncul dalam portal berita kriminal adalah pembunuhan.

***Kata Kunci**: Berita, Latent Dirichlet Allocation, Topic Modelling, Preprocessing.*

Abstract

News portals provide very diverse information, but news titles cannot be used as the main reference in determining the topic of article as a whole because news titles are hyperbolic to attract readers. Therefore, this study proposes a news article topic identification system using topic modeling with Latent Dirichlet Allocation (LDA) algorithm. The research stage begins with automatic data retrieval from the detik.com and tempo.co websites with web scrapping, then preprocessing the data is carried out. There are 4 stages of preprocessing, namely tokenization, case folding, stopword removal, and stemming. The last stage is topic modeling with LDA algorithm. Topic modeling is a statistical model for determining the core or topics in a document set. Topic identification with the LDA algorithm is based on the probability of occurrence of a word in documents set. This study resulted in the topic that most often appears in news portals is murder.

***Keywords**:News, Latent Dirichlet Allocation, Topic Modelling, Preprocessing.*

PENDAHULUAN

Saat ini, portal berita digital telah menjadi salah satu sumber berita terpenting bagi penguguna internet [1]. Hampir setiap media nasional maupun internasional memiliki portal berita digital disamping media konvensional seperti koran dan majalah. Walaupun pada awalnya media *online* hanya berperan sebagai tempat alternatif publik

Indonesia untuk menyalurkan opininya khususnya pada era pemerintahan Soeharto [2], namun portal berita *online* memiliki keunggulan yaitu lebih cepat dan luas dalam menyebarkan berita. Pada portal berita terdapat informasi yang berasal dari daerah, dalam negeri dan juga luar negeri dengan banyak berbagai jenis topik yang diulas.

Dengan banyaknya berita yang diulas pada portal berita, diperlukan suatu metode untuk identifikasi topik yang paling sering muncul dan menjadi *trend* pada berita secara otomatis. Identifikasi topik dalam kumpulan dokumen dapat dilakukan dengan *topic modelling*. *Topic Modelling* adalah teknik untuk menemukan pola kata dalam kumpulan dokumen menggunakan model probabilistik hierarkis [3]. *Latent Dirichlet Allocation* (LDA) adalah salah satu algortima yang menonjol untuk *topic modelling* [4]. Setiap kata pada algortima LDA dianggap memiliki tingkat kepentingan yang sama [5].

Banyak penelitian yang telah dilakukan mengenai *topic modelling* dengan LDA. Beberapa penelitian *topic modelling* dengan LDA pada domain data berbeda, antara lain Putra (2017) melakukan analisis topik informasi publik media sosial di Surabaya menggunakan LDA [6]. Alfanzar, Khalid dan Ronzar (2020) melakukan *topic modelling* skripsi menggunakan LDA [7]. Sotijohatmo dkk (2020) melakukan analisis LDA untuk klasifikasi dokumen tugas akhir berdasarkan pemodelan topik [8]. Chilmi (2021) melakukan identifikasi topik pembicaraan warganet twitter tentang *Omnibus Law* dengan LDA [9]. Wang dkk [10] melakukan *topic modelling* berdasarkan ukuran kesamaan teks dengan LDA dan *Labeled Latent Dirichlet Allocation* (LLDA) untuk dokumen keputusan pengadilan di Cina.

Detik.com dan Tempo.co adalah 2 situs berita yang ada di Indonesia. Berdasarkan *Alexa rank* [11], Detik.com adalah portal berita yang menempati peringkat 9 di Indonesia dengan total pengunjung 217.828 pengunjung/hari, sedangkan tempo.co menempati urutan ke 99 dengan total pengunjung 31.184 pengunjung/hari secara rata-rata dalam waktu 3 bulan. Situs Tempo dipilih karena merupakan media online pertama di Indonesia dan Detik.com adalah media *online* berupa portal berita pertama di Indonesia yang benar-benar menjual konten dan menerbitkan informasi secara *update* dan *real time* [12]. Penelitian ini bertujuan melakukan indentifikasi topik berita menggunakan *topic modelling* dengan LDA pada 2 situs berita tersebut. Penelitian ini menggunakan visualisasi hasil LDA dengan beberapa jumlah topik dan jumlah iterasi untuk mengetahui klaster topik yang terbentuk dan hal ini merupakan pembeda penelitian ini dengan penelitian yang telah ada sebelumnya.

METODE PENELITIAN

Tahapan penelitian diawali dengan pengambilan data secara otomatis dari situs web detik.com dan tempo.co dengan proses

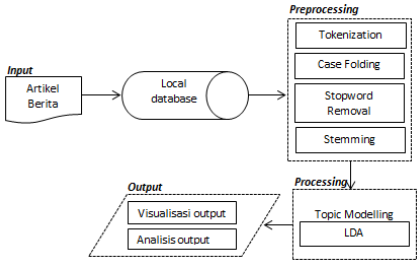
web scrapping kemudian dilakukan *preprocessing* data. Ada 4 tahap *prepocessing* yaitu *tokenization*, *case folding*, *stopword removal*, dan *stemming*. Tahap terakhir penelitian adalah *topic modelling* dengan algoritma LDA. Hasil *topic modelling* kemudian divisualisasikan dan dianalisis sehingga diperoleh kesimpulan topik yang paling sering dibahas atau sedang *trend*. Bagan Alur penelitian dapat dilihat pada Gambar 1. Metode *text mining* diterapkan dalam penelitian. Berikut penjelasan dari tiap tahapan yang dilakukan:

A. Pengambilan Data

Data pada penelitian ini diambil dari 2 situs berita nasional yaitu Detik.com dan Tempo.co. Penelitian ini menggunakan 20 artikel berita kriminal dari kedua situs ini. Pengambilan data menggunakan metode *scrapping* dengan *Beautiful Soup* yang digunakan untuk mengekstrak semua teks dan menyimpannya dalam *local database*. *Beautiful Soup* adalah *library* Phyton untuk menarik data dari berkas HTML dan XML [13]. Setelah *corpus* atau kumpulan artikel dari hasil *scrapping* disimpan dalam *database*, proses pembersihan data dilakukan untuk menghapus *HTML tags*, *double quotes*, *newline*, dan *multiple space* yang biasanya ada pada artikel berita *online*.

B. Preprocessing

Tahap *preprocessing* berguna sebagai persiapan data sebelum diproses menggunakan LDA. *Preprocessing* dilakukan dengan menggunakan aplikasi Jupyter-Notebook. Tahap ini dimulai dengan mengunduh dan mengaktifkan beberapa *library* yang dibutuhkan. Data pada tahap *preprocessing* data mengalami 4 proses, yaitu *tokenization*, *Case Folding*, *Stopword Removal* dan *Stemming/ Lemmatization*. *Tokenization* adalah proses untuk menghilangkan karakter lain selain huruf, seperti angka dan tanda baca [14]. *Case Folding* dibutuhkan untuk mengubah huruf kapital sehingga tidak ada huruf yang memiliki 2 bentuk yang berbeda, karena akan dianggap sebagai huruf yang berbeda oleh komputer [15]. Proses *Stopword removal* bertujuan untuk membuang kata yang tidak bermakna, seperti "yang", "di", "ke", "dari", "adalah", "dan", "atau" [16]. *Stemming* adalah proses untuk mengubah kata berimbuhan menjadi kata dasarnya [17], misalnya kata pembunuhan menjadi bunuh.



Gambar 1. Bagan Alur Penelitian

C. Topic Modelling

Setelah data selesai diproses pada tahap *preprocessing*, langkah berikutnya adalah *topic modelling*. Algortima LDA dipilih untuk *topic modelling* pada penelitian ini. Ide mendasar dari LDA adalah kumpulan dokumen direpresentasikan sebagai campuran dari topik-topik dan topik direpresentasikan sebagai campuran kata yang tersembunyi dan belum belum diketahui. Dalam bahasa statistika, kumpulan dokumen adalah *probability density function* (pdf) dari topik dan topik adalah *probability density function* (pdf) dari kata.

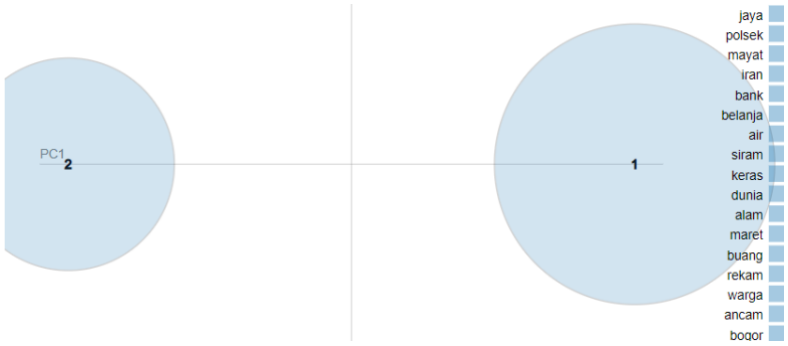
Pada *topic modelling*, kumpulan dokumen disebut *corpus* yang direpresentasikan sebagai *Document Term Matrix* (DTM) atau kadang disebut *document word matrix*. Baris DTM merepresentasikan dokumen sedangkan kolom merepresentasikan kata, jadi jika suatu penelitian menggunakan M buah dokumen dan N unik kata pada semua dokumen tersebut maka DTM berorde $M \times N$. Jika dokumen ke-i mengandung kata ke-j, maka elemen DTM pada baris i kolom j bernilai 1, namun sebaliknya jika dokumen i tidak mengandung kata ke-j maka elemen DTM pada baris i kolom j adalah 0. LDA mendekomposisi DTM menjadi dua matriks yaitu *document topic matrix* dan *topic word matrix*, dengan *document topic matrix* berorde $M \times K$ dan *topic word matrix* berorde $K \times N$, K adalah jumlah topik dalam *corpus* yang nilainya dimasukkan oleh pengguna. Tujuan akhir dari LDA adalah untuk menemukan representasi yang paling optimal dari *document topic matrix* dan *topic word matrix* untuk menemukan distribusi dokumen-topik dan distribusi topik-kata yang paling optimal.

HASIL DAN PEMBAHASAN

Pembuatan sistem identifikasi topik artikel berita pada penelitian ini menggunakan perangkat keras dengan spesifikasi sebagai berikut: *processor* Intel(R)Core(TM)i7-8750HCPU@2.20GHz 2.21GHz, RAM 8GB, dan *hard disk* SSD 256GB, sedangkan perangkat lunak yang digunakan adalah Microsoft Windows 10 Pro 64-bit, *Text Editor Jupyter Notebook* dan *Visual Studio*, Browser Google Chrome dan Anaconda Python 3.7. Gambar 2 adalah visualisasi pemodelan topik menggunakan algortima LDA 2 topik dan jumlah iterasi 500. Lingkaran berwarna biru pada Gambar 2 merepresentasikan topik yang terbentuk dan beberapa kata yang ada disebelah kanan pada Gambar 2 adalah kata-kata yang memiliki probabilitas tertinggi dalam corpus. Lingkaran biru pada Gambar 2 tidak beririsan berarti kedua topik yang terbentuk saling lepas atau tidak berkaitan. Pada algoritma LDA jumlah topik ditentukan oleh pengguna, sehingga dilakukan beberapa percobaan untuk menetukan jumlah topik yang sesuai. Tabel 1 adalah visualisasi hasil identifikasi topik model dengan LDA dari berbagai percobaan dengam jumlah iterasi dan jumlah topik yang beragam. Percobaan

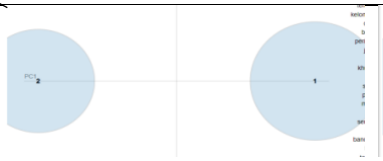
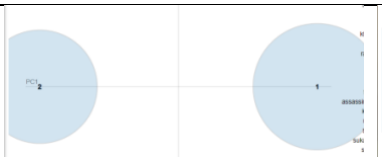
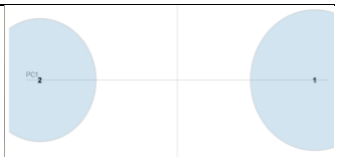
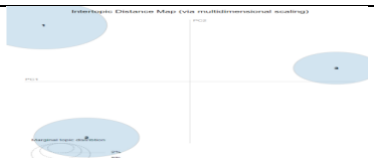
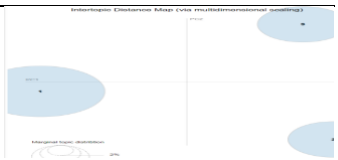
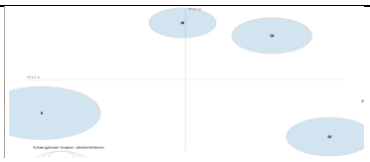
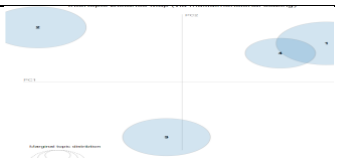
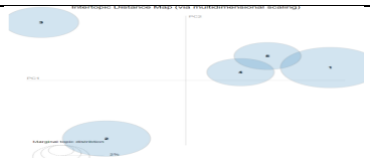
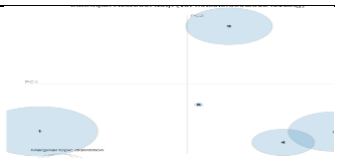
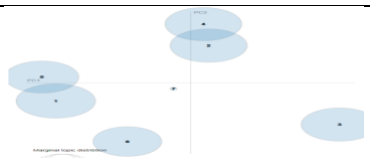
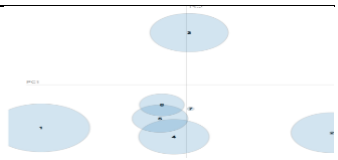
dilakukan dengan jumlah iterasi 500, 1000 dan 5000, sedangkan jumlah iterasi yang diuji cobakan adalah 2, 3, 4, 5, dan 7. Pada jumlah topik awal 2, dapat dilihat pada Tabel 1, terdapat 2 lingkaran berwarna biru yang artinya terbentuk 2 topik sesuai dengan jumlah topik yang diinginkan pengguna. Dua topik yang terbentuk tidak beririsan berarti kedua topik ini berbeda (tidak mengandung kata yang sama pada 10 kata yang memiliki probabilitas tertinggi) berapapun jumlah iterasinya.

Pada jumlah topik 2, jumlah iterasi tidak mempengaruhi jarak antara kedua lingkaran tersebut, berbeda halnya dengan jumlah topik 4. Pada jumlah iterasi 500 dan 1000 terdapat 4 lingkaran biru yang saling lepas, berarti ke-4 topik yang dihasilkan dengan LDA saling independen (tidak memiliki kata yang sama), sedangkan pada jumlah iterasi 5000 ada 2 lingkaran yang beririsan. Lingkaran yang beririsan ini menandakan bahwa ada topik yang sebenarnya tidak berbeda atau memiliki kesamaan kata. Dengan perkataan lain, dengan jumlah topik awal yang diinginkan pengguna 4, terbentuk 3 klaster topik pada jumlah iterasi 5000. Pada jumlah topik awal yang diberikan 5, untuk jumlah iterasi 500 terbentuk 4 klaster topik sedangkan di iterasi 1000 dan 5000 hanya terbentuk 3 klaster topik Pada jumlah topik awal yang diberikan 7, untuk semua iterasi yang diuji cobakan yaitu 500, 1000 dan 5000, terbentuk 4 topik yang saling bebas. Dari percobaan ini dapat disimpulkan bahwa jika jumlah topik diperbesar maka klaster topik yang terbentuk tidak akan lebih dari 4.



Gambar 2. Visualisasi LDA 2 Topik dan Iterasi 500

Tabel 1. Visualisasi Hasil LDA dengan Jumlah Iterasi dan Jumlah Topik Beragam

Jumlah Topik \ Iterasi	500	1000	5000
2			
3			
4			
5			
7			

Tabel 2. Sepuluh Kata dengan Probabilitas Tertinggi pada Setiap Topik

Topik 1	Topik 2	Topik 3
bunuh	Bunuh	bunuh
harry	Iran	mayat
lapor	Hukum	duga
fs	rumah	rumah
medan	perintah	periksa
rekam	duga	luka
tewas	louiker	jadi
saudi	tua	hubung
pm	bekas	jakarta
habis	tangkap	camat

Dari Tabel 1 disimpulkan pemodelan dengan jumlah topik 2, 3, 4, 5 dan 7 dengan iterasi 500, 1000 dan 5000 mengerucut bahwa jumlah topik 3 merupakan pemodelan topik yang paling sesuai. Tabel 2 adalah sepuluh kata dengan probabilitas tertinggi dari dalam setiap topik dengan iterasi 5000. Berdasarkan Tabel 2 dapat disimpulkan bahwa topik artikel

berita kriminal dari 2 portal berita Detik.com dan Tempo.co adalah mengenai pembunuhan, dengan topik 1, 2, serta 3 berbeda kasus, tempat kejadian dan pelaku.

KESIMPULAN DAN SARAN

Sistem identifikasi topik artikel berita menggunakan *topic modelling* dengan *Latent Dirichlet Allocation* (LDA) berhasil dibuat dan diperoleh topik dari kumpulan artikel kriminal yang diambil dengan *scrapping* pada 2 situs portal berita, Detik.com dan Tempo.co, adalah Pembunuhan. Tahapan penelitian dimulai dari pengambilan data. Data penelitian berupa artikel berita *online* berjumlah 20 artikel. Sebelum melakukan *topic modelling* dengan LDA, data terlebih dahulu dibersihkan untuk menghapus *HTML tags*, *double quotes*, *newline*, dan *multiple space*, kemudian dilakukan *preprocessing* data dengan 4 tahap, yaitu *tokenization*, *Case Folding*, *Stopword Removal* dan *Stemming/ Lemmatization*. Jumlah kemunculan setiap kata tersebut menjadi ukuran dalam LDA untuk dimodelkan. Dalam LDA, jumlah topik dan jumlah iterasi ditentukan di awal. Percobaan yang dilakukan dengan mengubah jumlah topik dan jumlah iterasi. Pada penelitian ini dilakukan percobaan sebanyak 3 uji iterasi dengan iterasi berbeda yakni: 500, 1000, dan 5000. Sedangkan terhadap setiap uji iterasi dimasukkan jumlah topik awal yang berbeda yaitu: 2, 3, 4, 5, dan 7. Hasil topik terbaik didapat pada jumlah topik 3. Hasil klaster tersebut dikarenakan tidak adanya topik yang beririsan dengan 3 klaster tersebut. Penulis berharap penelitian ini dapat dikembangkan dengan menambahkan jumlah sumber artikel dari situs berita *online* lainnya untuk melihat variasi topik yang ditemukan oleh sistem serta menambahkan lebih banyak artikel tindakan kriminal sehingga corpus dan distribusi topik semakin sempurna.

DAFTAR PUSTAKA

[1] Jamil, N. B. C. E, I. B. Ishak, F. Sidi, L. S. Affendy, A. Mamat, "A Systematic Review On The Profiling Of Digital News Portal For Big Data Veracity", *Procedia Computer Science*, Vol. 72, 2015, pp. 390-397.

[2] S. Haristya, Hersinta, F. Suwana dan I. Kurniana, "The Credibility Of News Portal In Indonesia: An Exploratory Study", 2012.

[3] R. Alghamdi, dan K. Alfalqi, "A Survey Of Topic Modelling In Text Mining", *International Journal of Advanced Computer Science and Applications*, vol. 6 no. 1, 2015, pp. 147-153.

[4] D. Blei, "Probabilistic Topic Models", *Communications of the ACM*, Vol 55, No.4, 2012.

[5] F. Martin, dan M. Johnson, "More Efficient Topic Modelling Through A Noun Only Approach", *In Proceedings of Australasian Language Technology Association Workshop*, 2015, pp. 111-115.

[6] K. B. Putra, dan R. P. Kusumawardani, "Analisis Topik Informasi Publik Media Sosial Di Surabaya Menggunakan Pemodelan Latent Dirichlet Allocation (LDA)", *Jurnal Tehnik ITS*, Vol. 6, No. 2, 2017.

[7] I. A. Alfanzar, Khalid, dan I. S. Rozas, "Topic modelling skripsi menggunakan metode Latent Dirichlet Allocation", *Jurnal Sistem Informasi*, Vol. 7, No. 1, 2020.

[8] U.T. Setijohatmo, S. Rachmat, T. Susilawati, Y. Rahman, "Analisis Metode Latent Dirichlet Allocation Untuk Klasifikasi Dokumen Laporan Tugas Akhir Berdasarkan Pemodelan Topik", In *Prosiding 11th Industrial Research Workshop and Natoonal Seminar (IRWNS)*, Vol. 11, No. 1, 2020.

[9] M. L. C. Chilmi, "Latent Dirichlet Allocation (LDA) Untuk Mengetahui Topik Pembicaraan Warganet Twitter Tentang Omnibus Law", *skripsi*, Universitas Islam Negeri Syarif Hidayatullah, Jakarta, 2021.

[10] Y Wang, J., Ge, Y. Zhou, Y. Feng, C. Li, Z. Li, X. Zhou, dan B. Luo, "Topic Model Based Text Similarity Measure for Chinese Judgment Document", *ICPCSEE*, 2017, pp. 42-54.

[11] https://ipsaya.com/alexarank.php diakses pada 23 November 2021.

[12] https:/www.wartaprima.com/sejarah-media-online-di-dunia-dan-di-indonesia diakses pada 20 November 2021.

[13]https://www.crummy.com/software/BeautifulSoup/bs4/doc/ diakses 7 Desember 2021.

[14] C. Fiarni, H. Maharani, and R. Pratama, "Sentiment Analysis System for Indonesia Online Retail Shop Review Using Hierarchy Naive Bayes Technique," in *International Conference on Information and Communication Technologies (ICoICT)*, 2016, pp. 212–217.

[15] F. C. Permana, Y. Rosmansyah, and A. S. Abdullah, "Naive Bayes as opinion classifier to evaluate students satisfaction based on student sentiment in Twitter Social Media," *J. Phys. Conf. Ser.*, vol. 893, no. 1, 2017, pp. 0–9.

[16] J. J. Stephen and P. Prabu, "Detecting the magnitude of depression in Twitter users using sentiment analysis," *Int. J. Electr. Comput. Eng.*, vol. 9, no. 4, 2019, pp. 3247–3255.

[17] D. D. Albesta, M. L. Jonathan, M. Jawad, O. Hardiawan, and D. Suhartono, "The impact of sentiment analysis from user on Facebook to enhanced the service quality," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 4, 2021, pp. 3424–3433.