Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.5, October 2025, pp. 2836~2844
e-ISSN: 2548-6861

Utilizing IndoBERT and BERTopic to Explore Public Opinion on BPS Instagram Posts

Ahmad Farhan Anugrah [1]*, Rendy Dwi Agatha [2]**
[1,2] Program Studi DIV Statistika, Politeknik Statistika STIS
Jalan Otto Iskandardinata No.64C Jakarta 13330, Indonesia
212212460@stis.ac.id [1], 212212840@stis.ac.id [2]

Article Info

Article history:

Received 2025-07-20
Revised 2025-09-10
Accepted 2025-09-20

Keyword:

Sentiment Analysis,
IndoBERT,
BERTopic,
Social Media,
Central Statistics Agency,
Public Opinion.

ABSTRACT

This study aims to analyze public sentiment and topics of opinion toward the Central Statistics Agency (BPS) through comments on the Instagram account @bps_statistics. A total of 3,075 comments collected from January 1 to July 24, 2025, were analyzed using the IndoBERT model for sentiment classification and BERTopic for topic modeling. The IndoBERT model was developed using a semi-supervised learning approach, achieving an 88% classification accuracy with high precision and recall across all sentiment categories. The analysis results show that neutral comments dominate (52.78%), followed by negative comments (31.54%) and positive comments (15.69%). Topic modeling on negative sentiment revealed two main issues: distrust of poverty data and preference for international institution indicators such as the World Bank. Positive sentiment reflects appreciation for the quality of statistical data and moral support for BPS. Neutral comments mostly contain informative discussions about socioeconomic conditions and access to digital services. These findings emphasize the importance of improving BPS public communication, particularly in bridging the gap in public perception of official data. The social media-based approach has proven effective as a complement to formal surveys in capturing public opinion in a broad and dynamic manner.

I. PENDAHULUAN

Badan Pusat Statistik (BPS) merupakan lembaga pemerintah non-departemen yang bertanggung jawab langsung kepada Presiden dan memiliki mandat untuk menyelenggarakan kegiatan statistik nasional. Mandat tersebut ditegaskan melalui Peraturan Presiden Republik Indonesia Nomor 86 Tahun 2007 [1], yang menetapkan peran strategis BPS dalam penyediaan data statistik resmi. Dalam menjalankan fungsinya, BPS menetapkan visi tahun 2020–2024, yaitu *"Penyedia Data Statistik Berkualitas untuk Indonesia Maju"*, yang menekankan pentingnya peran data dalam mendukung pembangunan nasional berbasis bukti. Dengan peran sentral ini, kualitas data menjadi landasan utama dalam mewujudkan tata kelola pemerintahan yang efektif dan responsif terhadap dinamika masyarakat.

Untuk memastikan kualitas data yang dihasilkan, BPS menerapkan pendekatan *Quality Gates* yang diatur dalam Peraturan Kepala BPS Nomor 117 Tahun 2023. Pendekatan ini menekankan enam dimensi mutu statistik, yaitu relevansi, akurasi, ketepatan waktu, koherensi, aksesibilitas, dan interpretabilitas. Masing-masing dimensi ini mencerminkan aspek-aspek penting dari keandalan suatu produk statistik dalam menjawab kebutuhan pengguna data. Namun demikian, dalam konteks meningkatnya tuntutan terhadap transparansi dan keterbukaan informasi publik, kepercayaan masyarakat terhadap data resmi yang diterbitkan pemerintah menjadi tantangan tersendiri yang perlu dijawab secara strategis.

Salah satu instrumen evaluasi yang digunakan BPS untuk mengukur kualitas dan kepuasan terhadap layanannya adalah Survei Kebutuhan Data (SKD), yang dilaksanakan secara berkala. Pada tahun 2024, SKD [2] mencatat bahwa sebanyak 98,16% responden menyatakan puas terhadap kualitas data BPS. Angka ini mencerminkan tingkat kepercayaan yang tinggi dari pengguna terhadap informasi statistik yang

dihasilkan BPS. Namun demikian, penting dicatat bahwa survei ini hanya mencakup responden yang menggunakan layanan statistik secara langsung melalui Pelayanan Statistik Terpadu (PST), sehingga belum mencerminkan aspirasi masyarakat luas yang mengakses data melalui media daring.

Keterbatasan jangkauan SKD dalam menjaring opini publik dari kanal digital seperti situs web dan media sosial menimbulkan kebutuhan akan pendekatan alternatif. Salah satu pendekatan alternatif yang relevan adalah memanfaatkan media sosial sebagai sumber data opini publik yang lebih luas, dinamis, dan langsung, mengingat karakteristik media sosial yang memungkinkan masyarakat mengekspresikan pandangan mereka secara spontan dan real-time [3]. Instagram, sebagai salah satu platform media sosial dengan jumlah pengguna yang besar di Indonesia, memberikan ruang bagi publik untuk menyampaikan opini mereka dalam bentuk komentar, tanggapan, atau diskusi terbuka. Hal ini menjadikan media sosial sebagai sarana yang relevan dalam mengukur persepsi publik terhadap lembaga negara, termasuk terhadap data statistik yang dirilis oleh BPS [4].

Media sosial telah berkembang menjadi arena diskusi publik yang memengaruhi persepsi masyarakat terhadap berbagai individual dan organisasi, termasuk lembaga negara dan institusi pemerintah. Dalam era digital yang kian berkembang, institusi pemerintah menggunakan media sosial untuk meningkatkan interaksi dengan masyarakat, menyediakan layanan publik, dan membangun citra positif yang dapat meningkatkan kepercayaan publik [5]. Dalam konteks BPS, keterbukaan dan responsivitas dalam komunikasi data di platform digital berpotensi meningkatkan trustworthiness lembaga, sedangkan miskomunikasi dapat memperlemah kepercayaan masyarakat. Oleh karena itu, integrasi media sosial dalam strategi komunikasi statistik BPS harus dipandang sebagai instrumen strategis dalam membangun dan memelihara kepercayaan publik di era digital.

Studi sebelumnya telah menunjukkan efektivitas media sosial dalam menangkap dinamika opini publik. Nabiilah et al. [6] menemukan bahwa media sosial di Indonesia memainkan peran penting dalam menyalurkan ekspresi publik, baik berupa dukungan maupun kritik terhadap isu-isu kebijakan. Amanda dan Nurmawati [7] juga memanfaatkan Twitter untuk menganalisis sentimen publik terhadap BPS dengan menggunakan model IndoBERT untuk klasifikasi sentimen dan LDA untuk pemodelan topik, dan menemukan bahwa opini netral mendominasi percakapan publik. Penelitian lain oleh Simanjuntak et al. [8] menggunakan algoritma BERTopic dan model IndoBERTweet untuk mengevaluasi komentar terhadap kendaraan listrik di YouTube, dan berhasil mengidentifikasi topik-topik utama dengan nilai koherensi dan keragaman topik yang tinggi.

Selain media sosial, studi oleh Adriansah dan Santoso [9] menunjukkan bahwa berita online juga dapat menjadi sumber valid dalam memahami sentimen masyarakat terhadap data statistik yang dirilis BPS. Mereka menganalisis lebih dari seribu artikel dari media seperti Detik, Kompas, dan Tempo, dan menemukan bahwa sebagian besar opini publik terhadap rilis data BPS bersifat positif dan sesuai dengan realitas lapangan. Hal ini menunjukkan bahwa data digital, baik dari media sosial maupun berita daring, dapat menjadi pelengkap yang kuat bagi survei konvensional dalam membangun citra dan kepercayaan publik. Partisipasi masyarakat dalam akses dan evaluasi data, termasuk melalui media daring, terbukti berkontribusi dalam membangun kepercayaan publik terhadap lembaga pemerintah [10].

Dalam konteks pemodelan topik, pendekatan BERTopic yang dikembangkan oleh Grootendorst [11] menawarkan keunggulan melalui kombinasi antara representasi teks menggunakan *transformer-based embeddings* dan pendekatan *class-based TF-IDF*. Dengan cara ini, BERTopic mampu mengidentifikasi dan menyajikan topik-topik utama dari dokumen dalam skala besar secara lebih akurat dan mudah diinterpretasikan. BERTopic telah terbukti efektif dalam berbagai konteks penelitian, termasuk bidang sosial, pemasaran, dan pemerintahan, karena kemampuannya dalam menyederhanakan kompleksitas data teks tidak terstruktur menjadi wawasan yang bermakna.

Berdasarkan uraian tersebut, penelitian ini bertujuan untuk mengeksplorasi opini publik terhadap data yang dipublikasikan BPS pada tahun 2025 melalui pendekatan analisis sentimen dan pemodelan topik berbasis media sosial Instagram. Penggunaan model IndoBERT memungkinkan klasifikasi sentimen yang kontekstual dan sesuai dengan karakteristik bahasa Indonesia, sedangkan pemanfaatan algoritma BERTopic memungkinkan ekstraksi topik-topik utama dari komentar publik secara otomatis. Dengan pendekatan ini, diharapkan BPS dapat memperoleh gambaran yang lebih komprehensif terhadap persepsi masyarakat, khususnya dari kelompok yang tidak terjangkau oleh survei formal, serta memperkuat strategi komunikasi statistik yang adaptif, transparan, dan berbasis data di era digital.

II. METODE

A. Pengumpulan Data

Pengumpulan dataset dilakukan dengan cara *scraping* dari komentar *Instagram* Badan Pusat Statistik dengan *username* @bps_statistics pada rentang waktu 1 Januari 2025 sampai dengan penelitian ini dilakukan, yaitu 24 Juli 2025. *Scraping* dilakukan menggunakan alat bantu *extension* di *google chrome* yang bernama *IG Comment Exporter* .Ketika komentar selesai diekstrak, maka akan terdapat sebanyak 3683 komentar dengan sentimen berlabel positif, negatif, dan netral.

Setelah proses *scrapping*, selanjutnya dilakukan *filtering* pada komentar-komentar pada postingan yang bersifat kuis kepada masyarakat dimana komentar pada postingan-postingan tersebut cukup banyak.. Selain itu filtering juga dilakukan pada komentar yang hanya berisi emot. Tindakan ini dilakukan atas dasar untuk mengurangi tingkat *noise* pada komentar yang dikumpulkan. Sehingga data komentar akhir yang siap dianalisis berjumlah 3075 komentar.

B. *Preprocessing*

Preprocessing teks pada tahap ini bertujuan untuk menyiapkan atau membersihkan dataset sebelum digunakan dalam proses training. Data yang digunakan untuk training perlu dalam kondisi bersih dan bebas dari gangguan atau *noise*. Dalam konteks komentar, *noise* dapat berupa kata-kata yang tidak relevan seperti stopwords, atau kata-kata yang hanya muncul satu atau dua kali dalam keseluruhan data. Untuk itu, proses *preprocessing* pada penelitian ini meliputi penghapusan URL, *mention*, dan *hashtag*, emoji, tanda baca, penghapusan kata-kata pengisi, penormalan kata, dan perubahan format penulisan menjadi huruf kecil. Penghapusan kata-kata yang tidak memberikan kontribusi signifikan menjadi langkah penting agar proses analisis menjadi lebih optimal, sebagaimana direkomendasikan dalam tahapan preprocessing teks untuk analisis sentimen media social [12].

Selain tahapan dasar tersebut, penelitian ini telah menerapkan *stemming* atau *lemmatization* guna menyatukan variasi kata ke dalam bentuk dasarnya, serta normalisasi kosakata informal atau slang yang kerap muncul dalam komentar media sosial. Proses tokenisasi *subword* berbasis metode *WordPiece* yang selaras dengan arsitektur BERT juga telah diterapkan dengan tujuan untuk meningkatkan representasi *embedding*. Penerapan *preprocessing* lanjutan ini diharapkan dapat memperkuat kualitas input bagi model IndoBERT maupun BERTopic sehingga hasil klasifikasi sentimen dan koherensi topik lebih optimal.

TABEL 1
HASIL PREPROCESSING KOMENTAR

Komentar	Komentar "Cleaned"	Sentimen
Wah keren banget videonya! Terus berkarya ya, ditunggu konten-konten selanjutnya 💪	keren video karya konten konten	Positif
Keren banget infonya kak 🔥	keren info	Positif
Kereeeen👏👏 Semangat selalu pejuang data🙌🔥🔥	keren semangat pejuang	Positif
Ini kita mau percaya siapa..versi luar apa versi dalam negeri😂	percaya siapa versi versi negeri	Negatif
Mengurangi kemiskinan dengan menurunkan standar kemiskinan KOCAK LU @bps_statistics 🤡🤡	mengurangi kemiskinan menurunkan standar kemiskinan kocak	Negatif
20 ribu dapet apaan kocak, bensin aja 2 liter 26	20 ribu kocak bensin 2 liter 26	Negatif
Kalo mau jadi mitra statistik gimana caranya	mitra statistik gimana cara	Netral
Kapan loker mitra dibuka lagi min?	loker mitra dibuka	Netral
Kapan pendaftaran sensus ekonomi di buka kembali kak?	pendaftaran sensus ekonomi buka	Netral

C. *Analisis Sentimen: BERT*

Analisis sentimen, atau *opinion mining*, merupakan aplikasi *text mining* yang bertujuan untuk mengekstrak opini dari data tekstual terkait peristiwa atau topik tertentu. Pendekatan *machine learning* dengan jenis analisis *supervised* digunakan dalam analisis sentimen ini, khususnya dengan memanfaatkan model *Bidirectional Encoder Representations from Transformers* (BERT). Berbeda dengan model bahasa lainnya, BERT dirancang sebagai model prapelatihan (*pre-trained model*) yang telah dilatih secara mendalam dan bidireksional menggunakan data teks tanpa label. Konsep BERT pertama kali diperkenalkan oleh Jacob Devlin pada tahun 2018 untuk optimasi mesin pencari Google, agar hasil pencarian lebih relevan dengan konteks masukan pengguna. Saat ini, berbagai varian model BERT pra pelatihan tersedia untuk berbagai bahasa di dunia, termasuk bahasa Indonesia dengan adanya IndoBERT [13]. Setelah melalui tahap *preprocessing*, data komentar yang sudah bersih akan diambil sampel secara *random* sebanyak 10%. Random sample ini akan dilakukan pelabelan secara manual. Pengambilan sampel komentar 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. Sampel yang telah dilabeli secara manual kemudian akan dibagi menjadi dua set data: data pelatihan (*training*) dan data pengujian (*testing*) dengan proporsi 70% data *training* dan 30% data *testing*. Data *training* berfungsi untuk mengembangkan model dan data *testing* digunakan untuk menguji dan menilai akurasi model. Hasil pelabelan akhir menjadi dasar untuk membangun *confusion matrix* guna menghitung performa model dalam hal akurasi, presisi (*precision*), *recall*, dan *f1 score*. *Learning rate* dan *batch size* yang optimal dari proses ini selanjutnya akan diterapkan pada model IndoBERT untuk melakukan analisis sentimen terhadap seluruh data yang belum berlabel, menggunakan pendekatan *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.

D. *Pemodelan Topik: BERTopic*

Preprocessing teks pada tahap ini bertujuan untuk menyiapkan atau membersihkan dataset sebelum digunakan dalam proses training. Data yang digunakan untuk training perlu dalam kondisi bersih dan bebas dari gangguan atau *noise*. Dalam konteks komentar, *noise* dapat berupa kata-kata yang tidak relevan seperti *stopwords*, atau kata-kata yang

hanya muncul satu atau dua kali dalam keseluruhan data. Untuk itu, proses *preprocessing* pada penelitian ini meliputi penghapusan URL, *mention*, dan *hashtag*, emoji, tanda baca, penghapusan kata-kata pengisi, penormalan kata, dan perubahan format penulisan menjadi huruf kecil. Penghapusan kata-kata yang tidak memberikan kontribusi signifikan menjadi langkah penting agar proses analisis menjadi lebih optimal.

Pemodelan topik berfungsi untuk mengungkap makna semantik yang tersembunyi dalam dokumen berukuran besar, mengidentifikasi topik-topik laten, dan mengekstraksi informasi dari teks tidak terstruktur. Dalam penelitian ini, proses pemodelan topik dilaksanakan dalam dua tahapan: pertama, menentukan jumlah topik optimal dengan mencari nilai koherensi tertinggi; kedua, melakukan pemodelan topik berdasarkan jumlah topik yang telah ditentukan tersebut.

BERTopic merupakan jenis analisis *unsupervised* yang beroperasi dengan mengubah dokumen menjadi representasi numerik yang disebut *embeddings*. Transformasi numerik ini memungkinkan data untuk diproses oleh algoritma pengelompokan dan pemodelan topik. Proses ini secara efektif mengubah kalimat menjadi kumpulan vektor yang dapat digunakan untuk mengidentifikasi kesamaan semantik antar kalimat. Model *embedding* yang dimanfaatkan adalah IndoBERT, yang merupakan model *embedding* Bahasa Indonesia yang dikembangkan melalui metode *transfer learning* dari model BERT. Dalam BERTopic, algoritma *dimensionality reduction* diterapkan untuk mengurangi jumlah dimensi atau fitur dalam data. Ini dilakukan untuk menghindari masalah yang timbul saat bekerja dengan ruang berdimensi tinggi, dengan tujuan mengurangi kompleksitas data dan menghilangkan informasi yang tidak relevan agar visualisasi data menjadi lebih mudah.

Representasi topik dalam BERTopic dibangun menggunakan pendekatan *class-based TF-IDF* (c-TF-IDF), yang merupakan modifikasi dari metode TF-IDF klasik. Pendekatan ini menggabungkan semua dokumen dalam satu klaster menjadi dokumen tunggal, lalu menghitung bobot setiap istilah berdasarkan frekuensinya dalam klaster tersebut dan distribusinya di seluruh klaster. Bobot kata $W_{t,c}$ dihitung dengan rumus:

$$W_{t,c} = tf_{t,c}.log\left(1 + \frac{A}{tf_t}\right)$$

di mana $tf_{t,c}$ adalah jumlah kemunculan kata *t* dalam klaster *c*, tf_t adalah frekuensi kata tersebut di seluruh klaster, dan *A* adalah rata-rata jumlah kata per klaster. Pendekatan ini memungkinkan identifikasi kata-kata yang paling penting dan representatif untuk setiap topik secara keseluruhan, bukan hanya per dokumen [11]. Penggunaan c-TF-IDF dalam BERTopic telah terbukti meningkatkan interpretabilitas topik terutama pada analisis media sosial seperti komentar pengguna di forum daring dan platform digital lainnya [14],[15]. Setelah representasi terbentuk, BERTopic melakukan penggabungan topik-topik yang kurang umum dengan topik yang paling mirip secara semantik, untuk mengoptimalkan jumlah topik yang dihasilkan sesuai kebutuhan pengguna. Relevansi pendekatan BERTopic dalam konteks data statistik di Indonesia juga diperkuat oleh penelitian Simbolon et al. [16] yang mengimplementasikan analisis sentimen dan pemodelan topik terhadap opini publik terkait BPS. Setelah representasi terbentuk, BERTopic melakukan penggabungan topik-topik yang kurang umum dengan topik yang paling mirip secara semantik, untuk mengoptimalkan jumlah topik yang dihasilkan sesuai kebutuhan pengguna. Setelah representasi terbentuk, BERTopic melakukan penggabungan topik-topik yang kurang umum dengan topik yang paling mirip secara semantik, untuk mengoptimalkan jumlah topik yang dihasilkan sesuai kebutuhan pengguna.

III. HASIL DAN PEMBAHASAN

A. Analisis Sentimen

Model sentimen yang dilakukan dengan metode IndoBERT menggunakan *hyperparameter* dengan rincian *epoch* yang berjumlah 3, dan *batch size* untuk *training* 16 dan untuk validasi 64. Selanjutnya pada 10% data yang dilabeli secara *manual* akan dibangun model IndoBERT untuk prediksi sentimen pada 40% komentar yang belum dilabeli.

TABEL 2
METRIK PELATIHAN MODEL PADA SAMPEL 10%

Epoch	***Training Loss***	***Validation Loss***
1	1.1848	1.0619
2	1.0663	0.9423
3	1.0224	0.8448

Output yang diperlihatkan pada tabel 2 menunjukkan tren positif. Baik *training loss* maupun *validation loss* secara konsisten menurun selama 3 *epoch*. Hal ini menunjukkan bahwa model secara efektif belajar dari data dan juga menggeneralisasi dengan baik untuk data baru yang belum terlihat, tanpa tanda-tanda *overfitting* yang signifikan. Kemudian pada tabel 3, model menunjukkan akurasi keseluruhan 59%, dengan performa bervariasi antar kelas: kelas "Positif" memiliki presisi 0.74 dan recall 0.54; kelas "Netral" menunjukkan recall sangat tinggi 0.86 namun presisi rendah 0.52; sementara kelas "Negatif" mencapai presisi sempurna 1.00 namun recall sangat rendah 0.08, mengindikasikan kegagalan model dalam mendeteksi sebagian besar kasus "Negatif" yang sebenarnya, yang menjadi kelemahan utama model ini.

TABEL 3
METRIK KLASIFIKASI SENTIMEN PADA SAMPEL 10%

Ukuran	**Positif**	**Negatif**	**Netral**
Precision	0.74	1.00	0.52
Recall	0.54	0.08	0.86
F1-score	0.62	0.15	0.65
Akurasi	**0.59**		

Hasil prediksi klasifikasi sentimen disimpan pada *data frame,* kemudian model prediksi IndoBERT juga dibangun berdasarkan *data frame* gabungan tersebut (10% data sampel sebelumnya dan 40% sampel baru) untuk memprediksi klasifikasi sentimen pada 50% sisa data yang ada.

TABEL 4
METRIK PELATIHAN MODEL PADA DATA GABUNGAN 50%

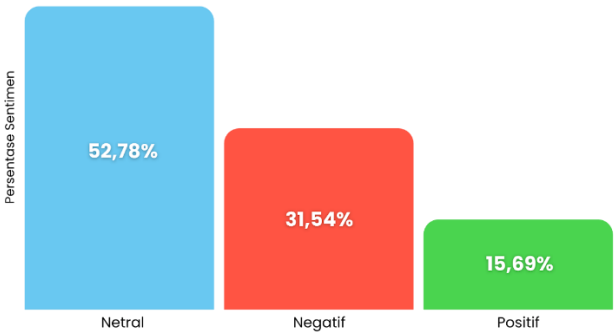
Epoch	*Training Loss*	*Validation Loss*
1	0.7364	0.6209
2	0.3251	0.3314
3	0.2589	0.3139

Secara keseluruhan, proses pelatihan model ini berjalan sangat baik. Model belajar dengan cepat dan efektif, ditunjukkan oleh penurunan konsisten pada *Training Loss* dan *Validation Loss* di setiap *epoch*. Perbedaan yang moderat antara kedua *loss* pada akhir pelatihan menunjukkan bahwa model bergeneralisasi dengan baik dan tidak menunjukkan tanda-tanda *overfitting* yang signifikan. Dibandingkan dengan model prediktif sebelumnya, nilai *loss* yang jauh lebih rendah dan tren yang stabil pada model ini mengindikasikan bahwa model ini memiliki performa prediktif yang jauh lebih unggul. Ini artinya metode *semi-supervised learning* yang diaplikasikan pada penelitian ini terbukti efektif dalam memprediksi sentimen publik.

TABEL 5
METRIK KLASIFIKASI SENTIMEN PADA DATA GABUNGAN 50%

Ukuran	**Positif**	**Negatif**	**Netral**
Precision	0.84	0.87	0.91
Recall	0.90	0.87	0.87
F1-score	0.87	0.87	0.89
Akurasi	**0.88**		

Model klasifikasi sentimen ini menunjukkan kinerja yang luar biasa pada "Data Gabungan 50%". Dengan akurasi 88% dan nilai presisi, recall, serta F1-Score yang tinggi (semua di atas 0.84) untuk setiap kelas, model ini sangat efektif dalam mengklasifikasikan sentimen menjadi positif, negatif, atau netral. Dibandingkan dengan model sebelumnya, peningkatan ini sangat signifikan, terutama dalam mengatasi masalah *recall* yang rendah pada kelas "Negatif" dan presisi yang rendah pada kelas "Netral". Hal ini mengindikasikan bahwa penambahan data atau metode *semi-supervised learning* yang diaplikasikan telah berhasil secara dramatis meningkatkan kemampuan generalisasi dan akurasi model di seluruh spektrum sentimen.



Gambar 1. Persentase Klasifikasi Sentimen Keseluruhan

Berdasarkan hasil analisis terhadap 3.075 komentar yang ditinggalkan publik pada akun Instagram Badan Pusat Statistik (BPS) selama periode 1 Januari hingga 24 Juli 2025, diperoleh distribusi sentimen yang ditunjukkan pada Gambar 1 dan Tabel 5. Hasil ini memberikan gambaran umum mengenai kecenderungan opini publik terhadap konten yang dipublikasikan oleh BPS di media sosial. Gambar 1 menunjukkan bahwa sebagian besar komentar tergolong dalam kategori netral, yaitu sebesar 52,78% dari total komentar. Selanjutnya, komentar bernuansa negatif mencakup 31,54%, sedangkan komentar positif hanya mencapai 15,69%. Proporsi ini mencerminkan bahwa mayoritas publik memberikan tanggapan yang bersifat informatif atau tidak memuat ekspresi emosional yang kuat, sementara komentar negatif masih cukup signifikan dan melampaui jumlah komentar positif.

(a)

(b)

(c)

Gambar 2. Word Cloud
(a) Sentimen Positif, (b) Sentimen Negatif, (c) Sentimen Netral

Hasil visualisasi berupa *Word cloud* yang dihasilkan pada masing-masing kategori sentimen memperkuat temuan analisis sebelumnya. Pada sentimen negatif, dominasi kata seperti "gaji", "orang", "kerja", dan "kemiskinan" menunjukkan adanya ketidakpuasan publik terhadap isu kesejahteraan dan ketidakpercayaan terhadap validitas data kemiskinan yang dipublikasikan BPS, bahkan sering dibandingkan dengan indikator internasional *seperti World Bank*. Sementara itu, *word cloud* sentimen netral menampilkan kata "mitra", "sensus", "kemiskinan", dan "ekonomi", yang mengindikasikan bahwa komentar publik cenderung bersifat informatif dan deskriptif terkait kegiatan statistik maupun kondisi sosial ekonomi tanpa ekspresi emosional yang kuat. Sebaliknya, pada sentimen positif, kemunculan kata "keren", "mantap", "semangat", dan "alhamdulillah" merefleksikan dukungan moral serta apresiasi terhadap kualitas data statistik yang disediakan BPS. Dengan demikian, visualisasi word cloud ini tidak hanya menguatkan hasil analisis IndoBERT dan BERTopic yang menunjukkan dominasi opini netral (52,78%), diikuti sentimen negatif (31,54%) dan positif (15,69%), tetapi juga menegaskan bahwa isu kemiskinan menjadi titik sensitif yang memicu kritik, sementara apresiasi publik lebih banyak diarahkan pada transparansi data dan citra kelembagaan BPS.

TABEL 6
KATA SENTIMEN YANG PALING SERING DIUCAPKAN

Negatif		Positif		Netral	
Kata	*Freq*	Kata	*Freq*	Kata	*Freq*
gaji	144	keren	61	mitra	100
orang	114	kasih	28	kemiskinan	86
kalau	78	terima	26	sensus	82
kerja	78	semangat	25	semoga	71
kemiskinan	68	mantap	22	ekonomi	69

Tabel 6 mendukung temuan ini dengan menampilkan kata-kata yang paling sering digunakan dalam setiap kategori sentimen. Pada sentimen negatif, kata "gaji", "orang", dan "kemiskinan" menjadi kata yang paling dominan. Hal ini menunjukkan adanya kekhawatiran atau kritik publik terhadap isu-isu sosial dan ekonomi yang berkaitan dengan kesejahteraan masyarakat. Kata-kata seperti "kalau" dan "kerja" juga mencerminkan bentuk opini atau saran yang mengandung penilaian kritis. Di sisi lain, komentar positif menampilkan kata seperti "keren", "kasih", dan "terima", yang mengindikasikan adanya apresiasi dan dukungan publik terhadap informasi atau kebijakan yang disampaikan. Kata-kata seperti "semangat" dan "mantap" juga memperlihatkan adanya dukungan emosional dari sebagian pengguna terhadap BPS. Untuk kategori netral, kata "mitra", "kemiskinan", dan "sensus" mendominasi. Hal ini menandakan bahwa banyak komentar yang bersifat informatif, menyebut institusi atau topik spesifik tanpa menyiratkan sikap emosional. Kata-kata seperti "ekonomi" dan "semoga" menunjukkan bahwa sebagian publik menggunakan kolom komentar untuk menyampaikan harapan atau sekadar menanggapi informasi dengan nada netral.

Secara keseluruhan, temuan ini menunjukkan bahwa komunikasi publik yang dilakukan BPS melalui Instagram berhasil menjangkau berbagai respons dari masyarakat. Meskipun sentimen netral mendominasi, adanya proporsi komentar negatif yang cukup besar menandakan pentingnya evaluasi terhadap persepsi publik, khususnya terkait isu kesejahteraan dan kebijakan sosial ekonomi. Porsi komentar positif yang lebih rendah menunjukkan bahwa BPS dapat meningkatkan pendekatan komunikasi publiknya agar lebih menggugah dukungan dan apresiasi dari masyarakat luas.

B. Pengelompokan Topik Sentimen

Analisis kemudian dilanjutkan ke tahap pemodelan topik berbasis sentiment mengenai komentar dan opini publik terhadap postingan *Instagram* di BPS yang dibedakan berdasarkan klasifikasi sentimen positif, negatif, dan netral. Pemodelan topik ini dilakukan dengan pendekatan metode *BERTopic*, dengan penghitungan bobot kontribusi kata dalam suatu dokumen menggunakan ukuran statistik c-TF-IDF (*Class based Term Frequency-Inverse Document Frequency*).

TABEL 7
PEMODELAN TOPIK (NEGATIF)

No	Kata Kunci	Interpretasi
1	0.064*"gaji" + 0.048*"kemiskinan" + 0.042*"orang" + 0.041*"indonesia" + 0.041*"juta" + 0.034*"makan" + 0.032*"sehari" + 0.031*"hidup" + 0.028*"angka" + 0.028*"negara"	Angka Kemiskinan
2	0.476*"bank" + 0.219*"world" + 0.175*"dunia" + 0.143*"akal" + 0.128*"pakai" + 0.127*"parameter" + 0.119*"asn" + 0.113*"masuk" + 0.113*"kocak" + 0.087*"percaya"	Skeptisisme Statistik

Tabel 7 menyajikan hasil pemodelan topik yang diklasifikasikan dalam kategori sentimen negatif. Berdasarkan hasil analisis tersebut, berhasil diidentifikasi dua topik utama yang masing-masing memiliki sekumpulan kata kunci representatif yang mencerminkan isu-isu krusial dalam persepsi publik. Topik pertama ditandai oleh kemunculan kata-kata seperti "gaji", "kemiskinan", "orang", dan "Indonesia". Kemunculan istilah-istilah ini mengindikasikan

bahwa narasi yang berkembang dalam topik ini berkaitan erat dengan kondisi kemiskinan dan kesejahteraan masyarakat. Respon publik tampaknya menunjukkan ketidakpuasan terhadap angka kemiskinan yang dipublikasikan, yang dianggap tidak sesuai dengan kondisi nyata di lapangan. Hal ini menunjukkan adanya kesenjangan antara data statistik resmi dan persepsi masyarakat terhadap realitas sosial-ekonomi yang mereka alami sehari-hari. Topik kedua memunculkan kata-kata seperti "world", "bank", dan "parameter", yang merefleksikan perbandingan publik antara data yang dirilis oleh Badan Pusat Statistik (BPS) dengan data dari lembaga internasional seperti *World Bank*. Narasi dalam topik ini menunjukkan bahwa sebagian masyarakat menganggap parameter atau indikator kemiskinan yang digunakan oleh World Bank lebih logis dan dapat dipercaya dibandingkan dengan parameter yang digunakan oleh BPS.

Meskipun muncul sebagai kelompok dokumen yang berbeda, kedua topik ini memiliki keterkaitan tematik yang kuat. Keduanya mengekspresikan keraguan dan kritik masyarakat terhadap validitas dan reliabilitas data kemiskinan yang dirilis oleh BPS. Puncak diskursus ini terjadi pada tanggal 22 Januari 2025, saat BPS mempublikasikan data garis kemiskinan melalui akun resmi Instagram-nya. Publik merespons unggahan tersebut dengan membandingkan data nasional tersebut dengan data versi *World Bank*, serta menyuarakan opini bahwa data dari BPS dinilai kurang mewakili kondisi riil di masyarakat. Temuan ini menyoroti pentingnya membangun kepercayaan publik terhadap data resmi, serta memperkuat komunikasi publik agar informasi statistik dapat lebih mudah dipahami dan diterima oleh masyarakat luas.

TABEL 8
PEMODELAN TOPIK (POSITIF)

No	Kata Kunci	Interpretasi
1	0.133*"gaji" + 0.126*"lihat" + 0.106*"statistik" + 0.099*"data" + 0.077*"berkualitas" + 0.077*"dokumen" + 0.077*"informatif" + 0.077*"rilis" + 0.066*"kakak" + 0.066*"langsung"	Apresiasi Data Statistik
2	0.093*"keren" + 0.066*"kasih" + 0.063*"terima" + 0.061*"semangat" + 0.057*"semoga" + 0.051*"mantap" + 0.045*"banget" + 0.037*"selamat" + 0.035*"informasi" + 0.034*"bermanfaat"	Dukungan Moral

Pemodelan topik dengan sentimen positif yang dihasilkan pada tabel 8 juga menghasilkan dua topik utama yang menunjukkan aspek positif dalam persepsi publik terhadap konten yang dianalisis. Topik pertama, yang ditandai oleh kata-kata seperti gaji, lihat, statistik, data, berkualitas, dokumen, informasi, dan rilis, mengindikasikan penghargaan publik terhadap keterbukaan dan penyediaan data statistik yang berkualitas. Topik ini dapat diinterpretasikan sebagai apresiasi data statistik, mencerminkan respon positif terhadap ketersediaan informasi yang dianggap bermanfaat dan relevan. Topik kedua mencerminkan ekspresi emosional positif seperti keren, kasih, terima, semangat, selamat, bermanfaat, dan mantap. Dominasi kata-kata yang bersifat emotif dan mendukung menunjukkan bahwa publik juga mengekspresikan bentuk penghargaan moral atau dukungan sosial terhadap pihak yang dianggap berhasil menyampaikan informasi dengan baik. Oleh karena itu, topik ini diinterpretasikan sebagai dukungan moral.

TABEL 9
PEMODELAN TOPIK (NETRAL)

No	Kata Kunci	Interpretasi
1	0.062*"kemiskinan" + 0.059*"ekonomi" + 0.059*"sensus" + 0.059*"indonesia" + 0.047*"garis" + 0.043*"pdb" + 0.041*"statistik" + 0.033*"gaji" + 0.032*"orang" + 0.029*"survey"	Statistik Sosial Ekonomi dan Kemiskinan
2	0.189*"website" + 0.151*"video" + 0.114*"tabel" + 0.099*"diakses" + 0.096*"akses" + 0.091*"postingan" + 0.075*"google" + 0.072*"halo" + 0.071*"penjelasan" + 0.067*"maaf"	Layanan Informasi dan Komunikasi Digital

Tabel 9 menunjukkan pemodelan kategori sentimen netral yang menghasilkan dua topik utama, dimana kedua topik tersebut merefleksikan konten informatif dan deskriptif tanpa kecenderungan emosional kuat. Topik pertama diidentifikasi melalui kata kunci seperti kemiskinan, ekonomi, sensus, gaji, data, statistik, dan survey. Topik ini mengarah pada diskusi objektif terkait kondisi sosial ekonomi dan indikator statistik. Oleh karena itu, interpretasi topik ini adalah "Statistik Sosial Ekonomi dan Kemiskinan, yang menunjukkan bahwa komentar dalam kategori ini bersifat informatif dan faktual tanpa ekspresi emosional. Topik kedua ditandai oleh kata-kata seperti *website*, video, tabel, akses, postingan, penjelasan, dan maaf, yang secara umum menunjukkan interaksi netral publik terhadap media atau platform informasi digital. Keberadaan kata seperti halo dan maaf juga menandakan penggunaan bahasa sopan dan netral dalam komunikasi daring.

Dengan demikian, topik ini diinterpretasikan sebagai Layanan Informasi dan Komunikasi Digital, yang menunjukkan bahwa sebagian komentar bersifat administratif atau berisi permintaan dan pertanyaan teknis.

Evaluasi pemodelan topik melalui *coherence score* pada tabel 10 menunjukkan variasi tingkat koherensi antar sentimen. Pada sentimen negatif, topik yang berkaitan dengan isu gaji dan kemiskinan memperoleh coherence score tertinggi (0.4768), memperkuat temuan sebelumnya bahwa kritik publik lebih konsisten diarahkan pada isu kesejahteraan dibandingkan pada perbandingan dengan lembaga internasional. Pada sentimen positif, nilai koherensi rata-rata relatif lebih rendah (0.3808), namun tetap mencerminkan ekspresi apresiasi publik yang terwujud melalui kata kunci seperti keren, semangat, dan terima kasih.

TABEL 10
NILAI COHERENCE SCORE PEMODELAN TOPIK BERDASARKAN SENTIMEN

Sentimen	Topik	Kata Kunci Dominan	Coherence Score
Positif	1	gaji, lihat, statistik, data	0,3417
	2	keren, kasih, terima, semangat	0,4768
Rata-rata (Tanpa Outlier)			0,4092
Negatif	1	gaji, kemiskinan, orang, indonesia	0,4044
	2	bank, world, dunia, akal	0,3573
Rata-rata (Tanpa Outlier)			0,3808
Netral	1	kemiskinan, ekonomi, sensus, indonesia	0,3007
	2	website, video, tabel, diakses	0,5296
Rata-rata (Tanpa Outlier)			0,4152

Sementara itu, sentimen netral menunjukkan koherensi paling tinggi pada topik yang bersifat informatif dan administratif (0.5296), menegaskan bahwa diskusi publik yang bernuansa netral cenderung lebih terstruktur dan konsisten dibandingkan ekspresi positif maupun kritik negatif. Dengan demikian, coherence score ini tidak hanya menilai kualitas topik yang dihasilkan, tetapi juga memperkuat kesimpulan bahwa isu kesejahteraan menjadi titik utama ketidakpuasan publik, sementara informasi teknis lebih banyak mendominasi percakapan netral. Dengan demikian, *coherence score* yang diperoleh mengindikasikan bahwa kualitas pemodelan topik terbaik terdapat pada sentimen netral, sementara sentimen negatif menghasilkan topik yang cukup representatif meskipun lebih beragam, dan sentimen positif cenderung menampilkan ekspresi apresiasi yang kurang terstruktur secara tematik.

IV. KESIMPULAN

Penelitian ini berhasil mengintegrasikan pendekatan machine learning dalam analisis sentimen dan pemodelan topik untuk memahami opini publik terhadap Badan Pusat Statistik (BPS) melalui platform media sosial Instagram. Dengan menggunakan model IndoBERT untuk analisis sentimen dan BERTopic untuk pemodelan topik, penelitian ini memberikan kontribusi signifikan dalam memahami persepsi masyarakat secara lebih luas, khususnya di luar cakupan responden formal seperti pengguna Layanan Statistik Terpadu (PST).

Hasil analisis sentimen menunjukkan bahwa dari 3.075 komentar yang dianalisis, komentar netral mendominasi (52,78%), diikuti oleh komentar negatif (31,54%), dan positif (15,69%). Penerapan metode *semi-supervised learning* menunjukkan keberhasilan signifikan dalam meningkatkan performa model, dengan akurasi akhir mencapai 88% dan metrik klasifikasi (presisi, *recall*, dan *f1-score*) yang tinggi untuk semua kategori sentimen. Hal ini menegaskan bahwa pendekatan IndoBERT yang diadaptasi melalui pelabelan *semi-supervised* memiliki kapabilitas yang kuat dalam memahami dinamika sentimen publik terhadap lembaga pemerintah. Melalui pemodelan topik, penelitian ini mengungkapkan dua topik utama pada masing-masing kategori sentimen. Sentimen negatif didominasi oleh topik terkait angka kemiskinan dan literasi keuangan, yang mencerminkan ketidakpuasan publik terhadap data kemiskinan BPS dan preferensi terhadap indikator dari lembaga internasional seperti *World Bank*. Sebaliknya, sentimen positif mengangkat topik apresiasi terhadap kualitas data statistik dan dukungan moral terhadap BPS. Sementara itu, sentimen netral cenderung membahas statistik sosial-ekonomi dan layanan informasi digital, menunjukkan adanya respons yang bersifat informatif dan administratif dari masyarakat.

Meskipun opini netral mendominasi, keberadaan opini negatif terhadap data kemiskinan tetap signifikan sehingga perlu ada langkah strategis dalam komunikasi publik BPS. Ketidakpercayaan publik terhadap data resmi berpotensi mengurangi efektivitas statistik sebagai dasar perumusan kebijakan. Oleh karena itu, BPS disarankan untuk meningkatkan transparansi metodologi, memperkuat strategi komunikasi berbasis data, serta memperluas kanal umpan balik publik melalui media sosial agar persepsi masyarakat dapat lebih terakomodasi secara komprehensif. Pemanfaatan analisis big data berbasis media sosial juga penting sebagai sistem pemantauan opini publik berkelanjutan untuk mendeteksi secara dini isu-isu strategis serta memahami dinamika sentimen masyarakat secara real time. Selain itu, pengembangan dashboard monitoring opini publik yang terintegrasi berisi metrik sentimen, topik dominan, intensitas diskusi, dan tren temporal akan mendukung pengambilan keputusan strategis sekaligus menjadi instrumen komunikasi krisis berbasis data. Untuk memperkuat pemahaman masyarakat, BPS juga perlu mengintensifkan kampanye literasi statistik melalui media sosial dengan penyampaian narasi data yang sederhana, transparan, dan kontekstual.

Lebih lanjut, strategi komunikasi publik BPS perlu diarahkan pada pendekatan yang lebih adaptif. Beberapa langkah implementatif yang dapat dipertimbangkan meliputi: (1) pemanfaatan visualisasi data dalam bentuk infografik interaktif atau dashboard digital untuk menyederhanakan informasi statistik yang kompleks; (2) penyelenggaraan dialog publik melalui forum daring atau siaran langsung di media sosial untuk memperkuat transparansi metodologi serta menjawab keraguan publik; dan (3) pengembangan program edukasi statistik yang dikemas dalam bentuk literasi singkat yang mudah dipahami masyarakat awam. Pendekatan-pendekatan ini diyakini mampu meningkatkan pemahaman dan apresiasi publik terhadap data statistik, memperkuat kredibilitas BPS, serta meneguhkan posisinya sebagai lembaga statistik nasional yang adaptif dan relevan di era digital.

DAFTAR PUSTAKA

[1] Badan Pusat Statistik, Peraturan Kepala BPS Nomor 117 Tahun 2023 tentang Quality Gates. BPS RI, 2023.

[2] Badan Pusat Statistik, Analisis hasil Survei Kebutuhan Data (Vol. 16). BPS RI, 2024.

[3] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger, "Social media analytics–Challenges in topic discovery, data collection, and data preparation," Int. J. Inf. Manage., vol. 39, pp. 156–168, 2018, doi: 10.1016/j.ijinfomgt.2017.12.002.

[4] M. Harahap, F. Firman, and R. Ahmad, "Penggunaan social media dan perubahan sosial budaya masyarakat," Edukatif: J. Ilmu Pendidik., vol. 3, no. 1, pp. 135–143, 2021, doi: 10.31004/edukatif.v3i1.252.

[5] W. Kencana, "Peran dan manfaat komunikasi pembangunan pada aplikasi pelacak covid-19 sebagai media komunikasi kesehatan", Commed : Jurnal Komunikasi Dan Media, vol. 5, no. 1, p. 83-95, 2020. https://doi.org/10.33884/commed.v5i1.2495

[6] A. S. Nabiilah and D. R. Rahman, "Classification of Indonesia false news detection using IndoBERT and syntactic features," Procedia Comput. Sci., vol. 219, pp. 433–439, 2023, doi: 10.1016/j.procs.2023.01.189.

[7] M. Amanda and R. Nurmawati, "Analisis sentimen dan pemodelan topik opini publik terhadap Badan Pusat Statistik melalui media sosial Twitter," J. Teknol. Sist. Komput., vol. 11, no. 2, pp. 98–106, 2023.

[8] K. A. Simanjuntak, M. Koyimatu, and Y. P. Ervanisari, "Analisis perubahan opini publik terhadap kendaraan listrik di Indonesia melalui komentar YouTube: Pendekatan topic modeling BERTopic," J. Inov. Kewirausahaan, vol. 1, no. 3, pp. 1–12, 2024, doi: 10.37817/jurnalinovasikewirausahaan.v1i3.

[9] R. Adriansah and I. Santoso, "Analisis sentimen Badan Pusat Statistik berdasarkan media online," in Seminar Nasional Official Statistics 2019: Pengembangan Official Statistics dalam mendukung Implementasi SDG's, 2019, pp. 217–225.

[10] J. W. Campbell, "Public participation and trust in government: Results from a vignette experiment," J. Policy Stud., vol. 38, no. 2, pp. 23–31, 2023, doi: 10.52372/jps38203.

[11] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," arXiv preprint arXiv:2203.05794, 2022. [Online]. Available: https://arxiv.org/abs/2203.05794

[12] O. Martínez-Cámara, M. T. Martín-Valdivia, L. A. Ureña-López, and A. Montejo-Ráez, "Evaluating the effectiveness of text pre-processing in sentiment analysis," Appl. Sci., vol. 12, no. 17, Art. no. 8765, 2022, doi: 10.3390/app12178765.

[13] B. Wilie et al., "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," in Proc. 1st Conf. Asia-Pacific Chapter Assoc. Comput. Linguistics and 10th Int. Joint Conf. Natural Lang. Process., 2020, pp. 843–857. [Online]. Available: https://aclanthology.org/2020.aacl-main.84B.

[14] M. Khodeir and A. Elghannam, "Efficient topic identification for urgent MOOC forum posts using BERTopic and traditional topic modeling techniques," Educ. Inf. Technol., 2024, doi: 10.1007/s10639-024-13003-4.

[15] D. Maier, A. Waldherr, and M. Eismann, "A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to support social science research," Front. Sociol., vol. 7, Art. no. 886498, 2022, doi: 10.3389/fsoc.2022.886498.

[16] S. D. Simbolon, R. N. Syah, and H. Hartono, "A sentiment analysis and topic modelling of the socio-economic registration 2022," in Proc. Int. Conf. Data Sci. Official Stat. (ICDSOS), 2023. [Online]. Available: https://proceedings.stis.ac.id/icdsos/article/view/301