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

Pemodelan Topik Pada Ulasan Hotel Menggunakan Metode BERTopic Dengan Prosedur c-TF-IDF

I Komang Tryana Mertayasa[a1], I Dewa Made Bayu Atmaja Darmawan[a2]

[a]Program Studi Informatika, Universitas Udayana
Jimbaran, Badung, Bali, Indonesia
[1]tryanamertayasa@student.udayana.ac.id
[2]dewabayu@unud.ac.id (Corresponding author)

Abstract

User review data on travel guidance services can be useful textual data for other users. By knowing what topics are discussed in user reviews in hotel products, travel guidance service providers can group these reviews based on the topics discussed. In grouping textual data into several topics, the use of topic modeling methods can be done. In this study, the author uses the BERTopic method in modeling topics on user review data related to hotel products on one of the TripAdvisor travel guidance services. This study uses secondary data in the form of hotel reviews on the TripAdvisor site. Topic modeling with BERTopic begins with document embedding, dimensionality reduction (UMAP), clustering (HDBSCAN), and c-TF-IDF. Topic modeling using the BERTopic method resulted in 78 topics with a topic coherence value of 0.07287 and a topic diversity of 0.496154. The lower the number of topics to be generated, the value of topic coherence and topic diversity decreases.

Keywords: *TripAdvisor, BERTopic, UMAP, HDBSCAN, c-TF-IDF*

1. Pendahuluan

Sektor pariwisata merupakan salah satu sektor unggulan di Indonesia. Sebagai salah satu sektor unggulan, sektor pariwisata terpengaruh oleh perkembangan teknologi dimana terdapat berbagai layanan digital yang memudahkan kegiatan berwisata. Adapun beberapa layanan digital tersebut berupa OTA (*Online Travel Agent*), *travel guidance,* HMS (*Hotel Management System*)*,* dan lain-lain. Banyak layanan *travel guidance*, seperti TripAdvisor, Lonely Planet, dan Google Maps, yang telah menjadi lebih umum dalam penelitian dan praktik, yang mengarah pada proliferasi studi tentang ulasan online dan mengamati penggunaan analisis baru [1]. Fitur ulasan yang terdapat pada layanan *travel guidance* seperti TripAdvisor membantu pengguna dalam menyampaikan suatu ulasan terkait dengan beberapa produk wisata yang ada, sehingga hal ini tentunya dapat menjadi pertimbangan bagi pengguna lainnya dalam melakukan pemesanan produk wisata. TripAdvisor sebagai salah satu situs *travel guidance* terbesar, pada bulan Agustus tahun 2022 memiliki 186,5 juta pengunjung bulanan [2].

Data ulasan pengguna pada layanan *travel guidance* dapat menjadi data tekstual yang berguna bagi pengguna lainnya. Saat ini, keputusan pemesanan hotel semakin dipengaruhi oleh ulasan pelanggan dimana pengguna tidak hanya melihat peringkat bintang konvensional tetapi juga melihat ulasan pengguna sebelumnya. Dengan mengetahui apa saja topik yang dibahas pada ulasan pengguna dalam produk hotel, penyedia layanan *travel guidance* dapat mengelompokan ulasan tersebut berdasarkan topik yang dibahas.

Dalam pengelompokan data tekstual menjadi beberapa topik, penggunaan metode pemodelan topik dapat dilakukan. Adapun beberapa metode yang sering digunakan dalam melakukan pemodelan topik adalah *Latent Dirichlet Allocation* (LDA) dan *Non-Negative Matrix Factorization* (NMF). Tentunya metode LDA dan NMF memiliki batasan karena metode tersebut mengabaikan hubungan semantik di antara kata-kata. Salah satu metode yang dapat mengatasi batasan tersebut adalah metode BERTopic. BERTopic memperluas proses pemodelan topik dengan mengekstraksi representasi topik yang koheren melalui pengembangan variasi TF-IDF berbasis kelas. Pada penelitian yang dilakukan Yunanto pada tahun 2021 terkait dengan pemodelan topik

pada ulasan hotel dimana pada penelitian tersebut menggunakan data ulasan pada pada *booking hotel platform* Pegipegi dengan metode *Latent Dirichlet Allocation* (LDA) yang menghasilkan lima topik [3]. Penelitian lainnya dilakukan Darell pada tahun 2021 mengenai pemodelan topik pada *customer service chat* dimana BERTopic memperoleh hasil nilai evaluasi yang lebih tinggi dibandingkan *baseline model* yaitu *Latent Dirichlet Allocation* (LDA) [4].

Pada penelitian ini, penulis menggunakan metode BERTopic dalam melakukan pemodelan topik pada data ulasan pengguna terkait produk hotel pada situs layanan *travel guidance* TripAdvisor. Dengan penelitian ini diharapkan mampu mengelompokkan data tekstual ulasan pengguna terkait produk hotel melalui pemodelan topik. Sehingga nantinya hasil dari pemodelan topik ini dapat bermanfaat bagi pihak layanan *travel guidance*, para pengguna, serta pihak hotel sendiri dalam menganalisis data tekstual ulasan konsumen.

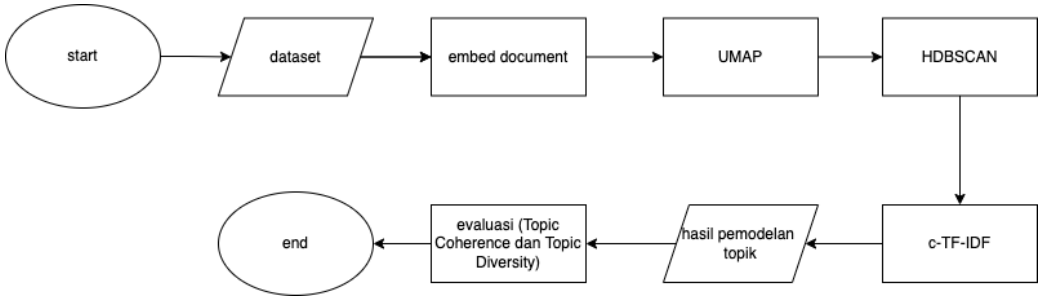
2. Metode Penelitian

2.1. Data Penelitian

Penelitian ini menggunakan jenis data sekunder. Dataset diperoleh dari penelitian yang dilakukan Alam, M. H., Ryu, W.-J., Lee, S., pada tahun 2016 dengan judul penelitian *Joint multi-grain topic sentiment: modeling semantic aspects for online reviews* [5]. Dataset tersebut diperoleh dengan cara melakukan *crawling* pada situs TripAdvisor. Adapun dataset ini memiliki data ulasan pengguna sebanyak 20.000 lebih data ulasan. Dataset ini memiliki format *csv* dan memiliki dua kolom yaitu kolom “Review” yang berisikan data ulasan pengguna dan juga kolom “Rating” yang berisikan penilaian pengguna dari skala satu sampai dengan lima. Pada penelitian tersebut membahas mengenai *topic sentiment* dengan menggunakan metode *Joint Multi-grain Topic Sentiment* (JMTS) *model* dimana bertujuan untuk mengekstrak aspek rata-rata berorientasi sentimen dari ulasan online. Berbeda dengan penelitian yang dilakukan sebelumnya dimana dataset digunakan untuk *topic sentiment,* pada penelitian yang penulis lakukan, dataset tersebut akan digunakan untuk pemodelan topik dengan metode BERTopic.

2.2. Metode Penelitian

Pada tahap ini akan dijelaskan mengenai desain penelitian secara keseluruhan. Penelitian diawali dengan pencarian dataset, dimana dataset yang digunakan adalah dataset ulasan hotel pada layanan TripAdvisor pada penelitian yang dilakukan oleh Alam, M. H., Ryu, W.-J., Lee, S., pada tahun 2016. Selanjutnya akan dilakukan proses pemodelan topik dengan menggunakan metode BERTopic. BERTopic adalah metode pemodelan topik yang memanfaatkan *embedding* BERT dan c-TF-IDF untuk membuat *dense cluster* yang memungkinkan topik dengan mudah ditafsirkan sambil menyimpan kata-kata penting dalam deskripsi topik [6]. Dalam melakukan pemodelan topik, metode BERTopic memiliki tiga tahapan yaitu melakukan *document embedding,* melakukan *cluster* ke dalam bentuk *semantic similar cluster,* lalu membuat representasi topik dari masing-masing *cluster.* Proses diawali dengan melakukan *document embedding* agar memperoleh representasi pada ruang *vector*. Selanjutnya, menggunakan *Uniform Manifold Approximation and Projection* (UMAP) untuk mengurangi dimensional *vector*, lalu dilakukan proses *clustering* menggunakan *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN). Setelah pemodelan topik berhasil dilakukan, selanjutnya dilakukan tahap evaluasi. Tahap evaluasi dilakukan dengan menggunakan *topic coherence* dan *topic diversity*.



Gambar 1. Alur Desain Penelitian

2.3. *Document Embedding*

Proses *document embedding* dilakukan untuk merepresentasikan suatu kata atau kalimat ke dalam bentuk *dense vector.* Jika suatu dokumen memiliki semantik yang sama dengan dokumen lainnya, maka dapat diasumsikan bahwa dokumen tersebut memiliki topik yang sama. Pada penelitian ini, proses *document embedding* menggunakan *sentence-transformer. Sentence-transformer* sendiri terdapat operasi *pooling*, dimana operasi tersebut dilakukan agar nantinya ukuran dari *sentence-embedding* menjadi tetap. Adapun *input* data pada proses *document embedding* ini yaitu teks ulasan yang terdapat pada dataset. Setelah dilakukan proses *document embedding,* maka *output* data yang dihasilkan berupa data *vector embedding.* Hal ini dilakukan dengan tujuan mengubah data teks ulasan menjadi *numerical* data.

	0	1	2	3	4	5	6	7	8	9	...
0	0.019152	0.025172	0.077539	0.043416	-0.084842	0.056074	0.082910	-0.041291	-0.037210	-0.061863	...
1	0.013609	0.018621	0.075413	0.056367	-0.000859	0.057643	0.049825	-0.066042	-0.023772	-0.061398	...
2	0.077366	0.003373	0.046953	0.009228	-0.022182	0.014700	0.063607	-0.038185	-0.002556	-0.035796	...
3	0.017363	-0.003071	0.056516	0.052276	0.014113	0.033408	0.090919	-0.030463	-0.027905	-0.075787	...
4	0.002439	-0.025305	0.056496	0.001316	-0.082974	0.030920	0.088570	-0.033171	-0.051499	-0.018673	...
5	0.035034	-0.016818	0.084592	0.085277	0.044474	0.001155	0.063051	-0.033511	-0.049756	-0.074421	...
6	0.004106	0.025019	0.074884	0.111210	-0.000917	0.028222	0.038566	-0.059216	-0.055597	-0.061574	...
7	-0.008810	-0.033733	0.121630	0.020636	-0.091407	0.049938	0.029530	-0.040932	-0.141578	-0.038496	...
8	0.067704	-0.049052	0.004333	0.097545	0.008608	0.072784	0.060265	-0.059933	-0.036879	-0.042546	...
9	0.034898	-0.011066	0.011715	0.083125	-0.047884	0.052017	0.043366	-0.019935	-0.015285	-0.045089	...

Gambar 2. *Output* Proses *Document Embedding*

2.4. UMAP

Hasil dari *document embedding* akan meningkatkan dimensi data, sehingga perlu dilakukan *dimensionality reduction*. UMAP menunjukkan lebih banyak fitur lokal dan global dari data berdimensi tinggi dalam dimensi yang diproyeksikan lebih rendah [7]. UMAP dapat digunakan di seluruh model bahasa dengan ruang dimensi yang berbeda. Adapun *input* data pada proses *dimensionality reduction* menggunakan metode UMAP ini yaitu *vector embedding* yang berupa *numerical* data hasil dari dataset teks ulasan sebelumnya. Setelah dilakukan proses *dimensionality reduction* menggunakan metode UMAP*,* maka *output* data yang dihasilkan berupa data *vector embedding* yang dimensi datanya sudah tereduksi.

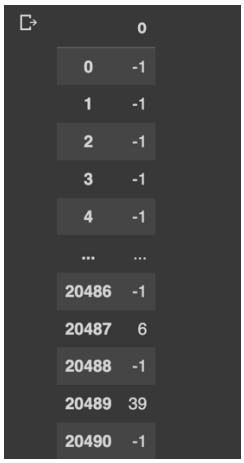
	0	1	2	3	4
0	6.281369	4.037725	4.443634	8.274034	6.193143
1	6.424414	3.681332	4.066862	7.672165	5.084017
2	6.871318	3.490070	4.795259	7.797281	5.782706
3	6.579267	4.863573	4.107368	9.038058	6.736990
4	6.897512	2.984293	4.428236	8.406915	5.841094
...	...	...	...	...	...

Gambar 3. *Output* Proses *Dimensional Reduction* Menggunakan UMAP

2.5. HDBSCAN

Pada pemodelan topik menggunakan metode BERTopic, proses *clustering* dari hasil *document embedding* yang telah melalui proses *dimensionality reduction* menggunakan *Uniform Manifold Approximation and Projection* (UMAP), dilakukan dengan metode *Hierarchical Density- Based*

Spatial Clustering of Applications with Noise (HDBSCAN). Penggunaan UMAP dalam mengurangi dimensionalitas *embedding* terbukti dapat meningkatkan kinerja algoritma HDBSCAN, baik dari segi akurasi *clustering* maupun waktu [8]. Metode HDBSCAN sendiri menggunakan pendekatan *soft clustering,* dimana *noise* dimodelkan sebagai *outlier* sehingga dokumen yang tidak terkait tidak akan dimasukkan ke dalam *cluster.* Hal ini tentunya akan meningkatkan representasi topik yang dihasilkan nantinya. Adapun *input* data pada proses *clustering* menggunakan metode HDBSCAN ini yaitu *vector embedding* yang dimensinya sudah tereduksi menggunakan metode UMAP. Setelah dilakukan proses *clustering* menggunakan metode HDBSCAN*,* maka *output* data yang dihasilkan berupa *clustering label* dari setiap data teks ulasan pada dataset.



	0
0	-1
1	-1
2	-1
3	-1
4	-1
...	...
20486	-1
20487	6
20488	-1
20489	39
20490	-1

Gambar 4. *Output* Proses *Clustering* Menggunakan HDBSCAN

2.6. c-TF-IDF

Dari hasil *cluster* yang diperoleh, setiap *cluster* akan direpresentasikan oleh satu topik. TF-IDF berbasis kelas digunakan untuk membuat representasi topik pada setiap *cluster.* Penggunaan c-TF-IDF akan menghasilkan distribusi topik kata untuk setiap *cluster* dokumen karena metode ini memodelkan pentingnya kata dalam *cluster* dibandingkan dengan dokumen individual. Adapun *input* data pada proses representasi *cluster* menggunakan TF-IDF berbasis kelas ini yaitu *cluster* data dari proses HDBSCAN sebelumnya. Setelah dilakukan proses representasi *cluster* menggunakan TF-IDF berbasis kelas*,* maka *output* data yang dihasilkan berupa kata-kata yang merepresentasikan setiap topik atau *cluster*.

$$W_{x,c} = tf_{x,c} \times \log(1 + \frac{A}{f_x}) \quad (1)$$

Keterangan:

$tf_{x,c}$ = frekuensi dari kata *x* pada kelas *c*

f_x = frekuensi dari kata *x* pada keseluruhan kelas

A = Rata-rata jumlah kata pada setiap kelas

2.7. Evaluasi

Dalam melakukan evaluasi dari hasil pemodelan topik dengan metode BERTopic, penelitian ini menggunakan evaluasi *topic coherence* dan *topic diversity.* Nilai koherensi menunjukkan tingkat keterpaduan kata-kata dalam suatu topik yang dihasilkan dari analisis perbedaan atau kesamaan semantik antara kata-kata dalam topik tersebut. Digunakan juga *Normalized Pointwise Mutual Information* (NPMI) yang akan menemukan seberapa sering dua kata muncul bersamaan dalam dokumen tertentu. NPMI dihitung dengan membagi probabilitas munculnya kedua kata dengan probabilitas kemunculan setiap kata secara terpisah. $P(\omega i)$, adalah probabilitas satu kata yang ada dalam dokumen 'd', 'i' dan 'j' adalah kata-kata yang diambil sebagai kata teratas yang ada dalam topik 't', $\Theta wj,d$ adalah kata yang ada dalam dokumen [9].

$P(\omega_i), = \sum_d \Theta_{i,d}$ (2)

$$P(\omega_i, \omega_j) = \sum_d \Theta_{w_i,d} * \Theta_{j,d} \quad (3)$$

$$NPMI(\omega_i, \omega_j) = \frac{ln\, P(\omega_i), + \ln P(\omega_j)}{lnP(\omega_i, \omega_j)} - 1 \quad (4)$$

3. Hasil dan Pembahasan

Dalam melakukan pemodelan topik menggunakan metode BERTopic, pada penelitian ini penulis menggunakan *bertopic* yang merupakan *open-source python package* [10]. OCTIS (*Optimizing and Comparing Topic models is Simple*), yang merupakan *open-source python package*, digunakan untuk menjalankan eksperimen, memvalidasi hasil, dan memproses data sebelumnya [11]. Eksperimen dalam menjalankan pemodelan topik dilakukan dengan penggunaan beberapa parameter dengan nilai terdapat pada tabel 1.

Tabel 1. Parameter BERTopic

Parameter	*Value*
embedding_model	"all-mpnet-base-v2"
diversity	None
min_topic_size	15
n_gram_range	[1, 2]
verbose	True
calculate_probabilities	False

Eksperimen dijalankan dengan menjalankan BERTopic dengan parameter *default* sesuai dengan tabel 1 serta eksperimen lainnya dengan parameter tambahan berupa jumlah topik yang akan dibuat secara manual, dimana nilai ini dimulai dari 10 hingga 100 topik dengan kelipatan 10. Dari hasil eksperimen akan menghasilkan evaluasi dari pemodelan topik yang dihasilkan.

3.1. Pemodelan Topik Pada Ulasan Hotel

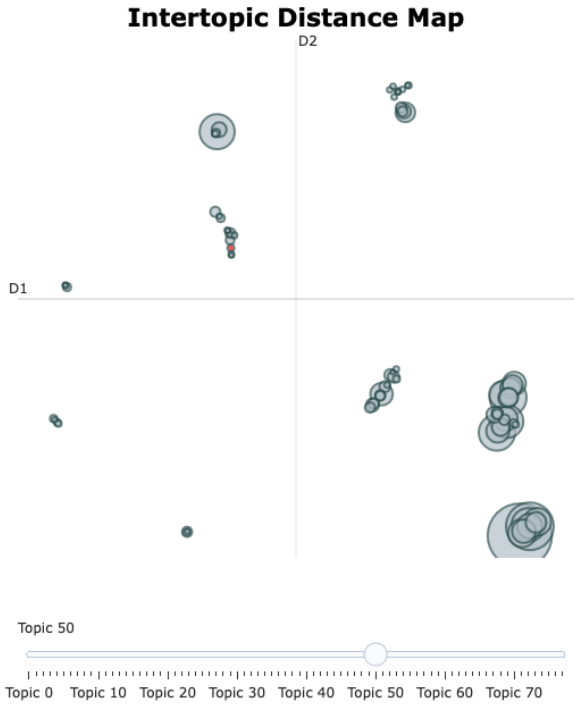
Jumlah topik yang dihasilkan dari penggunaan metode BERTopic adalah sebanyak 78 topik yang yang ditunjukkan pada tabel 2. Topik yang dihasilkan terbilang cukup banyak mengingat besaran dari jumlah dataset yang digunakan pada penelitian ini sebesar lebih dari 20.000 data ulasan. Gambar 5 menunjukan visualisasi dari sebaran pemodelan topik yang dihasilkan.

Tabel 2. Hasil Pemodelan Topik

Topic	*Count*	*Topic Representation*
0	2880	0_resort_beach_food_did
1	1585	1_location_hotel_great_staff
2	945	2_paris_metro_hotel_eiffel
3	918	3_barcelona_hotel_ramblas_metro
4	856	4_florence_hotel_duomo_italy
5	784	5_room_told_hotel_desk
6	764	6_amsterdam_hotel_room_canal
7	559	7_juan_san_puerto_rico
8	531	8_york_nyc_square_new
9	494	9_sydney_harbour_darling_quay
10	434	10_seattle_downtown_pike_needle
11	410	11_london_tube_bridge_hotel
12	382	12_waikiki_hawaii_beach_honolulu
13	374	13_bali_ubud_villa_kuta

14	372	14_ramblas_las_hotel_location
15	355	15_orleans_quarter_new_french
16	297	16_boston_hotel_copley_room
17	291	17_hong_kong_hk_kowloon
18	284	18_singapore_raffles_hotel_orchard
19	269	19_francisco_san_sf_union
20	258	20_tokyo_shinjuku_japan_station
21	243	21_beijing_china_chinese_forbidden
22	220	22_madrid_plaza_sol_hotel
23	180	23_berlin_bahn_hotel_breakfast
24	153	24_venice_canal_ponte_al
25	135	25_inn_castle_francisco_wharf
26	117	26_toronto_eaton_cambridge_suites
27	97	27_miami_south_pool_beach
28	92	28_cruise_beach_hotel_great
29	88	29_bahn_hotel_breakfast_location
30	85	30_phoenix_scottsdale_desert_clarendon
31	80	31_frankfurt_airport_hotel_station
32	80	32_chancellor_francisco_san_cable
33	76	33_hollywood_beverly_elan_hills
34	75	34_dallas_dfw_hyatt_palomar
35	72	35_great_hotel_room_breakfast
36	72	36_mexico_city_reforma_zona
37	71	37_riu_bambu_macao_palace
38	71	38_casablanca_york_rick_cheese
39	70	39_berlin_bahn_hotel_breakfast
40	69	40_argonaut_wharf_alcatraz_cable
41	65	41_ritz_carlton_club_service
42	50	42_sofitel_york_nyc_new
43	48	43_europa_florence_duomo_gassim
44	43	44_room_hotel_fab_breakfast
45	43	45_jazz_ramblas_modern_las
46	42	46_pike_market_parking_downtown
47	35	47_union_square_regis_great
48	34	48_airport_shuttle_flight_free
49	34	49_melia_caribe_tropical_resort
50	32	50_omni_francisco_san_sf
51	31	51_affinia_50_nyc_suite
52	28	52_dumont_affinia_nyc_york
53	28	53_stag_warwick_lads_groups
54	27	54_lincoln_rooms_hotel_miami
55	26	55_casci_florence_paolo_pierpaolo
56	26	56_rex_francisco_san_union
57	25	57_moon_east_york_tenement

58	24	58_shangri_wing_la_singapore
59	24	59_mela_nyc_square_york
60	23	60_orchard_union_francisco_san
61	22	61_vieques_hix_island_house
62	22	62_bugs_bed_bites_bug
63	22	63_sofitel_dc_washington_white
64	21	64_wharf_fisherman_cable_car
65	21	65_nadia_amsterdam_hotel_frank
66	21	66_milano_juan_san_old
67	21	67_watertown_university_uw_carts
68	19	68_gold_excellent_staff_sons
69	19	69_41_york_414_new
70	19	70_adagio_union_square_cortez
71	18	71_swan_white_inn_fireplace
72	18	72_swissotel_berlin_ku_station
73	18	73_muse_york_square_nyc
74	17	74_riu_palace_punta_cana
75	16	75_langham_kong_hong_kok
76	15	76_needle_pioneer_space_parking
77	15	77_juan_san_rollaway_condado



Gambar 5. Sebaran Topik

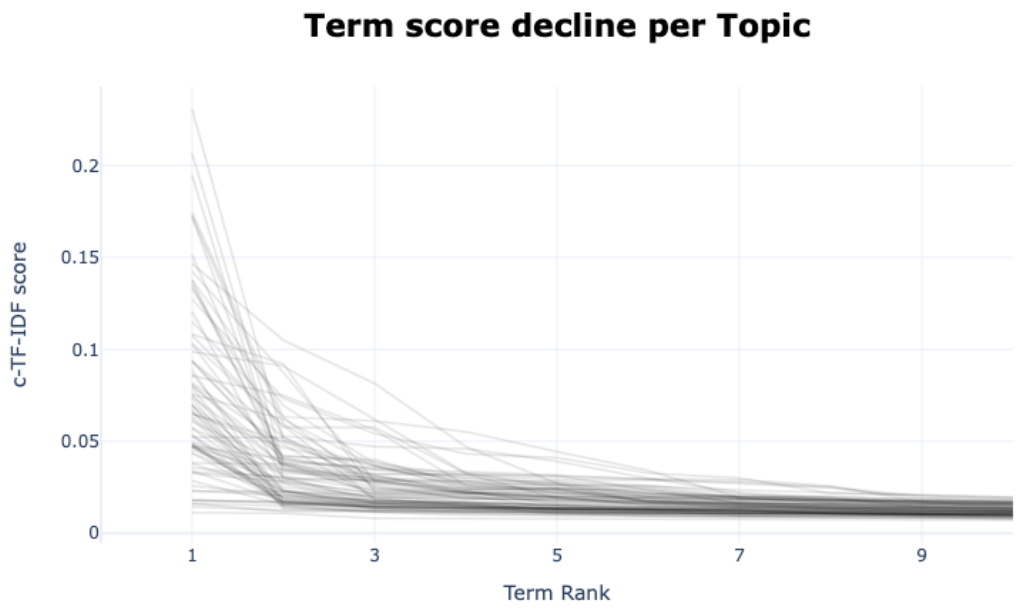
Pada gambar 6 menunjukan visualisasi berupa *topic word scores* pada beberapa istilah yang terdapat pada beberapa topik, dimana pada visualisasi ini penulis mengambil lima topik untuk divisualisasikan. Pada “topic 0” terdapat beberapa *terms* seperti *resort, beach, food, did,* dan *pool.* Pada “topic 0”, adapun topik yang dibahas adalah mengenai hotel bertipe *resort,* dimana hotel ini biasanya berada pada dekat pantai serta memiliki kolam renang yang menghadap

langsung ke pantai. Pada “topic 1” terdapat beberapa *terms* seperti *location, hotel, great, staff,* dan *room.* Pada “topic 1”, adapun topik yang dibahas adalah mengenai pelayanan serta lokasi dari hotel tersebut. Pada “topic 2” terdapat beberapa *terms* seperti *paris, metro, hotel, eiffel,* dan *location.* Pada “topic 2”, adapun topik yang dibahas adalah mengenai lokasi hotel yang berada di daerah Paris. Pada “topic 3” terdapat beberapa *terms* seperti *barcelona, hotel, ramblas, metro,* dan *city.* Pada “topic 3”, adapun topik yang dibahas adalah mengenai lokasi hotel yang berada di daerah Barcelona. Pada “topic 4” terdapat beberapa *terms* seperti *florence, hotel, duomo, italy,* dan *breakfast.* Pada “topic 4”, adapun topik yang dibahas adalah mengenai lokasi hotel yang berada di daerah Italia.

Ulasan konsumen sendiri dianggap lebih kredibel dibandingkan dengan deskripsi produk dari vendor atau pemilik produk karena ulasan konsumen berisi perspektif pengguna dengan skenario penggunaan yang berbeda. Hal tersebut tentunya menjadi pertimbangan bagi pengguna lainnya dalam keputusan pemesanan suatu produk dalam hal ini produk wisata berupa akomodasi hotel. Pemodelan topik yang dihasilkan menggunakan metode BERTopic ini tentunya dapat membantu para pengguna dalam menganalisis ulasan secara otomatis nantinya. Sebagai contoh pada “topic 1” membahas mengenai pelayanan serta lokasi dari hotel tersebut, dengan mengelompokan data ulasan suatu hotel berdasarkan “topic 1” tersebut maka memudahkan pengguna layanan *travel guidance* dalam melihat dan menganalisis ulasan hotel dari segi pelayanan serta lokasi dari hotel tersebut karena ulasannya sudah dikelompokkan pada topik tersebut. Selain itu, pada “topic 28” juga membahas mengenai hotel dekat pantai yang memberikan pengalam bagus. Tentunya hal ini akan membantu pengguna yang ingin memesan hotel dekat pantai dalam melihat serta menganalisis ulasan tersebut secara otomatis. Pemodelan topik yang dihasilkan akan memudahkan pengguna layanan *travel guidance* dalam melihat dan menganalisis ulasan suatu hotel, sehingga hal ini akan mempengaruhi keputusan mereka dalam memesan suatu hotel.

Gambar 6. *Topic Word Scores*

Setiap topik yang dihasilkan oleh pemodelan topik dengan BERTopic akan diwakli oleh sejumlah kata dengan kata representatif terbaik. Setiap kata tersebut memiliki skor c-TF-IDF, dimana semakin tinggi skornya, semakin representatif sebuah kata untuk topik tersebut. Sebagai contoh, pada “topic 1” *term* yang memiliki skor tertinggi adalah *location*, sedangkan untuk *terms* selanjutnya diikuti oleh *hotel, great, staff,* dan *room* secara berurutan serta skornya semakin menurun. Skor tersebut akan membantu dalam proses representasi topik, dimana pada “topic 1” sendiri topik yang dibahas adalah mengenai pelayanan serta lokasi dari hotel tersebut yang didapat dari *term* berdasarkan urutan skor sebelumnya. Seperti yang diketahui bahwa setiap kata-kata yang ada pada topik diurutkan berdasarkan skor c-TF-IDF, tentunya skor perlahan-lahan menurun dengan setiap kata yang ditambahkan. Akan terdapat titik dimana dengan menambahkan suatu kata untuk merepresentasikan topik hanya sedikit meningkatkan skor total c-TF-IDF dan tidak akan bermanfaat untuk representasinya. Gambar 7 menunjukan visualisasi dari skor c-TF-IDF pada setiap topik. Terlihat bahwa terdapat penurunan skor c-TF-IDF saat menambahkan kata pada representasi topik.



Gambar 7. Penurunan Skor c-TF-IDF Pada Topik

3.2. Evaluasi Pemodelan Topik Menggunakan Metode BERTopic

Dari data eksperimen pada hasil Tabel 3, menunjukan bahwa jumlah topik yang dihasilkan pada pemodelan topik menggunakan BERTopic menghasilkan 78 topik dengan nilai *topic coherence* sebesar 0,07287 serta *topic diversity* sebesar 0,496154. *Topic coherence* sendiri memiliki nilai rentang -1 sampai dengan 1, dimana nilai 1 menunjukan asosiasi yang sempurna. *Topic diversity* memiliki nilai rentang 0 sampai dengan 1, dimana 0 menunjukkan topik yang *redundant* dan 1 menunjukkan lebih banyak topik yang bervariasi. Penulis juga melakukan eksperimen dengan mengganti parameter jumlah topik yang akan dibuat secara manual. Dimana didapatkan hasil bahwa semakin rendah jumlah topik yang ingin dihasilkan, nilai dari *topic coherence* dan *topic diversity* menjadi berkurang.

Tabel 3. Evaluasi BERTopic

Jumlah Topik	*Topic Coherence*	*Topic Diversity*	*Computation Time (s)*
Auto (78)	0,07287	0,496154	35,96054
10	0,030192	0,44	37,50018
20	0,056643	0,465	34,02165
30	0,06101	0,446667	33,79216
40	0,066321	0,4675	33,30897
50	0,073314	0,472	34,77169
60	0,072731	0,458333	33,85231
70	0,070536	0,494286	33,52127
80	0,063349	0,484932	29,92879
90	0,078367	0,489552	30,12689
100	0,065468	0,504918	30,81338

4. Kesimpulan

Berdasarkan paparan penelitian yang telah dilakukan sebelumnya, adapun beberapa hal yang dapat disimpulkan adalah sebagai berikut:

1. Jumlah topik yang dihasilkan dari penggunaan metode BERTopic adalah sebanyak 78 topik. Setiap topik yang dihasilkan oleh pemodelan topik dengan BERTopic akan diwakli oleh sejumlah kata dengan kata representatif terbaik. Terdapat penurunan skor c-TF-IDF saat menambahkan kata pada representasi topik. Dengan pemodelan topik yang dihasilkan akan memudahkan pengguna layanan *travel guidance* dalam melihat dan menganalisis ulasan suatu hotel, sehingga hal ini akan mempengaruhi keputusan mereka dalam memesan suatu hotel.
2. Pemodelan topik menggunakan BERTopic menghasilkan 78 topik dengan nilai *topic coherence* sebesar 0,07287 serta *topic diversity* sebesar 0,496154. Pemodelan topik dilakukan dengan penggunaan parameter beberapa parameter *default* dan eksperimen lainnya dilakukan dengan mengganti jumlah topik yang ingin dibuat secara manual. Semakin rendah jumlah topik yang ingin dihasilkan, nilai dari *topic coherence* dan *topic diversity* menjadi berkurang.

Daftar Pustaka

[1] Cheng, X., Fu, S., Sun, J., Bilgihan, A., & Okumus, F., "An Investigation on Online Reviews in Sharing Economy Driven Hospitality Platforms: A Viewpoint Of Trust" *Tourism Management*, vol. 71, p. 366-377, 2019.

[2] Anonim. "Estimated Total Number of Visits To The Travel and Tourism Website Tripadvisor.Com Worldwide From August 2020 To August 2022". 31 August 2022. [Online]. Available: https://www.statista.com/statistics/1215473/total-visits-to-tripadvisor-website/ [Accessed on 24 September 2022]

[3] Putranto, Y., Sartono, B., dan Djuraidah, A., "Topic Modelling And Hotel Rating Prediction Based on Customer Review in Indonesia" *International Journal of Management and Decision Making*, vol. 20, no. 3, p. 282-307, 2021.

[4] Hendry, D., Darari, F., Nurfadillah, R., Khanna, G., Sun, M., Condylis, P. C., dan Taufik, N., "Topic Modeling for Customer Service Chats" *International Conference on Advanced Computer Science and Information Systems (ICACSIS),* p. 1-6, 2021.

[5] Alam, M. H., Ryu, W.-J., Lee, S., "Joint Multi-Grain Topic Sentiment: Modeling Semantic Aspects for Online Reviews" *Information Sciences*, vol. 339, p. 206–223, 2016.

[6] Grootendorst, M., "BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure" *arXiv preprint arXiv:2203.05794*, 2022.

[7] McInnes, L., Healy, J., & Melville, J., "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction" *arXiv preprint arXiv:1802.03426,* 2018.

[8] Allaoui, M., Kherfi, M. L., dan Cheriet, A., "Considerably improving clustering algorithms using umap dimensionality reduction technique: A comparative study" *International Conference on Image and Signal Processing*, p. 317–325, 2020.

[9] George, Shini, "Comparison of LDA and NMF Topic Modeling Techniques for Restaurant Reviews" *Indian Journal of Natural Sciences,* vol. 10, no. 6, p. 28210-28216, 2020.

[10] Grootendorst, M., "BERTopic", 11 September 2022. [Online]. Available: https://github.com/MaartenGr/BERTopic [Accessed on 20 September 2022]

[11] Terragni, S., Fersini, E., Galuzzi, B. G., Tropeano, P., dan Candelieri, A., "OCTIS: Comparing and Optimizing Topic Models is Simple!" in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2021, p. 263–270.