Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.2, April 2026, pp. 1707~1718
e-ISSN: 2548-6861

Forecasting Topic, Word, and Hashtag Popularity on X (Twitter) Using LightGBM for Digital Marketing Optimization

Deannisa Syafira Putri [1]*, Amri Muhaimin [2]*, Mohammad Idhom [3]*
* Sains Data, Universitas Pembangunan Nasional "Veteran" Jawa Timur
22083010062@student.upnjatim.ac.id [1], amri.muhaimin.stat@upnjatim.ac.id [2], idhom@upnjatim.ac.id [3]

Article Info

Article history:

Received 2026-01-11
Revised 2026-02-25
Accepted 2026-04-08

Keyword:

BERTopic,
LightGBM,
Optuna,
Time Series Forecasting,
Social Media Analytics.

ABSTRACT

This study presents a machine learning, based framework to forecast the popularity of topics, words, and hashtags on platform X (Twitter) for data-driven digital marketing optimization. Using one year of Indonesian promotional tweets, BERTopic was applied for topic modeling, while LightGBM optimized with Optuna was used to forecast temporal dynamics based on engineered time series features. Evaluation results using RMSE, MAE, and RMSSE, complemented by comparisons with an ARIMA baseline, indicate that the proposed LightGBM model achieves competitive and consistently superior performance. Despite challenges in predicting word-level spikes caused by noise and event-driven behavior, the model effectively captures underlying trend patterns. The proposed approach supports improved campaign timing, content planning, and ROI, with the full implementation publicly available at Github including detailed documentation of the dataset, preprocessing steps, and experimental pipeline to support reproducibility and further research.

I. PENDAHULUAN

Di era digital, perkembangan teknologi informasi dan internet telah mengubah cara masyarakat berinteraksi dan mengonsumsi informasi, termasuk dalam kegiatan pemasaran. Platform digital, khususnya media sosial, menjadi ruang strategis bagi perusahaan untuk menjangkau audiens secara cepat dan luas [1] [2]. Salah satu platform media sosial yang banyak dimanfaatkan dalam aktivitas pemasaran digital adalah platform X (Twitter), yang memiliki karakteristik penyebaran informasi yang cepat melalui mekanisme *retweet* dan *trending topic* [3]. Berdasarkan data The Global Statistic, platform X digunakan oleh 58,30% pengguna internet di Indonesia, sehingga menempatkannya sebagai platform media sosial keempat paling populer setelah Instagram, Facebook, dan TikTok [4]. Tingginya tingkat penggunaan tersebut menunjukkan potensi besar platform X sebagai sarana pemasaran digital yang efektif.

Perkembangan tersebut turut mengubah strategi *digital marketing* secara signifikan melalui penyebaran konten singkat berbasis teks serta interaksi pengguna yang mencerminkan tingkat keterlibatan audiens [5] [6]. Dalam konteks pemasaran digital, relevansi konten serta perencanaan topik, kata kunci, dan *hashtag* yang tepat berperan penting dalam meningkatkan visibilitas dan jangkauan distribusi konten, sehingga berdampak langsung pada efektivitas strategi *content marketing* [7]. Namun, dinamika percakapan dan tren yang berubah sangat cepat di platform X menyebabkan *digital marketer* mengalami kesulitan dalam mengantisipasi topik, kata, dan *hashtag* yang berpotensi populer di masa mendatang, sehingga strategi pemasaran yang diterapkan cenderung bersifat reaktif dan kurang optimal [8]. Dalam konteks ini, popularitas secara implisit merujuk pada tingginya frekuensi kemunculan suatu topik, kata, atau *hashtag* dalam percakapan pengguna, yang mengindikasikan bahwa isu tersebut sedang menjadi tren dan banyak diperbincangkan, sehingga pemanfaatannya dalam konten yang dipublikasikan berpotensi meningkatkan visibilitas merek serta memperluas jangkauan audiens.

Sejumlah penelitian sebelumnya menunjukkan bahwa analisis tren dan popularitas topik di media sosial dapat memberikan wawasan penting untuk pengambilan keputusan strategis. Penelitian [9] menunjukkan bahwa popularitas *hashtag* dan topik memiliki pola temporal yang dapat dimodelkan dan diprediksi. Selain itu, studi terbaru juga menegaskan bahwa pendekatan *machine learning* mampu

menangkap pola nonlinier dalam data *time series* media sosial dengan lebih baik dibandingkan metode statistik konvensional. Algoritma berbasis *gradient boosting*, seperti LightGBM, terbukti unggul dalam hal efisiensi komputasi dan akurasi prediksi pada data berskala besar dan berdimensi tinggi, termasuk data teks dan frekuensi kemunculan kata [10].

Beberapa penelitian juga mengombinasikan model prediksi dengan teknik optimasi *hyperparameter* untuk meningkatkan performa model. Optuna, sebagai *framework hyperparameter tuning* berbasis Bayesian optimization, telah banyak digunakan dalam penelitian terkini karena kemampuannya mencari konfigurasi parameter optimal secara efisien [11]. Studi [12] menunjukkan bahwa penggunaan Optuna secara signifikan dapat meningkatkan kinerja model *machine learning*, termasuk LightGBM, pada berbagai tugas prediksi. Namun demikian, penelitian yang secara khusus memanfaatkan LightGBM dengan *hyperparameter tuning* Optuna untuk melakukan *forecasting* frekuensi topik, kata, dan *hashtag* pada platform X masih relatif terbatas, terutama dalam konteks penerapannya untuk optimalisasi strategi *digital marketing*.

Berdasarkan *gap* pada penelitian sebelumnya tersebut, penelitian ini bertujuan untuk melakukan prediksi (*forecasting*) jumlah frekuensi topik, kata, dan *hashtag* pada masa depan di platform X (Twitter). Prediksi ini dilakukan dengan memanfaatkan data historis *tweet* guna menangkap pola perubahan popularitas secara temporal. Dengan demikian, penelitian ini diharapkan mampu mengidentifikasi elemen konten yang berpotensi menjadi populer sebelum dipublikasikan, sehingga dapat digunakan sebagai dasar pengambilan keputusan dalam strategi *digital marketing* yang lebih proaktif dan berbasis data.

Dalam penelitian ini, metode yang digunakan untuk melakukan *forecasting* adalah *Light Gradient Boosting Machine* (LightGBM). LightGBM merupakan algoritma *machine learning* berbasis *gradient boosting decision tree* yang dirancang untuk menangani data berskala besar dengan efisiensi komputasi yang tinggi [13]. Keunggulan LightGBM terletak pada kemampuannya dalam menangkap hubungan nonlinier, menangani fitur dalam jumlah besar, serta memiliki performa prediksi yang lebih baik dibandingkan metode *boosting* konvensional [14] [15]. Oleh karena itu, LightGBM banyak digunakan dalam berbagai penelitian prediksi time series dan analisis data media sosial yang bersifat dinamis dan kompleks [10].

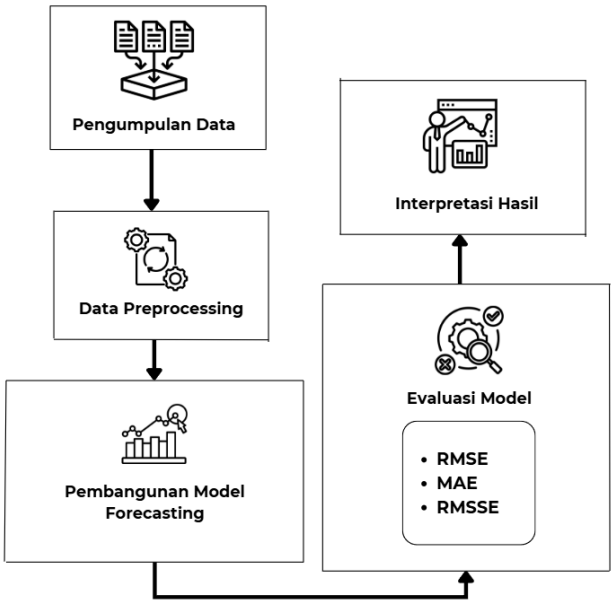
Untuk meningkatkan performa model LightGBM, penelitian ini menerapkan proses *hyperparameter tuning* menggunakan Optuna. Optuna merupakan framework optimasi hyperparameter berbasis Bayesian optimization yang mampu mencari kombinasi parameter terbaik secara adaptif dan efisien [11]. Dibandingkan dengan metode pencarian grid atau random search, Optuna mampu mengurangi waktu komputasi sekaligus meningkatkan akurasi model [16]. Penggunaan Optuna dalam penelitian ini diharapkan dapat menghasilkan konfigurasi LightGBM yang optimal dalam memprediksi frekuensi topik, kata, dan *hashtag* pada platform X.

Dengan mengombinasikan LightGBM dan Optuna, penelitian ini diharapkan dapat memberikan kontribusi metodologis dalam pengembangan model prediksi (*forecasting*) frekuensi topik, kata, dan *hashtag* pada konten media sosial berbasis *machine learning*. Selain itu, hasil penelitian ini diharapkan dapat menjadi referensi bagi penelitian selanjutnya yang berfokus pada *forecasting* tren media sosial, sekaligus memberikan solusi praktis bagi pelaku *digital marketing* dalam mengoptimalkan strategi konten di platform X secara lebih efektif dan terukur.

Manfaat dari penelitian ini tidak hanya bersifat akademis, tetapi juga praktis. Secara praktis, hasil penelitian dapat dimanfaatkan oleh praktisi *digital marketing* sebagai alat bantu pengambilan keputusan dalam menentukan topik, kata, dan *hashtag* yang optimal sebelum melakukan posting *tweet*, sehingga potensi *engagement* dapat dimaksimalkan. Secara akademis, penelitian ini diharapkan dapat memperkaya literatur terkait *forecasting* popularitas konten media sosial dengan pendekatan LightGBM dan Optuna, serta menjadi referensi bagi penelitian selanjutnya yang berfokus pada analisis tren dan optimasi strategi pemasaran digital berbasis data.

II. METODE

Dalam penelitian ini, terdapat beberapa tahapan yang harus diikuti agar proses penelitian dapat berjalan dengan lancar dan sistematis. Gambar 1 menunjukkan diagram alir proses penelitian yang dilakukan.



Gambar 1. Diagram Alir Proses Penelitian

A. Pengumpulan Data

Berdasarkan Gambar 1, penelitian ini dimulai dengan pengumpulan data penelitian yang dikumpulkan melalui *web*

scraping/crawling platform X menggunakan *tools* Tweet Harvest dengan kata kunci "promo OR sale OR diskon OR flash sale" pada *tweet* berbahasa Indonesia dalam periode 27 November 2024–27 November 2025. Proses *scraping* menghasilkan 3.243 *tweet* yang memuat informasi berupa teks *tweet*, tanggal publikasi, serta metrik interaksi pengguna yang meliputi jumlah *retweet*, *like*, balasan, dan kutipan. Seluruh proses pengumpulan dan penggunaan data dilakukan dengan memperhatikan aspek etika penelitian, di mana data yang digunakan bersifat publik, tidak melibatkan identitas pribadi pengguna, serta mematuhi kebijakan dan ketentuan penggunaan data yang berlaku pada platform X.

Setelah tahap pengumpulan data, dilakukan proses rekayasa fitur (*feature engineering*) dengan menambahkan variabel *hashtag list*, yaitu daftar *hashtag* yang muncul pada setiap *tweet*. Variabel ini dibentuk melalui proses ekstraksi *hashtag* dari teks *tweet* dan disimpan dalam bentuk daftar untuk setiap entri data. Hashtag list tersebut kemudian digunakan untuk menghitung frekuensi kemunculan masing-masing *hashtag* pada setiap periode waktu tertentu, yang selanjutnya menjadi dasar dalam proses *forecasting* popularitas *hashtag* di masa depan.

Adapun fitur utama yang digunakan dalam penelitian ini meliputi teks *tweet*, tanggal publikasi, dan *hashtag list*, di mana teks *tweet* dimanfaatkan untuk proses analisis topik dan kata, tanggal digunakan sebagai komponen temporal dalam pembentukan deret waktu, serta *hashtag list* digunakan untuk membangun fitur frekuensi *hashtag* yang akan diprediksi menggunakan model *machine learning*. Dengan kombinasi fitur tersebut, penelitian ini mampu mengintegrasikan informasi tekstual dan temporal secara komprehensif dalam memodelkan dinamika tren topik, kata, dan *hashtag* pada platform X.

B. Data Preprocessing

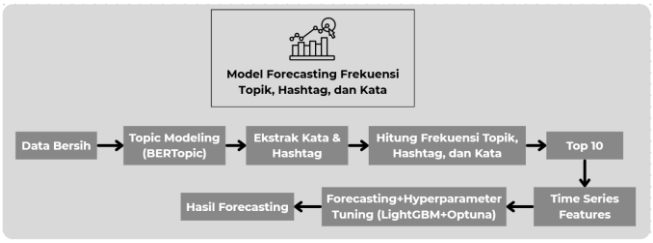
Tahap selanjutnya adalah *preprocessing data* yang bertujuan untuk membersihkan dan mempersiapkan data teks agar siap diproses oleh model. Proses ini diawali dengan pemilihan kolom yang relevan, kemudian dilakukan konversi format tanggal ke dalam tipe *datetime* agar dapat dimanfaatkan pada proses *forecasting* berbasis deret waktu. Selanjutnya, dilakukan pembersihan teks dengan menghapus URL, menormalisasi emoji, serta menghilangkan simbol, angka, dan karakter non-alfabet yang tidak memiliki makna semantik dalam analisis teks.

Pada tahap ini juga dilakukan penyaringan bahasa dengan menghapus *tweet* yang tidak berbahasa Indonesia dan tidak berbahasa Inggris. *Tweet* berbahasa Inggris tetap dipertahankan karena dalam konteks promosi dan pemasaran digital, banyak konten promosi yang menggunakan istilah atau frasa berbahasa Inggris, seperti *sale*, *discount*, atau *flash sale*, yang relevan dengan tujuan penelitian. Untuk menjaga konsistensi bahasa, *tweet* berbahasa Inggris selanjutnya diterjemahkan ke dalam bahasa Indonesia.

Selanjutnya, dilakukan penghapusan kata kunci "promo", "sale", "diskon", dan "flash sale", serta kata dengan panjang kurang dari tiga huruf yang umumnya tidak memiliki makna signifikan, dari teks *tweet*. Penghapusan ini dilakukan karena kata-kata tersebut digunakan sebagai kata kunci pada tahap pengumpulan data, sehingga berpotensi mendominasi representasi teks dan menimbulkan bias pada model. Dengan menghilangkan kata kunci tersebut, model diharapkan mampu menangkap variasi topik dan kata lain yang muncul secara lebih alami dalam konten promosi. Tahapan *preprocessing* juga mencakup pemeriksaan serta penanganan data duplikat dan *missing value* untuk memastikan kualitas dan konsistensi data. Setelah seluruh proses *preprocessing* selesai, jumlah data akhir yang digunakan dalam penelitian ini adalah sebanyak 3.114 *tweet*.

C. Pembangunan Model Forecasting

Tahap pembangunan model *forecasting* frekuensi topik, kata, dan *hashtag* dilakukan dengan pendekatan *time series forecasting* berbasis algoritma LightGBM yang dioptimalkan menggunakan Optuna. Gambar 2 menunjukkan alur kerja dari model yang dibangun.



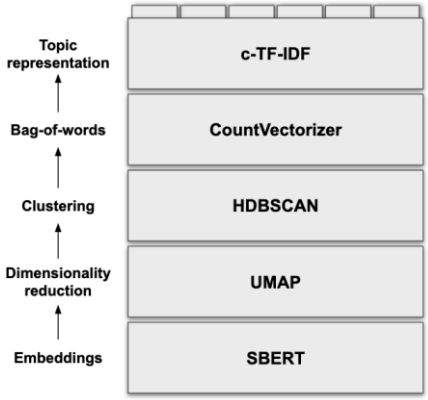
Gambar 2. Alur Kerja Model yang Dibangun

1. Topic Modeling

Berdasarkan Gambar 2, data teks yang telah melalui tahap *preprocessing* dibersihkan kembali untuk memastikan kebersihan dan konsistensi data. Proses ini meliputi normalisasi huruf ke bentuk *lowercase* serta penghapusan kata-kata umum (*stopword*) Bahasa Indonesia. Setelah data dipastikan bersih dan konsisten, dilakukan proses *topic modeling* berbasis BERTopic. BERTopic merupakan metode *topic modeling* modern yang mengombinasikan *pre-trained language* model berbasis BERT dengan teknik *clustering* berbasis kepadatan untuk mengekstraksi topik secara lebih kontekstual. Berbeda dengan pendekatan konvensional seperti LDA yang masih bergantung pada representasi *bag-of-words*, BERTopic mampu mempertimbangkan urutan kata dan makna semantik teks melalui *sentence embeddings*, sehingga lebih sesuai untuk menangani karakteristik data media sosial yang bersifat pendek, informal, dan bervariasi [17].

Secara teknis, BERTopic bekerja dengan menghasilkan representasi vektor dari setiap dokumen menggunakan model *embedding* berbasis BERT. Selanjutnya, vektor tersebut direduksi dimensinya menggunakan *Uniform Manifold Approximation and Projection* (UMAP) untuk meningkatkan efisiensi komputasi serta kualitas proses *clustering*. Selanjutnya, pengelompokan topik dilakukan menggunakan algoritma *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN), yang mampu menentukan jumlah topik secara otomatis tanpa memerlukan

penetapan jumlah topik di awal. Setelah klaster topik terbentuk, representasi setiap topik diekstraksi menggunakan *class-based Term Frequency–Inverse Document Frequency* (c-TF-IDF) guna mengidentifikasi kata-kata yang paling merepresentasikan masing-masing topik. Alur kerja BERTopic yang digunakan dalam penelitian ini ditunjukkan pada Gambar 3. Pendekatan ini dinilai efektif dalam menghasilkan topik yang koheren, khususnya pada data yang bersifat *noisy* dan tidak terstruktur seperti *tweet* [18].



Gambar 3. Alur Kerja BERTopic

Dalam penelitian ini, BERTopic memanfaatkan model *embedding* IndoBERTweet, yang secara khusus dilatih menggunakan data Twitter berbahasa Indonesia [19]. Pemilihan model ini bertujuan untuk menangkap konteks semantik, gaya bahasa informal, serta karakteristik linguistik khas *tweet* secara lebih akurat. Dalam implementasinya, model BERTopic dikonfigurasi dengan *CountVectorizer* menggunakan rentang *n-gram* satu hingga dua kata, ambang frekuensi minimum kemunculan kata, serta batas maksimum kemunculan kata untuk menghindari dominasi kata yang terlalu jarang maupun terlalu umum. Selain itu, parameter *minimum topic size* diterapkan untuk memastikan setiap topik memiliki jumlah dokumen yang memadai, sementara jumlah topik ditentukan secara otomatis. Setelah klaster topik terbentuk, representasi setiap topik diekstraksi menggunakan *class-based Term Frequency–Inverse Document Frequency* (c-TF-IDF) guna mengidentifikasi kata-kata yang paling merepresentasikan masing-masing topik. Setiap *tweet* kemudian diberikan label topik hasil pemodelan, sementara topik yang teridentifikasi sebagai *outlier* dikecualikan dari analisis lanjutan guna menjaga kualitas dan interpretabilitas hasil pemodelan topik.

Sebagai bentuk validasi kualitas topik yang dihasilkan, dilakukan analisis terhadap nilai probabilitas topik yang diberikan oleh BERTopic pada setiap dokumen. Nilai probabilitas ini merepresentasikan tingkat kepercayaan model dalam menetapkan suatu *tweet* ke dalam topik tertentu. Distribusi nilai probabilitas tersebut selanjutnya dianalisis untuk mengevaluasi tingkat konsistensi dan kejelasan pemisahan antar topik, di mana topik dengan nilai probabilitas yang relatif tinggi serta distribusi *confidence* yang stabil menunjukkan kualitas topik yang lebih baik dan lebih representatif terhadap data.

2. Ekstraksi Kata dan Hashtag

Proses selanjutnya, dilakukan ekstraksi kata dari teks *tweet* yang telah melalui tahap pembersihan dengan mengidentifikasi kata-kata yang muncul pada setiap *tweet*. Selain itu, dilakukan ekstraksi *hashtag* berdasarkan variabel *hashtag list* yang telah dibentuk pada tahap sebelumnya. Hasil ekstraksi kata dan *hashtag* tersebut selanjutnya digunakan sebagai dasar untuk analisis frekuensi kemunculan dan pemodelan lanjutan.

3. Perhitungan Frekuensi Kemunculan dan Pemilihan Top 10 Topik, Kata, dan Hashtag,

Selanjutnya, dilakukan perhitungan frekuensi kemunculan masing-masing topik, kata, dan *hashtag* dalam interval waktu harian. Perhitungan frekuensi ini dilakukan dengan mengagregasikan jumlah kemunculan pada setiap periode waktu tertentu sehingga diperoleh pola kemunculan yang bersifat temporal. Berdasarkan hasil perhitungan tersebut, dipilih sepuluh topik, sepuluh *hashtag*, dan sepuluh kata dengan frekuensi tertinggi yang dianggap paling representatif dalam mencerminkan dinamika percakapan selama periode pengamatan.

4. Time Series Feature

Pada tahap ini, deret frekuensi yang dihasilkan sebelumnya kemudian dibentuk menjadi deret waktu kontinu dengan penanganan tanggal kosong melalui pengisian nilai nol agar struktur data tetap konsisten untuk pemodelan deret waktu. Setelah data deret waktu terbentuk, dilakukan pembentukan fitur berbasis waktu (*time-based features*) untuk menangkap pola musiman dan periodik pada data. Fitur-fitur ini meliputi hari dalam seminggu (*day of week*), hari dalam bulan (*day of month*), minggu dalam tahun (*week of year*), bulan (*month*), serta indikator akhir pekan (*is weekend*). Penambahan fitur waktu ini bertujuan untuk membantu model mengenali pola keteraturan temporal, seperti perbedaan aktivitas pengguna pada hari kerja dan akhir pekan atau variasi tren antarbulan

Selanjutnya, dilakukan pembentukan fitur keterlambatan (*lag features*), yaitu fitur yang merepresentasikan nilai frekuensi pada beberapa waktu sebelumnya. Dalam penelitian ini digunakan beberapa nilai *lag*, yaitu satu, dua, tiga, tujuh, dan empat belas hari sebelumnya. Fitur lag ini memungkinkan model untuk mempelajari ketergantungan historis, di mana nilai frekuensi pada suatu waktu dipengaruhi oleh nilai pada periode sebelumnya, sehingga pola dinamika jangka pendek maupun menengah dapat ditangkap dengan lebih baik.

Selain fitur *lag*, ditambahkan pula fitur statistik bergulir (*rolling statistics*) untuk menangkap tren dan stabilitas data dalam jangka waktu tertentu. Pada penelitian ini, statistik bergulir yang digunakan meliputi nilai rata-rata, standar deviasi, minimum, dan maksimum pada jendela waktu 3, 7, dan 14 hari untuk menangkap pola jangka pendek hingga menengah. Statistik bergulir ini berfungsi untuk

merepresentasikan kecenderungan tren, tingkat fluktuasi, serta perubahan ekstrem dalam frekuensi kemunculan topik, kata, dan *hashtag*. Seluruh fitur hasil rekayasa tersebut kemudian digabungkan menjadi satu dataset fitur yang siap digunakan pada tahap pemodelan. Baris data yang mengandung nilai kosong akibat proses pembentukan fitur *lag* dan *rolling* dihapus untuk memastikan kualitas data yang digunakan.

5. Pelatihan Model

Pada tahap ini, data kemudian dibagi secara kronologis (*time-based split*), di mana sebesar 80% data awal digunakan sebagai data pelatihan dan 20% data terakhir digunakan sebagai data pengujian. Selanjutnya, dari data pelatihan tersebut, 20% bagian akhir kembali dialokasikan sebagai data validasi, sehingga proporsi akhir data menjadi sekitar 64% data latih, 16% data validasi, dan 20% data uji. Skema pembagian ini diterapkan untuk menjaga urutan temporal data serta menghindari data *leakage*. Selanjutnya, data dilatih menggunakan model *Light Gradient Boosting Machine* (LightGBM), yang dikombinasikan dengan proses *hyperparameter tuning* menggunakan Optuna untuk meningkatkan kinerja model melalui pencarian otomatis kombinasi parameter terbaik berdasarkan nilai kesalahan prediksi terendah. Proses *hyperparameter tuning* dilakukan menggunakan Optuna dengan pendekatan *Tree-structured Parzen Estimator* (TPE) untuk meminimalkan nilai *Root Mean Squared Error* (RMSE) pada data validasi. Model terbaik kemudian dilatih ulang menggunakan seluruh data historis dan digunakan untuk menghasilkan prediksi frekuensi kata dan *hashtag* selama 30 hari ke depan.

D. Evaluasi Model dan Interpretasi Hasil

Tahap selanjutnya adalah evaluasi kinerja model *forecasting* yang bertujuan untuk mengukur tingkat akurasi prediksi yang dihasilkan oleh model LightGBM. Evaluasi dilakukan menggunakan tiga metrik utama, yaitu *Root Mean Square Error* (RMSE), *Mean Absolute Error* (MAE), dan *Root Mean Squared Scaled Error* (RMSSE). RMSE digunakan untuk mengukur besar kesalahan prediksi dengan memberikan penalti yang lebih besar pada kesalahan ekstrem, sehingga sensitif terhadap deviasi yang signifikan antara nilai aktual dan nilai prediksi [20]. Rumus RMSE dapat dilihat pada persamaan 1.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (1)$$

MAE digunakan untuk mengukur rata-rata kesalahan absolut secara langsung, yang memberikan gambaran umum mengenai selisih prediksi tanpa mempertimbangkan arah kesalahan [20]. Rumus MAE dapat dilihat pada persamaan 2.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

Sementara itu, RMSSE digunakan sebagai metrik evaluasi untuk mengukur tingkat kesalahan prediksi dengan melakukan penskalaan terhadap kesalahan model berdasarkan kesalahan metode peramalan naïf. Metrik ini memungkinkan evaluasi performa model yang lebih stabil dan adil, terutama pada data deret waktu yang memiliki skala berbeda atau mengandung nilai nol, yang sering menjadi kelemahan pada metrik berbasis persentase seperti MAPE [21]. Rumus RMSSE dapat dilihat pada persamaan 3.

$$\mathrm{RMSSE} = \sqrt{\frac{1}{h}\frac{\sum_{t=n+1}^{n+h}(y_t - \hat{y}_t)^2}{\frac{1}{n-1}\sum_{t=2}^{n}(y_t - y_{t-1})^2}} \quad (3)$$

Hasil evaluasi yang diperoleh dari ketiga metrik tersebut kemudian diinterpretasikan untuk menilai kemampuan model dalam menangkap pola temporal pada frekuensi topik, kata, dan *hashtag*. Nilai RMSE, MAE, dan RMSSE yang rendah menunjukkan bahwa model mampu menghasilkan prediksi yang mendekati nilai aktual. Untuk menilai performa model secara komprehensif, evaluasi ini juga dilakukan dengan membandingkan hasil prediksi dengan *baseline* model yaitu ARIMA. Model ARIMA (*Autoregressive Integrated Moving Average*) merupakan metode *time series* klasik yang memprediksi nilai masa depan berdasarkan nilai historis dengan memadukan komponen *autoregression*, *differencing*, dan *moving average* untuk menangkap ketergantungan temporal dalam data [22].

Perbandingan dengan model *baseline* dapat memberikan konteks terhadap keunggulan model yang dikembangkan, dan interpretasi hasil evaluasi ini digunakan sebagai dasar untuk membandingkan performa model antar topik, kata, dan *hashtag*, serta untuk menilai kelayakan model dalam memproyeksikan tren konten di masa mendatang. Dengan demikian, hasil evaluasi juga dapat digunakan sebagai landasan dalam menilai keandalan model sebagai alat pendukung pengambilan keputusan dalam strategi *digital marketing* berbasis data.

III. HASIL DAN PEMBAHASAN

Berdasarkan rangkaian proses penelitian yang dimulai dari pengumpulan data, tahap *preprocessing* teks, pembentukan topik dan fitur deret waktu, hingga pembangunan serta evaluasi model peramalan, bagian ini menyajikan hasil yang diperoleh beserta analisis terhadap kinerja model yang diusulkan.

A. Data

Berdasarkan proses pengumpulan data yang telah dijelaskan pada bagian sebelumnya, penelitian ini

memperoleh data *tweet* melalui proses *scraping* pada platform media sosial X dengan menggunakan kata kunci yang relevan dengan topik penelitian. Tabel I menyajikan ringkasan hasil dari proses pengumpulan data tersebut, yang mencakup jumlah data, rentang waktu pengambilan, serta karakteristik umum tweet yang menjadi dasar dalam analisis selanjutnya.

TABEL I
DATA HASIL SCRAPING

No.	Tanggal	Full Text	…	Verified Status
1.	2024-05-27 00:39:54	CALLING ALL TEUME!!!! Get ready UT Find Your TREASURE launches today! Psstt you can get it at https://t.co/eUz5qYB7up & UNIQLO APP at 9AM! https://t.co/JKqqi1Qona	…	1
2.	2024-05-27 00:39:58	Join our #FINDYOURTREASURE #UTTreasureID Contest and get a chance to meet your favorite TREASURE members live in concert!!! https://t.co/nvFDM66tMp	…	1
…	…	…	…	…
3242.	2025-11-27 09:00:04	Promo Payday TShop kembali! Upgrade smartphone-mu dengan bundling Halo+ bold yang super hemat. Dapat potongan Rp200 ribu plus berbagai benefit hanya di https://t.co/Y13gA9hJdB Promo ini hanya berlaku 25 Nov 5 Des 2025 aja buruan dapetin offer menarik ini! *S&K berlaku https://t.co/PvwlOpM9hn	…	1
3243.	2025-11-27 10:54:50	Mau top up tapi tetap hemat? Bisa banget! Top up di aplikasi Dunia Games atau di https://t.co/hHXoGGgbGy dan bayar pakai LinkAja kamu bisa langsung dapet cashback 5 ribu dan DG Poin! Lebih untung lebih seru~ Top up sekarang karena periode promo ini hanya sampia 30 November https://t.co/DDVEbHkG1k	…	1

Tabel 1 menunjukkan data hasil *scraping* yang dilakukan dalam rentang waktu satu tahun, yaitu mulai 27 November 2024 hingga 27 November 2025. Dari proses pengumpulan data tersebut berhasil dihimpun sebanyak 3.243 *tweet* yang relevan dengan kata kunci penelitian. Data mentah ini masih mengandung berbagai elemen yang tidak terstruktur, seperti *noise*, duplikasi, simbol, serta kata-kata Bahasa yang tidak relevan dengan penelitian. Oleh karena itu, data selanjutnya diproses melalui tahap *preprocessing* sebagaimana telah dijelaskan pada bagian sebelumnya. Hasil dari proses *preprocessing* tersebut disajikan pada Tabel II, yang menunjukkan data teks yang telah bersih dan siap digunakan pada tahap analisis lanjutan, seperti pemodelan topik, ekstraksi kata dan *hashtag*, serta proses *forecasting*.

TABEL II
DATA BERSIH

No.	Tanggal	*Final Tweet*	Hashtag List
1.	2024-05-27 00:39:54	UT CALLING ALL TEUME FREASURE YOUR REASATIONES today Psstt you can get it at amp UNIQLO APP at 9AM	[]
2.	2024-05-27 00:39:58	Bergabunglah dengan kami untuk mengikuti kontes TTREAsureID menemukan kami dan mendapatkan kesempatan untuk bertemu dengan anggota TREASREURE yang Anda sukai hidup dalam konser	['#FINDYOURTREASURE', '#UTTreasureID']
…	…	…	…
3113.	2025-11-27 09:00:04	Promo Payday TShop kembali Upgrade smartphonemu dengan bundling Halo bold yang super hemat Dapat potongan Rp200 ribu plus berbagai benefit hanya di Promo ini hanya berlaku 25 Nov 5 Des 2025 aja buruan dapetin offer menarik ini SampK berlaku	[]
3114.	2025-11-27 10:54:50	Mau top up tapi tetap hemat Bisa banget Top up di aplikasi Dunia Games atau di dan bayar pakai LinkAja kamu bisa langsung dapet cashback 5 ribu dan DG Poin Lebih untung lebih seru Top up sekarang karena periode promo ini hanya sampia 30 November	[]

Tabel II menunjukkan data bersih hasil *preprocessing* yang telah dilakukan pada data *tweet* hasil *scraping*. Pada tahap ini, setiap *tweet* telah melalui proses pembersihan teks, termasuk penghapusan URL, simbol, angka, serta karakter non-alfabet, normalisasi kata, dan penyesuaian bahasa. Kolom Tanggal menunjukkan waktu unggahan *tweet* yang telah dikonversi ke format *datetime* untuk mendukung analisis berbasis deret waktu. Kolom *Final Tweet* berisi teks *tweet* yang telah dibersihkan dan dinormalisasi sehingga lebih representatif terhadap makna asli dan siap digunakan pada tahap analisis lanjutan. Sementara itu, kolom *Hashtag List* memuat daftar *hashtag* yang berhasil diekstraksi dari setiap *tweet* dan disajikan dalam bentuk *list*. Data bersih ini menjadi dasar utama dalam proses pemodelan topik, ekstraksi kata dan *hashtag*, serta pembangunan model *forecasting* frekuensi pada tahap selanjutnya.

B. Hasil Topic Modeling

Sebelum dilakukan proses *forecasting* frekuensi, penelitian ini terlebih dahulu menerapkan *topic modeling* untuk mengidentifikasi topik utama pada setiap *tweet*. Hasil *topic modeling* ini digunakan sebagai dasar dalam penentuan label topik, yang selanjutnya dianalisis dan diprediksi tren frekuensinya untuk periode 30 hari mendatang. Pada bagian ini akan dijelaskan hasil *topic modeling* yang dilakukan menggunakan metode BERTopic. Tabel III, Tabel IV, dan Tabel V menyajikan hasil *topic modeling* yang menunjukkan jumlah topik yang terbentuk, distribusinya, dan statistika probabilitasnya dari data *tweet* yang dianalisis.

TABEL III
HASIL DISTRIBUSI TOPIK

Topik	Kata Kunci	Jumlah Tweet	Rata-rata Probabilitas
0	shopee, sekarang, cek	2265 (72.7%)	78.28%
Outlier	-	640 (20.6%)	0%
1	100rb gratis, rp9000, kupon	97 (3.1%)	44.39%
2	teri, hai, kuch	75 (2.4%)	47.97%
3	akbar, 1000, nanti	11 (0.4%)	73.12%
4	mini, tengah, alhamdulillah	9 (0.3%)	74.39%
5	sol, menit terakhir, terjual	9 (0.3%)	77.63%
6	dom, jungkook, jakarta	8 (0.3%)	82.01%

Berdasarkan hasil *topic modeling* pada Tabel III, model BERTopic berhasil mengidentifikasi 7 topik utama dengan dominasi mutlak pada Topic_0 (shopee, sekarang, cek) yang mencakup 2.265 *tweet* (72,7%). Dominasi ini, didukung oleh *average probability* 78,28%, menegaskan bahwa percakapan di platform X berpusat pada pesan *hard-selling* dengan urgensi tinggi. Dari perspektif digital marketing, ini adalah sinyal kuat untuk memprioritaskan konten berbasis *Call-to-Action* (CTA) yang repetitif namun taktis. Strategi konten harus fokus pada kata kunci "cek" dan "sekarang" untuk memicu psikologi kelangkaan (*scarcity*) di mata audiens, mengingat sebagian besar interaksi terjadi pada narasi transaksional yang bersifat langsung.

Selain topik dominan, terdapat Topic_1 yang berkaitan dengan kata kunci "100rb", "gratis", "rp9000", dan "kupon", dengan 97 *tweet* (3,1%) serta *average probability* 44,39%. Topik ini merepresentasikan strategi promosi berbasis insentif harga langsung, seperti diskon nominal dan kupon. Meskipun proporsinya relatif kecil, keberadaan topik ini menegaskan bahwa sebagian audiens merespons konten promosi yang menonjolkan penghematan secara eksplisit. Dari perspektif *digital marketing*, temuan ini menunjukkan pentingnya mengombinasikan pesan urgensi dengan penawaran harga yang konkret untuk meningkatkan daya tarik dan konversi.

Topik lainnya seperti Topic_2 dengan 75 *tweet* (2,4%) dan Topic_3 dengan 11 *tweet* (0,4%) mencerminkan dimensi interaksi sosial dan personalisasi bahasa dalam percakapan promosi. Meskipun volumenya terbatas, topik-topik ini menunjukkan bahwa promosi *e-commerce* di X tidak selalu disampaikan secara formal, melainkan sering dipadukan dengan gaya bahasa santai, sapaan personal, atau referensi figur tertentu. Dalam konteks *digital marketing*, hal ini mengindikasikan potensi *humanized content dan community engagement* untuk meningkatkan kedekatan emosional audiens terhadap merek.

Beberapa topik minor lainnya, seperti Topic_4, Topic_5, dan Topic_6, masing-masing memiliki proporsi di bawah 0,5%, namun menunjukkan *average probability* yang tinggi (di atas 74%). Kondisi ini menandakan bahwa meskipun topik-topik tersebut hanya mencakup sedikit *tweet*, struktur semantiknya sangat koheren dan merepresentasikan segmen audiens yang spesifik (*niche*). Dari sudut pandang pemasaran digital, topik-topik bernilai kecil namun koheren ini dapat dimanfaatkan untuk *micro-targeting*, kampanye berbasis komunitas, serta personalisasi pesan promosi yang lebih presisi dibandingkan pendekatan *mass marketing*.

Model juga mengidentifikasi 640 *tweet* (20,6%) sebagai *outlier* dengan *average probability* 0%, yang menunjukkan bahwa *tweet-tweet* tersebut tidak memiliki kedekatan semantik yang cukup kuat dengan topik mana pun. Keberadaan *outlier* dalam jumlah yang relatif besar merupakan fenomena yang umum pada penerapan *topic modeling* berbasis BERTopic, khususnya pada data X yang bersifat pendek (*short-text*), tidak terstruktur, dan memiliki keragaman konteks yang tinggi [17]. Selain itu, penggunaan HDBSCAN dalam BERTopic dapat mengklasifikasikan hingga lebih dari 70% dokumen sebagai *outlier* pada data teks pendek lintas domain, sehingga proporsi *outlier* yang tinggi tidak serta-merta menunjukkan kelemahan model, melainkan karakteristik alami dari data dan algoritma klasterisasi yang digunakan [23]. Dalam konteks *digital marketing*, kelompok *outlier* ini justru berpotensi merepresentasikan eksperimen narasi promosi, humor, atau komunikasi implisit yang dapat dieksplorasi lebih lanjut sebagai sumber inovasi konten.

TABEL IV
STATISTIKA PROBABILITAS

Statistik Probabilitas	Nilai
Rata-rata probabilitas (non-*outlier*)	76%
Median probabilitas (non-*outlier*)	92.21%
Probabilitas minimum	0.27%
Probabilitas maksimum	100%

TABEL V
DISTRIBUSI CONFIDENCE

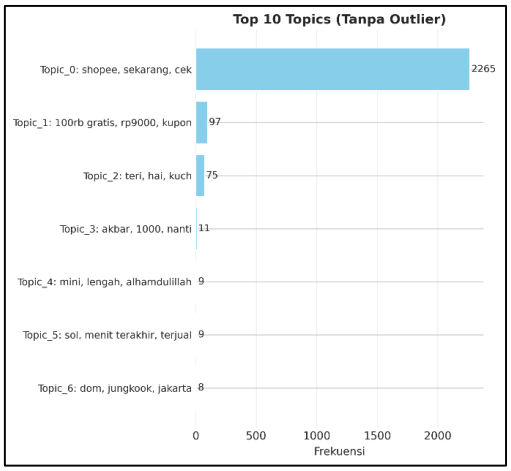
Tingkat *Confidence*	Rentang Probabilitas	Jumlah Tweet
High confidence	>50%	1948 (78.7%)
Medium confidence	30-50%	234 (9.5%)
Low confidence	<30%	292 (11.8%)

Dari sisi kualitas pemodelan, Tabel IV menunjukkan bahwa rata-rata probabilitas *tweet* non-*outlier* mencapai 76,00% dengan median 92,21%, yang menandakan tingkat kepercayaan klasifikasi yang sangat kuat. Distribusi

kepercayaan yang ditunjukkan pada Tabel V juga memperkuat hasil ini, dengan 78,7% *tweet* berada pada kategori *high confidence* (≥50%), sementara hanya 11,8% yang termasuk *low confidence* (<30%). Temuan ini menunjukkan bahwa struktur topik yang dihasilkan relatif stabil dan dapat diandalkan sebagai dasar optimalisasi strategi *digital marketing*, khususnya dalam menentukan fokus tema konten, penjadwalan promosi, serta perancangan pesan yang paling efektif untuk menarik perhatian dan mendorong konversi audiens.

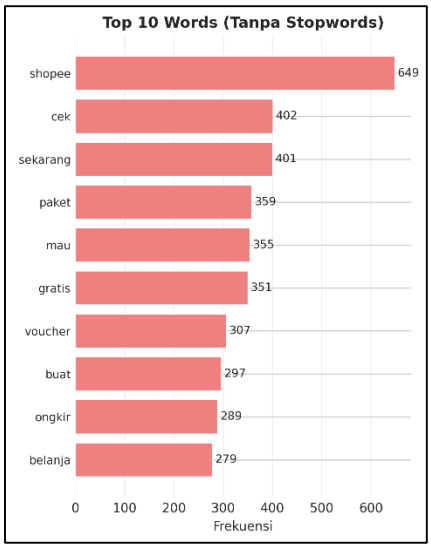
C. Hasil Pemilihan Top 10 Topik, Kata, dan Hashtag

Pada bagian ini disajikan hasil pemilihan *top* 10 Topik, Kata, dan *Hashtag* berdasarkan hasil *topic modeling* dan perhitungan frekuensi setiap kategori. Sepuluh topik, kata, dan *hashtag* dengan frekuensi tertinggi selama periode pengamatan dipilih sebagai representasi utama dinamika percakapan promosi di platform X. Visualisasi hasil analisis tersebut disajikan pada Gambar 4, Gambar 5, dan Gambar 6.



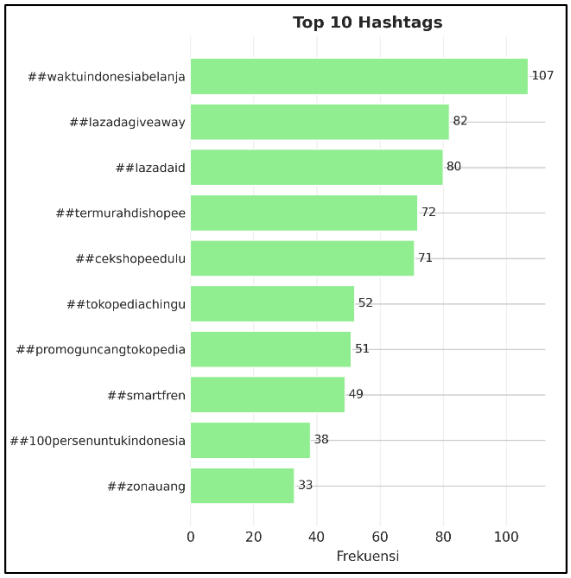
Gambar 4. *Top 10 Topics*

Berdasarkan Gambar 4, distribusi frekuensi topik menunjukkan ketimpangan yang sangat kuat, di mana Topic_0 dengan kata kunci “shopee”, “sekarang”, dan “cek” mendominasi secara signifikan dengan 2.265 *tweet*, jauh melampaui Topic_1 yang hanya mencakup 97 *tweet*, serta topik-topik lain yang seluruhnya berada di bawah 100 *tweet*. Pola ini menegaskan bahwa percakapan promosi di platform X sangat terpusat pada kampanye salah satu platform *e-commerce* yang bersifat transaksional dan berbasis urgensi waktu, sehingga strategi *digital marketing* yang menekankan *call-to-action*, momentum, dan dorongan instan terbukti paling efektif dalam menarik perhatian audiens. Di sisi lain, keberadaan topik-topik minor yang membentuk pola *long-tail* mencerminkan segmen audiens *niche* yang, meskipun volumenya kecil, memiliki koherensi tema yang jelas dan berpotensi dimanfaatkan untuk strategi *targeting*, personalisasi pesan, serta kampanye berbasis komunitas, sehingga optimalisasi pemasaran digital dapat dicapai melalui kombinasi fokus pada topik dominan dan eksploitasi topik minor secara lebih terarah.



Gambar 5. Top 10 Words

Berdasarkan Gambar 5, kata “shopee” muncul paling dominan dengan frekuensi 649, diikuti oleh “cek” (402), “sekarang” (401), “paket” (359), dan “mau” (355), yang mencerminkan gaya komunikasi promosi yang langsung, persuasif, dan berorientasi aksi di platform X. Selain itu, tingginya kemunculan kata “gratis” (351), “voucher” (307), dan “ongkir” (289) menunjukkan bahwa percakapan sangat menekankan nilai ekonomis melalui insentif harga dan kemudahan transaksi. Kombinasi penggunaan kata-kata tersebut menegaskan bahwa strategi *digital marketing* di X paling efektif ketika mengombinasikan penyebutan merek yang kuat, bahasa ajakan instan, dan penawaran konkret, sehingga tidak hanya meningkatkan *awareness*, tetapi juga mendorong klik, niat beli, dan potensi konversi secara lebih langsung.

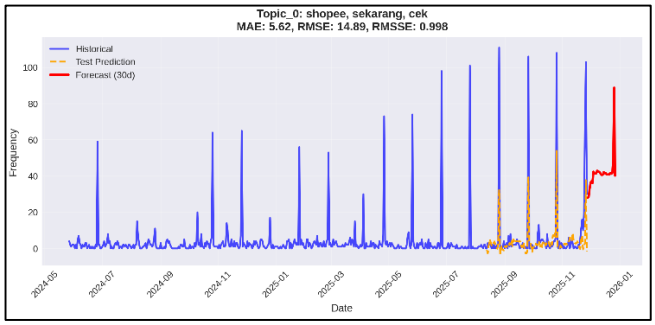


Gambar 6. *Top 10 Hashtags*

Berdasarkan Gambar 6, *hashtag* #waktuindonesiabelanja mendominasi dengan frekuensi tertinggi (107), menunjukkan kuatnya percakapan promosi yang dipicu oleh kampanye berbasis momentum waktu dan *event* nasional. *Hashtag* lain seperti #lazadagiveaway (82), #lazadaid (80), #termurahdishopee (72), dan #cekshopeedulu (71) mencerminkan intensitas aktivitas pemasaran sekaligus persaingan antar platform *e-commerce* dalam meningkatkan visibilitas konten. Sementara itu, kemunculan *hashtag* seperti #tokopediachingu (52), #promoguncangtokopedia (51), dan #smartfren (49) menunjukkan praktik *branding* dan kolaborasi lintas merek yang menyasar segmen audiens tertentu. Secara keseluruhan, temuan ini menegaskan bahwa pemilihan *hashtag* yang relevan, konsisten, dan selaras dengan kampanye promosi berkontribusi signifikan terhadap strategi *digital marketing* di platform X melalui peningkatan jangkauan, *engagement*, dan efektivitas distribusi konten.

D. Hasil Pembangunan Model

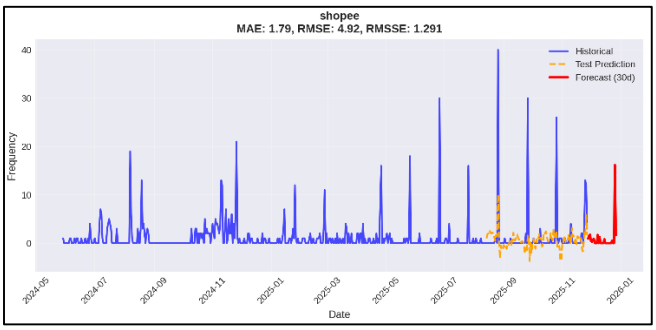
Berdasarkan hasil identifikasi *Top* 10 frekuensi topik, kata, dan *hashtag*, selanjutnya dibangun model *forecasting* menggunakan LightGBM yang dioptimalkan dengan Optuna, secara terpisah untuk masing-masing topik, kata, dan *hashtag* guna memprediksi tren frekuensi kemunculannya pada periode 30 hari mendatang. Namun, mengingat keterbatasan jumlah halaman, visualisasi dan pembahasan hasil *forecasting* ini difokuskan pada komponen dengan frekuensi tertinggi dari tiap kategori, yaitu Topic_0 (Gambar 7), kata "shopee" (Gambar 8), dan *hashtag* #waktuindonesiabelanja (Gambar 9).



Gambar 7. Hasil Forecasting Frekuensi Topic 0

Gambar 7 menampilkan hasil *forecasting* frekuensi Topic_0 menggunakan model LightGBM, yang menunjukkan pola historis dengan fluktuasi tajam dan lonjakan ekstrem pada waktu-waktu tertentu. Lonjakan berulang pada data historis (garis biru) mengindikasikan adanya aktivitas promosi atau kampanye *e-commerce* berskala besar yang bersifat periodik dan berbasis momentum, sehingga frekuensi topik ini memiliki volatilitas tinggi. Prediksi pada data uji (garis oranye) secara umum mampu mengikuti arah tren menjelang akhir periode pengamatan, sementara hasil peramalan 30 hari ke depan (garis merah) menunjukkan kecenderungan peningkatan frekuensi, yang mengindikasikan potensi berlanjutnya intensitas percakapan promosi *e-commerce* di platform X.

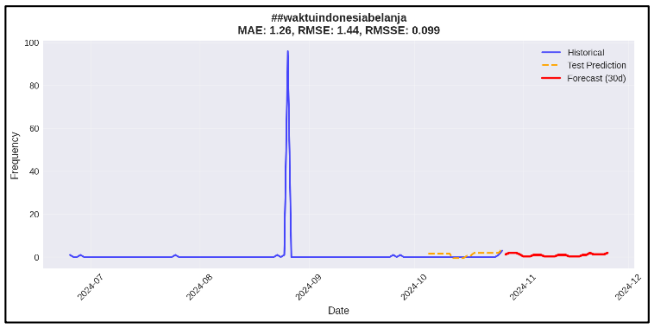
Berdasarkan evaluasi model, diperoleh nilai MAE sebesar 5,62, RMSE sebesar 14,89, dan RMSSE sebesar 0,998, yang menunjukkan bahwa kesalahan prediksi relatif dipengaruhi oleh keberadaan lonjakan ekstrem pada data. Meskipun demikian, nilai RMSSE yang mendekati satu mengindikasikan bahwa performa model masih kompetitif dibandingkan pendekatan *naive*, terutama dalam menangkap arah tren secara umum. Dari perspektif strategi *digital marketing*, hasil ini bermanfaat untuk perencanaan waktu kampanye dan alokasi konten promosi, dengan fokus pada antisipasi periode lonjakan percakapan, meskipun prediksi nilai absolut pada puncak ekstrem tetap perlu diinterpretasikan secara hati-hati.



Gambar 8. Hasil Forecasting Frekuensi Kata "shopee"

Gambar 8 menampilkan hasil *forecasting* frekuensi kata "shopee" menggunakan model LightGBM. Data historis (garis biru) menunjukkan volatilitas tinggi dengan lonjakan tajam (*spikes*) pada tanggal-tanggal tertentu yang berkorelasi dengan momentum promosi besar, seperti *Double Day* atau *Payday Sale*. Hasil prediksi pada data uji (garis oranye) menunjukkan model cukup sensitif dalam menangkap fluktuasi, didukung oleh nilai MAE 1,79 dan RMSE 4,92 yang menunjukkan deviasi prediksi masih dalam rentang wajar. Meskipun nilai RMSSE 1,291 mengisyaratkan adanya tantangan dalam memprediksi puncak lonjakan ekstrem, model ini tetap efektif dalam memetakan tren dasar percakapan publik di media sosial.

Untuk strategi *digital marketing*, pola lonjakan yang berulang namun temporer ini menunjukkan bahwa audiens sangat reaktif terhadap *event-driven marketing*. Tim pemasaran sebaiknya memusatkan anggaran iklan dan kampanye intensif hanya pada jendela waktu puncak (puncak garis biru) untuk memaksimalkan ROI (*Return on Investment*), daripada melakukan kampanye merata setiap hari. Peramalan 30 hari ke depan (garis merah) yang menunjukkan tren stabil dengan potensi kenaikan di akhir periode mengindikasikan perlunya persiapan stok konten yang relevan, dari segi kata kunci yang diprediksi, tepat sebelum puncak frekuensi terjadi. Pendekatan ini memastikan konten tidak hanya muncul di waktu yang tepat secara algoritma, tetapi juga memiliki relasi yang kuat dengan topik yang tengah hangat diperbincangkan, sehingga memperbesar peluang *engagement* organik.



Gambar 9. Hasil *Forecasting* Frekuensi *Hashtag* #waktuindonesiabelanja

Gambar 9 menyajikan hasil *forecasting* frekuensi *hashtag* #waktuindonesiabelanja menggunakan model LightGBM. Data historis (garis biru) menunjukkan karakteristik *sparse* dengan satu lonjakan ekstrem yang sangat tajam, mengonfirmasi bahwa *hashtag* ini bersifat *highly event-driven* dan bergantung pada momentum kampanye besar spesifik. Evaluasi model menunjukkan nilai MAE 1,26 dan RMSE 1,44, serta nilai RMSSE sebesar 0,099 yang mengindikasikan akurasi prediksi sangat tinggi karena jauh di bawah nilai satu. Hasil peramalan 30 hari ke depan (garis merah) menunjukkan estimasi tren dasar yang stabil di angka rendah, yang mencerminkan volume percakapan organik harian di luar periode promosi besar.

Dalam konteks strategi *digital marketing*, pola ini memberikan panduan akurat mengenai *timing* penggunaan *hashtag* sebagai instrumen pendorong visibilitas. Karena frekuensi diprediksi stabil rendah pada periode mendatang, tim pemasaran sebaiknya tidak menggunakan hashtag ini untuk konten harian (*always-on*), melainkan menyimpannya sebagai kata kunci utama untuk konten promosi taktis yang disinkronkan dengan jadwal belanja nasional guna memicu lonjakan interaksi. Fokus utama strategi konten harus beralih pada persiapan stok konten berkualitas tinggi yang siap dirilis tepat saat indikasi kenaikan frekuensi mulai terlihat, guna memastikan pesan merek tidak hanya muncul tetapi juga mendominasi ruang percakapan saat momentum *event* berlangsung.

Berdasarkan hasil pembangunan model *forecasting* pada masing-masing kategori yang telah dijelaskan sebelumnya, kesulitan dalam memprediksi lonjakan pada beberapa hasil disebabkan oleh karakteristik data media sosial yang dipengaruhi oleh *noise* linguistik, tren viral yang bersifat spontan, serta peristiwa eksternal yang tidak terjadwal, seperti kampanye mendadak atau isu publik. Pada level kata, suatu istilah dapat mengalami peningkatan frekuensi secara tiba-tiba akibat faktor eksternal yang tidak tercermin dalam pola historis, sehingga memicu fluktuasi ekstrem yang sulit dipelajari secara konsisten oleh model. Kondisi ini mencerminkan fenomena *concept drift*, yaitu perubahan cepat pada hubungan antara data historis dan pola di masa depan akibat event besar atau kampanye viral.

Meskipun demikian, pendekatan LightGBM yang memanfaatkan pembelajaran non-linear dan fitur deret waktu tetap mampu menyesuaikan diri terhadap perubahan tren secara bertahap, khususnya dalam menangkap arah pergeseran tren dasar setelah lonjakan terjadi. Di mana semakin diperkuat melalui penggunaan Optuna dalam proses *hyperparameter tuning* untuk menemukan konfigurasi model yang paling optimal dan adaptif terhadap karakteristik data yang volatil. Dengan demikian, meskipun prediksi nilai absolut pada puncak ekstrem masih menjadi tantangan, model tetap mampu memberikan informasi strategis yang bernilai untuk mendukung perencanaan konten dan respons adaptif terhadap dinamika percakapan di platform X.

E. Hasil Evaluasi Model

Pada bagian ini akan dibahas hasil evaluasi kinerja model peramalan yang telah dikembangkan. Evaluasi dilakukan menggunakan metrik *Root Mean Square Error* (RMSE), *Mean Absolute Error* (MAE), dan *Root Mean Squared Scaled Error* (RMSSE), serta melalui perbandingan dengan model *baseline*, yaitu ARIMA. Tabel VI menyajikan hasil evaluasi kinerja model peramalan yang diusulkan beserta model *baseline* untuk masing-masing kategori, yaitu topik, kata, dan *hashtag*.

TABEL VI
HASIL EVALUASI MODEL

Topik		
Metrik Evaluasi	**Model Utama**	**Model Baseline**
Rata-rata RMSE	2.779	9.24
Rata-rata MAE	0.966	6.25
Rata-rata RMSSE	0.749	0.47
Kata		
Metrik Evaluasi	**Model Utama**	**Model Baseline**
Rata-rata RMSE	3.561	13.475
Rata-rata MAE	1.53	9.347
Rata-rata RMSSE	1.91	2.261
Hashtag		
Metrik Evaluasi	**Model Utama**	**Model Baseline**
Rata-rata RMSE	0.549	1.038
Rata-rata MAE	0.371	0.811
Rata-rata RMSSE	0.757	0.68

Hasil evaluasi menunjukkan bahwa implementasi model LightGBM yang dioptimasi menggunakan *hyperparameter tuning* Optuna menghasilkan performa prediksi yang bervariasi namun tetap kompetitif pada seluruh kategori, yaitu Topik, Kata, dan *Hashtag*. Di antara ketiga kategori tersebut, *Hashtag* menunjukkan tingkat akurasi tertinggi dengan nilai rata-rata RMSE sebesar 0,549 dan MAE sebesar 0,371. Temuan ini mengindikasikan bahwa pola penggunaan *hashtag* dalam percakapan publik cenderung lebih terstruktur dan stabil secara temporal, sehingga dapat diprediksi dengan tingkat penyimpangan yang relatif rendah. Sebaliknya, kategori Kata mencatatkan nilai RMSE tertinggi sebesar 3,561, yang mencerminkan tingginya variabilitas penggunaan kata kunci dalam bahasa sehari-hari yang bersifat lebih dinamis dan fluktuatif dibandingkan *hashtag* maupun pengelompokan topik.

Analisis evaluasi lanjutan menggunakan RMSSE dilakukan untuk menilai keandalan model dibandingkan pendekatan naïve, yaitu metode sederhana yang memprediksi nilai masa depan berdasarkan nilai pada periode sebelumnya. Secara praktis, nilai RMSSE di bawah satu menunjukkan bahwa model mampu memberikan prediksi yang lebih baik daripada sekadar mengikuti pola historis, sementara nilai yang semakin kecil menandakan kemampuan model yang semakin efektif dalam menangkap pola tren yang bermakna.

Berdasarkan hasil evaluasi, nilai RMSSE pada kategori Topik (0,749) dan *Hashtag* (0,757) yang berada di bawah ambang batas satu mengindikasikan bahwa model LightGBM menghasilkan prediksi yang lebih akurat dan informatif dibandingkan metode naïve. Hal ini menunjukkan bahwa optimasi *hyperparameter* melalui Optuna berperan penting dalam meningkatkan kemampuan model untuk menangani data yang bersifat tidak teratur dan volatil. Sebaliknya, nilai RMSSE yang lebih tinggi pada kategori Kata (1,91) tidak dapat langsung diartikan sebagai kegagalan model, melainkan mencerminkan sifat alami data media sosial yang sangat fluktuatif. Dipengaruhi oleh peristiwa mendadak, tren viral sesaat, atau konteks yang tidak berulang, sehingga pola historisnya lemah dan sulit dipelajari secara konsisten oleh model.

Selain dibandingkan dengan metrik evaluasi, performa Model Utama juga dievaluasi terhadap model *baseline* ARIMA. Hasil perbandingan menunjukkan bahwa pendekatan berbasis *gradient boosting* secara konsisten unggul dalam menekan nilai kesalahan prediksi. Pada kategori Kata, misalnya, penerapan LightGBM mampu menurunkan nilai RMSE dari 13,475 pada model ARIMA menjadi 3,561, yang merepresentasikan peningkatan akurasi lebih dari 73%. Pola serupa juga terlihat pada kategori Topik, di mana nilai MAE model *baseline* mencapai 6,25, jauh lebih tinggi dibandingkan Model Utama yang hanya sebesar 0,966. Temuan ini menunjukkan bahwa model statistik tradisional seperti ARIMA, yang bergantung pada asumsi linearitas dan stasioneritas, cenderung kurang adaptif dalam menangani karakteristik data media sosial yang memiliki volatilitas tinggi dan pola non-linear.

Sebaliknya, fleksibilitas LightGBM dalam menangkap interaksi fitur yang kompleks, dikombinasikan dengan proses *hyperparameter tuning* menggunakan Optuna, terbukti lebih tangguh dalam memitigasi kesalahan prediksi, terutama saat terjadi lonjakan data secara tiba-tiba. Meskipun pada metrik RMSSE untuk kategori Topik dan *Hashtag* model *baseline* menunjukkan nilai yang relatif kompetitif terhadap skala naïve, selisih yang signifikan pada metrik RMSE dan MAE absolut menegaskan bahwa Model Utama merupakan pendekatan yang lebih reliabel untuk kebutuhan prediksi yang presisi dalam lingkungan data yang dinamis.

Dari perspektif strategi *digital marketing*, hasil *forecasting* ini memiliki implikasi praktis yang signifikan, khususnya dalam optimalisasi waktu publikasi dan relevansi konten. Tingginya akurasi prediksi pada kategori *Hashtag* memungkinkan tim pemasaran merencanakan waktu peluncuran kampanye (*campaign timing*) secara lebih presisi, sehingga *hashtag* yang diprediksi akan populer dapat digunakan sebelum mencapai puncak intensitas percakapan. Sebagai contoh, apabila model memprediksi peningkatan frekuensi *hashtag* #flashsale menjelang akhir bulan, maka *brand* dapat menjadwalkan kampanye promosi beberapa hari sebelumnya untuk memaksimalkan visibilitas dan jangkauan audiens.

Secara bersamaan, hasil peramalan pada kategori Topik dan Kata berperan sebagai panduan dalam penyusunan konten yang relevan secara kontekstual. Dengan mengetahui topik dan kata kunci yang diprediksi akan mengalami peningkatan, pemasar dapat menyiapkan materi konten, seperti teks promosi, visual, atau video, yang telah dioptimasi sejak awal. Misalnya, jika topik terkait "promo akhir tahun" diperkirakan meningkat, maka konten promosi dapat dirancang dengan narasi dan kata kunci yang selaras dengan tren tersebut. Pendekatan ini memungkinkan merek berada pada posisi yang lebih strategis dalam algoritma pencarian dan linimasa audiens saat tren terjadi, yang pada akhirnya berkontribusi pada peningkatan *engagement* dan *return on investment* (ROI).

IV. KESIMPULAN

Penelitian ini mengusulkan sebuah kerangka kerja berbasis *machine learning* untuk memprediksi tren popularitas topik, kata, dan *hashtag* pada platform X (Twitter) dengan memanfaatkan data *tweet* promosi berbahasa Indonesia. Dengan mengombinasikan pendekatan *topic modeling* menggunakan BERTopic dan pemodelan deret waktu menggunakan LightGBM yang dioptimasi melalui Optuna, penelitian ini mampu mengintegrasikan informasi tekstual dan temporal secara komprehensif dalam memodelkan dinamika percakapan publik di media sosial.

Hasil evaluasi menunjukkan bahwa model LightGBM memberikan performa prediksi yang kompetitif, terutama pada kategori topik dan *hashtag*. Dibandingkan dengan model *baseline* ARIMA, LightGBM mampu menurunkan nilai kesalahan secara signifikan, di mana pada kategori kata nilai RMSE berkurang lebih dari 70%, sementara pada kategori topik dan *hashtag* penurunan MAE dan RMSE juga terlihat sangat substansial. Selain itu, nilai RMSSE pada kategori topik (0,749) dan *hashtag* (0,757) yang berada di bawah satu menunjukkan bahwa model menghasilkan prediksi yang lebih baik dibandingkan pendekatan naïve. Secara praktis, hal ini mengindikasikan bahwa LightGBM efektif dalam menangkap pola tren yang bermakna pada data media sosial. Meskipun pada kategori kata, model masih menghadapi tantangan dalam memprediksi lonjakan ekstrem akibat sifat data yang sangat fluktuatif dan *event-driven*, model tetap mampu merepresentasikan arah tren dasar percakapan publik secara umum.

Dari perspektif *digital marketing*, hasil penelitian ini memberikan implikasi praktis yang signifikan. Prediksi tren *hashtag* dapat dimanfaatkan untuk menentukan waktu peluncuran kampanye secara lebih presisi, sementara hasil peramalan topik dan kata dapat digunakan sebagai dasar penyusunan konten yang relevan dan kontekstual. Dengan

demikian, pendekatan yang diusulkan tidak hanya mendukung pengambilan keputusan berbasis data, tetapi juga berpotensi meningkatkan efektivitas kampanye, *engagement* audiens, dan *return on investment* (ROI).

Namun demikian, penelitian ini memiliki beberapa keterbatasan. Pertama, ruang lingkup analisis dibatasi pada konteks konten promosi, sehingga pola dan temuan yang dihasilkan belum tentu merepresentasikan dinamika percakapan pada domain lain, seperti politik, hiburan, atau isu sosial. Kedua, penelitian ini hanya berfokus pada *tweet* berbahasa Indonesia, sehingga generalisasi hasil ke bahasa atau wilayah lain masih terbatas. Oleh karena itu, penelitian selanjutnya dapat memperluas cakupan domain dan bahasa, serta mengeksplorasi model berbasis *deep learning* atau *transformer time-series* untuk meningkatkan kemampuan menangkap pola nonlinier dan *event-driven spikes*.

DAFTAR PUSTAKA

[1] C. E. Sitanggang, D. A. Firda, R. Ramadhini, J. M. Panjaitan, Sofwan and M. Sholeh, " Studi Literatur: Penggunaan Media Sosial Sebagai Alat Promosi Usaha," *Jurnal Ilmiah Ekonomi Dan Bisnis Universitas Multi Data Palembang*, vol. 14, pp. 23-29, 2024.

[2] A. D. Ade, M. Rizan and I. Febrilia, "Pengaruh Aktivitas Pemasaran Media Sosial Media Sosial Terhadap Citra Merek, Loyalitas Merek, Dan Niat Beli Ulang Pada Social Commerce Tiktok Shop," *Jurnal Masharif al-Syariah: Jurnal Ekonomi dan Perbankan Syariah*, vol. 9, no. 4, pp. 2399-2416, 2024.

[3] M. Febiansyah, Jondri and Indwiarti, "Prediksi Retweet Berdasarkan Konten Dan Pengguna Dengan Metode Classifier Selection," *Smart Comp: Jurnalnya Orang Pintar Komputer*, vol. 14, no. 1, pp. 123-129, 2025.

[4] The Global Statistics, "The Global Statistics," The Data Expert, 12 Maret 2025. [Online]. Available: https://www.theglobalstatistics.com/indonesia-social-media-statistics. [Accessed 3 April 2025].

[5] H. Kwak, C. Lee, H. Park and S. Moon, "What is Twitter, a Social Network or a News Media?," in *WWW '10: Proceedings of the 19th international conference on World wide web*, North Carolina, 2010.

[6] P. A. Riyantoko and A. Muhaimin, "A Simple Data Sentiment Analysis using Bjorka phenomenon on Twitter," in *7st International Seminar of Research Month 2022*, Surabaya, 2023.

[7] K. Darvidou, "Content Marketing Strategy and Development," *Technium Business and Management (TBM)*, vol. 10, pp. 55-67, 2024.

[8] Y. WANG, J. CALLAN and B. ZHENG, "Should We Use the Sample? Analyzing Datasets Sampled from Twitter's Stream API," *ACM Transactions on the Web (TWEB)*, vol. 9, no. 3, pp. 13-35, 2015.

[9] S. S. S. Ramesh, C. Raghavaraju, S. L. P and A. T. Navis, "Exploratory Analysis and Predictive Modeling of Social Media Data by Decoding Twitter," Research Square, 2024.

[10] C.-C. Hsu, C.-M. Lee, X.-Y. Hou and C.-H. Tsai, "Gradient Boost Tree Network based on Extensive Feature Analysis for Popularity Prediction of Social Posts," in *MM '23: Proceedings of the 31st ACM International Conference on Multimedia*, Ottawa, 2023.

[11] A. Jafar and M. Lee, "Comparative Performance Evaluation of State-of-the-Art Hyperparameter Optimization Frameworks," *The Transactions of the Korea Institute of Electrical Engineers*, vol. 72, no. 5, pp. 607-619, 2023.

[12] K. Ng and P. Lei, "A Lightweight Method using LightGBM Model with Optuna in MOOCs Dropout Prediction," in *ICEMT '22: Proceedings of the 6th International Conference on Education and Multimedia Technology* , Guangzhou, 2022.

[13] R.-S. Constantin, A. A. Davidescu and E. M. Manta, "Time Series Forecasting with LightGBM under Data Scarcity: An Application to Romania's Inland Gas Consumption," *Proceedings of the International Conference on Business Excellence*, vol. 19, no. 1, pp. 1518-1531, 2025.

[14] Y. Chen, X. Xie, Z. Pei, W. Yi, C. Wang, W. Zhang and Z. Ji, "Development of a Time Series E-Commerce Sales Prediction Method for Short-Shelf-Life Products Using GRU-LightGBM," *Applied Science*, vol. 14, no. 2, pp. 1-16, 2024.

[15] M. Thoriqulhaq, M. Idhom and K. M. Hindrayani, "Implementasi Algoritma LightGBM untuk Prediksi Status Gizi Bayi dan Balita di Desa Doko Kabupaten Kediri," *J-TETA : Jurnal Teknik Terapan*, vol. 4, no. 1, pp. 65-73, 2025.

[16] T. Kee and W. K. Ho, "Optimizing Machine Learning Models for Urban Sciences: A Comparative Analysis of Hyperparameter Tuning Methods," *Urban Science*, vol. 9, no. 9, pp. 1-24, 2025.

[17] M. Mendonça and Á. Figueira, "Topic Extraction: BERTopic's Insight into the 117th Congress's Twitterverse," *Informatics*, vol. 11, no. 1, pp. 1-34, 2024.

[18] E. Zhu, "BERTopic-Driven Stock Market Predictions: Unraveling Sentiment Insights," arXiv, New York, 2024.

[19] F. Koto, J. H. Lau and T. Baldwin, "IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Punta Cana, 2021.

[20] Karnisih, Sunarno, Iqbal, Djuniadi and F. S. Pribadi, "Penerapan Algoritma Linear Regression dan Support Vector Regression dalam Prediksi Temperatur Udara di Malang," *Techno.COM*, vol. 24, no. 1, pp. 218-229, 2025.

[21] R. C. Ganzevoort and J. H. v. Vuuren, "Atwo-phasedcluster-basedapproachtowardsrankedforecast-model selection," *MachineLearningwithApplications*, vol. 13, pp. 1-14, 2023.

[22] X. Zhang and W. Cao, "Research on Time Series Forecasting Method Based on Autoregressive Integrated Moving Average Model with Zonotopic Kalman Filter," *Sustainability (MDPI)*, vol. 17, no. 7, pp. 1-18, 2025.

[23] M. d. Groot, M. Aliannejadi and M. R. Haas, "Experiments on Generalizability of BERTopic on Multi-Domain Short Text," arXiv, 2022.

[24] https://github.com/deannisasp/twitter-forecasting-lightgbm-