JURNAL SAINTIKA UNPAM : Jurnal Sains dan Matematika Unpam Vol. 3, No. 2 (2021)
p-ISSN 2621-7856 e-ISSN 2655-7312 16th article, page 161 - 177

PREDIKSI KASUS COVID-19 MELALUI ANALISIS DATA GOOGLE TREND DI INDONESIA: PENDEKATAN METODE *LONG SHORT TERM MEMORY* (LSTM)

CASE PREDICTION OF COVID-19 THROUGH GOOGLE TREND DATA ANALYSIS IN INDONESIA: THE LONG SHORT TERM MEMORY (LSTM) METHOD APPROACH

Lisa Widyarsi*, Ivana Yoselin Purba Siboro, Peterson Hamonangan Immanuel Sihotang, Satria Dirgantara, Yakobus Natanael Tarigan, Yuniar Putri Awaliyah Risky, dan Rani Nooraeni
Politeknik Statistika STIS, Jalan Otto Iskandardinata 64C, Bidara Cina Kecamatan Jatinegara Kota Jakarta Timur, DKI Jakarta 133301
211709790@stis.ac.id,

ABSTRACT

One of the factors that needed to reduce the number of COVID-19 cases is the high level of public attention. This can be seen by the intensity of public search for information about COVID-19 on an online platform called Google Trend. This paper aims to describe the condition of COVID-19 outbreak in the community by using Google Trend data and predicting COVID-19 cases with both nowcasting and forecasting methods by combining public attention data from Google Trend with official data on the growth of COVID-19 cases in Indonesia. The data used in the form of daily time series data from April 1st to September 30th 2020. The Multiple Linear Regression method is also used to compare the predicted results with LSTM. The result of time series regression yield RMSE 1060,80. In addition to the time series analysis method, prediction of additional COVID-19 cases were also carried out using the LSTM method with four scenarios, where the first scenario yield RMSE 526,59, the second scenario yield RMSE 528,81, the third yield RMSE 483,25 and the last scenario yield RMSE 482,21. The prediction using the LSTM method with the fourth scenario produces RMSE, so the LSTM method is the fourth method with a fairly good prediction

Keywords: **COVID-19, LSTM, Google Trend, PCA, Regression**

ABSTRAK

Salah satu faktor yang diperlukan untuk menekan angka kasus COVID-19 adalah tingginya perhatian atau atensi masyarakat. Hal tersebut terlihat dari intensitas pencarian informasi publik mengenai COVID-19 di platform online bernama Google Trend. Makalah ini bertujuan untuk mendeskripsikan kondisi wabah COVID-19 di masyarakat dengan menggunakan data Google Trend dan memprediksi kasus COVID-19 baik dengan metode nowcasting maupun forecasting dengan menggabungkan data atensi publik dari Google Trend dengan data resmi pertumbuhan COVID-19 di Indonesia. Data yang digunakan berupa data time series harian dari tanggal 1 April hingga 30 September 2020. Metode Regresi Linear Berganda juga digunakan untuk membandingkan hasil prediksi dengan LSTM. Hasil regresi time series menghasilkan RMSE 1060,80. Selain metode analisis time series, prediksi penambahan kasus COVID-19 juga dilakukan menggunakan metode LSTM dengan empat skenario, di mana skenario pertama menghasilkan RMSE 526,59, skenario kedua menghasilkan RMSE 528,81, skenario ketiga menghasilkan RMSE 528,81. RMSE 483,25 dan skenario terakhir menghasilkan RMSE 482,21.

Prediksi menggunakan metode LSTM dengan scnario keempat menghasilkan RMSE, sehingga metode LSTM merupakan metode keempat dengan prediksi yang cukup baik.
Kata kunci: COVID-19, LSTM, Google Trend, PCA, Regresi

1. PENDAHULUAN

COVID-19 (coronavirus disease 2019) adalah penyakit yang disebabkan oleh jenis coronavirus baru yaitu Sars-CoV-2, yang dilaporkan pertama kali di Wuhan, Tiongkok pada tanggal 31 Desember 2019. Menurut World Health Organization (WHO), per tanggal 17 September 2020, virus Corona telah menginfeksi 29.679.284 penduduk dunia dan sekitar 936.521 orang di antaranya dinyatakan meninggal dunia. Penularan COVID-19 dari manusia ke manusia yang dapat terjadi melalui kontak erat dan percikan cairan pada saat bersin dan batuk membuat wabah ini menyebar dengan cepat ke negara-negara lain, tanpa kecuali Indonesia. Di Indonesia kasus awal COVID-19 terjadi pada awal bulan Maret. Semenjak itu kasus pasien COVID-19 yang terkonfirmasi terus bertambah, bahkan berdasarkan Kementerian Kesehatan Republik Indonesia (Kemenkes RI) per tanggal 16 September 2020 kasus COVID-19 di Indonesia mencapai 228.993 total kasus terkonfirmasi dengan 164.101 pasien dinyatakan sembuh dan 9.100 dinyatakan meninggal yang berarti 55.792 pasien masih dalam perawatan. Peningkatan ini terus berusaha ditekan oleh pemerintah melalui kebijakan-kebijakan yang dibentuk, salah satunya ialah kebijakan Pembatasan Sosial Berskala Besar (PSBB) yang diatur dalam Pasal 13 Peraturan Menteri Kesehatan No. 9 Tahun 2020 tentang poin-poin PSBB.

Apabila ditinjau lebih lanjut, angka statistik yang dilaporkan diperoleh dari pelaporan hasil pemeriksaan antigen dengan metode Real Time PCR. Sementara itu, metode pencatatan untuk ODP dan PDP (suspek) dilakukan dengan menghimpun data dari setiap Dinas Kesehatan Daerah. Pencatatan dan pelaporan kasus COVID-19 di Indonesia dilaksanakan terkomputerisasi dengan cara online berbasis aplikasi, yaitu All Record TC-19 (https://allrecordtc19.kemkes.go.id) dan Sistem Online Pelaporan Harian COVID-19 (https://s.id/laporhariancovid) dengan melalui verifikasi Badan Litbangkes dan Kemenkes. Walaupun sudah terkomputerisasi, nyatanya metode ini masih tidak dapat mencakup masyarakat yang sebenarnya telah terkena COVID-19, namun belum melakukan tes baik secara Rapid Test maupun Real Time PCR. Selain itu, adanya lag waktu antara pemeriksaan Rapid Test dengan Real Time PCR mengakibatkan pemanfaatan data konvensional saja dalam meramalkan penyakit

menular seperti Covid-19 menjadi kurang fleksibel, khususnya dalam mengantisipasi dan pengambilan keputusan oleh pemerintah.

Di sisi lain, Indonesia telah memasuki era revolusi industri 4.0 yang ditandai dengan tingginya pergerakan pertukaran data dan informasi. Tingginya pergerakan pertukaran data dan informasi berdampak kepada segala aspek kehidupan manusia, termasuk dalam memenuhi kebutuhan informasi. Berdasarkan data BPS, dari tahun 2012 hingga 2018 persentase rumah tangga di Indonesia yang memiliki atau menguasai telepon seluler terus meningkat. Selain itu, perkembangan pengguna internet juga berkembang pesat. Berdasarkan data BPS, persentase rumah tangga yang pernah mengakses internet dalam 3 bulan terakhir pada tahun 2018 adalah 66,22 persen. Angka ini meningkat pesat bila dibandingkan pada tahun 2012 yang hanya 30,66 persen. Dampak dari perkembangan teknologi informasi tersebut adalah peningkatan kemudahan masyarakat dalam mendapatkan dan memberikan informasi. Hal ini juga berdampak pada penggunaan mesin pencari atau search engine di internet yang terlihat dari semakin banyaknya orang yang mencari informasi tentang apapun melalui mesin pencari.

Williams dan Sawyer (2011) mendefinisikan mesin pencari sebagai sebuah program yang memungkinkan pengguna untuk mengajukan pertanyaan atau menggunakan kata kunci untuk membantu mencari informasi pada web. Berdasarkan publikasi data Databoks tahun 2019 mesin pencari dengan jumlah pengguna terbanyak ialah Google dengan 64,4 persen pengguna. Jumlah pengguna yang besar serta intensitas pencarian yang tinggi akan memberikan kontribusi terhadap peningkatan jumlah data yang dihasilkan oleh mesin pencari. Hasil data yang berjumlah besar ini secara lebih lanjut disebut dengan Big Data. Apabila ditinjau lebih jauh, intensitas pencarian pengguna Google dirilis pertama kali pada tahun 2009 melalui antarmuka Google Trend. Mengingat bahwa seluruh kata kunci pencarian yang digunakan oleh pengguna merupakan refleksi dari atensi masyarakat terhadap suatu topik, maka dapat dikatakan bahwa Google Trend menjadi penyedia data yang cukup dalam menunjang penelitian-penelitian. Selain itu, fleksibilitas fitur pada Google Trend dalam memilih referensi waktu baik harian, mingguan, bulanan, maupun tahunan mulai dari tahun 2004 hingga saat ini menjadikan hasil data Google Trend lebih terkini.

Kemudahan dan kelebihan tersebut membawa Google Trend menjadi salah satu

solusi untuk meningkatkan akurasi prediksi dalam banyak bidang penelitian, tidak terkecuali pada bidang kesehatan. Dengan idenya yang cukup sederhana, Milinovich dkk. (2014) menjelaskan alasan utama dibalik kekuatan prediksi data online, yaitu orang yang mencurigai suatu penyakit cenderung mencari informasi online tentang gejalanya. Oleh karena itu, penggunaan data dari Google Trend dapat meningkatkan sensitivitas, ketepatan waktu deteksi kejadian kesehatan, dan dapat digunakan sebagai perluasan data konvensional. Hal ini juga didukung oleh hasil penelitian Sharma dan Sharma (2020), bahwa peningkatan kasus, ketakutan dan kekhawatiran COVID-19 dapat direfleksikan dengan baik melalui peningkatan tren pencarian masyarakat melalui Google Trend. Selaras dengan penelitian tersebut, Ayyoubzadeh, dkk. (2020) menyatakan bahwa penggunaan Google Trend dapat memprediksi tren penyebaran COVID-19 serta memberikan informasi yang lebih baik untuk mengatur krisis kesehatan yang disebabkan oleh COVID-19. Oleh karena itu, potensi internet di Indonesia dan ketersediaan data pada Google Trends dirasa akan mampu membantu pemerintah Indonesia untuk memprediksi kasus COVID-19 secara lebih baik dibanding model semula, yakni model peramalan dengan hanya menggunakan data konvensional.

Berdasarkan latar belakang tersebut, penelitian ini memiliki dua tujuan, yakni menggambarkan kondisi wabah COVID-19 di masyarakat dengan menggunakan data Google Trend dan memprediksi kasus COVID-19 *(forecasting)* dengan memadukan data atensi masyarakat dari Google Trends dengan data resmi pertumbuhan kasus COVID-19 di Indonesia. Data yang digunakan berupa data runtun waktu harian dengan rentang waktu dari tanggal 1 April 2020 hingga 30 September 2020. Dalam memprediksi kasus positif COVID-19 digunakan metode prediksi *Long Short Term Memory* (LSTM). Metode ini dipilih karena dapat memberikan hasil prediksi yang lebih akurat. Hal ini didukung oleh Shertinsky (2020) yang menyatakan bahwa LSTM adalah sebuah jaringan saraf tiruan berulang (RNN) yang merupakan model yang efektif untuk memprediksi data runtun waktu ketika datanya berurutan. Selain itu, digunakan pula metode analisis regresi runtun waktu sebagai pembanding hasil prediksi dengan LSTM.

2. METODOLOGI

2.1. Data dan Sumber Data

Data yang digunakan pada penelitian ini adalah data sekunder yang diperoleh dari website Kementerian Kesehatan RI dan Google Trends. Data yang dikumpulkan berupa data runtun waktu dengan rentang waktu tanggal 1 April 2020 hingga 30 September 2020. Data yang diperoleh dari website Satuan Tugas Penanganan COVID-19 adalah jumlah pertambahan kasus terkonfirmasi positif COVID-19 setiap harinya. Sedangkan data yang diperoleh dari Google Trends adalah subjek pencarian pada Google yang berkaitan dengan COVID-19. Pada penelitian ini digunakan 12 kata kunci utama dengan rincian sebagai berikut : "covid 19", "corona", "hand sanitizer", "masker", "PSBB", "cuci tangan", "disinfektan", "vaksin corona", "jaga jarak", "kebiasaan baru", "new normal", dan "di rumah aja". Kemudian, dari 12 kata kunci utama digunakan pula kata kunci yang terkait dengan 12 kata kunci tersebut, sehingga secara akumulasi diperoleh data runtun waktu dari 286 kata kunci terkait atensi masyarakat terhadap COVID-19. Data dari Google Trends tersebut digunakan sebagai *auxiliary variabel* untuk nowcasting dan forecasting pertumbuhan kasus terkonfirmasi positif COVID-19 di Indonesia.

2.2. Metode Analisis

Guna mencapai tujuan penelitian berupa nowcasting dan forecasting kasus terkonfirmasi COVID-19 di Indonesia digunakan metode *Long Short Term Memory* (LSTM) dan metode Analisis Regresi Deret Waktu. Metode LSTM digunakan dengan bantuan aplikasi Google Collabs dengan bahasa pemrograman Python digunakan module Tensorflow [2.3.0] dan RStudio Cloud. Sementara itu, metode Analisis Regresi Deret Waktu digunakan dengan bantuan aplikasi Eviews 10. Sebelum melakukan prediksi dengan menggunakan metode LSTM dan Analisis Regresi Deret Waktu, dilakukan pengumpulan dan *preprocessing* data.

2.2.1. Pengumpulan Data

Data dari penelitian ini diperoleh dari Google Trends dengan pencarian kata kunci utama dan kata kunci yang terkait yang berkaitan dengan COVID-19. Pada proses penambangan (*scrapping*) data Google Trends digunakan bantuan *packages* "gtrendsR" pada aplikasi R. Data yang dicari berada pada rentang tanggal 1 April 2020 hingga 30 September 2020, sehingga diperoleh 183 runtun waktu. Secara sederhana, proses penambangan data dilakukan dengan mencari setiap kata pada kata kunci utama. Kemudian, setiap pencarian kata kunci utama akan menghasilkan 50 kata kunci terkait

dengan 25 kata kunci kategori "Top" dan 25 kata kunci kategori "Rising". Selanjutnya diambil 25 kata kunci terkait dengan kategori "Top" dari setiap kata kunci utama untuk dicari kembali. Selain data dari Google Trends, dikumpulkan pula data penambahan kasus terkonfirmasi COVID-19 di Indonesia dengan cara mengunduh data dari laman situs Kementerian Kesehatan RI. Data yang diunduh memiliki rentang waktu yang sama, yakni 1 April 2020 hingga 30 September 2020.

2.2.2. Seleksi Kata Kunci

Data Google Trends yang masih berbentuk nilai indeks global tergolong sebagai data kasar, sehingga diperlukan serangkaian tahapan preprocessing hingga siap digunakan sebagai variabel penyerta dalam proses prediksi

a. Koefisien Korelasi

Kata kunci data Google Trends yang cukup banyak, yakni 278 kata kunci akan menyulitkan pada proses pengolahan data dan interpretasi hasil. Oleh karena itu, dilakukan reduksi data kata kunci (auxilliary variable). Reduksi dilakukan dengan mempertimbangkan korelasi antara atau hubungan antara tiap variabel independen (auxilary variable) dengan variabel dependen, yakni pertumbuhan kasus terkonfirmasi positif COVID-19 di Indonesia. Untuk mendapatkan nilai korelasi, digunakan Uji Korelasi Pearson. Korelasi pearson merupakan salah satu ukuran korelasi yang digunakan untuk mengukur kekuatan dan arah hubungan linier dari dua variabel. Nilai korelasi pearson disebut juga koefisien korelasi pearson. Koefisien korelasi pearson hanya dapat mengukur kekuatan hubungan linier dan tidak pada hubungan non linier. Korelasi pearson dirasa tepat karena data yang diperoleh berupa data rasio.

b. Stepwise Regression

Guna memilih variabel terbaik yang dapat dimasukkan dalam model, perlu dilakukan seleksi variabel. Salah satu metode yang dapat digunakan dalam seleksi variabel ialah metode regresi bertahap (*stepwise regression*). *Stepwise regression* merupakan metode regresi linear berganda, yang secara sekaligus menghapus variabel-variabel bebas yang tidak penting. Pada dasarnya *stepwise regression* menjalankan regresi berganda beberapa kali, setiap kali menghapus variabel berkorelasi lemah. Hingga pada akhirnya tersisa variabel-variabel yang

menjelaskan distribusi yang terbaik. Satu-satunya persyaratan adalah bahwa data tersebut berdistribusi normal dan tidak terdapat korelasi antar variabel independen.

Regresi bertahap memiliki dua jenis, yakni *stepwise regression* dan *backwards stepwise regression*. *Forward stepwise regression* dirancang untuk memilih dari sekelompok prediktor variabel, satu pada setiap tahap, yang memiliki semi parsial r-square terbesar dan karenanya membuat kontribusi terbesar r-square. Sedangkan, backwards stepwise regression bekerja secara sebaliknya, yaitu variabel yang secara statistik tidak signifikan, yang membuat kontribusi terkecil tidak digunakan

c. Metode *Principal Component Analysis (PCA*)

Tujuan metode *Principal Component Analysis* (PCA) untuk menyederhanakan variabel yang diamati dengan mereduksi dimensinya. Hal ini dilakukan dengan cara menghilangkan korelasi diantara variabel bebas melalui transformasi variabel bebas asal ke variabel baru yang tidak berkorelasi sama sekali atau sering disebut dengan principal component. Setelah beberapa komponen hasil PCA yang terbebas dari multikolinearitas diperoleh, maka komponen-komponen tersebut menjadi variabel bebas baru yang akan diregresikan atau dianalisa pengaruhnya terhadap variabel tak bebas (Y) dengan menggunakan analisis regresi.

2.2.3. Prediksi Penambahan Kasus COVID-19

2.2.3.1. *Long Short Term Memory (LSTM)*

Prediksi dengan metode *Long Short Term Memory* (LSTM) dilakukan melalui beberapa tahapan, yaitu mengumpulkan data, melakukan mekanisme pembaharuan bergulir (*rolling update mechanism*), jaringan struktur LSTM, dan evaluasi hasil prediksi. LSTM bekerja dengan menggunakan empat proses aktivasi (*gates unit*) yang terdiri dari *forget gate, input gate, cell gate*, dan *output gate*. Secara lebih rinci, berikut merupakan tahapan prediksi dengan metode LSTM:

a. Preprocessing Data

Hasil dari proses *preprocessing data* digunakan sebagai dataset input. Dari total dataset yang ada, persentase data training adalah 80% dari total dataset dan sisanya merupakan data testing. Tiap neuron dalam input layer mewakili vektor input yang melibatkan data training. Data training disimpan dalam bentuk file csv (*comma - separated values*).

Dataset input terdiri dari variabel independen dan dependen seperti yang telah dipaparkan di atas. Sebelum masuk pada tahapan selanjutnya, dataset input dilakukan proses normalisasi untuk menjaga agar keluaran jaringan sesuai dengan fungsi aktivasi yang digunakan dan memperkuat akurasi data. Proses normalisasi ialah proses mengubah data aktual menjadi data dengan nilai yang berada pada range interval [0,1]. Normalisasi dilakukan dengan metode min-max scaling, berikut merupakan formula yang digunakan:

$$X' = \frac{(X - min_x)}{(max_x - min_x)} \quad (1)$$

Dimana:

X : Data yang akan dinormalisasikan

X' : Data setelah normalisasi

min_x : Nilai minimum dari keseluruhan data

max_x : Nilai maksimum dari keseluruhan data

b. Mekanisme pembaharuan bergulir

Mekanisme pembaharuan bergulir bekerja dengan memperbaharui urutan sampel pelatihan sesuai dengan hasil prediksi saat ini untuk melatih model. Pengoptimalan bergulir bertujuan untuk mengoptimalkan kontrol koreksi umpan balik bergulir dalam waktu terbatas dengan menggunakan model prediktif dan data historis untuk melatih model secara iteratif. Pada tahap ini juga dilakukan inisialisasi paramater dasar berupa nilai learning rate, jumlah hidden layer, jumlah neuron pada hidden layer, target error berupa MSE, dan E-poch maksimum. Penelitian ini menggunakan *Adam Optimizer,* dengan *pseudo code* sebagai berikut:

α : learning rate

$\beta_1, \beta_1 \in [0,1]$: exponential decay rates for the moment esimates

$f(\theta)$: fungsi stokastik dengan parameter θ

θ_0 : inisial paramter vector

$m_0 \leftarrow 0$ (inisialisasi moment vector pertama)

$v_0 \leftarrow 0$ (inisialisasi moment vector kedua)

$t \leftarrow 0$ (inisialisasi timestep)

while θ_t not converged **do**

$t \leftarrow t + 1$

$g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$ (mendapatkan gradient dari fungsi stokastik pada timestep t)

$m_t \leftarrow \beta_1 . m_{(t-1)} + (1-\beta_1).g_t$ (*update* bias estimasi moment pertama)

$v_t \leftarrow \beta_1 . v_{(t-1)} + (1-\beta_1).{g_t}^2$ (*update* bias estimasi moment kedua)

$\widehat{m_t} \leftarrow m_t/\left(1-{\beta_1}^t\right)$ (*compute bias-corrected first* moment estimate)

$\widehat{v_t} \leftarrow v_t/\left(1-{\beta_2}^t\right)$ (*compute bias-corrected second raw* moment estimate)

$\theta_t \leftarrow \theta_{t-1} - \alpha.\widehat{m_t}/\left(\sqrt{\widehat{v_t}} + \varepsilon\right)$ (*update* parameters)

End while

Return θ_t (hasil parameter yang didapatkan)

c. Jaringan struktur LSTM

Berikutnya adalah penjelasan dari proses training pada jaringan LSTM Network yang diusulkan:

1) Menghitung semua fungsi gates unit pada setiap neurons. Secara berurutan fungsi gates yang akan dihitung adalah forget gates dengan persamaan, fungsi input gates dengan persamaan, fungsi cell gates dengan persamaan, dan yang terakhir fungsi output gates dengan persamaan
2) Menghitung fungsi aktivasi linear pada output layer dengan rumus $\varphi(x) = x$
3) Jika telah melakukan perulangan sebanyak epoch yang telah ditentukan, maka berhenti. Jika belum, akan dilakukan optimasi dengan *Adam Optimizer* dan memperbarui bobot dan bias pada sistem, kemudian kembali ke langkah dua.

d. Evaluasi hasil prediksi

Model yang telah didapatkan pada proses sebelumnya atau training akan diuji dengan menggunakan data testing yang telah didapat dari *preprocessing data*. Pada penelitian ini digunakan metode akurasi Root Mean Squared Error (MSE) dengan formula sebagai berikut:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2} \quad (2)$$

Dimana:

N: Jumlah data

f_i : Nilai ke-i yang didapat dari model

y_i : Nilai ke-i pada data aktual

Sebelum dilakukan penghitungan MSE, maka data hasil pemodelan dilakukan proses denormalisasi untuk mengembalikan data kepada bentuk sebelum

normalisasi atau nilai aslinya. Proses ini biasa disebut dengan *post processing data*. Berikut merupakan formula denormalisasi yang digunakan:

$$X = \frac{(f_i - 0{,}1)(max_x - min_x)}{0{,}8} + min_x \quad (3)$$

Dimana:

X : Data actual

f_i : Nilai ke-i yang didapat dari model

min_x : Nilai minimum dari keseluruhan data

max_x : Nilai maksimum dari keseluruhan data

2.2.3.2. Regresi Runtun Waktu

Pemodelan dengan menggunakan metode Regresi Deret Waktu, hubungan linier antara variabel dependen dengan variabel independen dibangun sesuai kondisi kestasioneran data sehingga perlu melalui beberapa tahapan, yaitu:

a. Uji Stasioneritas

Data stasioner adalah data yang menunjukkan mean dan varians konstan pada setiap periode waktu. Metode yang dapat dilakukan untuk pengujian stasioneritas data adalah metode akar-akar unit. Pada penelitian ini uji yang digunakan adalah Uji *Augmented Dicky-Fuller* (ADF). Uji ADF meregresikan ΔY_t, Y_{t-1}, dan komponen galat v_t. Yang menjadi hipotesis pada uji ini adalah koefisien Y_{t-1} hasil regresi yang disebut *rho* dengan hipotesis nol: *rho* = 0 atau data memiliki *unit root* yang berarti data tidak stasioner.

b. Pemodelan

Pemodelan hubungan linier dilakukan pada data yang sudah stasioner baik untuk data dependen maupun independen. Pada penelitian ini data stasioner pada *difference* pertama sehingga model yang diregresikan adalah:

$$D(Y_t) = C + D(Y_{t-1}) + D(GT_t) + u_t \quad (4)$$

Dimana:

Y : Penambahan Kasus COVID-19 per Hari

GT : indeks google trend

t : indeks waktu

u : komponen eror

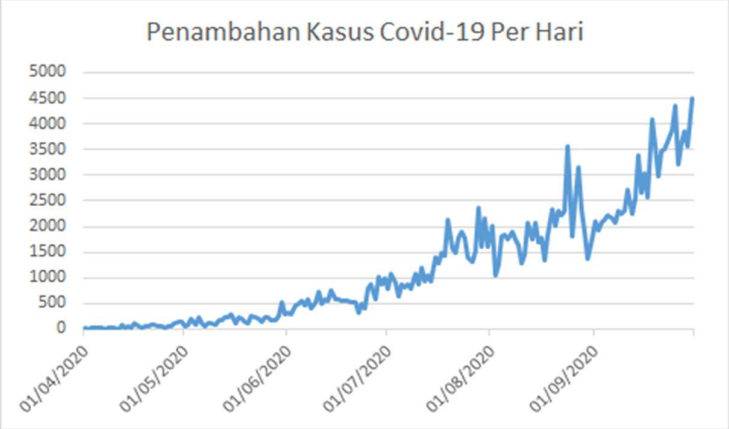
c. Peramalan

Sebelum dilakukan pemodelan data dibagi menjadi 80% data training dan 20% data testing dengan tujuan untuk melihat kesesuaian model yang terbentuk. Model yang dibangun dari data training selanjutnya akan dilakukan peramalan, lalu hasilnya akan dibandingkan dengan data asli pada data testing. Dari perbandingan ini akan didapatkan nilai *Root Mean Square Error* (RMSE) yang dapat mewakili tingkat kesalahan model.

3. PEMBAHASAN

3.1. Gambaran Umum Kasus COVID-19 di Indonesia dengan Data Google Trend

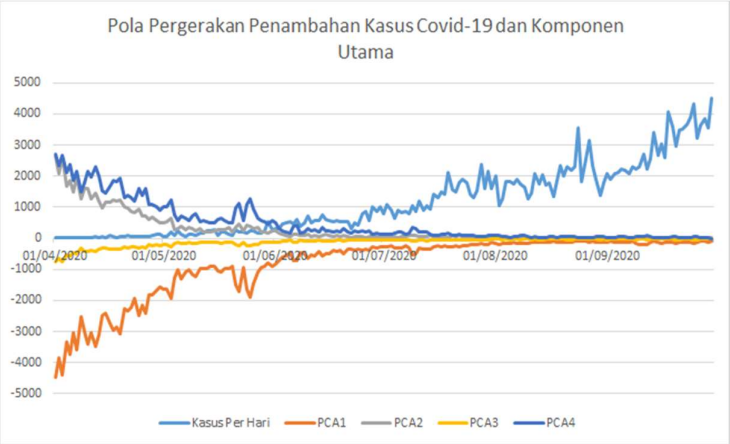
Penambahan kasus COVID-19 di Indonesia berdasarkan data yang dikeluarkan oleh Kementerian Kesehatan RI dari tanggal 1 April 2020 hingga 30 September 2020 terus mengalami kenaikan (Gambar 1). Penambahan jumlah kasus COVID-19 tiap harinya terus mengalami kenaikan dan belum menunjukkan adanya penurunan. Hal ini menunjukkan bahwa upaya-upaya yang telah dilakukan belum dapat menekan penambahan kasus COVID-19.



Gambar 1. Penambahan Kasus COVID-19 per Hari

Untuk meninjau lebih jauh terkait atensi masyarakat terhadap COVID-19 digunakan data Google Trends atas topik terkait. Berdasarkan hasil *pre-processing data*, bahwa kata kunci "Covid Indonesia" memiliki korelasi yang paling kuat dengan jumlah penambahan kasus COVID-19. Berdasarkan hasil PCA, diperoleh komponen utama dari 28 kata kunci paling relevan dengan variabel dependen. Hasil PCA menunjukkan bahwa komponen utama pertama merepresentasikan 61,17 persen dari keberagaman total, sedangkan empat komponen utama dapat merepresentasikan 81,59 persen keberagaman total. Komponen

utama ini bisa dinilai telah cukup menangkap struktur data. Komponen pertama merupakan atensi yang berhubungan negatif dengan COVID-19, komponen kedua adalah berita tentang COVID-19, komponen ketiga adalah pengetahuan tentang PSBB, dan komponen keempat adalah Pandemi COVID-19.



Gambar 2. Pola Pergerakan Penambahan Kasus COVID-19 dan Komponen Utama

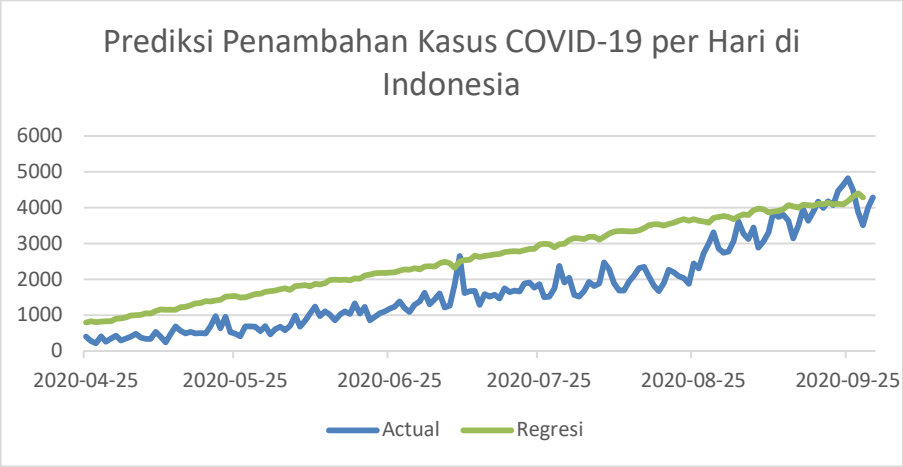
Pada Gambar 2, dapat dilihat bahwa terjadi fluktuasi pada tiap komponen utama. Tiap komponen utama menggambarkan perubahan atensi masyarakat dalam mencari setiap kata kunci yang berhubungan dengan COVID-19. Seriring bertambahnya penambahan kasus COVID-19 per hari, komponen atensi masyarakat terhadap kasus COVID-19 mengalami peningkatan pada awal pandemi hingga bulan Juli 2020, namun cenderung stabil pada masa setelahnya. Hasil ini kontradiktif dengan komponen berita tentang COVID-19 dan pengetahuan tentang PSBB semakin menurun dari waktu ke waktu. Hal ini menunjukkan bahwa atensi masyarakat terhadap kasus COVID-19 pada awal pandemi cenderung tinggi, tetapi atensi tersebut semakin berkurang seiring berjalannya waktu. Selanjutnya, komponen informasi terkait pandemi COVID-19 cenderung stabil dari waktu ke waktu.

3.2. Prediksi Penambahan Kasus COVID-19 dengan Regresi Runtun Waktu

Metode analisis regresi runtun waktu digunakan untuk memprediksi penambahan kasus COVID-19, di mana variabel terikatnya adalah jumlah penambahan kasus COVID-19 per hari dan variabel bebasnya adalah lag ke-1 dari variabel terikat dan data google trend dengan kata kunci yang paling berkorelasi kuat yaitu “Covid Indonesia”. Kemudian diperoleh model seperti pada persamaan (5).

$$\Delta\hat{y}_t = 27{,}95 - 0{,}17\,\Delta y_{t-1} + 3{,}19\Delta \mathrm{GT}_t \qquad (5)$$

Model prediksi di atas dapat digunakan untuk melihat kesadaran dan keterlibatan individu di situasi pandemi COVID-19 ini yang diwakili dengan adanya variabel indeks pencarian kata kunci "Covid Indonesia" pada Google. Berdasarkan model yang dibentuk, dilakukan *forecasting* untuk data pada tanggal 25 September hingga 30 september 2020. Data prediksi dengan regresi runtun waktu menghasilkan RMSE sebesar 1060,80. Hasil prediksi dengan menggunakan regresi time series memberikan hasil yang *overpredict* atau cenderung memprediksi penambahan kasus COVID-19 jauh lebih tinggi daripada kenyataanya. Hasil prediksi dengan menggunakan regresi time series bisa dilihat pada gambar 3.



Gambar 3. Prediksi Penambahan Kasus COVID-19 per Hari di Indonesia dengan Metode Regresi Data Deret Waktu

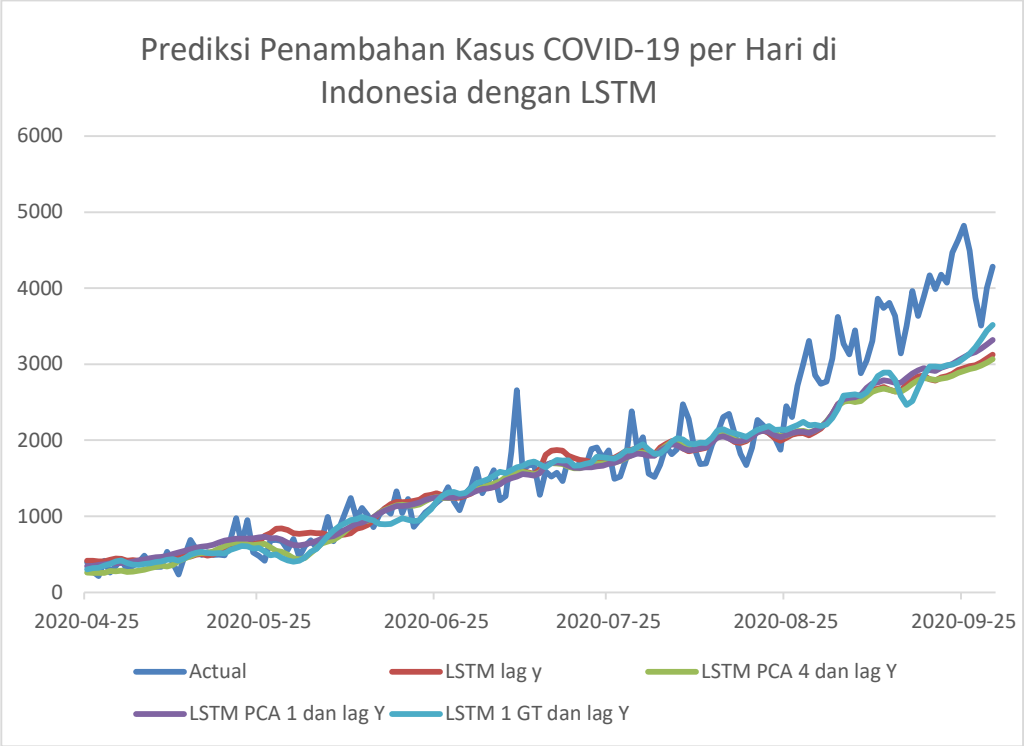
3.3. Prediksi Penambahan Kasus COVID-19 dengan Long-Short Term Memory (LSTM)

Selain dengan metode analisis runtun waktu, prediksi penambahan kasus COVID-19 juga dilakukan dengan menggunakan metode LSTM. Prediksi pembahan kasus COVID-19 dengan LSTM dirancang dalam beberapa skenario. Sebelum membentuk model, keseluruhan data dinormalisasi untuk menghindari dan menghilangkan redudansi data. Kemudian, dataset dibagi menjadi dua, yaitu terdiri dari 80 persen dataset pembelajaran (*training*) dan 20 persen dataset pengujian (*testing*). Skenario yang digunakan untuk memprediksi penambahan kasus COVID-19 dengan LSTM terdiri dari empat skenario, yaitu LSTM dengan menggunakan data lag dari variabel terikat sebagai prediktor, LSTM dengan menggunakan empat komponen utama dan lag dari variabel terikat sebagai prediktor, LSTM dengan menggunakan satu komponen utama dan lag dari variabel terikat

sebagai prediktor, serta LSTM dengan menggunakan satu indeks google trend dengan kata kunci "Covid Indonesia" dan lag dari variabel terikat sebagai prediktor.

Prediksi dengan menggunakan skenario LSTM pertama yang hanya menggunakan lag dari variabel terikat, tanpa memasukkan data Google Trends menghasilkan RMSE sebesar 526,59, sedangkan prediksi dengan menggunakan unsur lag dari variabel terikat dan memasukkan data Google Trends yang diwakili dengan 4 komponen utama memberikan RMSE yang jauh lebih tinggi yaitu 528,81. Hal ini mungkin saja terjadi karena prediktor yang digunakan terlalu banyak sehingga model yang dihasilkan menjadi overfit dan menurunkan tingkat akurasi dari hasil prediksi. Skenario ketiga yaitu prediksi dengan menggunakan LSTM dengan menggunakan 1 komponen utama indeks google trend dan lag variabel terikat sebagai prediktor memberikan RMSE sebesar 483.25. Skenario ketiga memberikan nilai RMSE yang jauh lebih kecil dibandingkan skenario sebelumnya. Sehingga prediksi dengan mengguankan 1 komponen utama indeks google trend dapat mengurangi tingkat kesalahan dan meningkatkan akurasi. Skenario terakhir yaitu prediksi dengan menggunakan LSTM dengan menggunakan 1 indeks google trend yang berkorelasi kuat dan lag dari variabel terikat memberikan RMSE sebesar 482,21. Skenario keempat ini memberikan nilai RMSE yang lebih kecil dibandingkan skenario lainnya. Akan tetapi nilai ini tidak terlalu berbeda jauh dengan skenario ketiga. Sehingga prediksi dengan menggunakan lag dari variabel terikat dan indeks google trend memberikan akurasi yang paling baik dan kesalahan yang paling kecil.

Pada Gambar 4, prediksi dengan LSTM dengan menggunakan unsur lag y dan indeks google trend dihasilkan proyeksi yang lebih baik dibandingkan dengan skenario LSTM lainnya dan regresi time series. Hal ini dapat dilihat dari pergerakan hasil prediksi menggunakan LSTM dengan indeks GT cenderung berhimpit dengan data penambahan kasus COVID-19 yang sebenarnya. Akan tetapi, semenjak bulan September, pergerakan hasil prediksi LSTM COVID-19 tidak berhimpit dengan data penambahan kasus COVID-19. Hal ini bisa menjadi indikasi awal bahwa atensi atau kekhawatiran individu terhadap Pandemi COVID-19 dengan melakukan pencarian yang berhubungan dengan COVID-19 di mesin pencari meningkat, tetapi individu mulai melanggar protokol kesehatan, sehingga atensi dan kekhawatiran individu sudah tidak sejalan lagi dengan penambahan kasus COVID-19.



Gambar 4. Prediksi Penambahan Kasus COVID-19 per Hari di Indonesia dengan LSTM

Hasil ini bisa merepresentasikan bahwa pemikiran, kekhawatiran, kondisi, dan kebutuhan masyarakat Indonesia dalam beberapa periode mampu membantu proses prediksi kasus COVID-19 di Indonesia. Secara lebih lanjut, hasil ini juga sejalan dengan penelitian sebelumnya yang dilakukan untuk memprediksi Influenza dan Zika oleh Santillana, dkk. (2015). Pada penelitian tersebut Santillana, dkk. (2015) mengajukan metode machine learning untuk memprediksi influenza di Amerika Serikat. Dalam penelitiannya, mereka menggunakan data pencarian google dan twitter, catatan kunjungan rumah sakit, serta sistem pengawasan. Hasil penelitian itu menunjukkan bahwa media sosial memberikan informasi yang efektif untuk memprediksi kasus influenza. Selain itu, penelitian McGough, dkk. (2017) juga mengusulkan prediksi Zika dengan menggunakan pencarian google tentang Zika, mikroblog twitter, dan sistem pengawasan digital. Penelitian tersebut juga menunjukkan bahwa sumber data berbasis internet berguna untuk memprediksi kasus mingguan Zika.

Hasil penelitian ini selaras dengan penelitian sebelumnya yang menunjukkan bahwa sumber daya internet bisa membantu dalam peramalan pandemi. Data pencarian, sebagai data yang mudah diperoleh, merupakan sumber yang lebih dinamis dan tersedia

sebagai perbandingan dengan sumber data tradisional. Ini bisa merepresentasikan pemikiran, kekhawatiran, kondisi, dan kebutuhan masyarakat Indonesia dalam beberapa periode.

Tabel 1. Perbandingan Nilai RMSE

Variabel Dependen	Variabel Independen	RMSE	Metode
(Jumlah penambahan kasus COVID-19)	Lag Y	526,59	LSTM
	4 PCA dan Lag Y	528,81	
	1 PCA dan Lag Y	483,25	
	1 GT dan Lag Y	482,21	
	1 GT dan Lag Y	1060,80	Regresi Runtun Waktu

4. SIMPULAN

Prediksi dengan metode analisis regresi deret waktu dan menghasilkan RMSE sebesar 1060,80. Hasil prediksi dengan menggunakan regresi time series memberikan hasil yang *overpredict* atau cenderung memprediksi penambahan kasus COVID-19 jauh lebih tinggi daripada kenyataanya. Selain dengan metode analisis runtun waktu, prediksi penambahan kasus COVID-19 juga dilakukan dengan menggunakan metode LSTM dengan 4 skenario, skenario pertama menggunakan lag dari variabel terikat, tanpa memasukkan data Google Trends menghasilkan RMSE sebesar 526,59, skenario kedua dengan menggunakan unsur lag dari variabel terikat dan memasukkan data Google Trends yang diwakili dengan 4 komponen utama memberikan RMSE sebesar 528,81, skenario ketiga dengan menggunakan 1 komponen utama indeks google trend dan lag variabel terikat sebagai prediktor memberikan RMSE sebesar 483.25, dan skenario terakhir dengan menggunakan 1 indeks google trend yang berkorelasi kuat dan lag dari variabel terikat memberikan RMSE sebesar 482,21. Dari semua hasil prediksi tersebut, prediksi menggunakan metode

LSTM dengan skenario keempat menghasilkan RMSE terkecil, sehingga metode LSTM dengan skenario keempat merupakan metode dengan prediksi yang cukup baik.

5. DAFTAR PUSTAKA

.

Covid19.go.id. Data Covid-19. Diakses pada 11 Oktober 2020, dari https://covid19.go.id/peta-sebaran.

IndoML.com. (2018, 13 April). Pengenalan Long Short Term Memory (LSTM) dan Gated Recurrent Unit (GRU)-RNN Bagian 2. Diakses pada 11 Oktober 2020, dari https://indoml.com/2018/04/13/pengenalan-long-short-term-memory-lstm-dan-gated-recurrent-unit-gru-rnn-bagian-2/.

McGough, S.F., et al.,. (2017). Forecasting Zika incidence in the 2016 Latin America outbreak combining traditional disease surveillance with search, social media, and news report data. PLoS neglected tropical diseases, 2017. 11(1): p. e0005295. Diakses pada 11 Oktober 2020, dari https://doi.org/10.1371/journal.pntd.0005295

Medium.com. (2016, 2 Juli). *What is Google Trends Data-and what does it mean?*. Diakses pada 11 Oktober 2020, dari https://medium.com/google-news-lab/what-is-google -trends-data-and-what-does-it-mean-b48f07342ee8.

Santillana, M., et al.,. (2015). Combining search, social media, and traditional data sources to improve influenza surveillance. PLoS computational biology, 2015. 11(10). Diakses pada 11 Oktober 2020, dari https://doi.org/10.1371/journal.pcbi.1004513

Sharma, Manik & Samriti Sharma. (2020). The Rising Number of COVID-19 Cases Reflecting Growing Search Trend and Concern of People: A Google Trend Analysis of Eight Major Countries. Journal of Medical Systems (2020) 44: 117. DOI: https://doi.org/10.1007/s10916-020-01588-5

WHO.int/indonesia. Pertanyaan dan Jawaban terkait Coronavirus di Indonesia. Diakses pada 11 Oktober 2020, dari https://www.who.int/indonesia/news/novel-coronavirus/qa-for-public.