Jurnal Sistem Informasi dan Informatika (Simika) *P-ISSN: 2622-6901*
Vol 6 No. 2 Tahun 2023 *E-ISSN: 2622-6375*

ANALISIS SENTIMEN DAN PEMODELAN TOPIK PADA *TWEET* TERKAIT DATA BADAN PUSAT STATISTIK

Adielia Amanda[1], Erna Nurmawati[2]
Program Studi Komputasi Statistik, Politeknik Statistika STIS
Jl. Otto Iskandardinata No.64C, Jakarta Timur
e-mail: *erna.nurmawati@stis.ac.id, [2]1221910874@stis.ac.id

Abstract

The SKD (Data Needs Survey), conducted by BPS every year to identify the level of satisfaction among data users, cannot cover all segments of the data user population. Hence, it becomes necessary to employ alternative methods that can reach data users beyond those who are registered in the PST (Perpustakaan Statistik Terpadu). This study aims to analyze the public sentiment towards BPS data, primarily on social media platform Twitter. To ascertain the distribution of topics being discussed within the community regarding BPS data indicators, a topic modeling approach was employed. The sentiment analysis process involves the utilization of the Indonesian BiDirectional Encoder Representations from Transformers (BERT) model, known as IndoBERT. For the topic modeling, the Latent Dirichlet Allocation (LDA) method was utilized, yielding a distribution of topics for each document through the application of the Dirichlet distribution. Based on the results of the sentiment analysis spanning the period from 2020 to 2022, it was observed that tweets related to BPS data predominantly conveyed a neutral sentiment. Meanwhile, the topic modeling procedure yielded a diverse set of topics each year. From 2020 to 2022, the topic most frequently discussed pertained to the statistical data from the 2020-2022 Data Needs Survey. Specifically, the emphasis lay in the variety of social statistical data, highlighting it as the most crucial type of data sought after in comparison to other data types.

Keyword: *BPS data, IndoBERT, LDA, Sentiment Analysis, Topic Modeling*

PENDAHULUAN

Badan Pusat Statistik (BPS) mempunyai visi yaitu “Penyedia Data Statistik Berkualitas untuk Indonesia Maju”. Dalam visi yang baru tersebut, berarti bahwa BPS berperan dalam penyediaan data statistik nasional maupun internasional, untuk menghasilkan statistik yang mempunyai kebenaran akurat dan menggambarkan keadaan yang sebenarnya, dalam rangka mendukung Indonesia Maju. Sesuai dengan visinya, maka BPS harus profesional dalam menyajikan data. Apalagi tuntutan masyarakat terhadap ketersediaan data dan informasi statistik yang beragam dan berkualitas, semakin hari semakin meningkat.

Respon dan opini masyarakat terkait data yang dikeluarkan oleh BPS merupakan salah satu aspek penting yang harus diperhatikan. Respon dan opini masyarakat dapat digunakan untuk mengetahui penilaian masyarakat terhadap data-data yang dikeluarkan oleh BPS, yang dapat dijadikan bahan untuk mengidentifikasi kepuasan masyarakat, indikasi masalah, serta untuk memperoleh bahan evaluasi guna meningkatkan kualitas data dan informasi statistik.

Setiap tahunnya, BPS rutin menyelenggarakan suatu survei yang bernama Survei Kebutuhan Data (SKD). Survei ini dilakukan untuk mengidentifikasi kebutuhan data dan tingkat kepuasan konsumen terhadap kualitas data yang dihasilkan BPS. Namun, responden dalam SKD hanya mencakup konsumen yang pernah menerima layanan dari unit Pelayanan Statistik Terpadu (PST) di BPS pada tahun tertentu. Artinya, SKD belum mampu menjangkau masyarakat secara keseluruhan, yaitu masyarakat yang memperoleh data BPS melalui *website* bps.go.id., serta media lain seperti portal berita digital dan cetak. Oleh karena itu, dibutuhkan metode lain untuk mengumpulkan data respon dan opini masyarakat, misalnya dengan media sosial.

Media sosial yang paling sesuai untuk mengumpulkan data respon dan opini masyarakat adalah media sosial Twitter. Twitter merupakan situs *microblogging* populer yang digunakan oleh banyak orang untuk saling mengungkapkan pendapat mereka terhadap topik yang sedang dibahas. Penggunaan Twitter yang masif dari waktu ke waktu menyebabkan volume *tweet* naik dan polanya semakin beragam, sehingga proses identifikasi dan analisis cukup sulit dan memakan banyak waktu.

Berdasarkan latar belakang tersebut, penelitian ini mencoba untuk menerapkan metode yang dapat membantu dalam mengidentifikasi ataupun analisis terhadap data respon dan opini dari kumpulan *tweet* yang cukup banyak. Data Twitter memungkinkan adanya keterlibatan publik, sebab di dalam Twitter mengandung informasi tentang persepsi, tren, dan tanggapan publik tentang data Badan Pusat Statistik berdasarkan topik indikator *strate*gis. Maka untuk melakukan analisis sentimen, penelitian ini menggunakan model *Bidirectional Encoder Representations from Transformers* (BERT) berbahasa Indonesia, yaitu IndoBERT. Sedangkan untuk melakukan pemodelan topik, menggunakan metode *Latent Dirichlet Allocation* (LDA).

Baik IndoBERT maupun LDA telah terbukti cocok untuk diterapkan dalam data yang berjumlah banyak. Penelitian yang dilakukan oleh Cindy, Adiwijaya, dan Said (2020) menggunakan model BERT-*base* untuk analisis sentimen terhadap data ulasan film dengan menggunakan *dataset* Cornelledu dari Pabo sebanyak 2000 data. Penelitian tersebut mendapatkan akurasi cukup baik sebesar 73,7%. Akurasi ini sudah terbukti cukup bagus dan cukup jauh dibandingkan dengan penggunaan algoritma Naïve Bayes untuk proses klasifikasi yang hanya memiliki akurasi 48%. Penelitian Albalawi, Yeap, dan Benyoucef (2020), melakukan perbandingan berbagai metode pemodelan topik dengan menerapkannya pada data media sosial. Hasil kesimpulan penelitian tersebut adalah LDA mampu menghasilkan topik dengan makna yang lebih logis dan konsisten, dibandingkan dengan metode *data mining* lainnya yaitu metode *non-negative matrix factorization* (NMF).

Penelitian ini berfokus pada *tweet* dalam rentang tahun 2020-2022. Periode tersebut dipilih, sebab pada tahun tersebut sedang terjadi pandemi COVID-19 di seluruh dunia, termasuk di Indonesia. Artinya, masyarakat akan semakin masif menggunakan media sosial, sebab terdapat himbauan untuk membatasi aktivitas di luar rumah. Hasil dari penelitian ini dapat membantu mengetahui penilaian masyarakat terhadap data-data yang dikeluarkan oleh BPS, guna dijadikan bahan untuk mengidentifikasi kepuasan masyarakat, indikasi masalah, serta memperoleh bahan evaluasi guna meningkatkan kualitas data dan informasi statistik.

METODE PENELITIAN

Metode yang digunakan untuk melakukan analisis terkait respon dan opini masyarakat tentang data-data yang dikeluarkan oleh Badan Pusat Statistik (berdasarkan topik indikator st*rate*gis data BPS) pada media sosial Twitter., antara lain: analisis sentimen dengan *Bidirectional Encoder Representations from Transformers* (BERT) berbahasa Indonesia yaitu IndoBERT, serta pemodelan topik menggunakan *Latent Dirichlet Allocation* (LDA). Tahapan-tahapan dalam metode penelitian, antara lain :

1. **Pengumpulan Data**

   Pengumpulan data Twitter dengan cara *scraping* menggunakan *library* SNscrape pada Python. Pengumpulan dilakukan secara tahunan pada periode 2020 hingga 2022. Data *tweet* yang telah terkumpul, disimpan menggunakan format csv. Atribut data yang diambil dalam proses *scraping*, antara lain: ID, *datetime*, *username*, dan *tweet*. Data Twitter dikumpulkan berdasarkan 12 (dua belas) kata kunci mengenai data yang dikeluarkan oleh Badan Pusat Statistik. Kata kunci tersebut antara lain : Data BPS, Data Badan Pusat Statistik, Data bps_statistics, Laporan BPS, Laporan Badan Pusat Statistik, Laporan bps_statistics, Menurut BPS, Menurut Badan Pusat Statistik, Menurut

bps_statistics, Publikasi BPS, Publikasi Badan Pusat Statistik, dan Publikasi bps_statistics.

Pemilihan kata kunci didasari oleh penyebutan Badan Pusat Statistik di Twitter, yaitu menyebut dengan akronim/singkatan BPS, menyebut dengan kepanjangan Badan Pusat Statistik, atau menyebut dengan *username* bps_statistics. Selain itu, pemilihan kata kunci juga didasari oleh sinonim kata "data", yaitu "laporan" dan "publikasi", serta kata "menurut" yang bermakna "sesuai/berdasarkan".

Penelitian ini tidak mencakup akun Twitter dengan *username* yang mengandung kata: bps, data, dan kementerian, serta akun-akun yang me-re*tweet* postingan bps, guna menghindari keberpihakan akun-akun tersebut terhadap data BPS. Namun, *quote* (komentar) pada *retweet* tetap digunakan dalam penelitian ini. Jumlah *tweet* berdasarkan kata kunci dapat dilihat pada Tabel 1.

Tabel 1. Jumlah *tweet* berdasarkan kata kunci

Kata Kunci	Tahun		
	2020	**2021**	**2022**
Data BPS	7.279	6.325	9.051
Data Badan Pusat Statistik	808	638	994
Data bps_statistics	2.701	1.945	2.364
Laporan BPS	238	236	322
Laporan Badan Pusat Statistik	32	44	89
Laporan bps_statistics	45	42	46
Menurut BPS	739	702	1.131
Menurut Badan Pusat Statistik	150	163	361
Menurut bps_statistics	81	101	90
Publikasi BPS	228	335	333
Publikasi Badan Pusat Statistik	15	22	46
Publikasi bps_statistics	79	172	109
TOTAL	**12.395**	**10.725**	**14.936**

Setelah proses *scraping* dan penghapusan duplikat, selanjutnya dilakukan penghapusan *tweet* pada akun twitter dengan *username* yang mengandung kata : bps, data, dan kementerian, serta akun-akun yang me-re*tweet* postingan bps, guna menghindari keberpihakan akun-akun tersebut terhadap data BPS. Namun, *quote* (komentar) pada re*tweet* tetap digunakan dalam penelitian ini. Jumlah *tweet* berdasarkan tahun (setelah *filtering* akun) dapat dilihat pada Tabel 2.

Tabel 2. Jumlah *tweet* setelah *filtering* akun

Tahun	**2020**	**2021**	**2022**	**TOTAL**
Hasil *Filter*	9.258	6.535	9.928	**25.721**

2. *Preprocessing* Data

Tahap pertama pada *preprocessing* data, yaitu *cleaning* data yang bertujuan untuk membersihkan data yang tidak diperlukan dalam analisis dan membuat data lebih terstruktur, sehingga data siap digunakan dalam proses analisis selanjutnya. Beberapa tahap pada *cleaning* data, antara lain: menghapus URL, menghapus tanda *hashtag* dan *retweet* pada *tweet*, menghapus simbol (seperti simbol @) dan emoji, menghapus tanda baca (seperti tanda koma, tanda titik, tanda tanya).

Langkah kedua pada *preprocessing* data, yaitu *transformation* data, dengan tahapan diantaranya : *case folding*, yaitu mengubah seluruh teks kedalam huruf kecil; *convert Word*, yaitu memperbaiki adanya *slang word* yang terkandung dalam *tweet*, dengan cara mengkonversi kata *slang* menjadi kata bahasa Indonesia yang lebih formal. Proses

konversi menggunakan bantuan kamus *Colloquial Indonesian Lexicon* oleh (Salsabila et al., 2018); *delete stopword*s, yaitu menghapus kata yang sering muncul, konjungsi, dan tidak memiliki makna. Penghapusan *stopword*s menggunakan bantuan kamus *library* NLTK. Namun, dalam penelitian ini melakukan kustomisasi pada kamus tersebut, yaitu menghapus *stopword*s kata negasi "tidak", "bukan", "belum", dan "jangan". Kustomisasi ini dilakukan untuk mempertahankan konteks sentimen pada *tweet*; dan *stemming*, yaitu mengubah kata kedalam bentuk kata dasar.

Setelah melalui tahap *preprocessing*, masih ditemukan duplikasi *tweet*. Hal ini disebabkan karena terdapat *tweet* yang memiliki isi sama, namun tidak terdeteksi duplikasi sebelum *preprocessing* akibat adanya perbedaan pada bagian spasi dan simbol. *Tweet* dengan isi yang sama akan dihapus duplikasinya, guna menghindari redudansi (pengulangan) data *tweet*. Selain itu, ditemukan datum yang mengandung nilai NaN akibat pembersihan isi *tweet* pada saat *preprocessing*. Datum ini akan dihapus sebab nilai NaN tidak dapat digunakan dalam proses analisis. Jumlah *tweet* akhir yang akan digunakan untuk tahap analisis selanjutnya dapat dilihat pada tabel 3.

Tabel 3. Jumlah *tweet* akhir

Tahun	2020	2021	2022	TOTAL
Banyaknya data	5.700	4.955	7.384	**18.039**

3. Analisis Sentimen

Analisis sentimen atau *opinion mining* adalah salah satu bentuk dari pengaplikasian *text mining* yang digunakan untuk mengetahui opini dari sekumpulan data tekstual mengenai peristiwa atau topik tertentu. Sentimen dianalisis menggunakan pendekatan berbasis *machine learning* dengan menggunakan model *Bidirectional Encoder Representations from Transformers* (BERT). Berbeda dengan model bahasa lainnya, BERT dibuat sebagai *pre-trained* model yang sudah di *training* secara *deep bidirectional* dari data teks yang tidak berlabel. Pada tahun 2018, Jacob Devlin pertama kali memperkenalkan BERT untuk Search Engine Optimization pada Google agar hasil pencarian sesuai dengan konteks yang di input oleh user. Hingga kini, model *pre-trained* BERT sudah tersedia dalam berbagai bahasa di dunia, termasuk bahasa Indonesia. IndoBERT merupakan *pre-trained* model bahasa Indonesia yang dibangun berdasarkan BERT (Bryan Wilie dkk., 2020).

Setelah melalui tahap *preprocessing*, data *tweet* yang sudah bersih akan diambil sampel secara *random* sebanyak 10% setiap tahunnya. *Random sample* ini akan dilakukan pelabelan secara manual. Pengambilan sampel *tweet* sebanyak 10% bertujuan untuk efisiensi waktu karena total data yang cukup banyak, serta untuk menghindari adanya inkonsistensi bila proses pelabelan manual dilakukan terhadap keseluruhan data. Ukuran *random sample* masing-masing tahun dapat dilihat pada Tabel 4.

Tabel 4. Ukuran *random sample*

Tahun	2020	2021	2022	TOTAL
Ukuran *Random Sample*	570	496	738	**1.804**

Sampel yang sudah diberi label secara manual akan dibagi menjadi 3 jenis data, yaitu data *training*, data *validation*, dan data *testing*. Data *training* digunakan untuk pengembangan model, data *testing* digunakan untuk menguji dan melihat keakuratan model, sedangkan data *validation* digunakan untuk memvalidasi kinerja model dan meminimalisir *overfitting*. Pembagian ketiga data ini memakai metode *stratified*, agar pembagian sesuai dengan proporsi label sentimen pada data tersebut. Pembagian

menggunakan rasio 6:2:2. Rasio ini dipilih berdasarkan penelitian menggunakan BERT model yang dilakukan oleh Zhancheng Ren (2021). Scikit learn merupakan *library* python yang sering digunakan untuk pembagian data. Tabel 5 menunjukkan pembagian data sampel yang telah dilabelkan manual.

Tabel 5. Pembagian data *sample*

Data *Training*	Data *Validation*	Data *Testing*
1.082	361	361

Data tersebut digunakan untuk melakukan beberapa skenario, guna memilih kombinasi *learning rate* dan *batch size* yang optimal untuk membangun model IndoBERT.

Pembagian data tersebut digunakan untuk melakukan beberapa skenario, guna memilih kombinasi *learning rate* dan *batch size* yang optimal untuk membangun model IndoBERT. *Hyperparameter* yang direkomendasikan untuk mendapatkan performa yang optimal adalah menggunakan *batch size* 16 atau 32, *learning rate* sebesar 5e-5, 3e-5, 2e-5 dengan *optimizer* adam dengan *epoch* 2,3, dan 4 (Devlin, 2019). Namun, penelitian ini telah ditetapkan *epoch* sebanyak 5, sebab untuk melihat gambaran iterasi yang lebih luas. Skenario kombinasi *learning rate* dan *batch size* dapat dilihat pada Tabel 6.

Tabel 6. Skenario kombinasi *learning rate* dan *batch size*

Skenario	*Learning Rate*	*Batch Size*
Skenario 1	2e-5	16
Skenario 2		32
Skenario 3	3e-5	16
Skenario 4		32
Skenario 5	5e-5	16
Skenario 6		32

Hasil pelabelan akhir menjadi dasar untuk membentuk tabel *confusion matrix* dalam penghitungan performa nilai akurasi, *precision*, *recall*, dan f1 *score*. *Learning rate* dan *batch size* optimal tersebut akan digunakan dalam model IndoBERT, untuk melakukan analisis sentimen pada seluruh data yang tidak berlabel, secara semi *supervised learning*. Dimana, 10% data yang sudah berlabel akan digunakan untuk melakukan pemodelan terhadap 50% data yang belum berlabel. Selanjutnya, 60% data yang sudah berlabel (10% data berlabel manual ditambah 50% data berlabel dari pemodelan) disebut sebagai *pseudolabel* digunakan untuk memodelkan 40% data yang belum berlabel.

4. Pemodelan Topik

Pemodelan topik berguna untuk menemukan gambaran makna dari dokumen secara semantik berupa topik-topik yang tersembunyi pada teks berukuran besar dan menemukan informasi pada teks yang tidak berstruktur. Pada penelitian ini, dilakukan 2 tahap pemodelan topik. Pertama, mencari nilai *coherence* yang terbesar untuk menentukan jumlah topik. Kedua, melakukan pemodelan topik berdasarkan jumlah topik yang terpilih.

Algoritma model yang digunakan dalam penelitian adalah *Latent Dirichlet Allocation* (LDA). LDA merupakan metode pemodelan topik yang menerapkan distribusi *Dirichlet* untuk menghasilkan distribusi topik tiap dokumen. Kelebihan metode LDA adalah mampu mengidentifikasi topik pada data yang besar dengan tidak

perlu bergantung pada pelabelan secara manual atau tidak memerlukan data latih sebelumnya (Oman Somantri, 2019).

HASIL DAN PEMBAHASAN

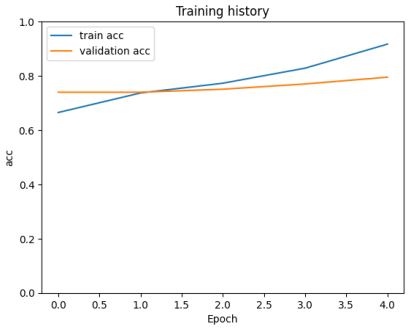
1. Analisis Sentimen

Dari masing-masing skenario kombinasi *learning rate* dan *batch size*, dibuat *confusion matrix* guna menghitung performa nilai akurasi, *precision*, *recall*, dan f1*score*. Skenario yang terbaik adalah skenario yang memiliki akurasi yang tinggi. Selain itu, diperlukan juga nilai *precision* dan *recall* yang tinggi karena dapat menggambarkan *True Positive* dan *True Negative* yang tinggi.

Tabel 7. Performa skenario

Skenario	Data *Training*	*Accuracy*	*Precision*	*Recall*	F1 *score*
Skenario 1	***Testing***	0,76	0,65	0,48	0,49
	Validation	0,78	0,81	0,49	0,49
Skenario 2	***Testing***	0,75	0,60	0,53	0,56
	Validation	0,80	0,71	0,55	0,60
Skenario 3	***Testing***	0,76	0,62	0,49	0,51
	Validation	0,77	0,73	0,49	0,51
Skenario 4	***Testing***	0,73	0,59	0,53	0,51
	Validation	0,72	0,75	0,51	0,48
Skenario 5	***Testing***	0,75	0,58	0,48	0,51
	Validation	0,76	0,60	0,46	0,49
Skenario 6	***Testing***	0,76	0,61	0,47	0,50
	Validation	0,79	0,78	0,48	0,51

Performa masing-masing skenario dapat dilihat pada Tabel 7 dimana skenario yang memiliki nilai accuracy, *precision*, dan *recall* yang tertinggi adalah pada Skenario 2. *Learning* curve dari Skenario 2 dapat dilihat pada Gambar 1 yang terlihat bahwa model sudah *good* fit sebab akurasi data *training* dan data *validation* sudah cukup tinggi, serta terbentuk garis yang mendekati.



Gambar 1. *Learning* curve data sampel

H*yperparameter* yang digunakan pada skenario 2 dapat dilihat pada tabel 8.

Tabel 8. *Hyperparameter* yang digunakan

No.	*Hyperparameter*	Ukuran
1.	*Batch Size*	32
2.	*Learning Rate*	2e-5
3.	*Epoch*	5

Selanjutnya, dilakukan pemodelan IndoBERT dengan menggunakan *hyperparameter* yang terpilih. Model ini diterapkan pada data yang tidak berlabel untuk mendapatkan hasil klasifikasi sentimen secara semi *supervised learning*. Dimana, 10% data yang sudah berlabel akan digunakan untuk melakukan pemodelan terhadap 50% data yang belum berlabel. Selanjutnya, 60% data yang sudah berlabel (10% data berlabel manual ditambah 50% data berlabel dari pemodelan) disebut sebagai *pseudolabel* digunakan untuk memodelkan 40% data yang belum berlabel. Hasil evaluasi metrik terhadap keseluruhan data dan hasil klasifikasi sentiment per tahun dapat dilihat pada Tabel 9 dan 10.

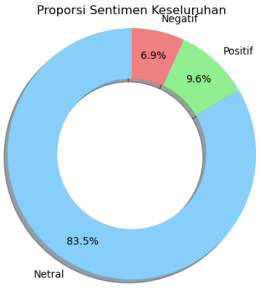
Tabel 9. Hasil evaluasi metrik keseluruhan data

Skenario	Data *Training*	*Accuracy*	*Precision*	*Recall*	F1 *score*
Skenario 2	***Testing***	0,91	0,83	0,74	0,77
	Validation	0,91	0,84	0,73	0,77

Tabel 10. Hasil klasifikasi sentimen per tahun

Sentimen	Tahun			TOTAL
	2020	2021	2022	
Positif	577	467	688	1.732
Netral	4.753	4.114	6.201	15.068
Negatif	370	374	495	1.239

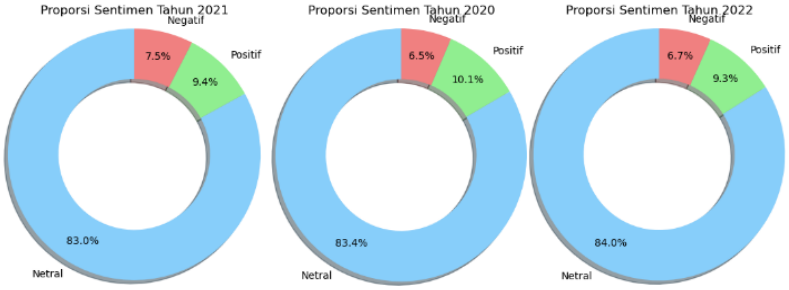
Berdasarkan visualisasi persentase hasil klasifikasi sentimen keseluruhan pada Gambar 2, *tweet* bersentimen netral memiliki kemunculan lebih besar dibandingkan *tweet* bersentimen positif dan negatif. Sentimen netral muncul lebih banyak karena sebagian besar *tweet* mengenai data BPS membahas tentang kegiatan pengumpulan data BPS dan hasil statistik data BPS. Visualisasi persentase hasil klasifikasi sentimen per tahun dapat dilihat pada Gambar 3.



Gambar 2. Persentase klasifikasi sentimen keseluruhan

Secara umum, tahun 2020 hingga 2022 memiliki sentimen netral yang lebih banyak dibandingkan sentimen positif dan negatif. Namun, ada hal yang menarik yaitu sentimen negatif pada tahun 2021 lebih tinggi dibandingkan sentimen negatif tahun 2020 dan 2022. Tingginya sentimen negatif pada tahun 2021 dapat dilihat dari 374 tweet bersentimen negatif pada tahun 2021, sebanyak 109 tweet berkaitan dengan pidato kontroversial dari Presiden Joko Widodo mengenai tidak adanya impor beras sejak 3 tahun. Pidato ini bertentangan dengan data BPS yang menyatakan bahwa pada tahun 2019-2021 terdapat impor beras. Perbedaan ini menimbulkan sentimen negatif dari masyarakat mengenai data BPS.

Pada tahun 2020 memiliki sentimen netral yang lebih banyak dibandingkan sentimen positif dan negatif. Pada bulan Februari 2020 memiliki kenaikan jumlah *tweet* dibandingkan bulan lainnya dan sentimen positif lebih banyak muncul dibandingkan sentimen negatif.



Gambar 3. Persentase klasifikasi sentimen per tahun

Dari 4.753 tweet bersentimen netral pada tahun 2020, sebanyak 1.247 tweet membahas tentang kegiatan BPS yaitu Sensus Penduduk yang dilaksanakan pada bulan Februari 2020. Dari 577 tweet bersentimen positif pada tahun 2020, sebanyak 292 tweet juga membahas tentang Sensus Penduduk pada bulan Februari 2020. Tingginya sentimen positif menunjukkan antusiasme masyarakat terhadap kegiatan pengumpulan data BPS melalui SP2020. Kemudian, dari 370 tweet bersentimen negatif pada tahun 2020, sebanyak 60 tweet membahas tentang perilisan data BPS pada bulan Mei 2020 yang terjadi penurunan pada banyak sektor akibat imbas dari pandemi COVID-19.

Pada tahun 2021 memiliki sentimen netral yang lebih banyak dibandingkan sentimen positif dan negatif. Pada bulan Maret 2021 memiliki kenaikan jumlah *tweet* dibandingkan bulan lainnya, lalu sentimen negatif lebih banyak muncul dibandingkan sentimen positif. Dari 4.114 tweet bersentimen netral pada tahun 2021, sebanyak 721 tweet berkaitan dengan kegiatan BPS yaitu Survei Sosial Ekonomi Nasional (SUSENAS) yang dilaksanakan pada bulan Maret 2021. Lalu, dari 467 tweet bersentimen positif pada tahun 2021, sebanyak 78 tweet membahas tentang ucapan kepada BPS terkait Hari Statistik Nasional pada tanggal 26 September setiap tahunnya. Kemudian, dari 374 tweet bersentimen negatif pada tahun 2021, sebanyak 109 tweet membahas tentang Presiden Joko Widodo yang pada bulan Maret 2021 menyampaikan pidato berbeda dengan hasil data BPS. Perbedaan ini menimbulkan sentimen negatif dari masyarakat mengenai data BPS.

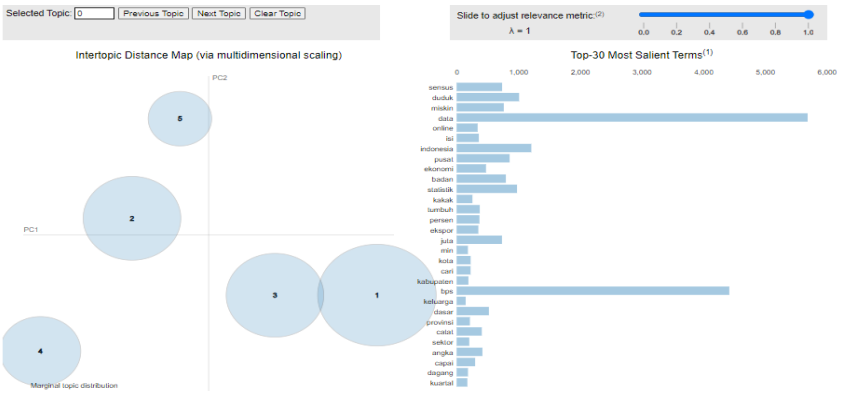
Pada tahun 2022 memiliki sentimen netral yang lebih banyak dibandingkan sentimen positif dan negatif. Kemunculan sentimen netral tertinggi terjadi pada bulan November, sedangkan kemunculan sentimen positif tertinggi terjadi pada bulan September. Dari 6.201 tweet bersentimen netral pada tahun 2022, sebanyak 1.252 tweet membahas tentang kegiatan Registrasi Sosial Ekonomi (Regsosek) yang dilaksanakan pada bulan November 2022. Lalu, dari 688 tweet bersentimen positif pada tahun 2022, sebanyak 89 tweet membahas tentang ucapan kepada BPS terkait Hari Statistik Nasional pada tanggal 26 September setiap tahunnya. Kemudian, dari 495 tweet bersentimen negatif pada tahun 2022, sebanyak 93 tweet membahas tentang pernyataan kontroversial dari salah satu media nasional yaitu Kompas. Pada edisi Senin 20 Juni 2022, Kompas mengabarkan bahwa berdasarkan data BPS DKI Jakarta, angka kemiskinan di Jakarta per September 2021 mencapai 4,67 persen atau mendekati situasi 15 tahun lalu, yaitu pada tahun 2007. Sentimen negatif muncul karena ketidakpercayaan masyarakat (khususnya Warga DKI Jakarta) terhadap data BPS terkait hal ini.

2. Pemodelan Topik

A. Tahun 2020

Nilai *coherence score* tertinggi pada tahun 2020 pada jumlah topik 5 yaitu sebesar 0,4334. Untuk mengetahui apakah jumlah topik sudah optimal, maka digunakan peta jarak intertopik yang dapat dilihat pada Gambar 4 dimana topik

telah menyebar di seluruh kuadran (tidak memusat di satu kuadran). Selain itu, tidak ada lingkaran topik yang saling beririsan/tumpang tindih. Maka, dapat disimpulkan bahwa jumlah topik sudah optimal.



Gambar 4. Visualisasi jarak topik tahun 2020

Tabel 11 menyajikan hasil 5 topik beserta bobot angka pada tiap kata. Bobot kata merupakan nilai probabilitas dari kata dalam dokumen. Semakin tinggi angkanya, maka semakin tinggi kata tersebut mewakili sebuah topik yang didiskusikan.

Tabel 11. Hasil pemodelan topik tahun 2020

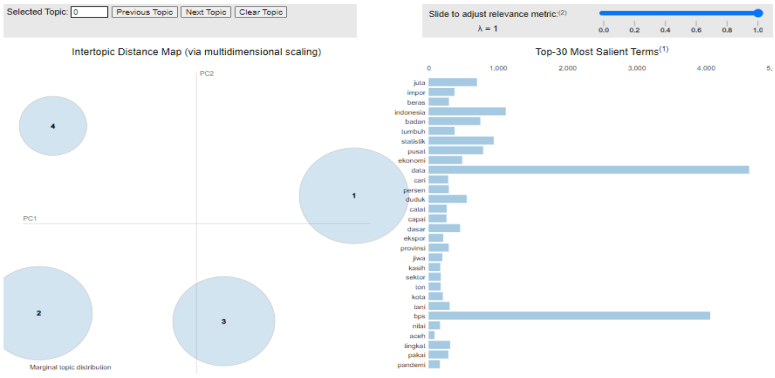
No.	Kata Kunci	Interpretasi
1.	'0.062*"data" + 0.054*"bps" + 0.038*"miskin" + 0.030*"duduk" + 0.022*"statistik" + 0.020*"pusat" + 0.019*"indonesia" + 0.017*"badan" + 0.012*"kota" + 0.010*"kabupaten"'	Indikator kemiskinan
2.	'0.042*"data" + 0.039*"bps" + 0.009*"angka" + 0.008*"miskin" + 0.008*"beras" + 0.007*"luas" + 0.007*"orang" + 0.007*"wisatawan" + 0.007*"lapor" + 0.007*"ngisi"'	Indikator kemiskinan
3.	'0.106*"data" + 0.073*"bps" + 0.012*"cari" + 0.011*"pakai" + 0.007*"lapor" + 0.007*"baca" + 0.006*"tau" + 0.006*"coba" + 0.006*"perintah" + 0.006*"beda"'	Indikator publikasi BPS
4.	'0.051*"bps" + 0.043*"data" + 0.030*"indonesia" + 0.021*"juta" + 0.019*"statistik" + 0.018*"pusat" + 0.017*"badan" + 0.017*"ekonomi" + 0.015*"dasar" + 0.014*"tumbuh"'	Indikator ekonomi
5.	'0.081*"data" + 0.049*"sensus" + 0.033*"duduk" + 0.027*"isi" + 0.027*"*online*" + 0.027*"bps" + 0.020*"kakak" + 0.014*"min" + 0.011*"keluarga" + 0.008*"lengkap"'	Indikator sensus penduduk

Pada topik 1, 2, dan 5, indikator kemiskinan dan indikator sensus penduduk berada pada ragam data Statistik Sosial. Indikator publikasi BPS pada topik 3, berada pada ragam data Metodologi dan Informasi Statistik (MIS). Pada topik 4, indikator ekonomi berada pada ragam data Statistik Distribusi dan Jasa.

B. Tahun 2021

Nilai *coherence score* tertinggi tahun 2021 pada jumlah topik 4 yaitu sebesar 0,4289. Gambar 5 menunjukkan bahwa topik telah menyebar di seluruh kuadran (tidak memusat di satu kuadran). Selain itu, tidak ada lingkaran topik yang saling beririsan/tumpang tindih. Maka, dapat disimpulkan bahwa jumlah topik sudah optimal. Tabel 12. menyajikan hasil 4 topik beserta bobot angka pada tiap kata. Bobot kata merupakan nilai probabilitas dari kata dalam dokumen. Semakin tinggi angkanya, maka semakin tinggi kata tersebut mewakili sebuah topik yang didiskusikan.

P-ISSN: 2622-6901
E-ISSN: 2622-6375



Gambar 5. Visualisasi jarak topik tahun 2021

Pada topik 1, indikator kemiskinan berada pada ragam data Statistik Sosial. Sedangkan indikator ekonomi dan indikator impor pada topik 2 dan 4, berada pada ragam data Statistik Distribusi dan Jasa. Lalu pada topik 3, indikator publikasi BPS berada pada ragam data Metodologi dan Informasi Statistik (MIS). Hasil pemodelan topik pada tahun 2021 menunjukkan bahwa ragam data Statistik Sosial merupakan topik yang paling sering dibicarakan pada *tweet* terkait data BPS.

Tabel 12. Hasil pemodelan topik tahun 2021

No.	Kata Kunci	Interpretasi
1.	'0.067*"data" + 0.052*"bps" + 0.019*"statistik" + 0.015*"duduk" + 0.014*"miskin" + 0.013*"indonesia" + 0.012*"provinsi" + 0.011*"pusat" + 0.011*"badan" + 0.009*"kota"'	Indikator kemiskinan
2.	'0.047*"bps" + 0.037*"data" + 0.033*"indonesia" + 0.030*"juta" + 0.024*"pusat" + 0.024*"statistik" + 0.023*"badan" + 0.019*"ekonomi" + 0.016*"tumbuh" + 0.015*"dasar"'	Indikator ekonomi
3.	'0.091*"data" + 0.069*"bps" + 0.013*"cari" + 0.010*"pakai" + 0.010*"lapor" + 0.008*"kasih" + 0.007*"lihat" + 0.006*"banget" + 0.006*"kayak" + 0.005*"coba"'	Indikator publikasi BPS
4.	'0.045*"bps" + 0.044*"data" + 0.024*"impor" + 0.023*"beras" + 0.010*"aceh" + 0.008*"indonesia" + 0.007*"presiden" + 0.006*"bilang" + 0.005*"orang" + 0.005*"jokowi"'	Indikator impor

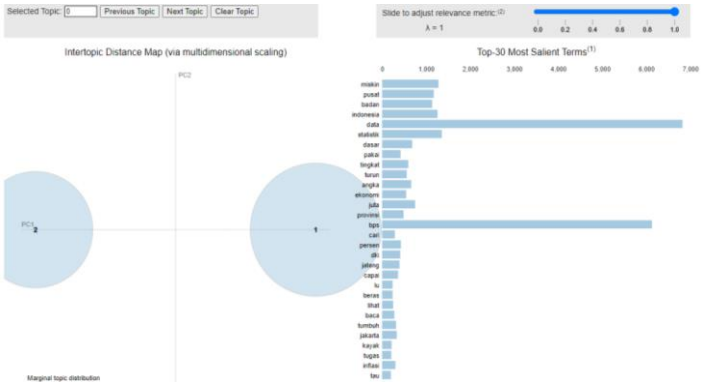
C. Tahun 2022

Nilai *coherence score* tertinggi tahun 2022 pada jumlah topik 2 yaitu sebesar 0,4336. Gambar 6 menunjukkan bahwa topik telah menyebar di seluruh kuadran (tidak memusat di satu kuadran). Selain itu, tidak ada lingkaran topik yang saling beririsan/tumpang tindih. Maka, dapat disimpulkan bahwa jumlah topik sudah optimal. Tabel 13. menyajikan hasil 2 topik beserta bobot angka pada tiap kata. Bobot kata merupakan nilai probabilitas dari kata dalam dokumen. Semakin tinggi angkanya, maka semakin tinggi kata tersebut mewakili sebuah topik yang didiskusikan.

Tabel 13. Hasil pemodelan topik tahun 2022

No.	Kata Kunci	Interpretasi
1.	'0.049*"bps" + 0.047*"data" + 0.021*"statistik" + 0.021*"miskin" + 0.020*"indonesia" + 0.019*"pusat" + 0.018*"badan" + 0.012*"juta" + 0.011*"dasar" + 0.011*"angka"'	Indikator kemiskinan
2.	'0.084*"data" + 0.067*"bps" + 0.009*"pakai" + 0.007*"orang" + 0.006*"cari" + 0.006*"negara" + 0.006*"baca" + 0.005*"lihat" + 0.005*"lu" + 0.005*"beras"'	Indikator pertanian

Pada topik 1, indikator kemiskinan berada pada ragam data Statistik Sosial. Sedangkan indikator pertanian pada topik 2, berada pada ragam data Statistik Produksi. Hasil pemodelan topik pada tahun 2022 menunjukkan bahwa ragam data Statistik Sosial merupakan topik yang paling sering dibicarakan pada *tweet* terkait data BPS.



Gambar 6. Visualisasi jarak topik tahun 2022

KESIMPULAN

Hasil analisis sentimen menggunakan model IndoBERT dengan *hyperparameter* terpilih yaitu *learning rate* sebesar 2e-5 dan *batch size* sebanyak 32, menunjukkan bahwa gambaran respon dan opini masyarakat terkait data BPS melalui Twitter selama periode 2020-2022, mengandung lebih banyak sentimen netral (83,5%) dibandingkan dengan sentimen positif (9,6%) dan negatif (6,9%). Sentimen netral muncul lebih banyak karena sebagian besar *tweet* mengenai data BPS membahas tentang kegiatan pengumpulan data BPS dan hasil statistik data BPS. Dimana hal ini tergolong sentimen netral karena berupa informasi, bukan sentimen respon tertentu. Hasil analisis sentimen pada paper ini memperoleh akurasi sebesar 91%, *precision* sebesar 83%, *recall* sebesar 74%, dan *F1 score* sebesar 77%.

Hasil pemodelan topik pada setiap tahun menghasilkan jumlah topik yang berbeda sesuai dengan nilai *coherence* tertinggi. Pada tahun 2020, jumlah topik yang dihasilkan adalah sebanyak 5 topik, dengan kemunculan topik tertinggi yaitu indikator kemiskinan dan indikator sensus penduduk yang berada pada ragam data Statistik Sosial. Pada tahun 2021, jumlah topik yang dihasilkan adalah sebanyak 4 topik, dengan kemunculan topik tertinggi yaitu indikator kemiskinan berada pada ragam data Statistik Sosial. Kemudian pada tahun 2022, jumlah topik yang dihasilkan adalah sebanyak 2 topik, dengan kemunculan topik tertinggi yaitu indikator kemiskinan berada pada ragam data Statistik Sosial. Hal ini sejalan dengan data statistik Survei Kebutuhan Data 2020-2022 yang menyatakan bahwa ragam data Statistik Sosial merupakan ragam data yang paling banyak dibutuhkan, yaitu masing-masing sebesar 35,51%; 39,13%; dan 43,26% (Badan Pusat Statistik, 2020,2021,2022).

SARAN

Penelitian selanjutnya diharapkan dapat mencari metode analisis sentimen dan evaluasi topik yang lebih baik, agar dapat meningkatkan hasil analisis. Kemudian, penelitian ini dapat dikembangkan dengan menghubungkan hasil penelitian dengan data statistik Survei Kebutuhan Data pada bagian Kepuasan konsumen data BPS berdasarkan dimensi kualitas. Penelitian selanjutnya dapat mempertimbangkan penggunaan emoji sebagai penguat dalam analisis sentimen. Penelitian lebih lanjut diharapkan membuat dashboard guna visualisasi analisis sentimen secara real time.

DAFTAR PUSTAKA

Adriansyah, R. (2020). *Sentiment Analysis tentang Badan Pusat Statistik Berdasarkan Media Online*. Bachelor's thesis, Jakarta: Politeknik Statistika STIS.

Albalawi, R., Yeap, T., and Benyoucef, M. (2020) *Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis*. Frontiers in Artificial Intelligence, vol. 3, no. 42

Aliyah Salsabila, N., Ardhito Winatmoko, Y., Akbar Septiandri, A dan Jamal, A. (2018). "*Colloquial Indonesian Lexicon*", 2018 International Conference on Asian Language Processing (IALP), pp. 226-229.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1(Mlm), 4171–4186.

Li, Z., Fan, Y., Jiang, B., Lei, T. and Liu, W. (2019). *A survey on sentiment analysis and opinion mining for social multimedia*. Multimedia Tools and Applications, 78(6), pp.6939-6967.

Musyarof, Z. (2019). "*Analisis Text Mining terhadap BPS (Badan Pusat Statistik) di Twitter Menggunakan R*" in Seminar Karya Tulis Ilmiah BPS Provinsi Kalimantan Selatan, Kalimantan Selatan.

Palen, L., & Vieweg, S. (2008). *The emergence of online widescale interaction in unexpected events: Assistance, alliance, & retreat*. Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW, 117-126. https://doi.org/10.1145/1460563.1460583

Putri, C., Adiwijaya, & Alfarabi, S. (2020). *Analisis Sentimen Review Film Berbahasa Inggris Dengan Pendekatan Bidirectional Encoder Representations from Transformers*. Jurnal Teknik Informatika dan Sistem Informasi, Vol. 6, No. 2, 181-193.

Ren, Z., Shen, Q., Diao, X., and Xu, H. (2021). "*A sentiment-aware deep learning approach for personality detection from text*". Information Processing & Management, vol. 58, no. 3, Article ID 102532.

Sahria, Y., and Fudholi, D.H. (2020). "*Analysis of Health Research Topics in Indonesia Using the LDA (Latent Dirichlet Allocation) Topic Modeling Method*", Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 4, no. 2, pp. 336-344.

Shalvee, & Sambhav, S. (2020). *Role of mass media and communication during pandemic*. Internasional Journal of Creative Re*search* Thoughts, 8(5), 3786-1790

Somantri, O. and Dairoh, D. (2019) "*Analisis Sentimen Penilaian Tempat Tujuan Wisata Kota Tegal Berbasis Text Mining*," J. Edukasi dan Penelit. Inform., vol. 5, no. 2, pp. 191–196.

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). *IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding*. Proceedings of the 1 st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistic and the 10th International Joint Conference on Natural Language Processing.

Zanini, N., & Dhawan, V. (2015). *Text Mining*. An Introduction to theory and some applications. Re*search* Matters, Re*search*gate.net, 38-44. https://www.re*search*gate.net/publication/304140500_*Text_Mining*_An_introduction_to_theory_and_some_applications