JNATIA Volume 1, Nomor 1, November 2022
Jurnal Nasional Teknologi Informasi dan Aplikasinya

Pemodelan Topik Teks Berita Menggunakan DistilBERT

I Made Anditya Mahesastra[a1], I Dewa Made Bayu Atmaja Darmawan[a2]

[a]Program Studi Informatika, Universitas Udayana
Jimbaran, Badung, Bali, Indonesia
[1]andimahesastra@email.com
[2]dewabayu@unud.ac.id

Abstract

Online newspapers are content that can be a source of information or entertainment for the audience. There are so many online newspapers on the internet and also from various publishers. This condition causes the available news to be very varied and with different structures. If in certain cases we want to group these online newspapers efficiently, then a technique will be needed that will be able to group these online newspapers efficiently into several groups so that the available online newspapers can be more structured to be enjoyed according to the needs and tastes of the readers. The technique that can be applied is topic modeling. In the case of modeling the topic of Indonesian online newspapers, currently LDA is one of the most widely applied algorithms. So, this study aims to determine whether the performance of using the DistilBERT model will be better or not when compared to commonly used algorithms such as LDA for topic modeling tasks in Indonesian online newspapers.

Keywords: *Natural Language Processing, Topic Modeling, DistilBERT, HDBSCAN, UMAP*

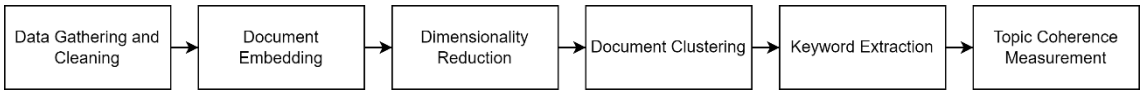
1. Introduction

Berita yang dapat dibaca secara daring menjadi salah satu konten yang menjadi sumber informasi dan hiburan bagi para penikmatnya, dengan saking banyaknya berita yang ada in internet dan dari penerbit yang beragam, tentunya akan diperlukan sebuah teknik yang akan dapat mengelompokkan berita tersebut secara efisien agar penyuguhan berita dapat menjadi lebih terstruktur untuk dapat dinikmati oleh pembaca sesuai dengan kebutuhan dan selera pembaca. Hal tersebut masuk ke dalam ranah pemrosesan bahasa alami, dalam pemrosesan bahasa alami ada yang disebut dengan pemodelan topik. Sampai saat ini, algoritma LDA masih menjadi algoritma yang paling sering digunakan dalam pemodelan topik [1], khususnya dalam kasus pemodelan topik berita berbahasa Indonesia. Saat ini pula, *Bidirectional Encoder Representation from Transformer* (BERT) menjadi *state-of-the-art language models* dikarenakan performanya yang terbukti unggul dalam berbagai kasus pemrosesan Bahasa alami [2]. Penelitian mengenai penggunaan model BERT dalam melakukan tugas pemodelan topik pada berita berbahasa Indonesia akan sangat berguna untuk mengetahui performa yang di dapat apakah dapat bersaing dengan algoritma yang lebih umum digunakan seperti LDA. Oleh karena itu, penelitian ini bertujuan untuk mencari fakta terkait hal tersebut. Hasil penelitian ini diharapkan dapat menjadi salah satu acuan pemilihan metode dalam kasus pemrosesan bahasa alami lainnya.

Ada pula beberapa penelitian lain yang meneliti kasus yang serupa dengan penelitian ini, contohnya [3], yaitu pemodelan topik pada teks berita berbahasa Indonesia yang datanya dikumpulkan mulai dari tanggal 10 sampai 23 April 2020 sebanyak 59.279 data, penelitian ini menggunakan metode LDA dengan nilai *topic coherence* sekitar 0.2. Penelitian serupa lainnya adalah pemodelan teks berita berbahasa Indonesia yang dikumpulkan mulai dari tanggal 1 Oktober 2019 hingga 31 Maret 2020 sebanyak 68.537 data, penelitian ini menggunakan metode LDA dengan nilai *topic coherence* sebesar 0.67 [4]. Penelitian serupa selanjutnya adalah analisis sentimen pada judul teks berita berbahasa Indonesia, penelitian ini menggunakan metode LDA dan LSTM dengan nilai akurasi sebesar 71.13% [5].

2. Reseach Methods

Penelitian ini diawali dengan tahap penumpulan dan pembersihan data, data dikumpulkan dengan menggunakan *web scraper* berbasis pustaka Selenium. Tahap kedua adalah membuat *document embedding* dari data yang telah tersedia untuk mendapatkan representasi dokumen dalam ruang vektor yang dilakukan menggunakan *pre-trained embedding model* yaitu DistilBERT. Tahap ketiga adalah melakukan *dimensionality reduction* yang dilakukan menggunakan UMAP untuk meningkatkan kinerja algoritma pengelompokan. Tahap keempat adalah melakukan *document clustering* untuk mengelompokkan dokumen yang serupa secara semantik yang dilakukan menggunakan HDBSCAN. Tahap kelima adalah *keyword extraction* yang dilakukan menggunakan c-TF-IDF untuk mencari representasi setiap kelompok/topik yang dihasilkan algoritma pengelompokan. Terakhir adalah tahap evaluasi metode menggunakan metrik *topic coherence* untuk menilai seberapa baik suatu topik didukung oleh korpus referensi. Gambaran alur penelitian ditunjukan pada Gambar 1.



Gambar 1. Alur Penelitian

2.1. Data Gathering

Dataset pada penelitian ini bersifat primer. *Dataset* yang digunakan adalah kumpulan teks berita berbahasa Indonesia. *Dataset* pada penelitian ini diambil dari 26 portal berita berbahasa Indonesia, di antaranya adalah Tribunnews.com, Detik.com, Kompas.com, Liputan6.com, Merdeka.com, Kapanlagi.com, Okezone.com, Tempo.co, Viva.co.id, Suara.com, JPNN.com, Sindonews.com, CNN Indonesia, CNBC Indonesia, Republika, IDN Times, TvOne News, ANTARA News, Portal Pekalongan, Tirto.ID, BBC Indonesia, Brilio.net, iNews.id, Kompasiana.com, dream.co.id, WowKeren.com. Data ini diambil dengan menggunakan *web scraper* berbasis pustaka Selenium. Proses pengambilan data dilakukan mulai dari tanggal 7 Mei 2022 sampai tanggal 18 Agustus 2022. Selama periode tersebut data yang telah terkumpul mencapai 6598 data.

2.2. Document Embedding

Document embedding diperlukan sebagai langkah untuk mendapatkan representasi dokumen dalam ruang vektor agar dokumen satu dengan dokumen lainnya dapat dibandingkan secara semantik [6], dengan asumsi dokumen yang memiliki topik yang sama juga akan serupa secara semantik [7]. Dalam melakukan *document embedding*, penelitian ini menggunakan model DistilBERT [8]. DistilBERT adalah model *embedding* yang telah dilatih sebelumnya, model tersebut digunakan untuk memberikan representasi dokumen dalam ruang vektor. Performa sangat baik di tunjukan oleh model DistilBert pada saat diuji dalam beberapa tugas seperti menjawab pertanyaan menggunakan *Stanford Question Answering Dataset* (SQuAD), dan analisis sentimen menggunakan *Internet Movie Database* (IMDb) [8]. *Document embedding* adalah langkah yang membedakan proses pemodelan topik lainnya yang menggunakan *Bag of Words* (BOW) seperti LDA atau LSA [9]. *Document embedding* kemudian digunakan untuk mengelompokan dokumen yang serupa secara semantik. Teknik *embedding* lainnya juga dapat diterapkan pada tahap ini.

2.3. Dimensionality Reduction

Dimensionality reduction diperlukan sebagai langkah untuk mengurangi dimensi dari *document embedding*. Dalam melakukan *dimensionality reduction*, penelitian ini menggunakan UMAP [10]. Penerapan UMAP dalam melakukan dimensionality reduction terbukti dapat meningkatkan kinerja algoritma pengelompokan seperti k-Means dan HDBSCAN baik dari segi akurasi maupun waktu pemrosesan [11]. UMAP dapat digunakan pada seluruh model *embedding* dengan ruang dimensi yang berbeda dikarenakan tidak memiliki batasan komputasi pada dimensi *embedding* [10].

2.4. Document Clustering

Document clustering dilakukan untuk mengelompokkan dokumen yang serupa secara semantik yang selanjutnya kumpulan dokumen yang di kelompokan akan di proses lebih lanjut untuk mencari kumpulan kata kunci yang berperan penting pada setiap kelompok dokumen. Dalam melakukan *document clustering*, penelitian ini menggunakan *Hierarchical Density Based Clustering* (HDBSCAN) [12]. HDBSCAN memodelkan kelompok menggunakan pendekatan *soft-clustering* yang memungkinkan *noise* dimodelkan sebagai *outlier* untuk mencegah dokumen yang tidak terkait dikelompokan kedalam kelompok manapun sehingga akan dapat meningkatkan representasi topik [13]. Selain itu, penerapan HDBSCAN dalam melakukan pengelompokan setelah penerapan UMAP dalam mengurangi dimensi *embedding* pada data terbukti dapat meningkatkan akurasi [11].

2.5. Keyword Extraction

Keyword extraction dilakukan untuk mencari representasi setiap kelompok/topik yang dihasilkan algoritma pengelompokan. Representasi dari suatu kelompok dimodelkan berdasarkan kumpulan dokumen penyusun kelompok tersebut. Setiap kelompok harus memiliki alasan mengapa kumpulan dokumen dapat di kelompokan ke dalam kelompok yang satu dan bukan kelompok lainnya. Dalam penelitian ini setiap kelompok akan di representasikan sebagai kumpulan kata kunci, kumpulan kata kunci tersebut di dapat dengan menggunakan algoritma *class-based* TF-IDF untuk menentukan seberapa relevan kumpulan kata tersebut dengan kelompok dokumen tertentu. Prosedur TF-IDF klasik menggabungkan dua statistik, yaitu *term frequency* (TF) dan *inverse document frequency* (IDF), dimana *term frequency* mengukur frekuensi *term* t dalam dokumen d, dan *Inverse document frequency* mengukur seberapa banyak informasi yang diberikan suatu *term* ke dokumen [14]. Dalam penelitian ini c-TF-IDF digunakan dengan maksud memperlakukan satu kelompok dokumen sebagai satu buah dokumen, di mana algoritma TF-IDF diterapkan untuk mendapatkan kumpulan kata kunci yang relevan terhadap satu kelompok dokumen.

2.6. Topic Coherence Measurement

Topic Coherence adalah metrik untuk menilai seberapa baik suatu topik (dalam hal ini adalah kumpulan kata kunci) didukung oleh kumpulan teks (korpus referensi). Perhitungan koherensi topik menggunakan statistik dan probabilitas untuk memberikan skor koherensi pada suatu topik dengan kisaran nilai 0 sampai 1, di mana nilai 0 berarti topik tidak relevan, sedangkan nilai 1 berarti sebaliknya. Nilai koherensi topik tidak hanya tergantung pada topik itu sendiri tetapi juga pada kumpulan data yang digunakan sebagai referensi (korpus referensi). *Topic coherence* adalah suatu *pipeline* yang menerima topik dan korpus referensi sebagai *input* kemudian mengeluarkan nilai koherensi dari topik tersebut sebagai *output*. Rincian dari *pipeline* tersebut di antaranya *segmentation*, *probability calculation*, *confirmation measure*, dan *aggregation*.

a. *Segmentation* berperan dalam membuat pasangan kata yang akan digunakan untuk menilai koherensi topik, berikut adalah rumus untuk mencari *segmentation*
$$S(W) = \{(W', W^*), W', W^* \in W\} \quad (1)$$
b. *Probability calculation* berperan dalam menghitung probabilitas dari tiap pasangan kata, penelitian ini menggunakan Pbd. Pbd menghitung $P(W')$ sebagai jumlah dokumen yang berisi kata W' dibagi dengan jumlah dokumen yang ada, dan $P(W'W^*)$ sebagai jumlah dokumen yang berisi kata W' dan W^* dibagi dengan jumlah dokumen yang ada
c. *Confirmation measure* berperan dalam menghitung nilai konfirmasi dengan menggunakan $P(W')$ dan $P(W'W^*)$. Terdapat beberapa rumus dalam menghitung nilai konfirmasi. Penelitian ini menggunakan rumus berikut:
$$m_c(S_i) = \frac{P(W'W^*)}{P(W^*)} \quad (2)$$
d. *Aggregation* adalah nilai rata-rata dari seluruh nilai yang didapat pada tiap *segment*

Keterangan:
$S(W)$: Himpunan pasangan kata dari himpunan kata
W : Himpunan Kata
W' : kata 1
W^* : kata 2
$m_c(S_i)$: confirmation measure dari (W', W^*)

3. Result and Discussion

3.1. Data Gathering

Dataset pada penelitian ini diambil dari 26 portal berita berbahasa Indonesia, di antaranya adalah Tribunnews.com, Detik.com, Kompas.com, Liputan6.com, Merdeka.com, Kapanlagi.com, Okezone.com, Tempo.co, Viva.co.id, Suara.com, JPNN.com, Sindonews.com, CNN Indonesia, CNBC Indonesia, Republika, IDN Times, TvOne News, ANTARA News, Portal Pekalongan, Tirto.ID, BBC Indonesia, Brilio.net, iNews.id, Kompasiana.com, dream.co.id, WowKeren.com dengan rincian yang ditunjukkan pada Tabel 1.

Tabel 1. Rincian Data

Penerbit	Jumlah Berita
CNN Indonesia	513
Sindonews.com	498
Detik.com	483
TvOne News	455
ANTARA News	430
Kompas.com	388
CNBC Indonesia	378
Republika	347
Tribunnews.com	301
Tempo.co	289
Okezone.com	243
Merdeka.com	243
JPNN.com	233
Suara.com	231
dream.co.id	182
BBC Indonesia	165
Liputan6.com	149
iNews.id	149
Tirto.ID	148
Portal Pekalongan	140
WowKeren.com	137
Viva.co.id	133
Kompasiana.com	131
Kapanlagi.com	106
IDN Times	99
Brilio.net	27
Total	6598

3.2. Document Embedding

Document embedding dilakukan dengan tujuan mendapatkan representasi dokumen dalam ruang vektor. Dalam melakukan *document embedding*, penelitian ini menggunakan model DistilBERT. Contoh penggunaan DistilBERT ditunjukan pada Tabel 2.

Tabel 2. Contoh Hasil dari Penggunaan DistilBERT dalam Melakukan *Documment Embedding*

Teks	*Embedding* dalam bentuk vektor sepanjang 768
Serangan Rusia ke Ukraina terus berlanjut dan kini Kota Mariupol mendapat gempuran bertubi-tubi.	[-2.77933143e-02 -7.33952671e-02 -9.79798436e-02 1.03335127e-01 2.59766459e-01 1.61224723e-01 -6.59341663e-02 1.98788896e-01 -4.11727317e-02 2.53799617e-01 1.94205835e-01 3.57450619e-02 -8.34005997e-02

	…]

3.3. Dimensionality Reduction

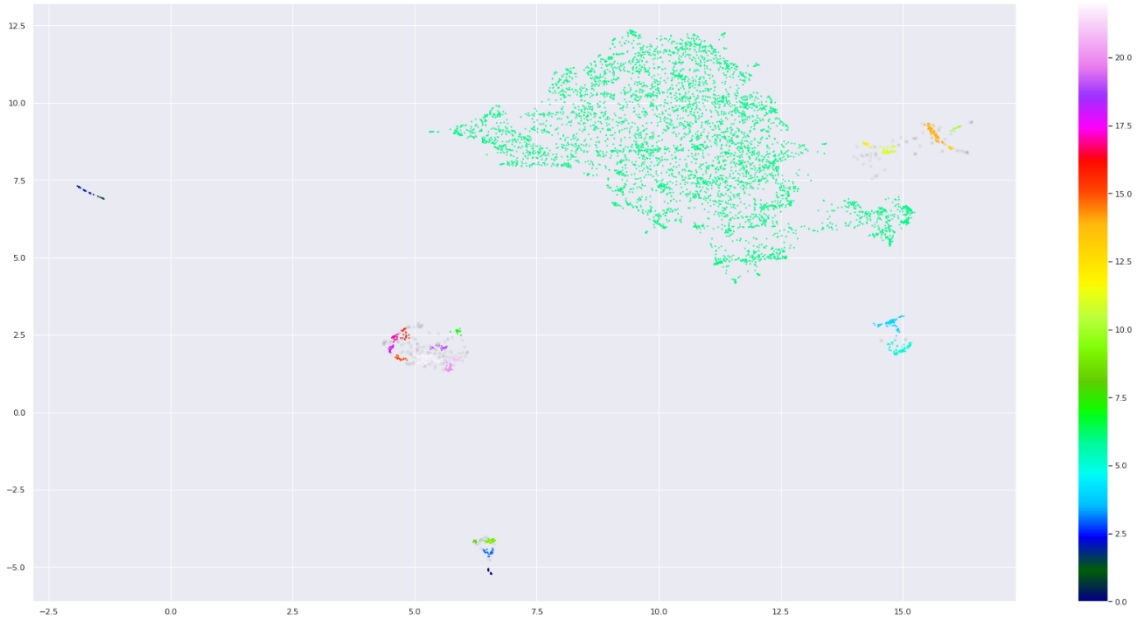
Dimensionality reduction diperlukan sebagai langkah untuk mengurangi dimensi dari *document embedding*. Dalam melakukan *dimensionality reduction*, penelitian ini menggunakan UMAP. Contoh penggunaan UMAP ditunjukan pada Tabel 3.

Tabel 3. Contoh Hasil dari Penggunaan UMAP dalam Melakukan *Dimensionality Reduction*

Embedding dalam bentuk vektor sepanjang 768	*Embedding* dalam bentuk vektor yang telah direduksi hingga sepanjang 2
[-2.77933143e-02 -7.33952671e-02 -9.79798436e-02 1.03335127e-01 2.59766459e-01 1.61224723e-01 -6.59341663e-02 1.98788896e-01 -4.11727317e-02 2.53799617e-01 1.94205835e-01 3.57450619e-02 -8.34005997e-02 …]	[15.094062 , 6.592128]

3.4. Document Clustering

Document clustering dilakukan untuk mengelompokan dokumen yang serupa secara semantik. Dalam melakukan *document clustering*, penelitian ini menggunakan *Hierarchical Density Based Clustering* (HDBSCAN). Hasil dari proses *document clustering* ditunjukan pada Gambar 2. Proses pengelompokan dokumen menghasilkan 23 Kelompok.

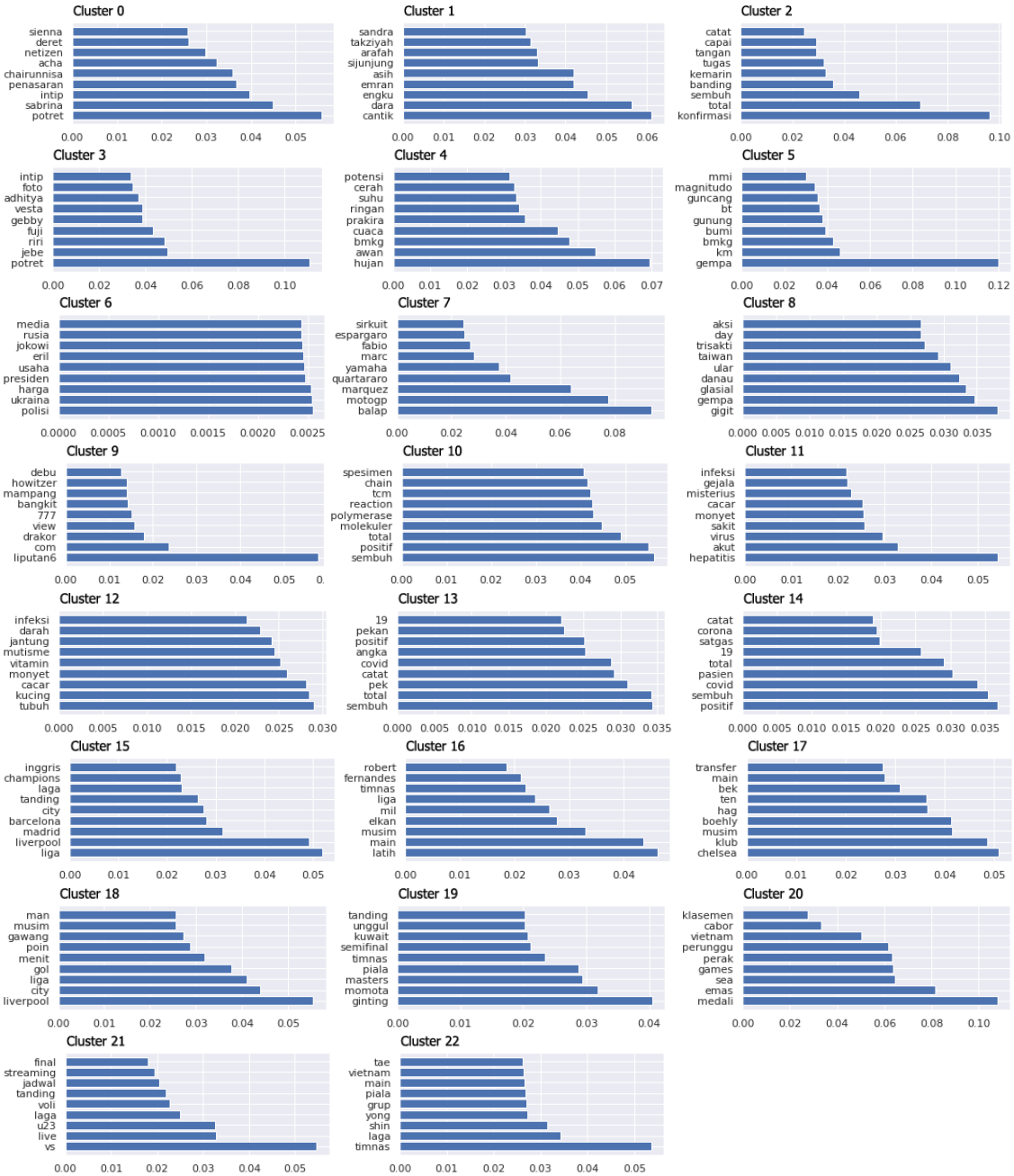


Gambar 2. Visualisasi Hasil *Document Clustering*

3.5. Keyword Extraction

Keyword extraction dilakukan untuk mencari representasi setiap klaster/topik yang dihasilkan algoritma pengelompokan. Dalam penelitian ini c-TF-IDF digunakan dengan maksud memperlakukan satu kelompok dokumen sebagai satu buah dokumen, dimana algoritma TF-IDF

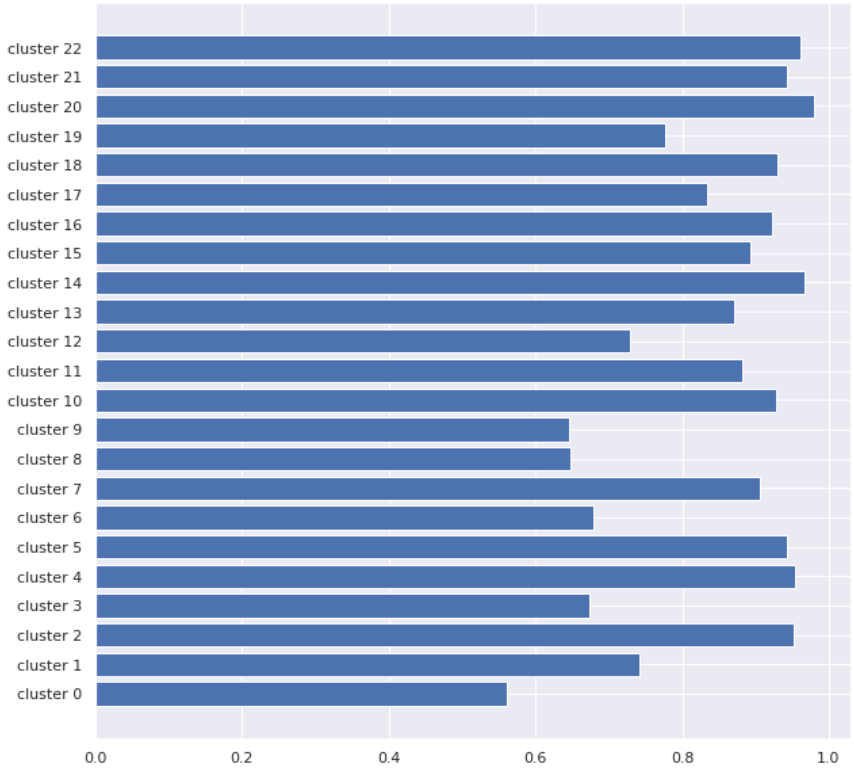
diterapkan untuk mendapatkan kumpulan kata kunci yang relevan terhadap satu kelompok dokumen. Kumpulan kata yang dihasilkan tiap klaster/topik ditunjukan pada Gambar 3.



Gambar 3. Representasi Tiap Klaster yang Dihasilkan Dari Penggunaan Metode DistilBERT

3.6. Topic Coherence Measurement

Hasil dari perhitungan *Topic Coherence* menggunakan korpus referensi data teks berita yang sama menunjukan bahwa penggunaan model *embedding* DistilBERT dapat memberikan hasil yang lebih unggul apabila dibandingkan dengan penggunaan metode yang lebih umum yaitu LDA. Hasil perhitungan *topic coherence* menggunakan DistilBERT dintunjukan pada Gambar 4.



Gambar 4. Hasil Perhitungan *Topic Coherence* pada Topik yang Dihasilkan Dari Penggunaan Metode DistilBERT

4. Conclusion

Jumlah klaster paling optimal yang didapat untuk data teks berita berbahasa Indonesia yang digunakan pada penelitian ini yaitu sejumlah 23 klaster. Hasil dari perhitungan *Topic Coherence* untuk 23 klaster yang dihasilkan dari penggunaan metode DistilBERT menggunakan korpus referensi data teks berita yang sama didapat nilai rata-rata sebesar 0.8402955343126479. Sedangkan hasil dari perhitungan *Topic Coherence* untuk klaster yang dihasilkan dari penggunaan metode LDA menggunakan korpus referensi data teks berita yang sama didapat nilai rata-rata sebesar 0.8389966817692823. Hasil tersebut menunjukan bahwa penggunaan model *embedding* DistilBERT dapat memberikan hasil yang lebih unggul apabila dibandingkan dengan penggunaan metode yang lebih umum yaitu LDA.

References

[1] C. E. Moody, “Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec,” Mei 2016, [Daring]. Available: http://arxiv.org/abs/1605.02019

[2] N. Reimers dan I. Gurevych, “Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation,” Apr 2020, [Daring]. Available: http://arxiv.org/abs/2004.09813

[3] Wahyudin, “APLIKASI TOPIC MODELING PADA PEMBERITAAN PORTAL BERITA ONLINE SELAMA MASA PSBB PERTAMA.”

[4] “Pemodelan Topik Berita pada Portal Berita Online Berbahasa Indonesia Menggunakan Latent Dirichlet Allocation (LDA),” *Jurnal Ilmiah Komputasi*, vol. 20, no. 2, Jun 2021, doi: 10.32409/jikstik.20.2.2719.

[5] C. Naury, D. H. Fudholi, dan A. F. Hidayatullah, “Topic Modelling pada Sentimen Terhadap Headline Berita Online Berbahasa Indonesia Menggunakan LDA dan LSTM,” *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 1, hlm. 24, Jan 2021, doi: 10.30865/mib.v5i1.2556.

[6] D. Jiang, Z. Chen, R. Lian, S. Bao, dan C. Li, “Familia: An Open-Source Toolkit for Industrial Topic Modeling,” Jul 2017, [Daring]. Available: http://arxiv.org/abs/1707.09823
[7] A. B. Dieng, F. J. R. Ruiz, dan D. M. Blei, “Topic Modeling in Embedding Spaces,” Jul 2019, [Daring]. Available: http://arxiv.org/abs/1907.04907
[8] V. Sanh, L. Debut, J. Chaumond, dan T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” Okt 2019, [Daring]. Available: http://arxiv.org/abs/1910.01108
[9] R. Hakim *dkk.*, “Topic Modeling Pada Abstrak Skripsi Menggunakan Metode Latent Semantic Analysis,” 2022. [Daring]. Available: http://digilib.uinsby.ac.id.
[10] L. McInnes, J. Healy, dan J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” Feb 2018, [Daring]. Available: http://arxiv.org/abs/1802.03426
[11] M. Allaoui, M. L. Kherfi, dan A. Cheriet, “Considerably improving clustering algorithms using umap dimensionality reduction technique: A comparative study,” dalam *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2020, vol. 12119 LNCS, hlm. 317–325. doi: 10.1007/978-3-030-51935-3_34.
[12] L. McInnes, J. Healy, dan S. Astels, “hdbscan: Hierarchical density based clustering,” *The Journal of Open Source Software*, vol. 2, no. 11, hlm. 205, Mar 2017, doi: 10.21105/joss.00205.
[13] M. Grootendorst, “BERTopic: Neural topic modeling with a class-based TF-IDF procedure,” Mar 2022, [Daring]. Available: http://arxiv.org/abs/2203.05794
[14] T. Joachims, “A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization.”