ISSN: 2614-6754 (print)
ISSN: 2614-3097(online)

Halaman 8043-8055
Volume 7 Nomor 2 Tahun 2023

Analisa Pemodelan Topik Berita Daring Menggunakan Semi-Supervised dan Fully Unsupervised Latent Dirichlet Allocation

Khoirunnisa Fi Nurdin[1], Taufik Edy Sutanto[2], Ary Santoso[3]

[1,2,3] Program Studi Matematika, Universitas Islam Negeri Syarif Hidayatullah Jakarta

e-mail: khoir.finur19@mhs.uinjkt.ac.id

Abstrak

Masyarakat milenial tidak membutuhkan pemberitaan yang aktual dan akurat saja akan tetapi juga kecepatan pemberitaan. Media massa yang mampu dalam memenuhi kebutuhan tersebut ialah media daring. *Paluposcom* ialah portal berita daring yang populer saat ini di Kota Palu, Sulawesi Selatan. Berita pada portal tersebut akan terus bertambah sehingga menyebabkan semakin bertumpuknya data berita yang ada. Pemodelan topik mampu membantu mengelompokkan informasi dan juga memetakan informasi tersebut ke suatu bidang tertentu dengan menggunakan metode *latent dirichlet allocation (LDA).* Pada penelitian ini dilakukan dua pendekatan pemodelan topik antara metode *Fully Unsupervised* dan *Semi Supervised*. Topik yang digunakan sebagai acuan topik umum sebanyak 8 yaitu Ideologi, Politik, Ekonomi, Sosial, Budaya, Pertahanan, Keamanan dan Olahraga. Dari hasil penelitian ini diperoleh nilai *loglikelihood* dan *coherence score Semi Supervised* lebih baik dibandingkan Fully *Unsupervised*, serta lebih mudah diinterpretasi oleh manusia dibandingkan dengan Fully *Unsupervised*. Sehingga memudahkan pengguna untuk menemukan berita yang sesuai dengan minat mereka.

Kata Kunci: Lda, Topik Modeling, Fully Unsupervised, Semi-Supervised, Webite Berita Online

Abstract

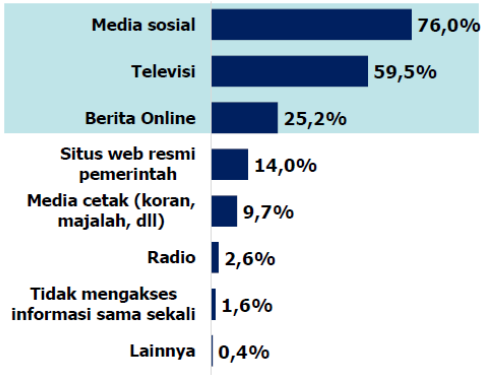
Millennials do not only need actual and accurate reporting, but also the speed of reporting. Mass media that is capable of meeting these needs is online media. Paluposcom is a popular online news portal currently in Palu City, South Sulawesi. The news on the portal will continue to grow, causing more and more piles of existing news data. Topic modeling is able to help classify information and also map that information to a particular field using the latent dirichlet allocation (LDA) method. In this research, two approaches to topic modeling were carried out, namely Fully Unsupervised and Semi Supervised methods. The topics used as a reference for general topics are 8, namely Ideology, Politics, Economics, Social, Culture, Defense, Security and Sports. From the results of this study, the log-likelihood and coherence scores of Semi Supervised were better than Fully Unsupervised, and more easily interpreted by humans than Fully Unsupervised. Making it easier for users to find news according to their interests.

Keywords: Lda, Modeling Topics, Fully Unsupervised, Semi-Supervised, Online News Websites.

PENDAHULUAN

Perkembangan teknologi informasi berkembang sangat pesat. Hal ini di tandai dengan munculnya media komunikasi yang semakin canggih sehingga memperkecil jarak antara pembicara dengan pendengar. Beragam manfaat yang diberikan teknologi tersebut membuat manusia disuguhkan beragam kemudahan untuk mendapatkan informasi baru serta mencari wawasan baru. Informasi terbaru dapat dimanfaatkan oleh manusia menjadi sebuah ilmu pengetahuan yang baru. Pada era digital ini sangatlah banyak informasi yang didapatkan dari

teknologi informasi, karena sudah begitu banyak menyediakan informasi yang dibutuhkan manusia. Kebutuhan atas ketersediaan jaringan komunikasi dan internet saat ini sangat tinggi dengan semakin meningkatnya ketergantungan manusia akan peranan Teknologi Informasi. Orang-orang menghabiskan lebih banyak waktu berinteraksi di media daring, karena kemudahan akses dari mana pun dan kapan saja. Media daring dianggap menjadi suatu kebutuhan hidup dan menjadi hal yang sangat digemari oleh masyarakat [1].



Gambar 1. Sumber Informasi yang Diakses Masyarakat

Berdasarkan data Kominfo pada Gambar 1, berita online merupakan salah satu dari tiga sumber informasi utama yang menjadi rujukan masyarakat Indonesia dalam mengakses informasi, yaitu sebesar 25,2%[2] . Di Indonesia sudah terdapat banyaknya portal berita daring. Data yang dimuat dalam laman resmi Menteri Komunikasi dan Informatika (Menkominfo) mengatakan bahwa terdapat kurang dari 100 portal media daring yang terverifikasi oleh Badan Pers dari sebanyak 43.000 [3]. Data dengan jumlah besar tersebut menjadi tantangan tersendiri untuk dapat diolah. Salah satu bentuk pengolahan data-data besar tersebut ialah dengan melakukan ekstraksi topik dari data teks berita dengan pemodelan topik Latent Dirichlet Allocation (LDA) agar data-data tersebut dapat dikelompokkan berdasarkan topik.

Beberapa penelitian terkait dengan pemodelan topik yang pernah dilakukan diantaranya yaitu penerapan LDA untuk klasterisasi pada Judul Berita Online Detikcom [4], penerapan LDA untuk klasterisasi cerita berbahasa Bali [5], pemodelan topik untuk menemukan topik pembangunan di Indonesia melalui berita online [6]. Berdasarkan uraian di atas, penelitian ini bertujuan untuk mengelompokkan berita-berita berdasarkan topik-topik yang ditujukan melalui pendekatan pemodelan topik yaitu *Fully Unsupervised* dan *Semi Supervised*.

METODE

Data pada penelitian ini menggunakan data sekunder yaitu data judul beserta isi berita online yang diambil pada website berita Kota Palu, Sulawesi Tengah dengan halaman website https://www.palupos.com periode Bulan November 2022 sebanyak 1962 data. Variabel dalam penelitian ini adalah Ideologi, Politik, Ekonomi, Sosial, Budaya, Pertahanan, Keamanan dan Olahraga.

Metode yang digunakan dalam penelitian ini menggunakan pendekatan *Fully Unsupervised* dan *Semi-Supervised Latent Dirichlet Allocation* dengan menggunakan software python dan terdiri dari beberapa tahapan. Adapun tahapannya sebagai berikut :

1. Pengambilan data judul berita beserta isi berita pada portal Palu Pos. Pengumpulan data tersebut dilakukan melalui teknik metode web scraping.
2. Melakukan input data menggunakan Python dan beberapa modul serta packagenya.
3. Melakukan Preprocessing data dengan tujuan menyiapkan data menjadi lebih terstruktur yaitu meliputi tokenizing, stop word, Lemmatization.
4. Pembobotan kata menggunakan Term Frequency digunakan untuk pembobotan kata yang awalnya data teks menjadi data numerik.
   Proses pembobotan kata akan memberikan nilai pada setiap kata ke bentuk unik dan

melihat kata mana saja yang paling sering muncul dengan frekuensi sehingga dapat diketahui bahwa kata tersebut penting. Selain itu, untuk mengurangi kata yang kurang memberikan informatif akan dibatasi bobot nilainya. Batasan pada setiap kata yang muncul, yaitu max_df = 90% dan min_df = 5. Artinya, penghapusan kata yang muncul lebih dari 90% dari jumlah data dan kata yang jarang muncul kurang dari 5 pemunculan dari jumlah data maka akan diabaikan. Persamaan *term frequency* adalah sebagai berikut :

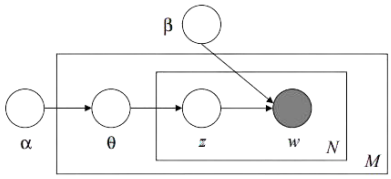
$$w = 0.5 + \left(\frac{0.5\ tf}{\max tf}\right)$$

Nilai maksimum *term frequency* yang digunakan pada penelitiaan ini yaitu 5. Teknik maksimum merupakan teknik normalisasi pada proses term frequency yang sering digunakan.

5. Melakukan pembentukan model topik menggunakan *Fully Unsupervised* Dan *Semi-*Supervised *Latent Dirichlet Allocation*.

Fully Unsupervised

Setelah dilakukannya proses pembobotan kata pada suatu dokumen, kata tersebut akan mengelompokkan dirinya dan berdistribusi ke dalam daftar topik yang berbeda. Salah satu metode yang tepat untuk mendapatkan distribusi topik dari dokumen yang berukuran besar yaitu menggunakan Latent Dirichlet Allocation (LDA). LDA termasuk salah satu metode soft clustering yang digunakan untuk menganalisis data teks. Setiap dokumen akan direpresentasikan sebagai campuran acak atas topik yang tersembunyi (laten), dengan mengidentifikasi setiap topik memiliki karakter yang didalamnya terdapat kata-kata tertentu di dalamnya. Setiap topik merupakan distribusi diskrit atas kosa kata dalam corpus. Sehingga LDA akan mengidentifikasi struktur suatu topik dan menghasilkan daftar topik berdasarkan probabilistik data yang diamati. Berikut adalah model grafis dari LDA sebagai model probabilitas.



Gambar 2. Model grafis LDA

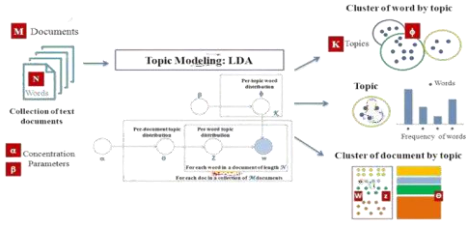
Menurut Blei, terdapat tingkatan pada pemodelan LDA. α dan β adalah parameter distribusi topik pada tingkatan corpus yang dimana merupakan kumpulan dari M dokumen. Sedangkan k ialah variabel sebagai penentuan jumlah topik. Nilai parameter α dan β adalah bilangan real positif tidak lebih dari 1 atau $0 \leq \alpha, \beta \leq 1$. Parameter α sebagai penentuan distribusi topik, semakin besar nilai dari α maka terdapat beberapa topik yang akan dibahas pada suatu dokumen. Parameter β menentukan distribusi kata pada topik, semakin besar nilai dari β maka terdapat banyak kata-kata yang ada dalam suatu topik. Jika nilai dari β kecil, maka dalam suatu topik mempunyai kata-kata yang lebih spesifik [7].

Distribusi topik pada suatu dokumen, dinotasikan sebagai variabel θ. Dimana jika nilai dari θ besar, banyak topik yang akan dibahas sedangkan jika nilai dari θ kecil maka terdapat suatu pembahasan topik yang lebih spesifik. Variabel dari W sebagai kata yang berhubungan dengan topik tertentu dalam suatu dokumen. Pada variabel Z_n dan W_n sebagai variabel tingkat kata (N). Semua kata dalam M dokumen dikelompokkan ke dalam Z topik untuk setiap topik $Z \in 1,2,\ldots,k$, contoh distribusi kata $\phi_k \sim Dirichlet\ (\beta)$:

1. Pilih $N \sim Poisson\ (\xi)$.
2. Pilih distribusi topik $\theta m \sim Dirichlet\ (\alpha)$.
3. Untuk setiap dokumen kata $w_{m,n}$:
   a. Pilih topik dari kata $Z_{m,n} \sim Multinomial\ (\theta_m)$.

b. Pilih kata $w_{m,n}$~ $Multinomial$ ($\phi_{Z_{m,n}}$). Probabilitas multinomial pada topik $Z_{m,n}$.
Keterangan :
α : Distribusi topik per dokumen
β : Distribusi kata per topik
K : Jumlah topik
W : Jumlah token dalam dokumen keseluruhan

Kata yang masuk ke setiap topik dapat diambil berdasarkan kata unik dalam suatu dokumen yang diperoleh dari hasil *term frequency*, selanjutnya untuk pemetaan dapat dilakukan setiap kata dalam setiap dokumen secara acak. Sehingga memperoleh suatu topik. Berikut adalah gambaran umum mengenai topic modelling *Fully Unsupervised* LDA :



Gambar 3. Gambaran Umum Fully Unsupervised LDA

Fully Unsupervised LDA masih memiliki kebebasan dalam mengidentifikasi model topik yang tersembunyi pada suatu dokumen sehingga belum dapat menghasilkan suatu topik berdasarkan kata-kata yang dikelompokkan karena tidak ada label yang membuat hasil topik menjadi informatif, topik yang diinginkan oleh manusia tidak ada dan saling bertumpukkan antar kata dalam suatu topik. Dengan pendekatan *Semi Supervised* LDA dapat mengatasi hal tersebut karena dalam *Semi Supervised* LDA terdapat seed yang disusun sebagai label acuan dalam topik sehingga dapat memodelkan topik lebih baik dan sesuai yang diinginkan.

Semi Supervised LDA

Semi Supervised dapat mengelompokkan suatu dokumen dan datanya lebih terstruktur daripada Fully Unsupervised. Hal ini dikarenakan Semi Supervised akan memasukan variabel input dan memiliki nilai khusus target output yang disebut sebagai dengan seed.

Proses Semi Supervised akan menetapkan seed pada masing-masing topik. Diberikan kumpulan suatu dokumen, $D = \{d_1, d_2, \dots, d_M\}$ yang setiap dokumen d_i terdiri dari suatu kumpulan kata $w_i = \{w_{i,1}, w_{i,2}, \dots, w_{i,N\ i}\}$. Untuk setiap kata pada $w_{i,j}$, jika ada topik yang ditetapkan maka $w_{i,j}$ dibatasi untuk menjadi bagian dari sekumpulan label seed yang sudah ditentukan $Z_i = \{Z_{i_1}, Z_{i_2}, \dots, Z_{i_k}\}$. Jika $w_{i,j}$ diberikan sekumpulan label dan topik baru untuk semua kata yang tidak berlabel dibatasi untuk seluruh domain topik $Z = \{Z_1, Z_2, \dots, Z_k\}$ [8].

Dalam implementasi Semi Supervised, selama pada proses pengambilan Gibbs sampling, untuk kata $w_{i,j}$ di dalam dokumen d_i, Semi Supervised terlebih dahulu mengecek apakah terdapat topik $Z_{i,j}$ yang ditetapkan untuk kata ini, jika topik $Z_{i,j}$ ditetapkan ke $w_{i,j}$, topik baru dibatasi untuk dijadikan sampel dalam $T_{i,j}$. Jika tidak, maka akan dicek apakah ada topik Z_i yang ditetapkan untuk dokumen d_i. Jika iya, maka topik Z_i ditugaskan kedalam di, topik baru untuk $w_{i,j}$ dibatasi dalam pengambilan sampel topik Zi. Jika tidak, topik baru untuk $w_{i,j}$ dapat diambil sampelnya dari seluruh domain topik Z.

Pengambilan topik yang baru sangat memerlukan menormalisasi parameter. Untuk menjamin bahwa pelabelan diambil dengan valid dan dibatasi oleh seed yang telah ditetapkan, persamaan untuk pengambilan sampel sebagai berikut :

$$(Z_i = j|w_i, d, z_{-i}) = \frac{n_{-i,j}^{w_i} + \beta w_i}{\sum_{w_i} n_{-i,j}^{w_i} + sum\_\beta} \times \frac{n_{-i,j}^{d_i} + \alpha_j}{\sum_j n_{-i,j}^{d} + sum\_\alpha}$$

Dengan:
sum_α = $\sum j\ ada\ didalam\ topik\ a_j, kendala\ Z_{i,j}, Zi, atau\ Z.$
sum_β = $\sum j\ ada\ didalam\ topik\ \beta w_i, kendala\ Z_{i,j}, Zi, atau\ Z.$

Dalam implementasi Semi Supervised, sudah ditetapkan kata-kata apa saja yang akan masuk ke dalam topik $Z\ Z = \{Z_1, Z_2, \dots, Z_k\}$ disebut dengan seed topik. Di dalam seed topik terdapat parameter seed confidence yang bernilai antara 0 dan 1 untuk mengontrol seed yang akan bertugas ke kata dan juga dilakukan pemetaan kedalam suatu topik. Dengan seed confidece sebesar 0,1 tersebut akan melakukan sebuah pemetaan kata sebesar 10% ke arah topik yang diunggulkan sesuai seed yang ditentukan. Dengan Semi Supervised dapat memisahkan topik yang memiliki representasi lebih kecil dalam corpus sehingga hasil topik lebih intepretatif dan mengelompokkan dokumen cukup lebih baik daripada Fully Unsupervised.

Semi Supervised akan menghasilkan sebuah topik ke arah yang lebih sesuai dan menyatu dengan menggunakan seed. Seed akan mengarahkan seluruh kata dokumen dengan kata-kata yang berkaitan dengan bidang tertentu berdasarkan informasi kata dan mewaliki makna simantiknya. Dengan menggunakan tidak lebih banyak kata pada seed akan tetapi cakupan dokumen dari kata seed untuk semua bidang sudah sesuai dalam pengelompokan. Terdapat suatu batasan dalam beberapa seed yang akan menjamin bahwa informasi berkaitan dengan seed dan sebagai proses pengelompokan kata berdasarkan probabilitasnya. Pada penelitian ini, seed topik yang digunakan berdasarkan kata yang sering muncul dan relevan bidang topik tersebut pada Ideologi, Politik, Ekonomi, Sosial, Budaya, Pertahanan, Keamanan dan Olahraga [9].

Menentukan evaluasi model dengan nilai Loglikehood dan Topic Coherence

Evaluasi model dilakukan untuk mengetahui seberapa baiknya model yang sudah dibentuk dan memvalidasi nilai dari hasil model tersebut sehingga topik mudah diinterpretasikan. Namun pada dasarnya nilai akurasi pada evaluasi model tidak selalu berkorelasi dengan pemahaman manusia bahkan tidak berkorelasi [10] sehingga untuk mengetahui seberapa baiknya topic modelling bukan dilihat berdasarkan nilai akurasi evaluasi model akan tetapi berdasarkan pada keberhasilan interpretasi yang dipahami oleh manusia. Terdapat banyak cara dalam mengevaluasi topic modelling. Dalam penelitian ini menggunakan evaluasi loglikelihood dan coherence score.

Loglikelihood menggunakan dataset uji yang dikeluarkan berupa matrik. Matrik yang dikeluarkan merupakan hasil dari perhitungan kumpulan corpus. Matrik akan merepresentasikannya dengan mudah dan membandingkan nilai loglikelihood dari berbagai model atau pilihan parameter topik [10]. Semakin tinggi nilai yang terdapat di loglikehood, semakin baik juga model yang dibangun [11].

Kedua, UMass coherence terbukti sangat cocok dengan penilaian manusia tentang kualitas topik [12]. Matriks UMass akan menentukan skor dari jumlah corpus yang digunakan pada dataset uji model topik. Semakin kecil nilai yang dihasilkan oleh UMass coherence maka semakin baik model yang dibangun.

Melakukan visualisasi data menggunakan word cloud, dan word link.

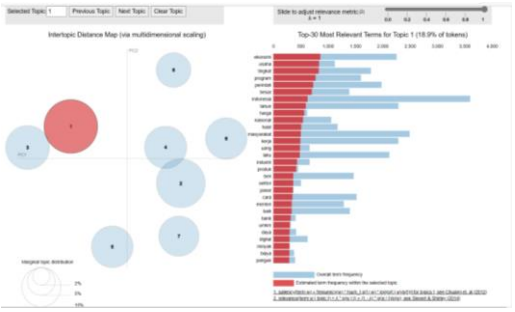
Visualisasi digunakan untuk menampilkan keyword yang muncul dan merepresentasikan dari data teks. Terdiri dari kata-kata tunggal, ukuran font dan warna yang berbeda dengan menunjukan signifikansinya masing-masing. Semakin besar ukuran font yang akan ditampilkan maka semakin besar juga frekuensi kata pada data teks tersebut.

Melakukan hasil interpretasi.

Pada tahapan ini, dilakukan interpretasi terhadap topik yang didapatkan dari pemodelan. Interpretasi hasil dilakukan secara keseluruhan untuk mendapatkan topik yang merepresentasikannya dengan menggunakan LDAvis.

LDAvis merupakan sebuah sistem untuk mengeksplorasi secara fleksibel hubungan antar topik kata untuk lebih memahami model LDA yang sudah dibentuk dengan fitur tambahan yang memberikan perspektif pada model

Hal yang terpenting dalam LDAvis yaitu untuk memilih topik berdasarkan ungkapan istilah pada topik tersebut



Gambar 4. Ilustrasi pada pyLDAvis

Dalam pyLDAvis pada gambar 4 terdapat barchart kanan dengan panjang batang abu-abu sebagai frekuensi corpus dari tiap kata dan panjang batang merah menggambarkan frekuensi topik yang spesifik dari tiap istilah. Peneliti dapat menentukan apakah istilah tersebut sesuai atau tidak dengan topik yang dipilih berdasarkan rasio antara batang abu-abu dan merah berdasarkan probabilitasnya.

Untuk menghitung relevan pyLDAvis tersebut dapat memisalkan ϕ_{kw} sebagai probabilitas untuk istilah $w \in \{1,\dots,V\}$ dan $topik\ t \in \{1,\dots,T\}$ dimana V adalah jumlah istilah dalam kosakata sedangkan pw ialah probabilitas dari istilah w di corpus [14]. Definisi istilah w dengan topik t yang diberikan bobot parameter λ dimana ($0 \le \lambda \le 1$) :

$$r(w,k|\lambda) = \lambda \log(\phi_{kw}) + (1-\lambda)\log(\frac{\phi_{kw}}{p_w})$$

Pada panel disebelah kiri, terdapat suatu lingkaran yang sebanding dengan topik dalam corpus. Panel tersebut menunjukkan area lingkaran topik dari tiap topik dan perbedaan antar topik. Untuk menghitung jarak antar topik tersebut dengan menggunakan metode Kullback-Leibler Divergence. Suatu kata w dihitung probabilitasnya P(T|w). Kemungkinan kata yang diamati w dihasilkan oleh topik tersembunyi (laten) T [14]. Kemudian hitunglah probabilitas P(T). Kemungkinan serupa bahwa setiap kata dipilih secara acak w dihasilkan dari topik T. Pendefinisian kekhasan pada suatu kata w berdasarkan Kullback-Leibler Divergence yaitu antara P(T|w) dan P(T) sebagai berkut :

$$w = \sum_T P(T|w)\log\frac{P(T|w)}{P(T)}$$

HASIL DAN PEMBAHASAN

Visualisasi EDA

Analisis data teks pada penelitian ini diawali dengan membuat visualisasi untuk melihat kata-kata yang dapat dijadikan sebagai ciri-ciri kandidat kata kunci sehingga lebih mudah untuk mengetahui isu yang dibicarakan. Seperti pada gambar dibawah ini, yang menampilkan hasil exploratory data analysis dari data yang membahas tentang Ideologi, Politik, Ekonomi, Sosial, Budaya, Peratahanan, Keamanan, dan Olahraga. Visualisasi data menggunakan wordcloud dengan hasil yang disajikan pada gambar di bawah ini.

ISSN: 2614-6754 (print)
ISSN: 2614-3097(online)

Halaman 8043-8055
Volume 7 Nomor 2 Tahun 2023

Gambar 5. Wordcloud

Gambar 5 memberikan informasi tentang 80 kata yang memiliki frekuensi kemunculan paling besar. Kata “Indonesia” merupakan kata yang memiliki frekuensi paling besar yaitu dengan total frekuensi mencapai lebih dari 7100 kali. Berdasarkan gambar tersebut juga dapat dilihat bahwa kata-kata lain yang memiliki frekuensi kemunculan lebih besar dibandingkan dengan kata-kata yang lainnya adalah masyarakat, dan ekonomi.

Visualisasi data menggunakan wordlink dengan hasil yang disajikan pada gambar di bawah ini.

Gambar 6. Wordlink

Gambar 6 memberikan informasi tentang 80 kata yang memiliki frekuensi kemunculan paling besar. Kata “Indonesia” merupakan kata yang memiliki frekuensi paling besar yaitu dengan total frekuensi mencapai lebih dari 7100 kali. Berdasarkan gambar tersebut juga dapat dilihat bahwa kata-kata lain yang memiliki frekuensi kemunculan lebih besar dibandingkan dengan kata-kata yang lainnya adalah masyarakat, dan ekonomi.

Evaluasi

Dalam penelitian ini, topik yang diambil berdasarkan delapan pilar utama yaitu Ideologi, Politik, Ekonomi, Sosial, Budaya, Pertahanan, Keamanan, dan Olahraga. Topik umum tersebut dianggap sebagai seed yang akan dijadikan acuan lexicon pada corpus dalam implementasi LDA. Sehingga dalam penelitian tersebut menggunakan n topik = 8. Seed yang digunakan berdasarkan kata umum yang dibahas dalam topik tersebut. Sebelumnya data sudah di preprocessing sehingga data yang diolah lebih mudah untuk dikelompokkan sesuai dengan topik atau seed yang sudah ditentukan. Hasil dari preprocessing tersebut akan terbentuk dictionary yaitu kata unik yang terdapat dalam dokumen dan corpus merupakan kata yang berbentuk BoW reference dari dokumen yang digunakan untuk pembentukan model.

Dalam pemodelan antara Semi-Supervised dan Fully Unsupervised LDA menggunakan CountVectorizer yaitu pembobotan kata menggunakan term frequency dengan matrix 30 (n sample data) dan jumlah kata sesuai dengan masing-masing bidang. Hal ini lebih dikenal dengan Vector Space Model (VSM) yang berbentuk sebuah vektor dengan nilai baris menunjukan banyaknya dokumen atau jumlah isi berita dalam penelitian ini menggunakan 1349 isi berita. Dan kolom menunjukkan jumlah kata yang muncul sesuai masing-masing berita. Tujuan menggunakan CountVectorizer ini untuk mengetahui probabilitas setiap kata unik yang muncul (term frequency). Penelitian ini menggunakan evaluasi nilai loglikelihood dan UMass coherence karena mudah diimplementasikan.

Dari data yang dikumpulkan dan dimasukkan ke dalam model LDA, kemudian peneliti

akan menentukan nilai evaluasinya. Berikut pada tabel 4.1 terdapat hasil perbandingan nilai evaluasi antara loglikehood, coherence score dengan versi gensim dan UMass pada website berita daring dengan menggunakan package sklearn.

Tabel 1 Tabel Nilai Evaluasi Fully Unsupervised dan Semi-Supervised LDA

Metode	Nilai Evaluasi	
	Loglikehood	UMass coherence
Fully Unsupervised	-1780942.3	-52.0
Semi Supervised	-1865350.9	-52.36

Nilai loglikelihood pada Semi-Supervised masih sangat rendah dan bernilai negatif. Nilai tersebut menunjukkan bahwa model LDA yang dibentuk belum optimal. Hal ini dikarenakan oleh data teks yang diambil dari website berita Palu Pos memiliki banyak singkatan dan kalimat tidak baku dalam bahasa Indonesia. Kata-kata singkatan menjadi tidak tersaring dan mempunyai representasi numerik yang berdekatan sehingga membuat salah satu klaster mempunyai nilai yang besar. Akan tetapi, nilai UMass coherence pada Semi-Supervised lebih baik daripada Fully Unsupervised. Kemudian, akan didapatkan nilai probabilitas setiap topik ke seed untuk menentukan hasil pemetaan Topik ke Bidang di Seed dengan nilai α dan β merupakan bilangan ril positif yang tidak lebih dari 1 atau dapat dituliskan sebagai $0 \leq \alpha, \beta \leq 1$.

Dengan menggunakan n_gram = (1,1), min_df = 5 dan max_df = 0.9 yang mengartikan bahwa max_df akan mengabaikan istilah yang muncul lebih dari 90% dokumen dan menghapus istilah yang terlalu jarang muncul kurang dari 5 dokumen. Sehingga dalam website berita tersebut terdapat matrix (1349, 3391) yang berarti terdapat 1349 kolom sample data dan 3391 kata unik yang berada dalam kolom matrix.

Menggunakan library lda dari scikit-learn yaitu sklearn.decomposition maka LDA akan mendapatkan n topic = 8. Kemudian, untuk mepetakan 8 topik ke bidang pada seed topik yang telah ditentukan dengan menggunakan threshold yaitu ambang batas kata atau kata diseleksi [15] dengan menggunakan rumus $\frac{1}{n} = \frac{1}{8}$, dimana setiap kata corpus yang masuk ke dalam seed topik akan dihitung probabilitasnya dengan matrix (8, 3391) sehingga memperoleh hasil pada tabel 4.2

Tabel 2 Nilai probabilitas setiap topik di seed pada Fully Unsupervised

	Ideologi	Politik	Ekonomi	Sosial	Budaya	Petahanan	Keamanan	Olahraga
0	0.221651	0.214318	0.408225	0.141305	**0.170669**	**0.360538**	**0.379234**	0.056549
1	**0.205379**	**0.205379**	0.024843	**0.175794**	**0.175794**	**0.154212**	**0.213194**	0.093866
2	0.093866	0.069351	0.090524	0.090524	**0.162630**	0.091936	0.031556	**0.397153**
3	0.050418	0.050418	0.050418	0.094475	0.069489	0.076216	0.076216	0.076216
4	0.115298	0.115298	0.034717	0.049317	0.049317	0.064090	0.070462	0.041872
5	**0.184952**	**0.141460**	0.037494	0.057173	0.074855	0.102721	0.064452	0.118291
6	0.087670	0.087670	0.008215	**0.186739**	**0.202363**	0.099945	0.088348	**0.223151**

7	**0.223151**	**0.223151**	0.092511	**0.174858**	**0.171713**	0.050341	0.050327	0.050143

Berdasarkan hasil dari nilai probabilitas pada tabel 2 matrix (8, 3391) akan menghitung kata yang masuk ke dalam tiap bidang. Dimana topik dengan index 0-7 dapat mempetakan topik sesuai dengan kecenderungan nilai diatas ke bidang.

Pada topik 1, terdapat di bidang ekonomi, pertahanan dan keamanan. Topik 2 terdapat di bidang politik, keamanan, dan ideologi. Topik 3 terdapat di bidang olahraga dan budaya. Topik 4 terdapat di bidang ekonomi. Topik 5 terdapat di bidang politik. Topik 6 terdapat di bidang ideologi dan politik. Topik 7 terdapat di bidang olahraga, budaya, sosial. Topik 8 terdapat di bidang sosial dan budaya.

Tabel 3 Nilai probabilitas setiap topik di seed pada Semi-Supervised

	Ideologi	Politik	Ekonomi	Sosial	Budaya	Pertahanan	Keamanan	Olahraga
0	**0.235649**	**0.142123**	0.037761	0.111128	0.053590	0.115326	**0.129651**	0.059177
1	0.059177	0.059177	0.027811	0.078875	0.078875	0.085434	0.124104	0.070869
2	0.122923	0.094107	**0.623619**	0.114574	0.077657	0.086932	0.080716	0.027299
3	0.078891	0.069334	0.032555	**0.361928**	**0.162518**	0.094769	0.075099	0.052377
4	0.112281	0.075851	0.036081	0.093733	**0.308947**	0.124947	0.109507	0.058467
5	0.038402	0.038402	0.062149	0.105985	0.095627	0.089506	0.028399	0.119253
6	0.119253	**0.195642**	**0.167487**	0.095432	**0.165975**	**0.389953**	**0.447828**	0.028876
7	0.096181	0.033953	0.012537	0.038345	0.079434	0.013134	0.004695	**0.583683**

Pada topik 1, terdapat di bidang ideologi, politik, dan keamanan. Topik 2 terdapat di bidang politik dan ideologi. Topik 3 terdapat di bidang ekonomi. Topik 4 terdapat di bidang sosial dan budaya. Topik 5 terdapat di bidang budaya. Topik 6 terdapat di bidang olahraga. Topik 7 terdapat di bidang keamanan, pertahanan dan politik. Topik 8 terdapat di bidang olahraga.

Jika dijabarkan tiap kata pada topik dengan jumlah kata kunci pada tiap dokumen sebagai berikut :

Tabel 4 Topik yang dominan pada tiap dokumen Fully Unsupervised

	Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Dominant_topic
Doc0	0.000670	0.000670	0.084690	0.000669	0.000670	**0.176390**	**0.665214**	0.000670	**6**
Doc1	**0.205609**	0.000703	**0.712378**	0.000703	0.078497	0.000703	0.000704	0.000704	**2**
Doc2	0.000338	**0.166114**	0.000338	**0.640851**	0.000337	0.000337	0.080170	0.111516	**3**
Doc3	0.001108	0.119484	0.017571	0.001108	0.031702	0.001108	**0.410861**	**0.417058**	**7**
Doc4	0.000480	**0.600560**	0.050780	0.000479	0.000480	0.000480	0.000480	0.000480	**1**

Doc5	**0.393589**	0.001062	0.001061	0.001061	**0.600044**	0.001061	0.001061	0.001060	**4**
Doc6	0.001691	0.001692	**0.259158**	0.001690	0.001692	0.001692	**0.535829**	**0.196555**	**6**
Doc7	0.093436	0.000320	0.000320	0.110742	0.037297	**0.467998**	0.000320	**0.289567**	**5**

Berdasarkan tabel 4 menampilkan pemetaan dokumen terhadap topik. Dokumen 1 masuk ke dalam topik 7 sebesar 0.665214. Dokumen 2 masuk ke dalam topik 3 sebesar 0.712378. Dokumen 3 masuk ke dalam topik 4 sebesar 0.640851. Dokumen 4 masuk ke dalam topik 8 sebesar 0.417058. Dokumen 5 masuk ke dalam topik 2 sebesar 0.600560. Dokumen 6 masuk ke dalam topik 5 sebesar 0.600044. Dokumen 7 masuk ke dalam topik 7 sebesar 0.535829. Dokumen 8 masuk ke dalam topik 6 sebesar 0.467998.

Tabel 5 Topik yang dominan pada tiap dokumen Semi Supervised

	Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Dominant_topic
Doc0	0.072537	0.026046	0.001764	0.000470	**0.696801**	0.026650	0.073394	0.102340	**4**
Doc1	0.004088	0.000218	0.086772	0.000250	0.000071	0.000572	**0.340700**	**0.567330**	**7**
Doc2	0.065986	0.000045	**0.353119**	0.013816	0.077659	**0.484536**	0.004819	0.000021	**5**
Doc3	0.000876	0.097877	0.000257	**0.604340**	**0.179234**	0.021436	0.030383	0.065598	**3**
Doc4	**0.624221**	0.000503	0.066289	0.000553	0.001148	0.000203	0.006557	**0.300526**	0
Doc5	0.096433	0.097939	**0.753234**	0.000103	0.000114	0.000067	0.051903	0.000209	**2**
Doc6	0.123472	0.123472	0.000134	**0.663809**	0.000377	0.009940	0.000323	**0.201195**	**3**
Doc7	0.007513	0.000305	**0.190136**	0.000779	**0.654907**	0.080552	0.065767	0.000042	**4**

Berdasarkan tabel 5 menampilkan pemetaan dokumen terhadap topik. Dokumen 1 masuk ke dalam topik 5 sebesar 0.696801. Dokumen 2 masuk ke dalam topik 8 sebesar 0.567330. Dokumen 3 masuk ke dalam topik 6 sebesar 0.484536. Dokumen 4 masuk ke dalam topik 4 sebesar 0.604340. Dokumen 5 masuk ke dalam topik 1 sebesar 0.624221. Dokumen 6 masuk ke dalam topik 3 sebesar 0.753234. Dokumen 7 masuk ke dalam topik 4 sebesar 0.663809. Dokumen 8 masuk ke dalam topik 5 sebesar 0.654907.

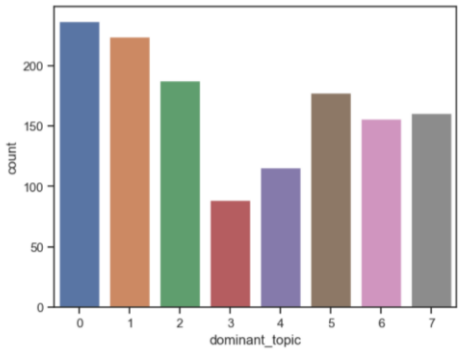
Berikut 10 kata pertama yang termasuk dalam masing-masing topik Fully Unsupervised :

Topic 1 : Indonesia ekonomi menko tingkat negara kerja tahun menteri perintah industri.
Topic 2 : partai milu kerja masyarakat politik golkar sama ketua Indonesia kib.
Topic 3 : program giat mahasiswa tingkat hasil didik siswa anak tahun guru.
Topic 4 : harga pangan perintah masyarakat minyak ekonomi menteri inflasi bbm tangan.
Topic 5 : Kota daerah presiden tahun ketua perintah nasional kerja dprd hasil.
Topic 6 : Sulteng laku hukum sulawesi guberbur tengah orang kasus buol pihak.
Topic 7 : Kabupaten sulteng desa sigi poso kota laku masyarakat sulawesi provinsi
Topic 8 : Desa masyarakat layan rumah warga bank haji usaha air jalan.

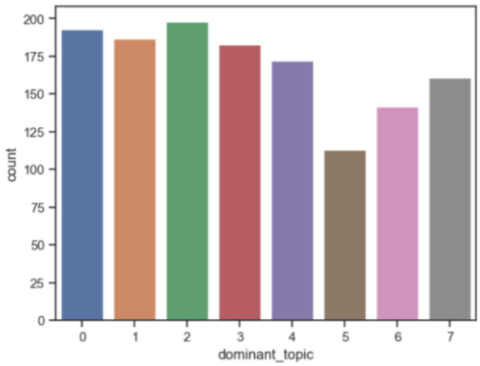
Berikut 10 kata pertama yang termasuk dalam masing-masing topik Semi Supervised:

Topic 1 : Sulteng kerja hukum milu masyarakat tengah sulawesi laku tahun daerah.
Topic 2 : partai golkar ketua politik kota daerah presiden kib umum Indonesia.

Topic 3 : ekonomi usaha tingkat program perintah besar Indonesia tahun harga nasional.
Topic 4 : masyarakat sigi kabupaten bencana kota daerah sehat giat sama rumah.
Topic 5 : desa poso sulteng laku kabupaten camat pihak masyarakat warga orang.
Topic 6 : layan telkomsel haji laku langgan masyarakat bbm program kota kendara.
Topic 7 : Indonesia ekonomi negara menko sama menteri kerja tahun global perintah
Topic 8 : Sulteng mahasiswa kota giat siswa agama guru didik tahun ketua.



Gambar 7. Distribusi topik di seluruh dokumen Fully Unsupervised



Gambar 8. Distribusi topik di seluruh dokumen Semi Supervised

Perceptual Map

Perceptual map merupakan hubungan antara objek yang dipersepsikan dan juga hubungan geometris antara titik-titik di alam ruang yang multidimensional koordinat. Setelah diperoleh model EDA menggunakan word cloud serta word link dan model Latent dirichlet allocation (LDA) maka model tersebut dapat dilihat juga dengan visualisasi Perceptual Map menggunakan PyLDAvis dan keterkaitan antar kata yang dihasilkan.

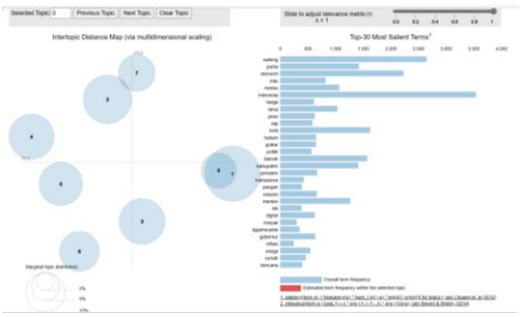
Dalam konteks berita daring, konsep perceptual map dengan menggunakan PyLDAvis dapat membantu untuk memahami dan memvisualisasikan persepsi pembaca terhadap topik-topik berita tertentu. Dengan menerapkan PyLDAvis pada berita, kita dapat mengidentifikasi topik utama yang muncul, menghubungkan topik-topik tersebut berdasarkan kesamaan antar kata, dan kemudian memvisualisasikan posisi relatif dari topik-topik tersebut dalam ruang topik.

Dengan demikian, konsep perceptual map menggunakan PyLDAvis pada topik umum berita daring dapat membantu dalam menganalisis dan memahami persepsi terhadap berita-berita tertentu, mengidentifikasi keterkaitan antara topik-topik tersebut, serta memvisualisasikan posisi relatif dari berbagai topik dalam persepsi pembaca.

Pada panel sisi kiri yaitu pemetaan jarak dari antar topik (intertopic distance map) yang terdapat juga cluster topik membentuk sebuah lingkaran. Sedangkan pada panel sebelah kanan terdapat 30 buah terminologi yang sangat relevan untuk suatu topik-topik tertentu.

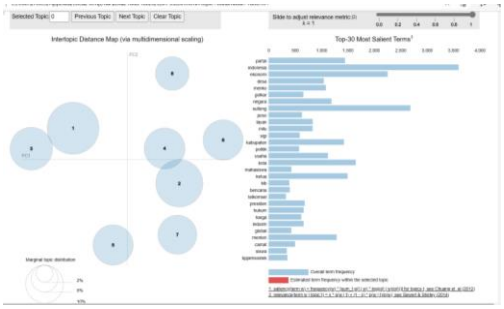
Selanjutnya, jika dipilih salah satu sebuah topik maka lingkarannya akan berubah warna menjadi merah dan kemudian pada bar chart panel disisi kanan akan berubah menjadi warna merah memperlihatkan estimasi kata dan term frequency pada topik yang dipilih. Terdapat juga bar berwarna biru yaitu kemunculan pada kata ada di seluruh dokumen.

Didapatkan hasil visualisasi menggunakan PyLDAvis pada fully unsupervised dan semi supervised sebagai berikut :



Gambar 9. PyLDAvis Fully Unsupervised

Topik 1 terdapat 19.7% of tokens yang berarti dari semua dokumen yang mewakili topik 1 sebesar 19.7% yang relevan dengan jumlah kemunculan kata pada topik 1, topik 2 sebesar 15% of tokens, topik 3 sebesar 13.5% of tokens, topik 4 sebesar 13% of tokens, topik 5 sebesar 12.1% of tokens, topik 6 sebesar 10.2% of tokens, topik 7 sebesar 8.6% of tokens, topik 8 sebesar 7.9%.



Gambar 10. PyLDAvis Semi Supervised

Topik 1 terdapat 18.9% of tokens yang berarti dari semua dokumen yang mewakili topik 1 sebesar 18.9% yang relevan dengan jumlah kemunculan kata pada topik 1, topik 2 sebesar 15.6% of tokens, topik 3 sebesar 12.7% of tokens, topik 4 sebesar 11.7% of tokens, topik 5 sebesar 11.4% of tokens, topik 6 sebesar 11.3% of tokens, topik 7 sebesar 10.6% of tokens, topik 8 sebesar 7.9%.

SIMPULAN

Berdasarkan hasil penelitian yang telah diperoleh menggunakan metode Fully Unsupervised LDA dan metode Semi-Supervised LDA, didapatkan sebanyak 8 jumlah topik (Ideologi, Politik, Ekonomi, Sosial, Budaya, Pertahanan, Keamanan, Dan Olahraga). Interpretasi pada penelitian ini, metode Semi-Supervised LDA lebih baik hasilnya karena topik yang dihasilkan lebih terarah dalam pengelompokkan. Sedangkan pada metode Fully Unsupervised sudah cukup baik dalam memberikan sebuah interpretasi akan tetapi belum terarah pada masing-masing topik

Adapun saran pada penelitian ini adalah diharapkan pada penelitian selanjutnya menggunakan tahapan stemming pada praproses data diharapkan menggunakan portal berita daring yang lain serta dapat mengembangkan pemodelan topik dengan metode yang lainnya

DAFTAR PUSTAKA

Chuang, J., Manning, C. D., & Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. In Proceedings of the international working conference on advanced visual interfaces (pp. 74-77).

Dewi, M. S. R. (2019). Islam dan etika bermedia (kajian etika komunikasi netizen di media sosial instagram dalam perspektif islam). Research Fair Unisri, 3(1).

Adimanggala, D., Bachtiar, F. A., & Setiawan, E. (2021). Evaluasi Topik Tersembunyi Berdasarkan Aspect Extraction menggunakan Pengembangan Latent Dirichlet Allocation. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 5(3), 511-519.

ER, N. A. S. (2021). Implementasi Latent Dirichlet Allocation (LDA) untuk Klasterisasi Cerita Berbahasa Bali. Jurnal Teknologi Informasi dan Ilmu Komputer, 8(1), 127-134

Fernanda, J. W. (2021). PEMODELAN PERSEPSI PEMBELAJARAN ONLINE MENGGUNAKAN LATENT DIRICHLET ALLOCATION. Jurnal Statistika Universitas Muhammadiyah Semarang, 9(2), 79-85.

Matira, Y., & Setiawan, I. (2023). Pemodelan Topik pada Judul Berita Online Detikcom Menggunakan Latent Dirichlet Allocation. ESTIMASI: Journal of Statistics and Its Application, 53-63.

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

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012, July). Exploring topic coherence over many models and many topics. In Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning (pp. 952-961).

Hidayatullah, A. F., Ma'arif, M. R., Habibie, M., & Khomsah, S. (2021, February). Indonesia infrastructure development topic discovery on online news with latent Dirichlet allocation. In IOP conference series: materials science and engineering (Vol. 1077, No. 1, p. 012012). IOP Publishing

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. Advances in neural information processing systems, 22.

D'Andrea, E., Ducange, P., Bechini, A., Renda, A., & Marcelloni, F. (2019). Monitoring the public opinion about the vaccination topic from tweets analysis. Expert Systems with Applications, 116, 209-226

Center, K. I. (2020). Status Literasi Digital Indonesia: Survei di 34 Propinsi.

Sievert, C., & Shirley, K. (2014, June). LDAvis: A method for visualizing and interpreting topics. In Proceedings of the workshop on interactive language learning, visualization, and interfaces (pp. 63-70).