Jurnal Informasi dan Teknologi

https://jidt.org/jidt

2023 Vol. 5 No. 1 Hal: 265-275 e-ISSN: 2714-9730

Pemodelan Topik Menggunakan n-Gram dan Non-negative Matrix Factorization

Razief Perucha Fauzie Afidh[1✉], Syahrial[2]

[1]Jurusan Informatika, Universitas Syiah Kuala, Banda Aceh, Aceh, Indonesia
[2]MANN Research Center, Banda Aceh, Aceh, Indonesia

razief@usk.ac.id

Abstrak

Pemodelan topik merupakan teknik pembelajaran mesin yang digunakan untuk melihat topik dalam sekumpulan dokumen teks. Salah satu sumber data teks yang banyak digunakan adalah data teks yang berasal dari portal berita. Pada penelitian ini, pemodelan topic yang digunakan adalah *Non-Negative Matrix Factorization* (NMF) dengan mengimplementasikan n-gram. Proses pemodelan topik diawali dengan *pre-processing* data. Diantara *pre-processing* yang dilakukan dalam penelitian ini adalah penghilangan tanda baca, angka dan *stopword*. Proses *pre-processing* dilakukan dengan terlebih dahulu mengubah kata yang terdapat dalam artikel menjadi kata berhuruf kecil. Tanda baca dan angka tidak digunakan dalam pembentukan topik. Penerapan *stopword removal* menggabungkan penggunaan *library* Sastrawi dan daftar *stopword* yang tersedia dalam berkas .txt terpisah. Penelitian ini juga mengeksplorasi keefektifan penerapan unigram, bigram, dan trigram pada pemodelan topik. Pada penelitian ini menggunakan *coherence value* untuk menentukan jumlah topik terbaik yang dapat dibentuk. Data yang digunakan pada penelitian ini berjumlah 53.920 artikel berita yang bersumber dari portal berita RMOL.id dan BeritaSatu.com untuk periode Juli sampai Desember 2022. Visualisasi t-SNE digunakan untuk melihat distribusi pembentukan topik. Berdasarkan hasil penelitian yang dilakukan diperoleh bahwa jumlah topik yang dapat dibentuk dari RMOL.id untuk unigram adalah 15 topik dengan nilai coherence value 0.812748. Untuk implementasi bigram diperoleh 10 topik dengan nilai coherence value 0.835738. Sedangkan untuk implementasi trigram diperoleh 7 topik dengan nilai coherence value 0.830572. Pada portal berita BeritaSatu.com diperoleh 10 topik untuk unigram dengan nilai coherence value 0.799718. Untuk implementasi bigram diperoleh 15 topik dengan nilai coherence value 0.788762. Sedangkan pada implementasi trigram diperoleh 15 topik dengan nilai coherence value 0.801935.

Kata kunci: Pemodelan Topik**,** Non-Negative Matrix Factorization, N-Grams, Coherence Value, t-SNE.

1. Pendahuluan

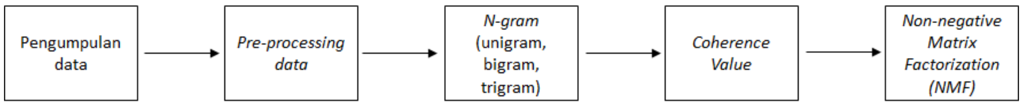
Pembelajaran mesin dapat dimanfaatkan dalam berbagai hal, diantaranya pada bidang kesehatan untuk mendeteksi kanker payudara dengan menggunakan Convolutional Neural Networks (CNN) dalam data berupa gambar[1]. Pembelajaran mesin juga dimanfaatkan untuk menjaga kelestarian budaya, sebagaimana yang diterapkan pada pendeteksian aksara Lota dengan menggunakan CNN [2]. Selain itu, Pembelajaran mesin dapat juga dimanfaatkan pada bentuk data lainnya, seperti pada data teks. Salah satu pemanfaatan data teks adalah pemodelan topik. Pemodelan topik merupakan salah satu kajian bidang pemrosesan bahasa alami atau *Natural Language Processing* (NLP). Pada kajian ini, data teks merupakan salat satu input utama dalam pengolahan data. Pemodelan topik menyediakan teknik untuk mengubah konten yang tidak terstruktur menjadi pengetahuan terstruktur dengan mengungkap tema dan pola yang terdapat dalam data teks. Data teks yang digunakan dapat berasal dari portal berita maupun sosial media seperti twitter dan facebook. Penelitian yang dilakukan oleh [3] menggunakan *Latent Dirichlet Allocation* (LDA) untuk melihat topik, kesamaan topik, dan visualisasi dari topik. Pada penelitian ini menggunakan data twitter untuk melihat 4 topik terkait dengan ekonomi, militer, olahraga, dan teknologi dalam bahasa Indoensia. Penelitian yang dilakukan oleh [4] [5] untuk mengetahui dan menginterpretasikan topik dari lirik lagu Indonesia. Pada penelitian ini menggunakan 200 lagu teratas pada Spotify untuk periode Januari 2017 – Januari 2018 dengan 193 lagu berbeda yang menggunakan bahasa Indonesia dalam liriknya. Pada penelitian yang dilakukan oleh [6] menggunakan Latent Semantics Analysis (LSA) adalah untuk mengekstrak informasi dari sebuah gambar yang mengandung teks, kemudian mengklasifikasikan teks tersebut menjadi kata-kata yang memiliki ujaran kebencian atau tidak mengandung ujaran kebencian. Pada penelitian ini menggunakan data yang bersumber dari twitter. Penelitian yang dilakukan oleh [7] [8] menggunakan LSA untuk membuat sistem peringkasan teks otomatis yang menghasilkan ringkasan singkat untuk dokumen legal. Pada penelitian ini menggunakan data terdiri dari keputusan hukum yang dikeluarkan oleh sistem peradilan India. Data Mahkamah Agung, Pengadilan Tinggi dan Pengadilan Negeri dikumpulkan dari situs resmi yang memiliki ekstensi .nic, .gov, dan lainnya. Penelitian yang dilakukan oleh [9] [1] membandingkan berbagai pendekatan pemodelan topik seperti

Diterima: 19-08-2023 | Revisi: 30-08-2023 | Diterbitkan: 02-09-2023 | doi: 10.60083/jidt.v5i1.385

LSA, LDA, *Probabilistic* LSA (PLSA) dan *Non-negative Matrix Factorization* (NMF) pada tweet berbahasa Urdu. Pada penelitian ini menggunakan n-Gram dan *Non-negative Matrix Factorization* (NMF) untuk melihat topik-topik yang terbentuk pada data teks. Diharapkan penelitian ini dapat memberikan gambaran penggunaan n-gram dan NMF sebagai alternatif dalam melihat topik-topik pada data teks [10] [11].

2. Metode Penelitian

Pada penelitian ini menggunakan *Non-negative Matrix Factorization* (NMF) dengan menerapkan n-Gram untuk melihat jumlah topik dalam suatu data teks yang bersumber dari portal berita. *Coherence Value* digunakan untuk menentukan jumlah topik yang dapat dibentuk. Metode penelitian dapat dilihat pada gambar x.x berikut ini:



Gambar 1. Metode Penelitian

2.1. Pengumpulan Data

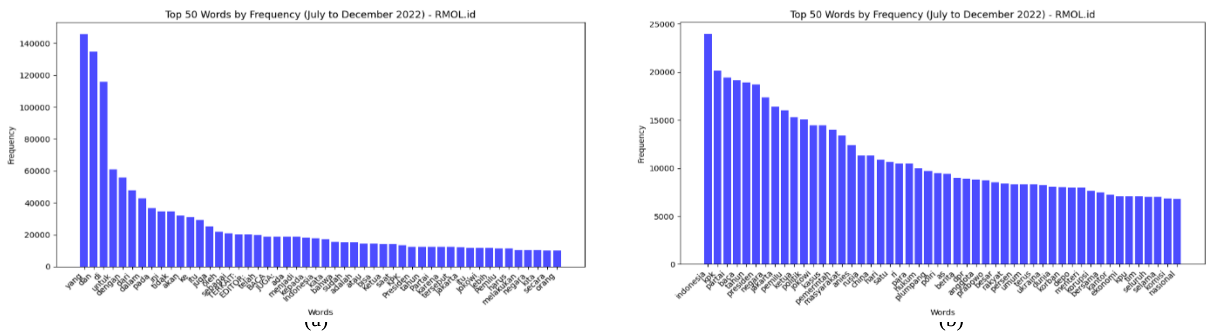
Data yang digunakan pada penelitian ini bersumber dari portal berita berbahasa Indonesia. Sebanyak 20.099 artikel dikumpulkan dari portal berita Republik Merdeka Online (RMOL.id) dan 33.821 artikel berita dikumpulkan dari portal berita BeritaSatu (BeritaSatu.com) untuk periode Juli sampai dengan Desember 2022. Rincian jumlah artikel yang dikumpulkan untuk masing-masing portal berita dapat dilihat pada tabel 1. berikut ini.

Tabel 1. Tabel Jumlah Artikel Berita

Portal Berita	Jumlah Artikel
RMOL	20.099
CNBC Indonesia	33.821
Total	53.920

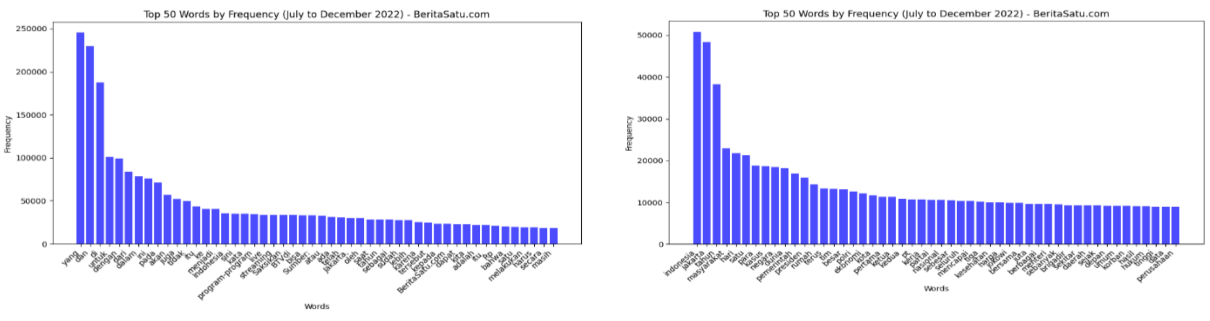
2.2. Pre-processing Data

Proses *pre-processing* data pada penelitian ini meliputi perubahan kata menjadi huruf kecil (*text-to-lower*), menghilangkan simbol, menghilangkan karakter selain kata (*non-word characters*), dan *non-whitespace character*, penghapusan *whitespace* dan angka. Pada tahapan ini juga dilakukan penghapusan *stopwords* dengan mengunakan *library* Sastrawi. Selain menggunakan daftar *stopwords* yang terdapat pada library Sastrawi, juga terdapat tambahan *stopwords* yang disimpan dalam format file .txt.



(a) (b)

Gambar 2. Distribusi Kata sebelum (a) dan setelah (b) *pre-processing* untuk portal berita RMOL.id



(a) (b)

Gambar 3. Distribusi Kata sebelum (a) dan setelah (b) *pre-processing* untuk portal berita BeritaSatu.com

Pada gambar 2 dapat dilihat perbedaan distribusi kata pada portal berita RMOL.id sebelum dan setelah dilakukannya *pre-processing* data. Pada gambar 2 (a) terlihat *stopwords* memiliki jumlah kemunculan kata yang signifikan. Setelah dilakukan penghapusan *stopwords* sebagaimana terlihat pada gambar 2 (b), didapati distribusi kata lebih baik dan memunculkan kata yang lebih bersesuaian.

Pada gambar 3 dapat dilihat distribusi kata pada portal berita BeritaSatu.com sebelum dan setelah dilakukannya *pre-processing* data. *Stopwords* memiliki kemunculan yang signifikan dibanding kata-kata lainnya.

Ikon 'Biasakan yang Tidak Biasa', New Man langsung berkeliling ke beberapa lokasi keramaian di Surabaya guna menyosialisasikan protokol kesehatan demi mencegah penyebaran Covid-19.BERITA TERKAIT: Kasus Aktif Covid-19 Turun Lagi, tapi Pasien Baru Terinfeksi di Atas 200 Pasien Kasus Aktif Covid-19 Masih Naik, Jumlah Tambahan Pasien Baru di Bawah 200 Orang Kasus Aktif Covid-19 Naik 10 Orang, Pasien Baru di Atas yang Sembuh Kasus Aktif Covid-19 Hari Ini Naik 9 Orang, Pasien Baru di Bawah 300 Orang Kasus Aktif Covid-19 Naik 124, Meninggal 9 Orang Tak hanya berkeliling di Kebun Binatang Surabaya (KBS), superhero yang diperankan langsung oleh Camat Sawahan M Yunus ini juga menyosialisasikan protokol kesehatan di pusat perbelanjaan Mal TP 5 Surabaya. Di mal besar yang ada di tengah kota itu, New Man yang identik dengan kepala plontos dan perut buncit, dibalut dengan pakaian hijau sembari memakai masker ini langsung menyosialisasikan protokol kesehatan dengan pengeras suaranya. Sosok New Man yang unik itu langsung menyita perhatian para pengunjung mal. Mereka pun tampak memperhatikan sosialisasi yang disampaikan Sang New Man. Ia langsung menuju area food court TP 5. Di tempat makan itu, ia tak henti-henti menyosialisasikan para pengunjung untuk tetap menjaga protokol kesehatan. Bahkan, ia meminta mereka untuk menutup maskernya ketika sudah selesai makan. "Ayo dipakai lagi maskernya kalau sudah selesai makan, jaga jaraknya juga. Mari kita bersama-sama lawan Covid-19 ini," ujar New Man seperti diberitakan Kantor Berita RMOLJatim , Jumat (1/1). Pada kesempatan itu, Sang New Man yang sekaligus Camat Sawahanini mengatakan, ikon New Man merupakan ide Bu Risma yang saat itu meminta jajarannya di Pemkot Surabaya membuat ikon supaya masyarakat selalu ingat kepada protokol kesehatan. Ide itu kemudian diterjemahkan oleh Wakil Sekretaris IV Satgas Percepatan Penanganan Covid-19 Surabaya yang sekaligus Kepala BPB Linmas, Irvan Widyanto, dengan membuat ikon New Man. "Kebetulan yang dipilih oleh Pak Irvan saya, karena perut saya buncit dan kepala saya botak. Mungkin ini unik, sehingga saya wakafkan diri saya ini untuk terus menyosialisasikan protokol kesehatan kepada masyarakat, supaya Covid-19 ini cepat selesai di Surabaya," jelasnya. Oleh karena itu, ia juga mengajak kepada warga Kota Surabaya untuk terus disiplin menjaga protokol kesehatan. Apalagi, saat ini sudah ada Perwali 67 yang memberlakukan denda bagi pelanggar protokol kesehatan sebesar Rp 150 ribu. "Jadi, bagi pengusaha dan masyarakat sebagai individu akan didenda Rp 150 ribu jika melanggar protokol kesehatan, dan itu tidak perlu sidang. Makanya ayo kita sama-sama disiplin," ujarnya. Sementara itu, Wakil Sekretaris IV Satgas Percepatan Penanganan Covid-19 Surabaya yang sekaligus Kepala BPB Linmas, Irvan Widyanto, mengaku sengaja beraksi bersama New Man ke KBS dan mal dalam momen tahun baru. "Rencananya nanti kami juga akan sasar pasar-pasar tradisional dan tempat kerumunan lainnya," tegas Irvan yang sejak awal mendampingi Sang New Man. BACA JUGA: Kasus Aktif Covid-19 Turun Lagi, Hari Ini Sebanyak 88 Orang Pasien Baru dan Sembuh di Bawah 200, Kasus Aktif Covid-19 Turun 158 Orang Mantan Kasatpol PP Surabaya ini juga memastikan bahwa selain sosialisasi protokol kesehatan, ia bersama jajaran Linmas juga membagi-bagikan sekitar 1.000 masker kepada para pengunjung KBS dan mal TP 5 Surabaya. "Menurut kami, vaksin terbaik adalah perubahan perilaku dengan biasakan yang tidak biasa, dengan cara itu, insyallah Covid-19 di Surabaya akan segera selesai," pungkasnya. EDITOR:

Gambar 4. Contoh data sebelum *pre-processing*

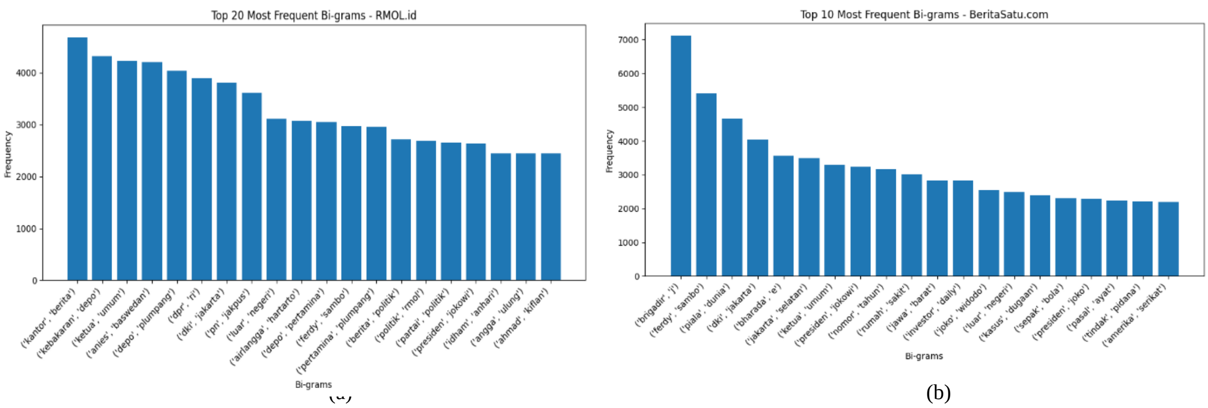
Implementasi penghapusan *stopwords* menjadi distribusi kata pada dataset portal berita BeritaSatu.com menjadi lebih baik. Kata-kata yang bersesuaian menjadi lebih terlihat jelas dibandingkan sebelumnya. Berdasarkan visualisasi data pada masing-masing portal berita tersebut diketahui bahwa jumlah kata / karakter yang muncul pada dataset didominasi oleh kata / karakter yang tidak memiliki makna yang signifikan. Sehingga, proses pre-processing, khususnya penghilangan *stopwords* merupakan tahapan penting yang harus dilakukan [12][13]. Contoh data teks sebelum dilakukan *pre-processing* dapat dilihat pada gambar 4. Sedangkan contoh data setelah *pre-processing*, dapat dilihat pada gambar 5 dibawah ini.

ikon biasakan biasa new man berkeliling lokasi keramaian surabaya guna menyosialisasikan protokol kesehatan demi mencegah penyebaran covidberita kasus aktif
aktif covid naik jumlah tambahan pasien bawah kasus aktif covid naik pasien sembuh kasus aktif covid hari naik pasien bawah kasus aktif covid naik meninggal
superhero diperankan camat sawahan m yunus menyosialisasikan protokol kesehatan pusat perbelanjaan mal tp surabaya mal besar kota new man identik kepala plc
sembari memakai masker menyosialisasikan protokol kesehatan pengeras suaranya sosok new man unik menyita perhatian para pengunjung mal tampak memperhatikan
area food court tp tempat makan hentihenti menyosialisasikan para pengunjung menjaga protokol kesehatan menutup maskernya selesai makan ayo dipakai maskerny
lawan covid new man diberitakan kantor berita rmoljatim jumat kesempatan sang new man camat sawahanini ikon new man ide bu risma jajarannya pemkot surabaya
kesehatan ide diterjemahkan wakil sekretaris iv satgas percepatan penanganan covid surabaya kepala bpb linmas irvan widyanto ikon new man kebetulan dipilih
wakafkan diri terus menyosialisasikan protokol kesehatan masyarakat supaya covid cepat selesai surabaya jelasnya mengajak warga kota surabaya terus disiplir
memberlakukan denda pelanggar protokol kesehatan sebesar ribu pengusaha masyarakat individu didenda ribu melanggar protokol kesehatan sidang makanya ayo sam
satgas percepatan penanganan covid surabaya kepala bpb linmas irvan widyanto mengaku sengaja beraksi bersama new man kbs mal momen tahun rencananya sasar pa
lainnya tegas irvan sejak awal mendampingi sang new man baca kasus aktif covid turun hari sebanyak pasien sembuh bawah kasus aktif covid turun mantan kasatp
protokol kesehatan bersama jajaran linmas membagibagikan sekitar masker para pengunjung kbs mal tp surabaya vaksin terbaik perubahan perilaku biasakan biasa
pungkasnya

Gambar 5. Contoh data setelah *pre-processing*

2.3. n-Gram

Dalam Pemrosesan Bahasa Alami (NLP), konsep "n-gram" digunakan untuk merujuk ke urutan n item yang berdekatan dalam sampel teks atau ucapan yang diberikan. Item-item ini dapat berupa kata, karakter, atau subkata, tergantung pada tingkat perincian yang diperlukan untuk analisis. "n" dalam n-gram menunjukkan jumlah item dalam urutan. N-Gram dapat digunakan untuk membantu klasifikasi teks, sebagaimana yang dilakukan oleh [14] dalam mengklasifikasi sentimen terhadap pembelian makanan ringan di Indonesia. [15] Membandingkan penggunaan bi-gram dan penghapusan *stopwords* pada klasifikasi menggunakan Naïve Bayes. N-Gram dapat pula digunakan untuk membantu pemodelan teks. Sebagaimana yang dilakukan oleh [16] dalam memodelkan topik pada data teks untuk mengetahui diskusi terkait Covid-19. Contoh n-gram pada dataset RMOL.id dan BeritaSatu.com dapat dilihat pada gambar 6 dibawah ini.



(a) (b)

Gambar 6. Contoh n-gram pada (a) RMOL.id dan (b) BeritaSatu.com

2.4. *Coherence Value*

Coherence value digunakan untuk mengevaluasi topik model. Ada 4 (empatt) tahapan dalam *coherence value*, yaitu *segmentation* yang merupakan pengelompokkan data menjadi pasangan kata. *Probability estimation* menghitung kemungkinan kata atau pasangan kata tersebut. C*onfirmation measure* menunjukkan seberapa kuat satu kumpulan kata mendukung yang lain, dan *aggregation* untuk melihat skor koherensi keseluruhan [17]. Dalam penentuan jumlah topik maka perlu dilihat *coherence value* dengan nilai terbaik.[18]

$$C = S \times M \times P \times \Sigma \tag{1}$$

Dimana:

C = Coherence Value
S = Segmentation
M = Confirmation Measure
P = Probaility Estimation
Σ = Agrregation

2.5. *Non-negative Matrix Factorization (NMF)*

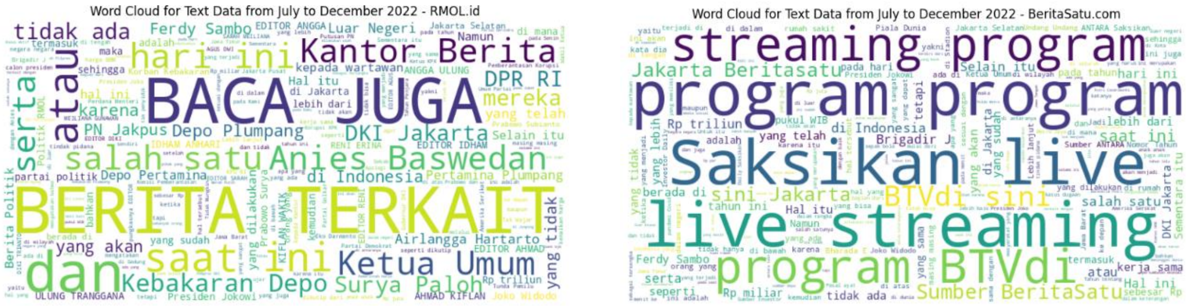
Faktorisasi Matriks Non-Negatif atau *Non-negative Matrix Factorization* (NMF) adalah teknik faktorisasi matriks yang dapat digunakan dalam pembelajaran mesin dan analisis data. NMF menguraikan matriks non-negatif menjadi dua matriks non-negatif, di mana produk matriks ini mendekati matriks aslinya. Semua elemen dalam matriks dan perkiraan yang dihasilkannya adalah non-negatif. NMF dapat digunakan dalam pemodelan topik dan pemrosesan bahasa alami (NLP) [19][20].

3. Hasil dan Pembahasan

Berdasarkan dataset dan algoritma yang digunakan, diperoleh hasil sebagaimana penjelasan berikut ini:

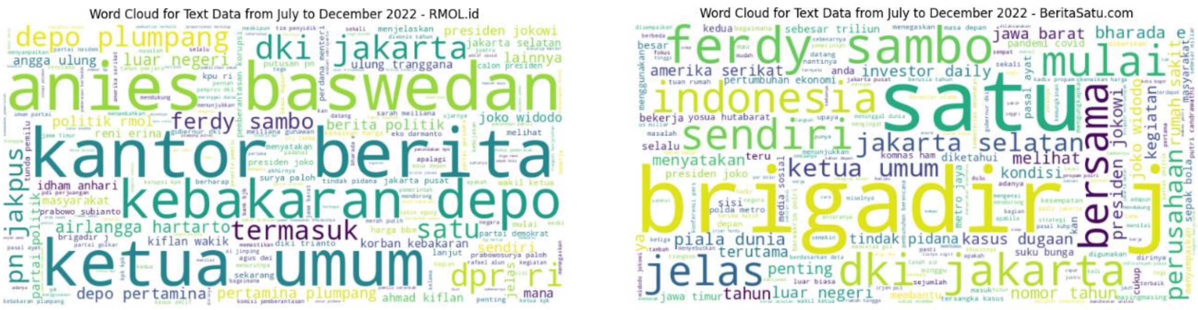
3.1. Visualisasi *Wordcloud*

Dalam penelitian ini diperoleh bahwa penghapusan *stopwords* memiliki pengaruh yang sangat signifikan dalam penentuan jumlah topik yang akan digunakan nantinya. Pada gambar 4 berikut dapat dilihat perbedaan kemunculan kata jika tidak mengimplementasikan penghapusan *stopword.*



Gambar 4. Visualisasi wordcloud sebelum penghapusan stopword pada (a) RMOL dan (b) BeritaSatu

Sedangkan pada gambar 5 dapat dilihat perbedaan kemunculan kata jika dimplementasikan penghapusan *stopwords*.



(a) (b)

Gambar 5. Visualisasi wordcloud setelah penghapusan stopword pada (a) RMOL dan (b) BeritaSatu

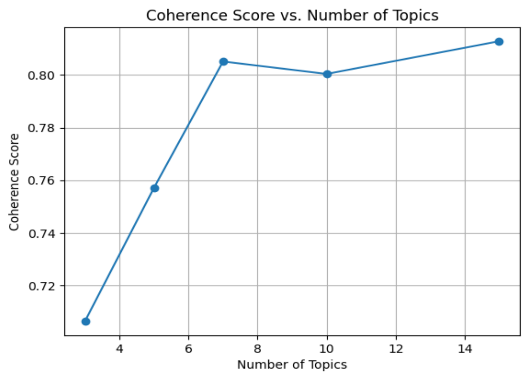
3.2. Coherence Value

Pada penelitian ini dilakukan perhitungan *coherence score* pada masing-masing data portal RMOL.id dan BeritaSatu.com. Pengujian juga dilakukan untuk n-gram kata yaitu unigram, bi-grams dan tri-grams. Hasil *coherence score* untuk masing-masing data dapat dilihat pada tabel 2 berikut ini.

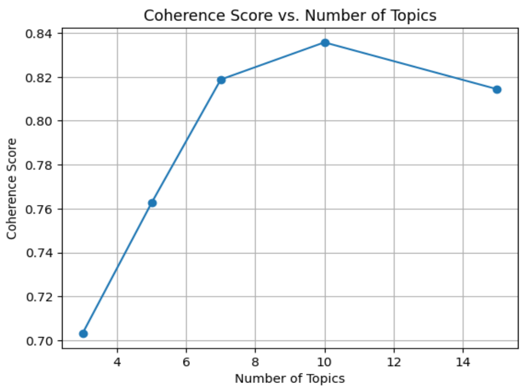
Tabel 2. Coherence Score dan jumlah topic untuk n-gram pada RMOL.id

Portal	N-Gram	Num of Topic	Coherence score (c_v)
RMOL.id	Unigram	3	0.706528
		5	0.757163
		7	0.805067
		10	0.800350
		15	0.**812748**
	Bigram	3	0.703162
		5	0.762690
		7	0.818811
		10	0.**835738**
		15	0.814412
	Trigram	3	0.690523
		5	0.783207
		7	0.**830572**
		10	0.818390
		15	0.825117

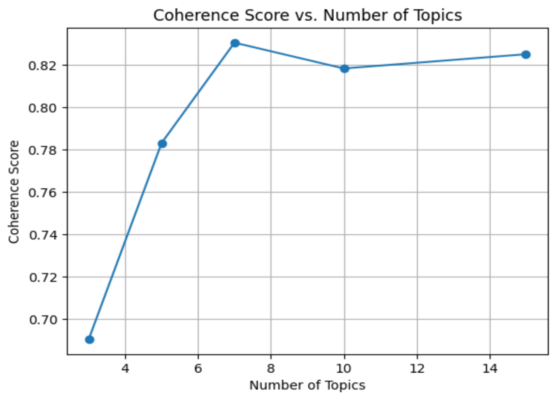
Untuk unigram diperoleh jumlah topik yang dapat dibentuk adalah 15 dengan nilai *coherence value* adalah 0.812748. Pada bigram diperoleh nilai coherence value terbaik adalah 0.835738 dengan jumlah topik 10. Sedangkan pada trigram diperoleh nilai *coherence value* terbaik adalah 0.830572 dengan jumlah topik 7. Visualisasi untuk masing-masing n-gram, *coherence value* dan jumlah topik dapat dilihat pada gambar 6 (a), 6 (b) dan 6 (c) berikut.



(a)



(b)



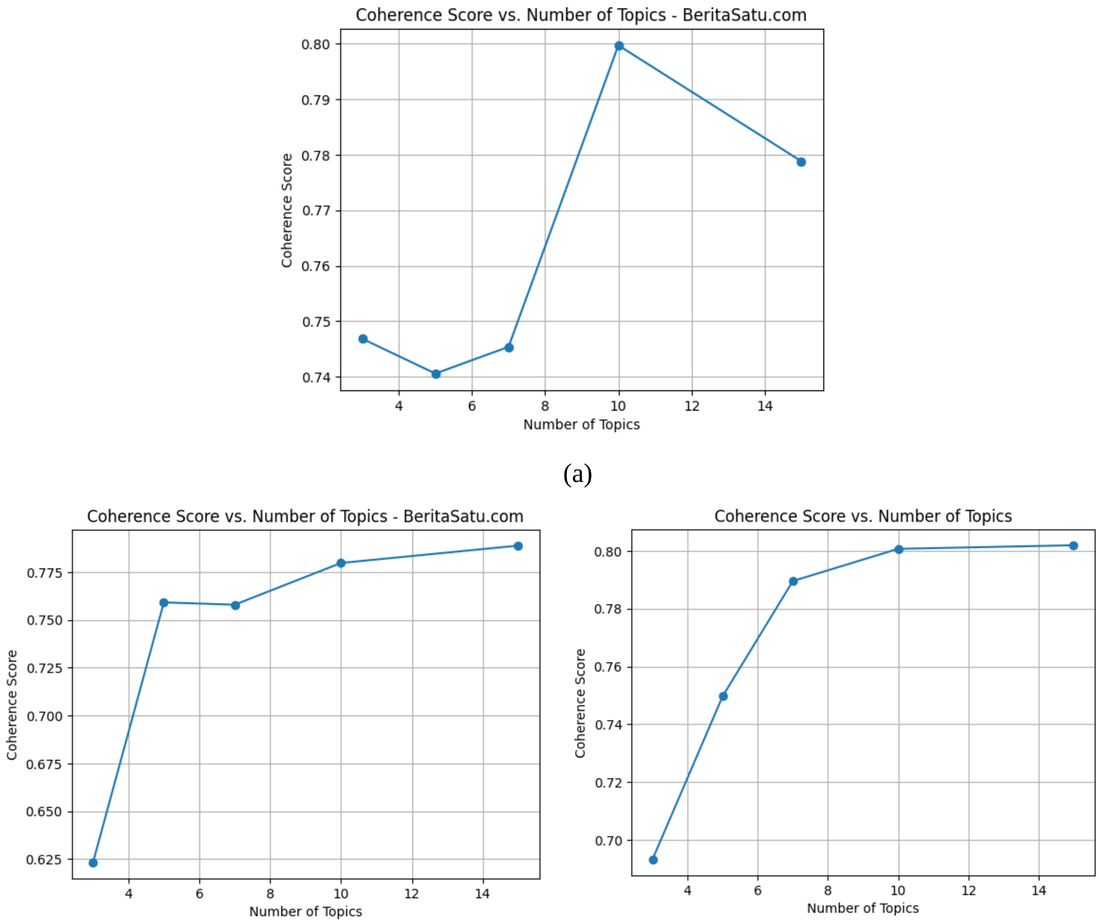
(c)

Gambar 6. Visualisasi Coherence Score dan Jumlah Topik untuk (a) Unigram, (b) bigrams dan (c) trigrams pada RMOL.id

Tabel 3. Coherence Score dan jumlah topic untuk n-gram pada BeritaSatu.com

Portal	N-Gram	Num of Topic	Coherence score (c_v)
BeritaSatu.com	Unigram	3	0.746840
		5	0.740562
		7	0.745374
		10	0.**799718**
		15	0.778819
	Bigram	3	0.623224
		5	0.759108
		7	0.757886
		10	0.779700
		15	0.**788762**
	Trigram	3	0.693200
		5	0.749699
		7	0.789545
		10	0.800696
		15	0.**801935**

Pada data portal BeritaSatu.com untuk unigram diperoleh jumlah topik yang dapat dibentuk adalah 10 dengan nilai *coherence value* adalah 0.799718. Pada bigram diperoleh nilai coherence value terbaik adalah 0.788762 dengan jumlah topik 15. Sedangkan pada trigram diperoleh nilai *coherence value* terbaik adalah 0.801935 dengan jumlah topik 15. Visualisasi untuk masing-masing n-gram, *coherence value* dan jumlah 270opic dapat dilihat pada gambar 6 (a), 6 (b) dan 6 (c) berikut.



(a)

(b) (c)

Gambar 6. Visualisasi Coherence Score dan Jumlah Topik untuk (a) Unigram, (b) bigrams dan (c) trigrams pada BeritaSatu.com

3.3. Distribusi Kata per Topik

Kata-kata yang membentuk topik untuk masing-masing portal RMOL.id dan BeritaSatu.com dapat dilihat pada tabel – tabel berikut ini. Distribusi kata-kata ini akan memberikan gambaran topik yang muncul pada masing-masing portal berita.

Pada tabel 4 dapat dilihat kata-kata yang saling terkait untuk masing-masing topik pada unigram untuk data RMOL.id

Tabel 4. Keywords dan Topik untuk unigram data pada RMOL.id

No Topik	Keyword Topik
1	jokowi, presiden, erick, depo, plumpang, ahok, thohir, bumn, indonesia, menteri
2	kpk, korupsi, firli, pemberantasan, cukai, bea, darmanto, eko, pejabat, perkara
3	china, rusia, as, ukraina, amerika, taiwan, negara, beijing, laut, olimpiade
4	kasus, covid, sebanyak, positif, pasien, aktif, provinsi, omicron, hari, total
5	ferdinand, hutahaean, gerindra, haris, cuitan, Allahmu, prabowo, polri, agama, knpi
6	afghanistan, taliban, perempuan, iran, italia, pbb, warga, unama, isis, berduka
7	airlangga, hartarto, golkar, survei, persen, harga, firli, goreng, minyak, partai
8	kazakhstan, rusia, tokayev, putin, kabur, kerusuhan, warga, protes, gas, csto
9	polri, bahar, habib, kebakaran, polda, sigit, plumpang, kapolri, pasal, korban
10	dpr, pemilu, ikn, ruu, kpu, ri, uu, pemerintah, tpks, komisi
11	vaksinasi, vaksin, booster, dosis, covid, omicron, varian, kesehatan, pemerintah, masyarakat
12	bekasi, selaku, rahmat, effendi, pepen, kpk, walikota, pemkot, rp, tersangka
13	ganjar, pdip, pan, megawati, puan, capres, maharani, pranowo, hasto, dukung
14	arteria, dahlan, sunda, bahasa, pdip, pdi, jabar, maaf, kajati, jawa
15	anies, demokrat, partai, jakarta, dki, baswedan, ketua, ahy, prabowo, pks

Pada tabel 5 dapat dilihat kata-kata yang saling terkait untuk masing-masing topik pada bigram untuk data RMOL.id

Tabel 5. Keywords dan Topik untuk bigram data pada RMOL.id

No Topik	Keyword Topik
1	anies, pemilu, partai, demokrat, dpr, jakarta, ketua, ikn, indonesia, pks
2	kpk, korupsi, bekasi, firli, pemberantasan, selaku, pemberantasan korupsi, uang, bea cukai, cukai
3	china, rusia, as, ukraina, amerika, taiwan, as china, negara, beijing, laut
4	ferdinand, ferdinand hutahaean, hutahaean, gerindra, polri, bahar, haris, smith, cuitan, bahar bin
5	kasus, covid, sebanyak, positif, kasus positif, kasus aktif, pasien, omicron, aktif, aktif covid
6	afghanistan, taliban, perempuan, iran, warga afghanistan, perempuan afghanistan, italia, taliban berkuasa, afghanistan kehilangan, warga
7	jokowi, plumpang, presiden, depo, depo plumpang, erick, erick thohir, thohir, bumn, presiden jokowi
8	airlangga, ganjar, airlangga hartarto, hartarto, golkar, survei, partai, pan, pdip, partai golkar
9	kazakhstan, rusia, warga rusia, tokayev, rusia kabur, kabur, kerusuhan, putin, ribu warga, warga
10	arteria, arteria dahlan, dahlan, sunda, pdip, bahasa, pdi, jabar, bahasa sunda, maaf

Pada tabel 6 dapat dilihat kata-kata yang saling terkait untuk masing-masing topik pada trigram untuk data RMOL.id

Tabel 6. Keywords dan Topik untuk trigram data pada RMOL.id

No Topik	Keyword Topik
1	jokowi, presiden, indonesia, partai, ganjar, plumpang, airlangga, depo, pdip, depo plumpang
2	kpk, korupsi, firli, bekasi, pemberantasan, selaku, pemberantasan korupsi, bea cukai, cukai, bea
3	china, rusia, as, ukraina, amerika, taiwan, as china, negara, beijing, qin
4	afghanistan, taliban, perempuan, warga afghanistan, iran, perempuan afghanistan, italia, taliban berkuasa, afghanistan kehilangan, warga
5	kasus, covid, sebanyak, positif, kasus positif, pasien, kasus aktif, omicron, aktif, aktif covid
6	ferdinand, hutahaean, ferdinand hutahaean, gerindra, polri, bahar, smith, haris, bahar bin, bahar bin smith
7	kazakhstan, warga rusia, rusia, tokayev, ribu warga rusia, warga rusia kabur, rusia kabur, kabur, kerusuhan, ribu warga

Pada tabel 7 dapat dilihat kata-kata yang saling terkait untuk masing-masing topik pada unigram untuk data BeritaSatu.com

Tabel 7. Keywords dan Topik untuk unigram data pada BeritaSatu.com

No Topik	Keyword Topik
1	indonesia, presiden, tahun, negara, pemerintah, ekonomi, jokowi, masyarakat, kerja, menteri
2	infografik, januari, covid, kematian, provinsi, positif, tingkat, tertinggi, data, kasus
3	kpk, bekasi, korupsi, ott, ali, effendi, gafur, rahmat, suap, tersangka
4	saham, turun, indeks, naik, melemah, ihsg, as, harga, menguat, perdagangan
5	kasus, omicron, covid, varian, positif, nadia, kesehatan, aktif, hari, transmisi
6	vaksinasi, vaksin, dosis, covid, juta, kesehatan, sasaran, lansia, ketiga, cakupan
7	hujan, berawan, wilayah, hari, cuaca, barat, angin, bmkg, jakarta, derajat
8	pemain, gol, pertandingan, menit, tim, laga, poin, piala, kemenangan, liga
9	pasien, sebanyak, jumlah, wisma, atlet, dirawat, rsdc, tidur, sakit, kemayoran
10	metro, polda, tersangka, pasal, korban, zulpan, jaya, polri, polisi, ayat

Pada tabel 8 dapat dilihat kata-kata yang saling terkait untuk masing-masing topik pada bigram untuk data BeritaSatu.com

Tabel 8. Keywords dan Topik untuk bigram data pada BeritaSatu.com

No Topik	Keyword Topik
1	indonesia, tahun, ekonomi, negara, digital, masyarakat, pemerintah, presiden, kerja, perusahaan
2	kasus, omicron, covid, varian, varian omicron, positif, kasus aktif, kesehatan, nadia, hari
3	kpk, bekasi, korupsi, ott, effendi, rahmat, kota bekasi, gafur, rahmat effendi, abdul gafur
4	infografik, covid januari, infografik kasus, kematian covid, januari, positif kematian, kasus positif, kematian, covid, positif
5	saham, indeks, turun, naik, ihsg, triliun, miliar, investor, melemah, menguat
6	berawan, hujan, wilayah, hujan ringan, hari, wilayah jakarta, jakarta, ringan, cuaca, derajat
7	vaksinasi, vaksin, dosis, covid, vaksinasi dosis, juta, kesehatan, sasaran, lansia, ketiga
8	pasien, sebanyak, wisma, jumlah, jumlah pasien, atlet, rsdc, wisma atlet, dirawat, tempat tidur
9	gol, pemain, pertandingan, menit, tim, laga, poin, piala, kemenangan, babak
10	provinsi januari, tingkat provinsi, infografik, tertinggi covid, covid tingkat, provinsi, tingkat, tertinggi, januari, infografik kematian
11	metro, polda, tersangka, pasal, metro jaya, zulpan, polda metro, jaya, korban, polisi
12	dolar, per dolar, dolar as, as, rupiah, melemah, per, menguat, poin mencapai, poin
13	perairan, laut, barat, angin, utara, kecepatan angin, kalimantan, selatan, sulawesi, kecepatan
14	minyak, goreng, minyak goreng, harga, liter, harga minyak, per, pasar, barel, per liter
15	ruu, ikn, seksual, dpr, tpks, kekerasan seksual, kekerasan, ruu tpks, presiden, rapat

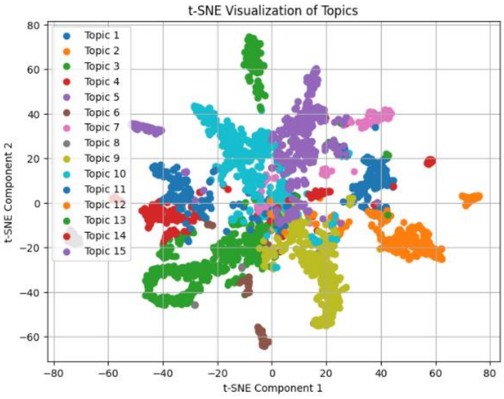
Pada tabel 9 dapat dilihat kata-kata yang saling terkait untuk masing-masing topik pada trigram untuk data BeritaSatu.com

Tabel 9. Keywords dan Topik untuk trigram data pada BeritaSatu.com

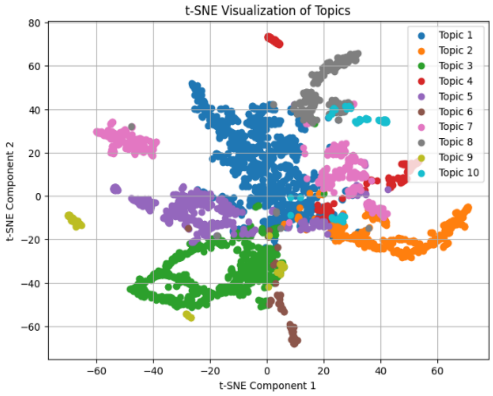
No Topik	Keyword Topik
1	indonesia, tahun, ekonomi, digital, negara, masyarakat, pemerintah, perusahaan, kerja, presiden
2	kasus, omicron, covid, varian, varian omicron, positif, kesehatan, kasus aktif, hari, kasus covid
3	kpk, bekasi, korupsi, effendi, ott, rahmat, kota bekasi, rahmat effendi, tersangka, gafur
4	infografik, covid januari, infografik kasus, infografik kasus positif, kematian covid, kasus positif kematian, positif kematian covid, positif kematian, kematian covid januari, januari
5	saham, indeks, turun, naik, ihsg, triliun, miliar, investor, melemah, menguat
6	berawan, hujan, wilayah, hujan ringan, hari, wilayah jakarta, jakarta, ringan, cuaca, derajat
7	pasien, sebanyak, wisma, jumlah pasien, jumlah, atlet, rsdc, wisma atlet, dirawat, tempat tidur
8	pemain, gol, pertandingan, menit, tim, poin, laga, kemenangan, piala, babak
9	vaksin, vaksinasi, dosis, covid, vaksinasi dosis, kesehatan, juta, lansia, sasaran, vaksin covid
10	provinsi januari, tingkat provinsi januari, tingkat provinsi, infografik, tertinggi covid tingkat, covid tingkat provinsi, tertinggi covid, covid tingkat, provinsi, tingkat
11	dolar, per dolar as, per dolar, dolar as, as, rupiah, melemah, per, menguat, poin mencapai
12	metro, polda, tersangka, metro jaya, polda metro, jaya, pasal, zulpan, polda metro jaya, korban
13	perairan, laut, barat, angin, utara, kecepatan angin, kalimantan, selatan, sulawesi, kecepatan
14	minyak, goreng, minyak goreng, harga, liter, harga minyak, per, pasar, barel, harga minyak goreng
15	ruu, ikn, dpr, seksual, tpks, kekerasan seksual, kekerasan, ruu tpks, presiden, pemilu

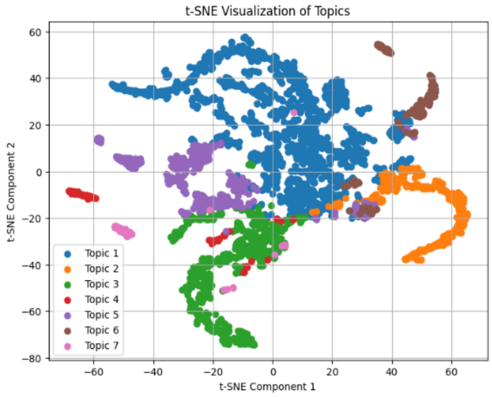
3.4. Visualisai *t-SNE*

Pada gambar 7 dapat dilihat visualisasi topik menggunakan t-SNE untuk unigram, bigram dan trigram pada data RMOL.id. Berdasarkan visualisasi tersebut dapat dilihat kedekatan antar topik yang terbentuk.



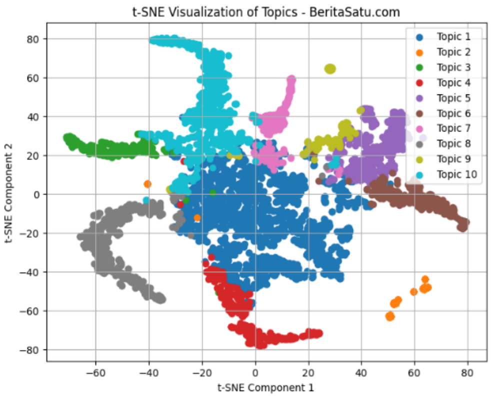
(a)



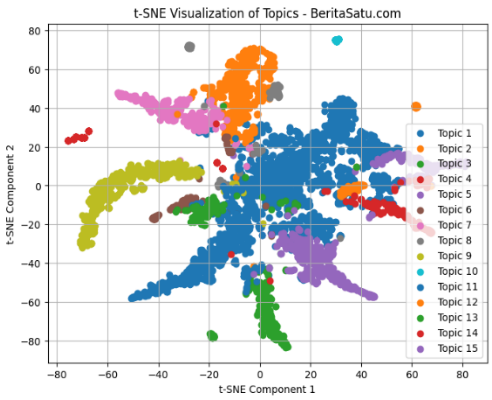


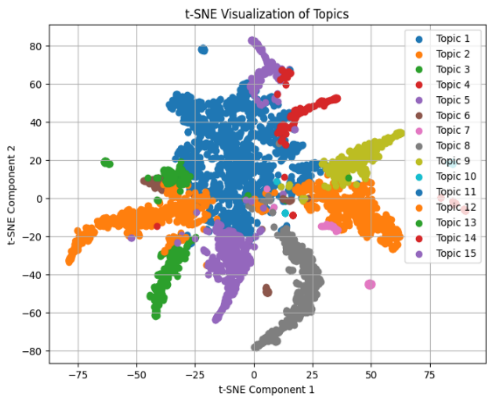
(b) (c)

Gambar 7. Visualisasi t-SNE jumlah topik pada RMOL.id untuk (a) unigram (b) bigram (c) trigram



(a)





(b) (c)

Gambar 8. Visualisasi t-SNE jumlah topik pada BeritaSatu.com untuk (a) unigram (b) bigram (c) trigram

Berdasarkan visualisasi pada gambar 7 diatas dapat dilihat bahwa pembentukan topik dengan menggunakan unigram tidak menghasilkan pembentukan topik yang baik. Ini dapat dilihat dari tidak terpisahnya dengan baik antar topik, seperti contoh yang terlihat pada topik 1, 3 dan 4. Untuk itu perlu dilakukan evaluasi terhadap topik yang terbentuk. Pada pembetukan topik menggunakan bigram, dapat dilihat pembentukan topik lebih baik dibandingkan pada unigram. Sedangkan pada trigram dapat dilihat dengan jelas pengelompokkan kata sehingga terlihat pemisahan topik dengan baik.
Pada gambar 8 dapat dilihat visualisasi topik menggunakan t-SNE untuk unigram, bigram dan trigram pada data BeritaSatu.com. Berdasarkan visualisasi gambar 8 diatas dapat dilihat bahwa terdapat pencilan kata-kata yang membentuk topik. Hal ini disebabkan diantaranya masih terdapat *stopword* yang belum dibersihkan dengan baik pada data. Untuk itu, perlu adanya evaluasi untuk setiap topik yang terbentuk dengan memastikan antar topik terpisah dengan baik.

4. Kesimpulan

Pada penelitian ini dapat dilihat bahwa penggunaan n-gram dalam membentuk topik dengan menggunakan *Non-negative Matrix Factorization* (NMF) dapat memberikan hasil yang baik. Hal ini harus didukung dengan proses *preprocessing* yang baik. Salah satu proses *preprocessing* yang sangat penting adalah penghilangan *stopwords.* Berdasarkan visualisasi topik menggunakan t-SNE dapat dilihat masih adanya sebaran kata pembentuk topik yang menjadi pencilan pada topik tersebut.

Daftar Rujukan

[1] S. Mutrofin, T. Wicaksono, and A. Murtadho, “Perbandingan Kinerja Algoritma Kmeans dengan Kmeans Median pada Deteksi Kanker Payudara,” *J. Inf. dan Teknol.*, vol. 5, no. 1, pp. 88–91, 2023, doi: 10.37034/jidt.v5i1.274.

[2] R. Aryanto, M. A. Rosid, and S. Busono, “Penerapan Deep Learning untuk Pengenalan Tulisan Tangan Bahasa Akasara Lota,” *J. Inf. dan Teknol.*, vol. 5, no. 1, pp. 258–264, 2023, doi: 10.37034/jidt.v5i1.313.

[3] E. S. Negara, D. Triadi, and R. Andryani, “Topic Modelling Twitter Data with Latent Dirichlet Allocation Method,” *ICECOS 2019 - 3rd Int. Conf. Electr. Eng. Comput. Sci. Proceeding*, pp. 386–390, 2019, doi: 10.1109/ICECOS47637.2019.8984523.

[4] E. Laoh, I. Surjandari, and L. R. Febirautami, “Indonesians’ Song Lyrics Topic Modelling Using Latent Dirichlet Allocation,” *Proc. - 2018 5th Int. Conf. Inf. Sci. Control Eng. ICISCE 2018*, pp. 270–274, 2019, doi: 10.1109/ICISCE.2018.00064.

[5] M. Savira and D. Abdullah, “Prototipe Aplikasi Pengukuran Efesiensi Produksi Air Mineral Dengan Metode DEA di Wilayah Aceh Utara Dan Kota Lhokseumawe,” *Ind. Eng. J.*, vol. 8, no. 2, Oct. 2019.

[6] I. M. Ahmad Niam, B. Irawan, C. Setianingsih, and B. P. Putra, “Hate Speech Detection Using Latent Semantic Analysis (LSA) Method Based on Image,” *Proc. - 2018 Int. Conf. Control. Electron. Renew. Energy Commun. ICCEREC 2018*, pp. 166–171, 2018, doi: 10.1109/ICCEREC.2018.8712111.

[7] K. Merchant and Y. Pande, “NLP Based Latent Semantic Analysis for Legal Text Summarization,” *2018 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2018*, pp. 1803–1807, 2018, doi: 10.1109/ICACCI.2018.8554831.

[8] S. Oktarian, S. Defit, and Sumijan, “Clustering Students’ Interest Determination in School Selection Using the K-Means Clustering Algorithm Method,” *J. Inf. dan Teknol.*, vol. 2, pp. 68–75, 2020, doi: 10.37034/jidt.v2i3.65.

[9] Zoya, S. Latif, F. Shafait, and R. Latif, “Analyzing LDA and NMF Topic Models for Urdu Tweets via Automatic Labeling,” *IEEE Access*, vol. 9, pp. 127531–127547, 2021, doi: 10.1109/ACCESS.2021.3112620.

[10] Z. Soares Lopes, F. Kurniawan, and J. Tistogondo, “Case Study of Public-Private Partnership on Infrastruc-ture Projects of Tibar Bay Port in Timor-Leste,” *Int. J. Eng. Sci. Inf. Technol.*, vol. 1, no. 3, 2021, doi: 10.52088/ijesty.v1i3.79.

[11] W. Febriani, G. W. Nurcahyo, and S. Sumijan, “Diagnosa Penyakit Rubella Menggunakan Metode Fuzzy Tsukamoto,” *J. Sistim Inf. dan Teknol.*, 2019, doi: 10.35134/jsisfotek.v1i3.4.

[12] A. Schofield, M. Magnusson, and D. Mimno, “Pulling out the stops: Rethinking stopword removal for topic models,” *15th Conf. Eur. Chapter Assoc. Comput. Linguist. EACL 2017 - Proc. Conf.*, vol. 2, pp. 432–436, 2017, doi: 10.18653/v1/e17-2069.

[13] A. W. Pradana and M. Hayaty, “The Effect of Stemming and Removal of Stopwords on the Accuracy of Sentiment Analysis on Indonesian-language Texts,” *Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control*, vol. 4, no. 3, pp. 375–380, 2019, doi: 10.22219/kinetik.v4i4.912.

[14] M. E. Purbaya, D. P. Rakhmadani, Maliana Puspa Arum, and Luthfi Zian Nasifah, “Implementation of n-gram Methodology to Analyze Sentiment Reviews for Indonesian Chips Purchases in Shopee E-Marketplace,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 7, no. 3, pp. 609–617, 2023, doi: 10.29207/resti.v7i3.4726.

[15] M. Hidayat, R. Hidayat, and D. Otik Kurniawati, “Comparison of the Use of Bigrams and Stopword Removal for Classification Using Naive Bayes (Case Study on Sentiment Analysis of By.U Internet Users),” *Proc. - 2021 Int. Conf. Softw. Eng. Comput. Syst. 4th Int. Conf. Comput. Sci. Inf. Manag. ICSECS-ICOCSIM 2021*, pp. 447–452, 2021, doi: 10.1109/ICSECS52883.2021.00088.

[16] R. P. Fauzie Afidh and Z. A. Hasibuan, “Indonesia’s News Topic Discussion about Covid-19 Outbreak using Latent Dirichlet Allocation,” *2020 5th Int. Conf. Informatics Comput. ICIC 2020*, pp. 1–6, 2020, doi: 10.1109/ICIC50835.2020.9288596.

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

[18] M. Röder, A. Both, and A. Hinneburg, “Exploring the space of topic coherence measures,” *WSDM 2015 - Proc. 8th ACM*

Int. Conf. Web Search Data Min., pp. 399–408, 2015, doi: 10.1145/2684822.2685324.

[19] A. Purpura, “Non-negative matrix factorization for topic modeling,” *CEUR Workshop Proc.*, vol. 2167, no. August, p. 102, 2018.

[20] A. Alfajri, D. Richasdy, and M. A. Bijaksana, “Topic Modelling Using Non-Negative Matrix Factorization (NMF) for Telkom University Entry Selection from Instagram Comments,” *J. Comput. Syst. Informatics*, vol. 3, no. 4, pp. 485–492, 2022, doi: 10.47065/josyc.v3i4.2212.