CENTIVE
VOL. 3, NO. 1, PP.475-484, DEC 2023
ISBN : 978-602-53004-0-0

Conference on Electrical Engineering, Informatics, Industrial Technology, and Creative Media 2023

Pengelompokan Teks Berita Utama dengan Metode LDA (*Latent Dirichlet Allocation*) melalui Pemahaman Pemodelan Topik

Ilham Fadhilah Akbar[1], Tri Ginanjar Laksana[2], Amalia Beladinna Arifa[3], Mario Rudy Silalahi*[4]

[1,2,3,4] *Teknik Informatika, Fakultas Informatika, Institut Teknologi Telkom Purwokerto*
Jln. DI Panjaitan No. 128, Purwokerto, Jawa Tengah, Indonesia

[1] 17102106@ittelkom-pwt.ac.id
[2] anjarlaksana@ittelkom-pwt.ac.id
[3] amalia@ittelkom-pwt.ac.id
[4] 20102018@ittelkom-pwt.ac.id

Received on 02-11-2023, revised on 14-11-2023, accepted on 15-11-2023

Abstrak

Berkembangnya teknologi informasi saat ini telah mengubah cara berita dan media disebarkan secara *online*, dengan perusahaan-perusahaan menggunakan teknologi ini untuk menyebarkan berita dan konten media. Namun, dengan *meningkatnya* jumlah berita di platform media online, masyarakat seringkali kesulitan memahami gambaran umum tentang topik utama dalam berita tersebut. Oleh karena itu, penelitian ini menerapkan konsep *Topic Modeling*, suatu metode yang memungkinkan ekstraksi konteks yang mewakili isi dokumen melalui analisis statistik pada kumpulan besar teks. Metode yang diterapkan pada penelitian ini adalah *Latent Dirichlet Allocation* (LDA), yang memungkinkan analisis topik yang tersembunyi dalam dokumen besar. Hasil analisis menunjukkan bahwa koherensi topik tertinggi adalah 0.57522 melalui berbagai percobaan dengan jumlah topik yang berbeda, dan model topik dengan lima topik telah ditentukan. Kelima topik ini masing-masing membahas pemain dalam klub sepak bola, sejarah pemain di Premier League, potret dramatis pemain dalam NBA dan sepak bola, pemain Liverpool dalam Liga Champions, dan pemain gelandang Jerman di klub Bayern Munchen. Temuan ini menggambarkan bahwa berita utama dalam kelompok teks ini fokus pada topik sepak bola dan basket, dengan penekanan lebih kuat pada sepak bola sebagai subkategori olahraga. Penggunaan berbagai batasan jumlah topik dalam analisis memberikan hasil yang optimal dan menghasilkan model topik yang lebih bermakna.

Kata Kunci: Pemodelan Topik, Topik Berita Utama, *Latent Dirichlet Allocation* (LDA)

Corresponding Author:
Mario Rudy Silalahi
Institut Teknologi Telkom Purwokerto
Jl. DI Panjaitan No.128, Kec. Purwokerto Sel., Kabupaten Banyumas, Indonesia 53147
Email: 20102018@ittelkom-pwt.ac.id

I. PENDAHULUAN

Melalui hasil survei yang dilaksanakan oleh lemaba APJII di tahun 2019-2020 tentang penetrasi pengguna Internet, terdapat peningkatan sebesar 8,9% dibandingkan dengan tahun 2018. Tingkat penetrasi Internet pada 2018 sebesar 64,8%, dan angka tersebut meningkat menjadi 73,7% pada tahun 2019-2020. Servey tersebut juga membahas mengenai konten internet berupa berita yang paling sering dikunjungi yaitu berita dengan kategori olahraga dengan nilai sebesar 7,9[1].

Saat ini terdapat banyak perusahaan memanfaatkan teknologi untuk menyebarluaskan informasi termasuk berita. Perusahaan membuat berita secara daring dan sering kali mengubah judul berita sesuai dengan topik yang sedang banyak diperbincangkan pada saat itu. Berita utama yang muncul di *headline*

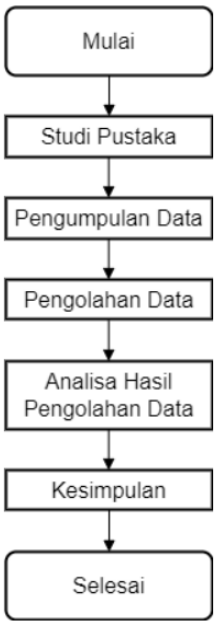
sebuah portal berita daring memiliki signifikansi khusus karena pembaca pertama kali melihat apa yang ada di sana untuk menentukan minat mereka[2].

Dikarenakan variasi data judul berita yang luas, terjadi kesulitan saat mencari berita yang berbeda namun memiliki topik yang sama. Oleh karena itu, untuk memudahkan navigasi, artikel berita perlu dikelompokkan berdasarkan topik yang serupa[3]. Dalam penelitian ini akan berfokus pada penggunaan *topic modeling* untuk menganalisis *headline news* dari situs-situs berita daring, diantaranya mencakup situs-situs berita daring milik liputan6, okezone, idntimes, kompas, detik, suara, dan kumparan menggunakan teknik *web scraping*. Data penelitian dikumpulkan dalam rentang waktu dari 3 Maret 2020 hingga 3 Maret 2021 dengan kata kunci "*sport*". *Topic modeling* adalah suatu metode otomatis untuk mengorganisasi topik dari data dengan menggunakan pemodelan, sehingga menghasilkan kelompok topik yang sesuai dengan konten data tersebut[4]. Ketika menggunakan pemodelan topik dalam *text mining*, evaluasi setiap topik dalam sebuah dokumen dilakukan berdasarkan sejauh mana kumpulan kata tersebut muncul dalam dokumen tersebut[5]. Dalam penelitian ini akan mengunakan metode yang popular dalam *topic modeling* yaitu Latent Dirichlet Allocation yang merupakan sebuah metode dalam analisis teks yang mengadopsi struktur hirarki tiga tingkat. Dalam konteks model jenis *Naïve Bayes* yang lebih sederhana, LDA dapat digambarkan sebagai proses pengelompokan kata-kata ke dalam kluster yang dikenal sebagai topik dalam suatu dokumen atau teks.[6]. Tujuan dilakukannya penelitian ini adalah untuk mendapatkan wawasan yang mendalam seputar analisis topik dan konsistensi topik dalam kerangka konteks berita olahraga.

II. METODE PENELITIAN

A. Alir Penelitian

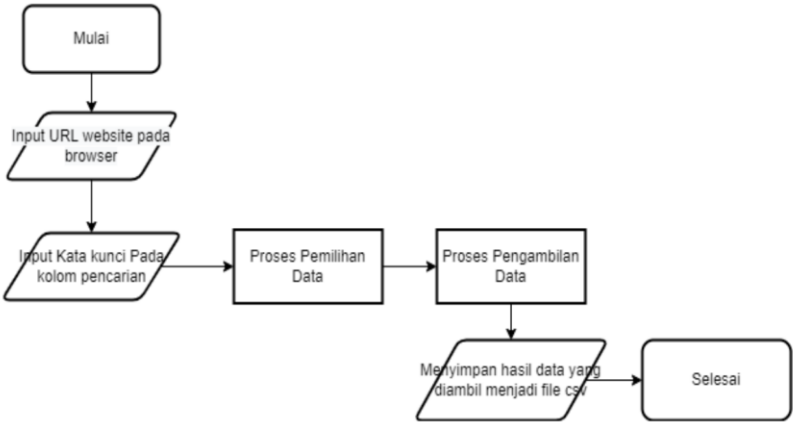
Terdapat sejumlah tahapan yang dilakukan selama penelitian ini. Diawali dengan studi Pustaka yang dilakukan untuk mengumpulkan informasi relevan tentang topik penelitian dan dasar teoritis yang mendukungnya. Lalu dilakukan proses pengumpulan data, kemudian data akan diolah , lalu hasilnya akan dianalisis, dan kemudian diambil kesimpulan. Tahapan penelitian dapat digambarkan melelui diagam penelitian dibawah ini :



Gambar 1. Diagram Alir Penelitian

B. Pengumpulan data

Setelah proses studi pustaka, dilakukan proses pengumpulan data dengan dengan mengunakan alat bantu berupa Data Scraper untuk mengambil judul dari berita utama pada daftar website seperti yang dijelaskan pada latar belakang. Prosses pengumpulan data mengunakan Data Scraper dapat digambarkan melalui diagram berikut :



Gambar 2. Proses Pengumpulan Data

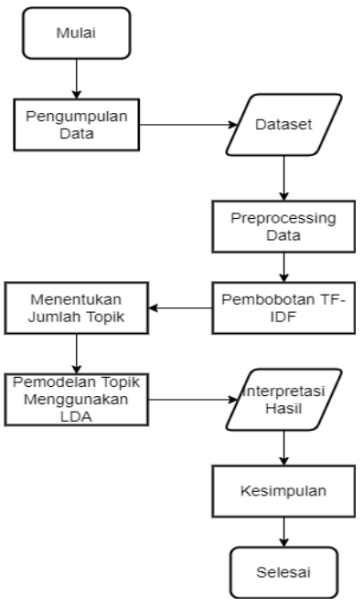
Contoh hasil dari pengumpulan data mengunakan *tool* Data Scraper dapat dilihat pada tabel berikut ini :

Tabel 1. Contoh Data Hasil *Scraping* dengan Data Scraper

No	Headline
1	Juventus Didepak Porto, De Ligt: Sulit Diterima
2	Karra Syam Si Presenter Olahraga Seksi yang Jago Main Bola
3	Mercedes-AMG CLA 45 S Siap Datang, Bakal Jadi Mobil Premium Tercepat di RI
…	…
4300	Unik 5 Derby Liga Inggris Ini Diberi Nama Berdasarkan Nama Jalan
4301	Di MotoGP 2021 Tim Satelit Bisa Lahirkan Juara Baru

C. Proses Pengolahan data

Proses pengolahan data yang dilaksanakan dalam penelitian ini akan mencakup tiga langkah utama yang diawali dengan proses pra-pemrosesan data, lalu dilanjutkan dengan mengunakan TF-IDF untuk pembobotan kata, dan yang dilakukan adalah pemodelan topik dengan metode LDA. Alir pengolahan data dapat digambarkan melalui diagram berikut :



Gambar 3. Diagram Pengolahan Data

1. Preprocessing Data

Dalam banyak kasus, dokumen memiliki susunan yang acak atau tidak teratur, oleh karena itu, diperlukan proses perubahan dan perbaikan data dari keadaan yang tidak terstruktur menjadi keadaan yang terstruktur, proses ini dikenal dengan preprocessing[7]. Penggunaan data dalam tahap preprocessing data dimaksudkan untuk meningkatkan tingkat akurasi dalam perhitungan

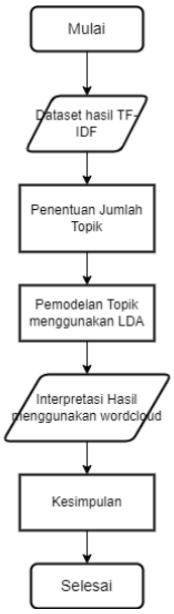
prediksi[8]. Pada penelitian ini, tahapan pra-pemrosesan data diadakan guna membersihkan teks yang tidak diperlukan. Proses pra-pemrosesan dalam penelitian ini dimulai dari prose *cleansing, lower casing,* kemudian dilakukan proses *remove punctuantion* , *stopword removal*, lalu yang terakhir melalui proses *tokenizing*.

2. Pembobotan TF-IDF

Pembobotan dengan TF-IDF akan menggabungkan 2 aspek penting, yaitu jumlah banyaknya kemunculan kata dalam dokumen tersebut dan jumlah banyaknya kebalikan dari dokumen yang didalamnya terdapat kata tersebut, sehingga memberikan bobot yang lebih baik pada kata-kata yang memiliki makna lebih besar dalam dokumen tersebut[9]. Pada penelitian ini, pembobotan melalui metode TF-IDF menghasilkan bobot nilai numerik pada setiap kata dalam dokumen atau dataset untuk persiapan penggunaan metode LDA.

3. Pemodelan Topik dengan LDA

Setelah itu, pemodelan topik dilakukan dengan menggunakan LDA. Dalam penelitian ini, LDA akan digunakan untuk meringkas, mengelompokkan, dan memproses data yang cukup besar, karena LDA menghasilkan susunan topik yang memiliki bobot untuk setiap teks atau dokumen[10]. Hasil dari proses ini akan dievaluasi melalui visualisasi menggunakan pyLDAvis. Visualisasi ini akan menyajikan hasil dari penelitian, termasuk penggunaan library *word cloud*. Proses ini dijelaskan lebih rinci melalui gambar berikut:



Gambar 4. Diagram Proses Pemodelan Topik dengan LDA

Dari diagram diatas, proses pemodelan topik LDA melibatkan penggunaan dataset yang telah dibobotkan dengan TF-IDF sebagai langkah awal. Jumlah topik yang dihasilkan ditentukan berdasarkan nilai koherensi tertinggi untuk memilih jumlah topik yang paling optimal. Setelah menentukan jumlah topik dan nilai koherensinya, langkah selanjutnya adalah menggunakan metode LDA untuk melakukan pemodelan topik dengan memanfaatkan library pyLDAvis untuk merancang visualisasi hasil pemodelan topik. Dalam pemodelan topik dengan LDA, penulis menggunakan probabilitas kata untuk memahami persebaran topik dan mengidentifikasi kata-kata yang signifikan. Hasil analisis ini digunakan untuk merumuskan kesimpulan penelitian.

III. HASIL DAN PEMBAHASAN

A. Hasil Pengolahan

1. Hasil *Preprocessing*
   a. *Cleansing* Data

Proses ini dilakukan untuk membersihkan data dari data yang duplikat atau sama dengan data yang lainnya. Hasil dari proses ini dapat dilihat melalui tabel berikut:

Tabel 2. Perbandingan Data Setelah Proses Pembesihan

Sebelum	Setelah
Juventus Didepak Porto De Ligt: Sulit Diterima Juventus Didepak Porto De Ligt: Sulit Diterima	Juventus Didepak Porto De Ligt: Sulit Diterima
Karra Syam Si Presenter Olahraga Seksi yang Jago Main Bola	Karra Syam Si Presenter Olahraga Seksi yang Jago Main Bola
Mercedes-AMG CLA 45 S Siap Datang Bakal Jadi Mobil Premium Tercepat di RI Mercedes-AMG CLA 45 S Siap Datang Bakal Jadi Mobil Premium Tercepat di RI	Mercedes-AMG CLA 45 S Siap Datang Bakal Jadi Mobil Premium Tercepat di RI

b. Lower Casing

Pada tahapan proses *lower casing*, semua huruf dijadikan sama menjadi huruf kecil semua. Hasil dari proses *lower casing* dapat dilihat melalui tabel berikut:

Tabel 3. Perbandingan Data Setelah Proses *Lower Casing*

Sebelum	Setelah
Juventus Didepak Porto De Ligt: Sulit Diterima	juventus didepak porto de ligt: sulit diterima
Karra Syam Si Presenter Olahraga Seksi yang Jago Main Bola	karra syam si presenter olahraga seksi yang jago main bola
Mercedes-AMG CLA 45 S Siap Datang Bakal Jadi Mobil Premium Tercepat di RI	mercedes-amg cla 45 s siap datang bakal jadi mobil premium tercepat di ri

c. Remove Punctuation

Proses *remove Punctuation* dilakukan dengan tujuan untuk menghapus tanda baca pada kaliimat. Hasil teks setelah melalui proses *remove Punctuation* dapat dilihat melalui tabel berikut:

Tabel 4. Perbandingan Data Setelah Proses *Remove Punctuation*

Sebelum	Setelah
juventus didepak porto de ligt: sulit diterima	juventus didepak porto de ligt sulit diterima
karra syam si presenter olahraga seksi yang jago main bola	karra syam si presenter olahraga seksi yang jago main bola
mercedes-amg cla 45 s siap datang bakal jadi mobil premium tercepat di ri	mercedes amg cla 45 s siap datang bakal jadi mobil premium tercepat di ri

d. Stopword Removal

Pada tahapan proses *stopword removal*, kata yang muncul namun tidak memiliki makna akan dihapus. Hasil dari proses *stopword removal* dapat diketahui melalui tabel berikut:

Tabel 5. Perbandingan Data Setelah Proses *Stopword Removal*

Sebelum	Setelah
juventus didepak porto de ligt sulit diterima	juventus porto de ligt sulit
karra syam si presenter olahraga seksi yang jago main bola	karra syam presenter olahraga jago main bola
mercedes amg cla 45 s siap datang bakal jadi mobil premium tercepat di ri RI	mercedes amg cla s mobil premium tercepat ri

e. Tokenizing

Tokenisasi adalah langkah untuk memisahkan deretan kata menjadi potongan kata atau token yang memiliki makna yang akan digunakan dalam pembentukan matriks dokumen selanjutnya.Hasil dari proses *tokenizing* dapat dilihat dari tabel berikut ini:

Tabel 6. Perbandingan Data Setelah Proses Tokenisasi

Sebelum	Setelah
juventus porto de ligt sulit	[‘juventus’, ‘porto’, ‘de’, ‘ligt’, ‘sulit’]
karra syam presenter olahraga jago main bola	['karra', 'syam', 'presenter', 'olahraga', 'jago', 'main', 'bola']
mercedes amg cla s mobil premium tercepat ri	['mercedes,’amg', 'cla', 's', 'mobil', 'premium', 'tercepat', 'ri']

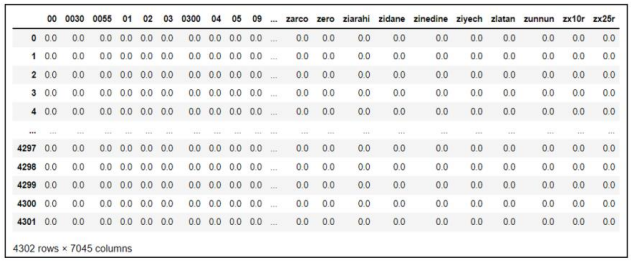
Setelah proses *preprocessing* data selesai, dataset ini telah disiapkan untuk digunakan dalam tahap penelitian selanjutnya. Langkah yang akan dilakukan setelah preprocessing adalah pembobotan kata, yang bertujuan mengkonversi kata-kata menjadi representasi data numerik. Di bawah ini adalah contoh tabel dataset yang telah melalui proses preprocessing.

Tabel 7. Contoh Dataset Hasil *Preprocessing*

Headline_text
"['juventus', 'porto', 'de', 'ligt', 'sulit']"
"['karra', 'syam', 'presenter', 'olahraga', 'jago', 'main', 'bola']"
"['mercedesamg', 'clas', 'mobil', 'premium', 'tercepat', 'ri']"

2. Hasil Pembobotan TF-IDF

Setelah preprocessing data, dataset diberikan pembobotan TF-IDF untuk mengonversi kata-kata menjadi data numerik. Hasil dari pembobotan TF-IDF melalui *softwere* python dapat dilihat melalui gambar berikut:



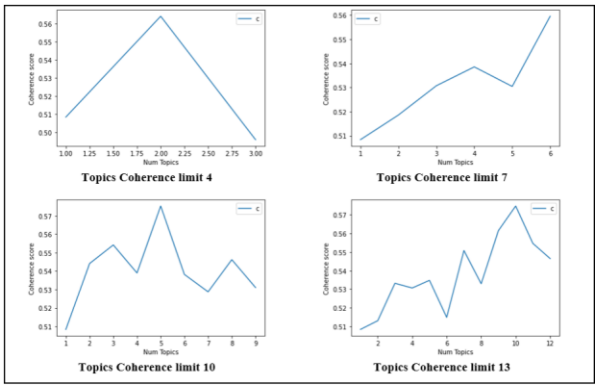
	00	0030	0055	01	02	03	0300	04	05	09	...	zarco	zero	ziarahi	zidane	zinedine	ziyech	zlatan	zunnun	zx10r	zx25r
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4297	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4298	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4299	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4300	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4301	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

4302 rows × 7045 columns

Gambar 5. Hasil Proses TF-IDF

3. Hasil Penentuan Jumlah Topik Berdasarkan *Topics Coherence*

Proses selanjutnya adalah menentukan jumlah topik dan melaksanakan pemodelan topik menggunakan LDA. Untuk menentukan jumlah model topik dalam *topic modeling*, visualisasi pada grafik koherensi topik dapat digunakan sebagai panduan. Dari gambar dibawah ini, peneliti melakukan eksperimen dengan menguji berbagai nilai limit, yaitu limit 4, limit 7, limit 10, dan limit 13, untuk menentukan jumlah topik yang optimal dalam pemodelan topik menggunakan metode LDA. Tujuan penggunaan limit adalah untuk mencari nilai tertinggi dalam grafik *topics coherence*, yang kemudian digunakan sebagai dasar untuk menentukan jumlah topik. Hasil dari visualisasi pada Gambar 6 menunjukkan bahwa nilai tertinggi dari *topics coherence* tercapai pada limit 10 dengan jumlah topik sebanyak 5.

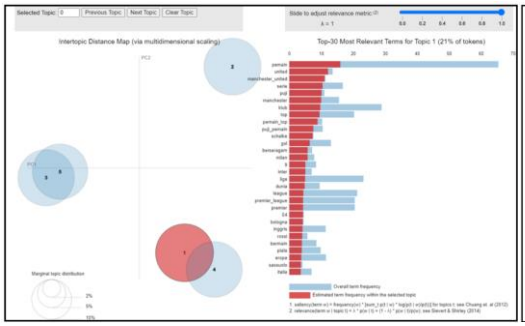


Gambar 6. Grafik *Topics Coherence*

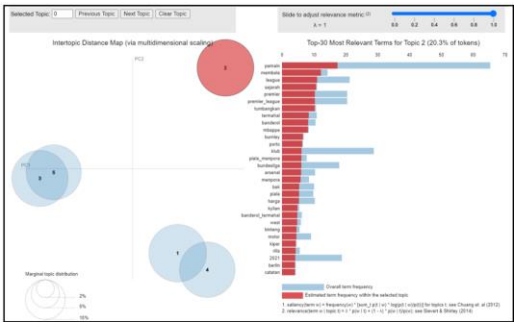
4. Hasil Pemodelan Topik dengan motode LDA

Dalam tahap ini, kita menggunakan pustaka pyLDAvis dan *word cloud* sebagai alat untuk visualisasi pemodelan topik. Ini menciptakan pemetaan jarak antar topik menggunakan multidimensional *scaling* dan mengidentifikasi kata-kata kunci dalam setiap topik. Gambar dibagi menjadi dua bagian. Ada 30 terminologi global yang paling relevan dengan topik-topik dalam dokumen, dengan grafik biru menunjukkan frekuensi total terminologi dalam dokumen, dan grafik

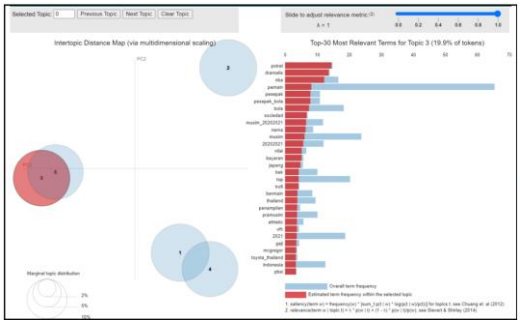
merah menggambarkan frekuensi terminologi terhadap topik tertentu dalam dokumen. Gambar-gambar dibawah ini menjelaskan distribusi margin, hasil model, dan visualisasi *word cloud* untuk setiap topik.



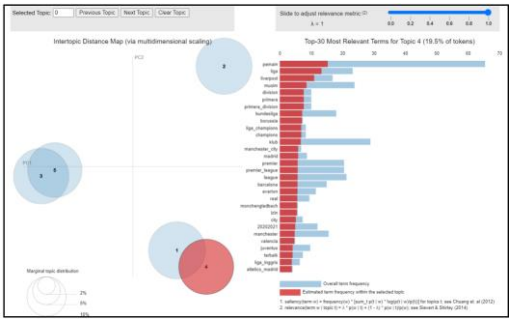
Gambar 7. Hasil Visualisai Topik 1



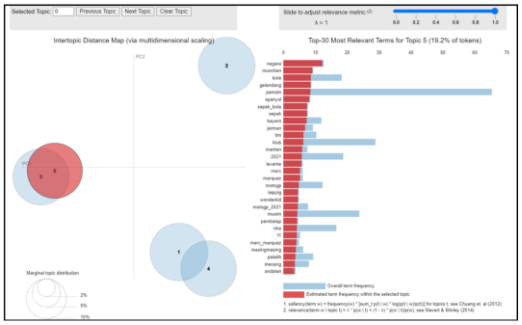
Gambar 8. Hasil Visualisai Topik 2



Gambar 9. Hasil Visualisai Topik 3



Gambar 10. Hasil Visualisai Topik 4



Gambar 11. Hasil Visualisai Topik 5

```
(4,
 '0.024*"pemain" + 0.018*"united" + 0.017*"manchester_united" + 0.016*"serie" '
 '+ 0.015*"puji" + 0.015*"manchester" + 0.015*"klub" + 0.014*"top" + '
 '0.013*"pemain_top" + 0.011*"puji_pemain"')]
```

Gambar 12. Hasil Model Topik 1

```
[(0,
  '0.027*"pemain" + 0.019*"membela" + 0.017*"league" + 0.017*"sejarah" + '
  '0.016*"premier" + 0.016*"premier_league" + 0.016*"tumbangkan" + '
  '0.013*"termahal" + 0.013*"banderol" + 0.013*"mbappe"'),
```

Gambar 13. Hasil Model Topik 2

```
(1,
 '0.023*"potret" + 0.022*"dramatis" + 0.019*"nba" + 0.013*"pemain" + '
 '0.013*"pesepak_bola" + 0.013*"pesepak" + 0.012*"bola" + 0.011*"sociedad" + '
 '0.010*"musim_20202021" + 0.010*"nama"'),
```

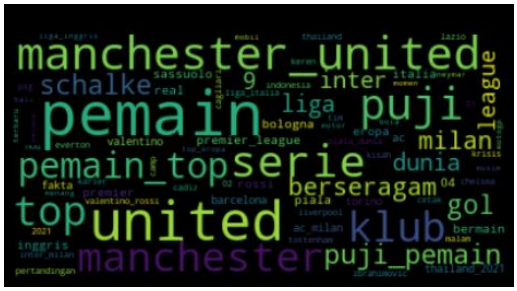
Gambar 14. Hasil Model Topik 3

```
(2,
 '0.025*"pemain" + 0.022*"liga" + 0.018*"liverpool" + 0.014*"musim" + '
 '0.012*"division" + 0.012*"primera_division" + 0.012*"primera" + '
 '0.012*"bundesliga" + 0.011*"borussia" + 0.011*"liga_champions"'),
```

Gambar 15. Hasil Model Topik 4

```
(3,
 '0.020*"negara" + 0.015*"munchen" + 0.014*"bola" + 0.014*"gelandang" + '
 '0.014*"pemain" + 0.013*"spanyol" + 0.012*"sepak_bola" + 0.012*"sepak" + '
 '0.012*"bayern" + 0.011*"jerman"'),
```

Gambar16. Hasil Model Topik 5



Gambar 17. *Word Cloud* Topik 1

Gambar 18. *Word Cloud* Topik 2

Gambar 19. *Word Cloud* Topik 3

Gambar 20. *Word Cloud* Topik 4

Gambar 21. *Word Cloud* Topik 5

Gambar-gambar topik 1 menjelaskan peta sebaran topik 1 dalam korpus, model topik 1, dan visualisasi word cloud untuk topik 1. Peta sebaran menunjukkan pentingnya topik 1 dengan lingkaran merah, model topik 1 mengungkap empat kata kunci utama, yaitu "pemain," "united," "manchester_united," dan "serie," sementara *word cloud* menyoroti 30 kata terkait topik 1 yang paling relevan. Kesimpulannya, topik 1 membahas pemain klub sepakbola.

Gambar-gambar topik 2 menjelaskan peta sebaran, model topik, dan visualisasi word cloud untuk topik ini. Peta sebaran menunjukkan pentingnya topik 2 dengan lingkaran merah, sementara model topik 2 mengungkap empat kata kunci utama, yaitu "pemain," "membela," "league," dan "sejarah." Visualisasi *word cloud* pada gambar topik 2 merepresentasikan 30 kata yang paling relevan. Dari visualisasi tersebut, dapat disimpulkan bahwa topik 2 membicarakan latar bekalang pemain yang bermain klub di liga Premier League.

Gambar-gambar topik 3 menjelaskan peta sebaran, model topik, dan visualisasi word cloud untuk topik ini. Peta sebaran menunjukkan pentingnya topik 3 dengan lingkaran merah, sementara model topik 3 mengungkap empat kata kunci utama, yaitu "potret," "dramatis," "NBA," dan "pemain." Visualisasi *word cloud* pada gambar topik 3 merepresentasikan 30 kata yang paling relevan. Dari visualisasi tersebut, dapat disimpulkan bahwa topik 3 membicarakan gambar dramatis pemain NBA dan pesepakbola.

Gambar-gambar topik 4 menggambarkan peta sebaran, model topik, dan visualisasi word cloud untuk topik ini. Peta sebaran menunjukkan pentingnya topik 4 dengan lingkaran merah, sementara model topik 4 mengungkap empat kata kunci utama, yaitu "pemain," "liga," "Liverpool," dan "musim." Visualisasi *word cloud* pada gambar topik 4 merepresentasikan 30 kata yang paling

relevan. Dari visualisasi tersebut, dapat disimpulkan bahwa topik 4 membicarakan pemain Liverpool dalam konteks Liga Champions.

Gambar-gambar topik 5 menggambarkan peta sebaran, model topik, dan visualisasi word cloud untuk topik ini. Peta sebaran menunjukkan pentingnya topik 5 dengan lingkaran merah, sementara model topik 5 mengungkap empat kata kunci utama, yaitu "negara," "Munchen," "bola," dan "gelandang." Visualisasi *word cloud* pada gambar topik 5 merepresentasikan 30 kata yang paling relevan. Dari visualisasi tersebut, dapat disimpulkan bahwa topik 5 membicarakan klub Bayern Munchen dan pemain gelandang Jerman .

B. Hasil Analisis

Analisis ini bertujuan untuk mengetahui hasil dari penelitian pengelompokan teks dari berita utama dengan pemodelan topik menggunakan metode LDA. Proses analisis mengikuti alur preprocessing data yang telah dilakukan, termasuk pembobotan kata menggunakan metode TF-IDF. TF-IDF digunakan untuk mengubah data kata menjadi data numerik, memungkinkan vektorisasi yang diperlukan untuk penelitian ini, seperti yang ada pada gambar 5. Selanjutnya, penentuan jumlah topik dilakukan berdasarkan pengukuran *topics coherence* yang hasilnya tercantum dalam dibawah ini. Hasil kata kunci dari setiap topik dikelompokkan dan kemudian dianalisis. Dalam penelitian ini, ditemukan 5 pemodelan topik yang akan menjadi fokus pembahasan.

Tabel 8. Hasil Analisis Topik

Topik (kategori)	*Keyword*
Topik 1 (Sepak Bola)	pemain, united, manchester_united, serie, klub
Topik 2 (Sepak Bola)	pemain, membela, league, sejarah, premier, premier_league
Topik 3 (Basket, Sepak Bola)	potret, dramatis, nba, pemain, pesepak_bola
Topik 4 (Sepak Bola)	pemain, liga, Liverpool, musim, liga_champions
Topik 5 (Sepak Bola)	negara, munchen, bola, gelandang, jerman, bayern

Dari hasil analisis data berita utama dari berbagai sumber berita, terlihat bahwa kategori sepak bola memiliki porsi yang paling signifikan. Kategori ini memiliki dampak yang luas pada sebagian besar topik yang diidentifikasi dalam penelitian ini, dengan kata kunci "*sport*" yang berkaitan dengan kasus sepak bola tersebar di seluruh topik yang ada. Hasil analisis, seperti yang terlihat pada Tabel 8, telah berhasil mengelompokkan topik-topik sesuai dengan kategorinya dan sesuai dengan data yang diambil, menunjukkan bahwa pendekatan pemodelan topik yang digunakan efektif dalam mengidentifikasi topik-topik utama dalam berita.

IV. KESIMPULAN

Melalui pemahaman pemodelan topik dengan LDA dalam pengelompokan teks berita utama,beberapa kesimpulan dapat ditarik. Data berita utama cenderung menghasilkan berita yang berkaitan dengan sepak bola dan basket, namun dengan fokus yang lebih kuat pada sepak bola. Dalam penggunaan metode LDA, penentuan jumlah topik dengan mencoba beberapa nilai batas (limit) menunjukkan bahwa nilai topics coherence tertinggi diperoleh pada 5 topik. Dengan demikian, penelitian ini menghasilkan lima model topik yang dapat diidentifikasi. Topik 1 berbicara tentang pemain di klub sepak bola. Topik 2 berkaitan dengan *history* pemain dalam klub di liga Premier League. Topik 3 membahas gambaran dramatis dari pemain basket di NBA dan pemain sepak bola. Topik 4 berfokus pada pemain dari klub sepak bola Liverpool dalam Liga Champions. Terakhir yaitu topik 5 berfokus pada klub sepak bola Bayern Munchen dan pemain dengan posisi gelandang untuk negara Jerman. Dengan demikian, pemodelan topik ini berhasil mengorganisir berita utama ke dalam kelompok yang berbeda berdasarkan topik-topik yang relevan.

DAFTAR PUSTAKA

[1] Asosiasi Penyelenggara Jasa Internet Indonesia, “Laporan Survei Internet APJII 2019 – 2020,” vol. 2020, pp. 1–146, 2020, [Online]. Available: https://apjii.or.id/survei

[2] B. A. Romadhoni, “Meredupnya Media Cetak, Dampak Kemajuan Teknologi Informasi,” *An-Nida J. Komun. Islam*, vol. 10, no. 1, 2019, doi: 10.34001/an.v10i1.741.

[3] B. Subeno, “Optimization Number of Topic Latent Dirichlet Allocation,” 2017.

[4] J. C. Campbell, A. Hindle, and E. Stroulia, “Latent Dirichlet Allocation: Extracting Topics from

Software Engineering Data," *Art Sci. Anal. Softw. Data*, vol. 3, pp. 139–159, 2015, doi: 10.1016/B978-0-12-411519-4.00006-9.

[5] S. Yang, "Text Mining of Twitter Data Using a Latent Dirichlet Allocation Topic Model and Sentiment Analysis," *Int. J. Comput. Inf. Eng.*, vol. 12, no. 7, pp. 525–529, 2018.

[6] I. N. Kabiru and P. K. Sari, "Analisa Konten Media Sosial E-commerce Pada Instagram Menggunakan Metode Sentiment Analysis Dan Lda-based Topic Modeling (studi Kasus: Shopee Indonesia)," *eProceedings Manag.*, vol. 6, no. 1, pp. 12–19, 2019, [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/management/article/view/8498

[7] F. S. Jumeilah, "Penerapan Support Vector Machine (SVM) untuk Pengkategorian Penelitian," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 1, no. 1, pp. 19–25, 2017, doi: 10.29207/resti.v1i1.11.

[8] B. H. Prakoso, "Pengaruh Preprocessing Data pada Metode SVR dalam Memprediksi Permintaan Obat," *J. Sist. Teknol. Inf. Indones.*, vol. 2, no. 2, pp. 92–99, 2017.

[9] M. Z. Naf'an, A. Burhanuddin, and A. Riyani, "Penerapan Cosine Similarity dan Pembobotan TF-IDF untuk Mendeteksi Kemiripan Dokumen," *J. Linguist. Komputasional*, vol. 2, no. 1, pp. 23–27, 2019.

[10] B. H. Puspita, M. Muhajir, and H. Aliady, "Topic Modeling Using Latent Dirichlet Allocation (LDA) and Sentiment Analysis for Marketing Planning Tiket.com," vol. 474, no. Isstec 2019, pp. 16–22, 2020, doi: 10.2991/assehr.k.201010.004.