Akreditasi Sinta 5 KEMRISTEKDIKBUD No. 225/E/KPT/2022; E-ISSN 2723-7958

Metode *Latent Dirichlet Allocation* Dalam Pemodelan Topik *Headline* Berita *Online*

Mega J. Medea[1], Vivi P. Rantung[2], Olivia Kembuan[3]

Program Studi Teknik Informatika, Universitas Negeri Manado
Email: 17210062@unima.ac.id

***Abstract*— Online news is a report about an event that is presented online by a news publishing company through an online news portal. Manado News is a news portal that presents news online which is divided into several categories. This study aims to analyze the data on news titles on the Manado News online news portal, Legal and Criminal category, to find out the crimes that are rampant and reported on this news portal. The analysis is carried out by implementing one of the methods in topic modeling, namely Latent Dirichlet Allocation. This method is used to find the topic of discussion in a document with a large capacity and displays the words that appear the most in a topic. The data needed for modeling was obtained by scraping on the online news portal Manado News, with a span of five years, from April 2017 to April 2022. After the data was processed using the Latent Dirichlet Allocation method, one topic of discussion was found, namely conventional crime groups. The topic contains a number of words which are types of crime in the conventional crime category, namely, persecution, theft, murder, stabbing, motor vehicle theft, sexual abuse and liquor.**

***Keywords*— Latent Dirichlet Allocation, Topic Modeling, Web Scraping, Online News**

***Abstrak*— Berita online merupakan laporan tentang suatu peristiwa yang disajikan secara daring oleh perusahaan penerbit berita lewat portal berita online. Manado News merupakan portal berita yang menyajikan berita secara online yang dibagi ke dalam beberapa kategori. Penelitian ini bertujuan untuk menganalisis data judul berita pada portal berita online Manado News, kategori Hukum dan Kriminal, untuk mengetahui tindak kejahatan yang marak terjadi dan diberitakan pada portal berita ini. Analisis dilakukan dengan mengimplementasikan salah satu metode dalam pemodelan topik yaitu Latent Dirichlet Allocation. Metode ini digunakan untuk menemukan topik bahasan dalam suatu dokumen dengan kapasitas yang besar dan menampilkan kata-kata yang paling banyak muncul dalam suatu topik. Adapun data yang diperlukan untuk pemodelan diperoleh dengan cara scraping pada portal berita online Manado News, dengan rentang waktu lima tahun yaitu April 2017 sampai dengan April 2022. Setelah data diproses dengan metode Latent Dirichlet Allocation, maka ditemukan satu topik bahasan yaitu tentang golongan kejahatan konvensional. Adapun topik tersebut memuat sejumlah kata yang merupakan jenis kejahatan dalam golongan kejahatan konvensional yaitu, penganiayaan, pencurian, pembunuhan, penikaman, pencurian kendaraan bermotor, pencabulan dan minuman keras.**

***Kata Kunci*— Latent Dirichlet Allocation, Pemodelan Topik, Web Scraping, Berita Online**

I. PENDAHULUAN

Berita merupakan suatu cerita atau informasi tentang suatu peristiwa. Berita dibedakan menjadi dua jenis, yaitu yang diungkapkan secara lisan dan secara tertulis. Berita yang disiarkan di televisi atau radio merupakan contoh berita dari mulut ke mulut atau lisan. Sedangkan pesan yang disampaikan secara tertulis dapat ditemukan di media cetak maupun *online*. Berita *online* merupakan laporan tentang suatu peristiwa yang disajikan secara daring oleh perusahaan penerbit berita lewat portal berita *online*. Perusahaan penerbit yang mengelolah berita biasanya membagi berita ke dalam beberapa kategori untuk memudahkan pembaca dalam mencari berita yang hendak dibaca.

Manado News merupakan satu dari sejumlah portal berita *online* di Sulawesi Utara. Manado News berkomitmen untuk menyediakan informasi yang cepat dan dapat dipercaya oleh setiap pembaca terkait dengan pemenuhan kebutuhan informasi. Portal berita ini membagi berita yang disajikan ke dalam beberapa kategori diantaranya: hukum & kriminal, pemerintahan, ekonomi & bisnis, kesehatan, olahraga, pendidikan, budaya, politik dan advertorial. Berdasarkan pengamatan terhadap urutan kategori pada *website* www.manadonews.co.id, hukum dan kriminal menempati posisi pertama. Dalam kategori ini disajikan berita terkini terkait peristiwa yang terjadi akibat adanya praktek tindak kejahatan di Sulawesi Utara.

Banyaknya berita yang disajikan tentang tindak kejahatan yang termasuk dalam kategori berita hukum dan kriminal ini mendorong peneliti untuk melakukan analisis dalam bentuk pemodelan topik (*topic modelling*) terhadap judul berita yang terdapat pada portal berita *online* Manado News kategori hukum & kriminal untuk mengetahui secara ringkas mengenai praktek tindak kejahatan yang telah terjadi dan diberitakan.

Judul berita dipilih karena pesan pertama yang dibaca pembaca berita adalah judul berita sebelum mereka membaca informasi lengkap yang disediakan dalam artikel berita. Ringkasan informasi dari berita biasanya diringkas dalam judul. Konstruksi kata dalam judul dilakukan secara efektif, dengan menekankan unsur siapa dan apa yang terjadi. Kata kerja, kata benda, kata sifat dan kata keterangan dipilih yang singkat dan tajam untuk menyampaikan pesan kepada pembaca. Hal inilah

Akreditasi Sinta 5 KEMRISTEKDIKBUD No. 225/E/KPT/2022; E-ISSN 2723-7958

yang mendasari peneliti hingga akhirnya memilih judul berita sebagai bagian dari berita yang akan diolah [1].

II. TINJAUAN PUSTAKA

A. Kejahatan

Pengertian kejahatan dibagi menjadi dua aspek, yaitu aspek hukum dan aspek sosiologis. Dalam hukum, kejahatan berarti perbuatan melawan hukum. Secara sosiologis kejahatan dipahami sebagai suatu perbuatan yang selain merugikan korban, juga merugikan masyarakat berupa ketidakseimbangan, ketentraman dan ketertiban.

Peraturan KAPOLRI No. 7 Tahun 2009 tentang Sistem Laporan Gangguan Keamanan dan Ketertiban Masyarakat, membagi kejahatan ke dalam lima golongan [2]:

1) Kejahatan Konvensional*:* Kejahatan terhadap nyawa, harta benda dan kehormatan yang menyebabkan kerugian fisik dan psikologis terjadi di negara ini, baik yang dilakukan dengan cara konvensional atau dalam dimensi baru.

2) Kejahatan Transnasional: Kejahatan terorganisir, yang cakupannya mencakup banyak negara, mempengaruhi kepentingan politik, pemerintahan, sosial budaya dan ekonomi negara tersebut dan lingkungan global.

3) Kejahatan Terhadap Kekayaan Negara: Tindak pidana yang mengakibatkan kerugian negara, dilakukan oleh orang perseorangan, kolektif dan/atau korporasi (badan).

4) Kejahatan Berimplikasikan Kontinjensi*:* Kejahatan yang mengganggu aspek keamanan, politik, sosial dan ekonomi serta masyarakat yang bersifat mendadak dan sulit diprediksi.

5) Pelanggaran Hak Asasi Manusia (HAM): Setiap perbuatan oleh seseorang atau sekelompok orang, termasuk aparatur Negara, baik disengaja maupun tidak disengaja, atau kelalaian yang melanggar hukum, yang mengurangi, membatasi, membatasi dan/atau mengabaikan hak asasi manusia seseorang atau sekelompok orang yang dijamin oleh undang-undang dan tidak dapat memperoleh atau dikhawatirkan tidak memperoleh penyelesaian hukum yang adil dan baik berdasarkan mekanisme hukum yang ada.

B. *Web Scraping*

Web scraping adalah teknik untuk mendapatkan informasi dari *website* secara otomatis tanpa harus menyalinnya secara manual. Tujuan dari *web scraper* adalah untuk mencari informasi tententu dan kemudian mengumpulkannya dalam *web* yang baru. *Web scraping* berfokus dalam mendapatkan data dengan cara pengambilan dan ekstraksi. Manfaat dari *web scraping* ialah agar informasi yang dikeruk lebih terfokus sehingga memudahkan dalam melakukan pencarian sesuatu [3].

C. *Pre-processing*

Pre-processing adalah tahap pertama dari *text mining* untuk mempersiapkan data teks. Seperangkat teks (korpus) harus dibagi menjadi elemen yang lebih bermakna. Dokumen dapat dibagi lagi menjadi bab, sub bagian, paragraf, kalimat, kata dan suku kata. Berikut adalah tahapan dalam *pre-processing*:

1) *Case Folding*: Mengubah huruf dalam dokumen menjadi huruf kecil (*lowercase*). Data tekstual harus ditulis dengan huruf kecil dengan maksud agar kata yang sama tidak disalahartikan sebagai kata yang berbeda.

2) *Tokenizing*: Kegiatan atau proses memecah sebaris kata menjadi kalimat atau paragraf menjadi potongan-potongan kata.

3) *Filtering*: Mengekstraksi kata kunci dari hasil tokenisasi. Proses penyaringan menggunakan algoritma *stoplist* (menghapus kata-kata yang tidak penting). *Stopwords* adalah kata-kata umum yang biasanya muncul dalam jumlah besar dan dianggap tidak berarti. Menghapus *stopword* merupakan langkah penting karena tingginya frekuensi *stopword* dalam dokumen.

4) *Stemming*: Mengubah bentuk suatu kata menjadi kata dasar. *Stemming* bekerja dengan menghapus semua imbuhan. Data teks harus terdiri dari kata kunci dasar tanpa kata-kata yang sama tetapi berbeda karena imbuhan.

D. *Topic Modelling*

Pemodelan topik dapat memproses dokumen sesuai dengan topik yang mendasarinya dan memungkinkan kita untuk mengatur dan meringkas arsip elektronik pada skala yang tidak mungkin dijelaskan dengan penjelasan manusia secara manual [4].

E. *Latent Dirichlet Allocation*

Latent Dirichlet Allocation (LDA) adalah metode paling populer untuk pemodelan dan analisis topik. LDA dapat digunakan untuk meringkas, mengelompokkan, menghubungkan, dan memproses data dalam jumlah yang sangat besar [5]. LDA adalah model probabilistik generatif dari kumpulan teks yang disebut *corpus*. Ide dasar dari metode LDA adalah bahwa setiap dokumen direpresentasikan sebagai campuran acak dari topik tersembunyi, di mana setiap topik memiliki karakter yang ditentukan berdasarkan distribusi kata-kata yang ada [4].

Latent Dirichlet Allocation (LDA) bekerja dengan menerapkan *gibbs sampling*. *Gibbs Sampling* adalah algoritma yang menghitung distribusi probabilitas gabungan dengan mengambil sampel secara individual setiap variabel berdasarkan nilai variabel lain, atau sering disebut probabilitas bersyarat [6].

Probabilitas topik pada dokumen adalah nilai probabilitas topik dalam suatu dokumen. Nilai probabilitas tersebut dapat diperoleh dengan rumus:

$$\frac{Jumlah\ topik\ \boldsymbol{K}\ pada\ dokumen\ \boldsymbol{D} + alpha}{Panjang\ dokumen\ \boldsymbol{D} + jumlah\ topik * alpha}$$

Dimana:

Akreditasi Sinta 5 KEMRISTEKDIKBUD No. 225/E/KPT/2022; E-ISSN 2723-7958

K : Indeks Topik
D : Indeks Dokumen
α : Parameter *Dirichlet* (0,1)

Parameter *alpha* (α) digunakan untuk menentukan jumlah topik dalam dokumen. Semakin tinggi nilai alpha, semakin banyak topik yang dibahas dalam dokumen.

F. *Term Frequency-Inverse Document Frequency*

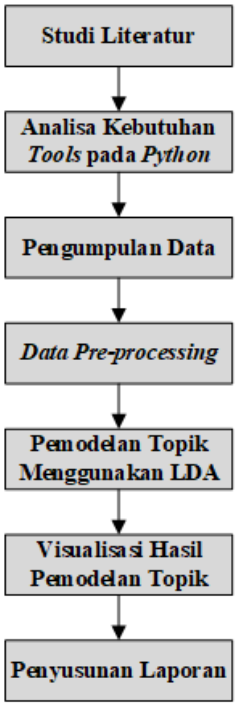
Term Frequency-Inverse Document Frequency (TF-IDF) merupakan perhitungan yang menentukan seberapa relevan sebuah kata dengan dokumen, dengan cara memberikan bobot pada setiap kata. TF-IDF terdiri dari dua konsep komputasi, yaitu perhitungan frekuensi kemunculan sebuah kata dalam sebuah dokumen dan kebalikan dari frekuensi dokumen yang berisi kata tersebut [7]. Saat memodelkan topik, perhitungan TF-IDF digunakan untuk menemukan jumlah kata dengan frekuensi kemunculan tertinggi di dalam topik.

G. *Python*

Python adalah bahasa pemrograman tingkat tinggi yang bertindak sebagai interpreter. *Python* bekerja dengan cara membaca tiap baris dalam program yang kemudian diproses untuk mendapatkan *output*. Dalam pengimplementasiannya, *python* membutuhkan banyak *library* dalam memproses setiap data dan memerlukan beberapa konfigurasi terlebih dahulu sebelum bisa dioperasikan [8].

III. METODE PENELITIAN

Tahapan penelitian yang dilakukan dapat dilihat pada Gambar 1 di bawah ini.



Gbr. 1 Diagram Alir Tahapan Penelitian

1) Studi Literatur: Studi literatur adalah proses pengumpulan data dan informasi dengan bahan-bahan seperti buku, dokumen, jurnal, majalah, dll. Dokumen-dokumen ini dapat ditemukan di perpustakaan atau di internet. Studi literatur yang dilakukan dalam penelitian ini adalah studi jurnal yang berkaitan dengan pemodelan topik dengan menggunakan metode LDA.

2) Analisa Kebutuhan *Tools* pada *Python*: Penentuan *tools* yang akan dipakai dalam bahasa pemrograman python untuk memproses data yang akan dimodelkan. Untuk dapat mengakses *python* dan *library* yang diperlukan, peneliti menggunakan paket distribusi *python* dari *anaconda*. Tabel I menunjukkan *tools* dalam *python* yang dipakai dalam pemodelan topik dengan metode LDA.

TABEL I. KEBUTUHAN *TOOLS*

Tools	**Kegunaan**
Anaconda3	Manajemen *tools* untuk bahasa pemrograman *python*.
Jupyter Notebook	Ruang kerja untuk mengimplementasikan bahasa pemrograman python.
Pandas	*Library* yang mendukung pembacaan dan penulisan data menggunakan media seperti *excel spreadsheet*, CSV, dan SQL untuk dijadikan *data frame*.
NLTK	*Natural Language Toolkit* merupakan *library* untuk *pre-processing* data.
Sastrawi	*Library* untuk mengubah kata bahasa Indonesia berimbuhan ke bentuk dasarnya.
Gensim	Gensim adalah *toolkit* pemodelan topik yang diimplementasikan dalam *python* sebagai modul untuk membangun model LDA.
Matplotlib	Digunakan untuk membuat grafik.
Wordcloud	Digunakan untuk presentasi visual data tekstual yang secara independen menggambarkan metadata.

3) Pengumpulan Data: Judul berita yang tersaji diambil dengan cara *web scraping* pada portal berita *online*, kemudian disimpan ke dalam format tertentu sesuai kebutuhan. Data yang diperoleh tersebut kemudian akan di sesuai dengan tujuan yang hendak dicapai. Data yang sudah siap kemudian akan di proses pada tahap *pre-processing*.

4) *Pre-processing*: Data yang diperoleh dari proses *scraping* disebut dengan data *input*. Data input tersebut kemudian akan diolah menggunakan bahasa perograman *python* dengan memanfaatkan *library* yang terdapat didalamnya. Terdapat empat tahapan *pre-processing* yang akan dilakukan yaitu, *case folding*, *tokenization*, *filtering*, dan *stemming*. Hasil yang diperoleh setelah melewati empat tahap *pre-processing* tersebut kemudian disebut dengan *data output*.

5) Pemodelan Topik Menggunakan LDA: Data yang diperoleh dari hasil *pre-processing* kemudian di *import* untuk dimodelkan. Selanjutnya, dengan menggunakan *library* yang terdapat dalam bahasa pemrograman *python*, akan dibuat

Akreditasi Sinta 5 KEMRISTEKDIKBUD No. 225/E/KPT/2022; E-ISSN 2723-7958

dictionary dan *corpus* dari data yang di *import*. Sebelum masuk pada proses pemodelan, akan ditentukan jumlah topik yang hendak dimunculkan. Untuk mendapatkan jumlah ideal dari berapa banyaknya topik, maka dilakukan pencarian nilai kohenrensi (*coherence score*) dan dibuktikan dengan melakukan perhitungan probabilitas kemunculan topik secara manual. Hasil dari nilai koherensi yang diperoleh kemudian dimodelkan ke menjadi topik sesuai banyak topik yang berhasil ditemukan dalam dokumen berdasarkan pengecekkan nilai koherensi.

6) Visualisasi Hasil Pemodelan Topik: Hasil dari pemodelan topik kemudian divisualisasikan ke dalam dua bentuk visualisasi, yaitu *wordcloud* dan diagram batang.

7) Penyusunan Laporan: Tujuan penyusunan laporan ini adalah untuk mendokumentasikan setiap tahapan penelitian dari tahap pertama hingga tahap terakhir.

IV. HASIL DAN PEMBAHASAN

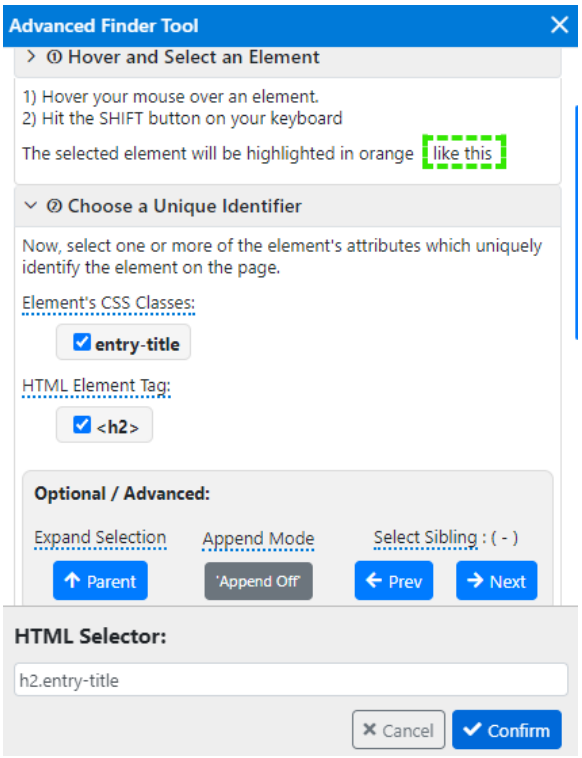
A. Pengumpulan Data

Pengumpulan data merupakan proses dimana data *headline* berita dari portal berita *online* dengan studi kasus Manado News diambil menggunakan *web scraping*. Selanjutnya, data *headline* berita yang diperoleh disaring untuk membuang data *headline* berita yang tidak mengandung unsur pemberitaan mengenai tindak kejahatan di wilayah Sulawesi Utara.

1) *Scraping Data*: Data yang akan diolah ke dalam pemodelan topik dikumpulkan dengan cara *scraping* pada halaman *web* portal berita Manado News (https://www.manadonews.co.id/category/hukum-kriminal/), dengan menitikberatkan pada kategori berita mengenai hukum dan kriminal.

Proses *scraping* dilakukan dengan menggunakan *data scraper* yang dipasang pada ekstensi *Google Chrome*. Saat proses *scraping* pada halaman *web* portal berita, data dipilah terlebih dahulu dengan fitur yang terdapat dalam data *scraper* untuk memperoleh hasil yang lebih spesifik. Data yang diperoleh kemudian disimpan dalam format CSV atau XLSX. Gambar 2 menunjukkan bahwa pada saat memilah data untuk di *scrap*, hanya diambil judul dari berita yang disajikan dalam portal berita dengan *Element's CSS Classes: entry-title dan HTML Element Tag: <h2>*. Setelah data telah siap, maka dilakukan proses *scraping* dan didapatkan data sebanyak 2070 baris data yang berisikan judul berita.

2) *Filtering Data*: Pada proses *scraping* yang dilakukan sebelumnya, data judul berita masih berupa data mentah. Hal ini disebabkan oleh data judul berita yang diperlukan masih tergabung dengan berita yang tidak membahas tentang kejadian kriminal, misalnya seperti kegiatan internal kepolisian yang tidak berkaitan dengan praktik tindak kejahatan. Pada tahap *filtering* ini, ditemukan 1.317 data judul berita yang tidak memiliki korelasi dengan kejadian kriminal. Adapun proses *filtering* dilakukan secara manual dengan cara memeriksa dan menghapus satu demi satu data yang tidak masuk dalam berita tentang kejadian kriminal. Tahapan ini menghasilkan 753 data yang siap di proses pada tahap selanjutnya yaitu *pre-processing*.



Gbr. 2 Pemilahan Elemem HTML Menggunakan Data Scraper

B. *Pre-processing*

1) *Case Folding*: Menggunakan fungsi *lower ()* pada *class series.str* dalam *library Pandas*. *Output* dari *case folding* dapat dilihat pada Tabel II.

TABEL II. *OUTPUT CASE FOLDING*

Data Ke-	Judul Berita	
	Sebelum	**Sesudah**
1	Korban Ditusuk dan Dipukul dengan Piring, Terduga Pelaku sudah Ditangkap Polisi	korban ditusuk dan dipukul dengan piring terduga pelaku sudah ditangkap polisi
3	Lagi, Kasus Kriminalitas Gara-gara Miras, Terduga Pelaku Diringkus Polisi	lagi kasus kriminalitas gara-gara miras terduga pelaku diringkus polisi
18	Polisi Tangkap Pelaku Pencurian Mobil Milik Dokter	polisi tangkap pelaku pencurian mobil milik dokter

Akreditasi Sinta 5 KEMRISTEKDIKBUD No. 225/E/KPT/2022; E-ISSN 2723-7958

2) *Tokenizing*: Proses *tokenizing* ini dijalankan dengan menggunakan fungsi *string* dan *re* pada *library* NLTK. Tabel III menunjukkan *output* dari proses *tokenizing*.

TABEL III. *OUTPUT TOKENIZING*

Data Ke-	Judul Berita	
	Sebelum	**Sesudah**
1	korban ditusuk dan dipukul dengan piring terduga pelaku sudah ditangkap polisi	korban ditusuk dan dipukul dengan piring terduga pelaku sudah ditangkap polisi
3	lagi kasus kriminalitas gara-gara miras terduga pelaku diringkus polisi	lagi kasus kriminalitas garagara miras terduga pelaku diringkus polisi
18	polisi tangkap pelaku pencurian mobil milik dokter	polisi tangkap pelaku pencurian mobil milik dokter

3) *Filtering*: Pada tahap ini digunakan *stopword* bahasa Indonesia yang diperoleh dari *library* NLTK. *Output* dari tahapan ini ditunjukkan pada Tabel IV.

TABEL IV. *OUTPUT FILTERING*

Data Ke-	Judul Berita	
	Sebelum	**Sesudah**
1	korban ditusuk dan dipukul dengan piring terduga pelaku sudah ditangkap polisi	korban ditusuk dipukul piring terduga pelaku ditangkap polisi
3	lagi kasus kriminalitas garagara miras terduga pelaku diringkus polisi	kasus kriminalitas miras terduga pelaku diringkus polisi
18	polisi tangkap pelaku pencurian mobil milik dokter	polisi tangkap pelaku pencurian mobil milik dokter

4) *Stemming*: Pada tahap ini digunakan fungsi *stemmer* dari *library Sastrawi* agar kata yang memiliki imbuhan dapat kembali ke bentuk dasarnya. Tabel V menunjukkan *output* dari proses *stemming*.

TABEL V. *OUTPUT STEMMING*

Data Ke-	Judul Berita	
	Sebelum	**Sesudah**
1	korban ditusuk dipukul piring terduga pelaku ditangkap polisi	korban tusuk pukul piring duga laku tangkap polisi
3	kasus kriminalitas miras terduga pelaku diringkus polisi	kasus kriminalitas miras duga laku ringkus polisi
18	polisi tangkap pelaku pencurian mobil milik dokter	polisi tangkap laku curi mobil milik dokter

Tahapan *pre-processing* selesai dan hasilnya disimpan ke dalam format *excel* untuk diolah kembali dalam pemodelan topik pada tahap selanjutnya.

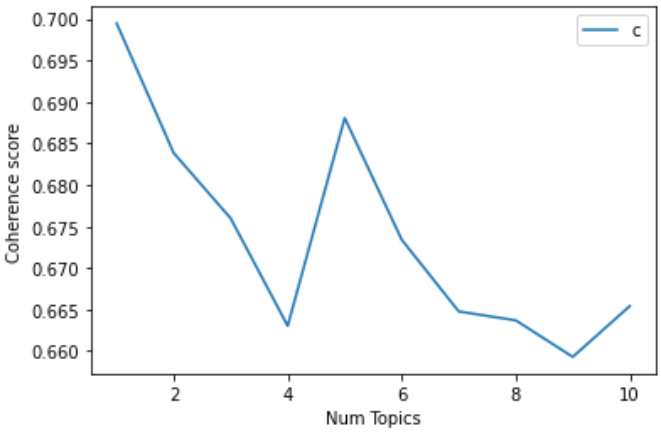
C. Pemodelan Topik Menggunakan LDA

1) *Import Text Data*: Data yang sudah di bersihkan dengan menjalankan *pre-processing* akan di *import* kembali ke dalam *Jupyter Notebook*. *Import text data* menghasilkan *output* berupa daftar data teks dari 753 baris data yang telah melalui tahapan *pre-processing*.

2) Membuat *Dictionary dan Corpus*: Dihasilkan *dictionary* berisikan 168 *unique tokens* dan *corpus* dengan notasi (81, 1) yang berarti kata dengan id 81 muncul sekali pada dokumen pertama.

3) Menentukan Jumlah Topik: Penentuan berapa banyak model topik dilakukan dengan cara memvisualisasikannya dalam bentuk grafik yang mengacu pada *dictionary* dan *corpus* yang dibuat sebelumnya. *Coherence score* merupakan ukuran yang digunakan untuk menemukan jumlah topik dengan skor koherensi tertinggi untuk mengevaluasi pemodelan topik. Hal ini dilakukan untuk mendapatkan model topik dengan jumlah peluang terbesar. Pada penelitian ini, penulis memberi rentang jumlah topik dari satu sampai sepuluh.

Gambar 3 merupakan visualisasi dalam bentuk grafik dari nilai koherensi yang diperlukan untuk penentuan jumlah topik paling ideal.



Gbr. 3 Grafik *Coherence Score*

Berdasarkan Gambar 3, jumlah topik yang paling baik untuk dimodelkan adalah sebanyak satu topik. Tabel VI menunjukkan nilai koherensi dari rentang satu sampai sepuluh dengan pengaturan dua angka dibelakang koma.

TABEL VI. *COHERENCE SCORE*

Number of Topics	***Coherence Score***
1	0.70
2	0.68
3	0.68
4	0.66
5	0.69
6	0.67
7	0.66
8	0.66
9	0.66
10	0.67

Akreditasi Sinta 5 KEMRISTEKDIKBUD No. 225/E/KPT/2022; E-ISSN 2723-7958

Berdasarkan grafik dan tabel nilai koherensi, dapat diketahui bahwa topik dengan nilai koherensi tertinggi berada pada rentang satu dengan nilai koherensi sebesar 0,70.

Untuk membuktikan bahwa jumlah topik ideal adalah sebanyak satu topik adalah benar, maka dilakukan perhitungan secara manual menggunakan rumus probabilitas topik pada suatu dokumen sebagai berikut:

$$\frac{Jumlah\ topik\ \boldsymbol{K}\ pada\ dokumen\ \boldsymbol{D} + alpha}{Panjang\ dokumen\ \boldsymbol{D} + jumlah\ topik * alpha}$$

Dimana:
K : Indeks topik
D : Indeks dokumen
α : Parameter dirichlet (0,1)

Diketahui:
Jumlah topik pada dokumen : 1
Panjang dokumen : 1
Nilai Alpha : 0,1

Penyelesaian:

$$\frac{1+0{,}1}{1+1*0{,}1} = \frac{1{,}1}{1+0{,}1} = \frac{1{,}1}{1{,}1} = 1$$

Berdasarkan perhitungan probabilitas topik pada suatu dokumen, maka jumlah ideal dari topik yang akan dimodelkan adalah sebanyak satu topik.

4) Pemodelan Topik: Setelah jumlah topik ideal diperoleh, langkah selanjutnya adalah mengaplikasikan pemodelan topik dengan LDA menggunakan *library Gensim*. *Output* yang diperoleh berupa daftar kata yang sudah diberi bobot dengan menggunakan perhitungan *Term Frequency-Inverse Document Frequency* (TF-IDF). Daftar kata yang paling banyak muncul berdasarkan perhitungan bobot kata menggunakan TF-IDF dapat dilihat pada Tabel VII berikut ini:

TABEL VII. DAFTAR KATA PENYUSUN TOPIK DAN NILAI TF-IDF

Word	***TF-IDF Score***
Aniaya	0,043
Curi	0,035
Bunuh	0,024
Tikam	0,022
Curanmor	0,020
Anak	0,016
Cabul	0,015
Miras	0,014
Edar	0,013
Cap_Tikus	0,012

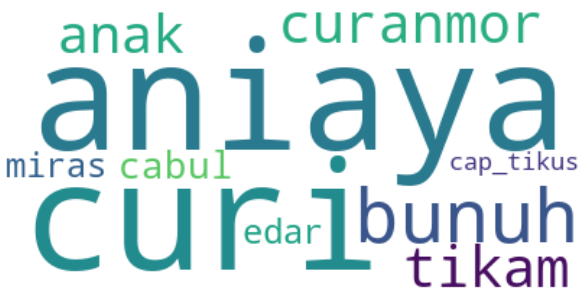
Berdasarkan hasil pembobotan kata yang tertera pada Tabel VI, topik yang berhasil dimodelkan ini adalah tentang tindak kejahatan yang masuk dalam golongan kejahatan konvensional/nasional berdasarkan Peraturan Kepolisian Negara Republik Indonesia Nomor 7 Tahun 2009 tentang Sistem Laporan Gangguan Keamanan dan Ketertiban Masyarakat [2].

Selain itu, diperoleh tujuh kata yang berkaitan dengan klasifikasi kejahatan yang dipublikasikan oleh Badan Pusat Statistik tentang Statistik Kriminal 2021 [9]. Tujuh kata tersebut adalah sebagai berikut:

- Aniaya (kejahatan terhadap fisik/badan)
- Curi (kejahatan terhadap hak milik/barang)
- Bunuh (kejahatan terhadap nyawa)
- Tikam (kejahatan terhadap nyawa)
- Curanmor (kejahatan terhadap hak milik/barang)
- Cabul (kejahatan terhadap kesusilaan)
- Miras (kejahatan terhadap ketertiban umum)

D. Visualisasi Hasil Pemodelan Topik

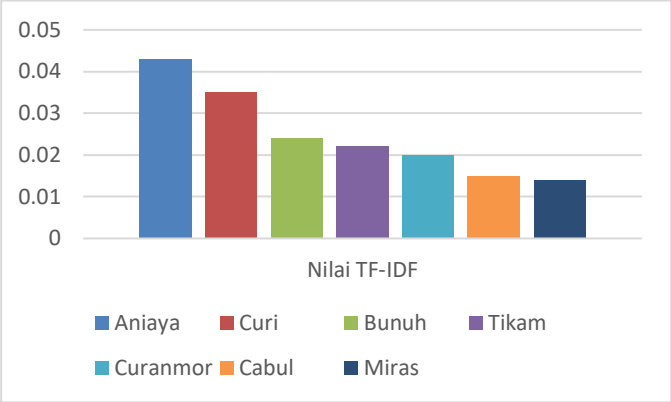
Berdasarkan hasil pemodelan topik dengan LDA di tahap sebelumnya, maka dapat dibuat visualisasi atau tampilan dari topik yang tersusun atas sepuluh kata berupa *wordcloud*.



Gbr. 4. *Wordcloud* Hasil Pemodelan Topik

Gambar 4 menunjukkan *wordcloud* dari topik yang dimodelkan. Dimana kemunculan tiap kata berdasarkan bobot yang dimiliki dalam topik menentukan besar kecilnya ukuran huruf dari kata tersebut. Semakin tinggi bobot dari suatu kata, maka ukuran hurufnya akan lebih besar. Sebaliknya, kata-kata dengan bobot paling kecil ditampilkan dengan ukuran font yang lebih kecil.

Selain dalam bentuk word cloud, visualisasi dari hasil pemodelan topik menggunakan LDA ini juga dapat disajikan dalam bentuk diagram batang yang menunjukkan daftar kata yang masuk dalam jenis kejahatan beserta bobot nilai dari perhitungan TF-IDF yang dibuat menggunakan pemrograman *python*.



Gbr. 5 Diagram Batang Jenis Kejahatan Berdasarkan Bobot TF-IDF

Gambar 5 menampilkan daftar jenis kejahatan yang diperoleh dari pengolahan data dalam dokumen menggunakan pembobotan TF-IDF. Terdaftar tujuh jenis kejahatan yaitu:

- Penganiayaan dengan bobot senilai 0,043
- Pencurian dengan bobot senilai 0,035
- Pembunuhan dengan bobot senilai 0,024
- Penikaman dengan bobot senilai 0,022
- Pencurian kendaraan bermotor dengan bobot senilai 0, 020
- Pencabulan dengan bobot senilai 0,015
- Minuman keras dengan bobot senilai 0,014.

V. KESIMPULAN

Berdasarkan hasil pemodelan topik, dapat diambil kesimpulan bahwa selama lima tahun terakhir, terhitung sejak April 2017 sampai dengan April 2022, jenis kejahatan yang marak terjadi dan diberitakan pada portal berita *online* Manado News dengan kategori berita hukum dan kriminal adalah penganiayaan, pencurian, pembunuhan, penikaman, pencurian kendaraan bermotor, pencabulan dan minuman keras. Penganiayaan menempati posisi paling tinggi dengan bobot nilai sebesar 0,043 berdasarkan perhitungan dengan menggunakan TF-IDF. Hal ini berarti selama lima tahun terakhir, tindak kriminal penganiayaan kerap terjadi di Sulawesi Utara dan diangkat pemberitaannya pada portal berita *online* Manado News. Adapun jenis-jenis kejahatan tersebut, termasuk dalam golongan kejahatan konvensional.

DAFTAR PUSTAKA

[1] bin Ladjamudin, A.-B. (2005). Analisis dan Desain Sistem Informasi. Yogyakarta: Graha Ilmu. Naury, C., Fudholi, D. H., & Hidayatullah, A. F. (2021). Topic Modelling Pada Sentimen Terhadap Headline Berita Online Berbahasa Indonesia Menggunakan LDA dan LSTM. JURNAL MEDIA INFORMATIKA BUDIDARMA, 24-33.

[2] Kepolisian Negara Republik Indonesia. (2009). Peraturan Kepala Kepolisian Negara Republik Indonesia Nomor 7 Tahun 2009 Tentang Sistem Laporan Gangguan Keamanan Dan Ketertiban Masyarakat. Jakarta: an Negara Republik Indonesia.

[3] Ayani, D. D., Pratiwi, H. S., & Muhardi, H. (2019). Implementasi Web Scraping untuk Pengambilan Data pada Situs Marketplace. Jurnal Sistem dan Teknologi Informasi, 257-262.

[4] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993-1022.

[5] Campbell, J. C., Hindle, A., & Stroulia, E. (2014). Latent Dirichlet Allocation: Extracting Topics from Software Engineering Data. 1-21.

[6] Rizky, R. (2018, Agustus 17). Topic Modelling Menggunakan Latent Dirchlect Allocation. Retrieved from Medium: https://medium.com/@ranggaantok/topic-modelling-menggunakan-latent-dirchlect-allocation-3fdf979ffd05

[7] Herwijayanti, B., Ratnawati, D. E., & Muflikhah, L. (2018). Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDF dan Cosine Similarity. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 306-312.

[8] Wibowo, W., Ulama, B., & Al Azies, H. (2020). Belajar Pemrograman Bahasa Python. Surabaya: ITS Press.

[9] Direktorat Statistik Ketahanan Nasional. (2021). Statistik Kriminal 2021. Jakarta: Badan Pusat Statistik.