MILESTONE 4 :
PEMBELAJARAN MESIN DALAM ANALISIS DATA GDELT UNTUK PERUBAHAN IKLIM

KELOMPOK 6 KELAS BD A
Muhamad Musta'in
Muzzammil Fadli
Nashrul Fatah
Novi Dwiasih
Nurul Hestiningtyas

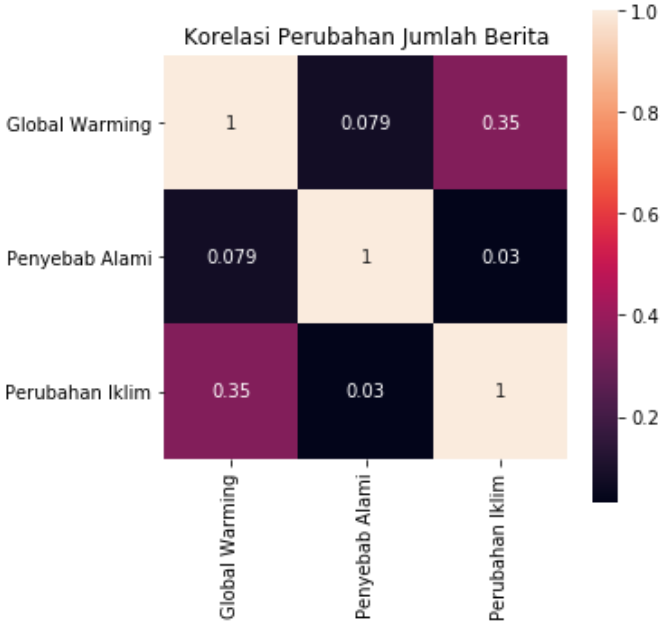
Model Pembelajaran Mesin

Sebagaimana yang telah diketahui, kolom “NumArticles” dalam dataset GDELT memuat jumlah pemberitaan yang dapat dijadikan tolok ukur untuk menentukan apakah suatu kejadian memiliki tingkat signifikansi yang tinggi. Karakteristik seperti yang tercantum dalam Codebook GDELT, semakin banyak pemberitaannya, maka berita tersebut semakin penting. Oleh karena itu, untuk pembelajaran mesin yang akan diterapkan dalam GDELT ini salah satunya memperhatikan jumlah pemberitaan tersebut. Data yang sudah dikelompokkan dalam label tertentu selanjutnya dianalisis. Metode pertama yang dilakukan adalah regresi linear sederhana, dengan memperhatikan jumlah pemberitaan mengenai “Perubahan Iklim” sebagai variabel dependen, dan global warming serta penyebab alami sebagai variabel independen.

Analisis Linear Regresi Sederhana

Sebelum dilakukan analisis regresi linear sederhana, dilakukan analisis korelasi untuk variabel independen dan variabel dependen yang akan dimasukkan dalam model regresi. Diperhatikan bahwa jumlah pemberitaan yang terdapat dalam semua kolom (label) tersebut merupakan data dalam runtun waktu. Dengan demikian, untuk mencari korelasinya, yang dibandingkan adalah percentage of change (perubahan level) dari jumlah pemberitaan tersebut.

Diperoleh, tabel grafik korelasi antar variabelnya adalah sebagai berikut.



Dengan memperhatikan tabel grafik tersebut, diperoleh kesimpulan bahwa ternyata antar variabel tidak berpengaruh sangat signifikan terhadap kejadian perubahan iklim. Korelasi paling tinggi adalah dari kejadian pemanasan global yang memberikan nilai

Pearson Correlation sebesar 0,35. Dengan demikian, sebetulnya dapat secara tidak langsung disimpulkan bahwa kejadian pemanasan global dan kejadian perubahan iklim berhubungan positif, meskipun tidak terlalu kuat.

Selanjutnya, dilakukan analisis regresi linear sederhana untuk menentukan model regresi serta melihat koefisien yang paling mempengaruhi kejadian perubahan iklim.

Dengan menggunakan library *statmodels* di Python, diperoleh keluaran untuk model regresinya adalah sebagai berikut.

Dep. Variable:	Perubahan Iklim	**R-squared:**	0.085
Model:	OLS	**Adj. R-squared:**	0.071
Method:	Least Squares	**F-statistic:**	6.223
Date:	Tue, 27 Nov 2018	**Prob (F-statistic):**	0.00260
Time:	13:01:21	**Log-Likelihood:**	-667.81
No. Observations:	137	**AIC:**	1342.
Df Residuals:	134	**BIC:**	1350.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	24.0002	3.827	6.272	0.000	16.431	31.569
Global Warming	0.1421	0.047	3.010	0.003	0.049	0.235
Penyebab Alami	0.0056	0.006	0.983	0.328	-0.006	0.017

Omnibus:	75.772	**Durbin-Watson:**	1.437
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	295.693
Skew:	2.095	**Prob(JB):**	6.18e-65
Kurtosis:	8.851	**Cond. No.**	748

Diperoleh bahwa koefisien dari kolom pemanasan global ternyata hanya berpengaruh sebesar 0,1421 terhadap perubahan iklim. Nilai R-squared dari model regresi ini sendiri ternyata terlampau kecil, yaitu hanya 0,085. Artinya, 8,5% kejadian perubahan iklim dipengaruhi oleh pemanasan global dan penyebab alami (bencana alam), sedangkan sisanya yaitu 91,5% dipengaruhi oleh faktor lain yang belum diperhatikan di dalam model.

Adapun model regresi linear sederhana yang diperoleh dari output tersebut adalah

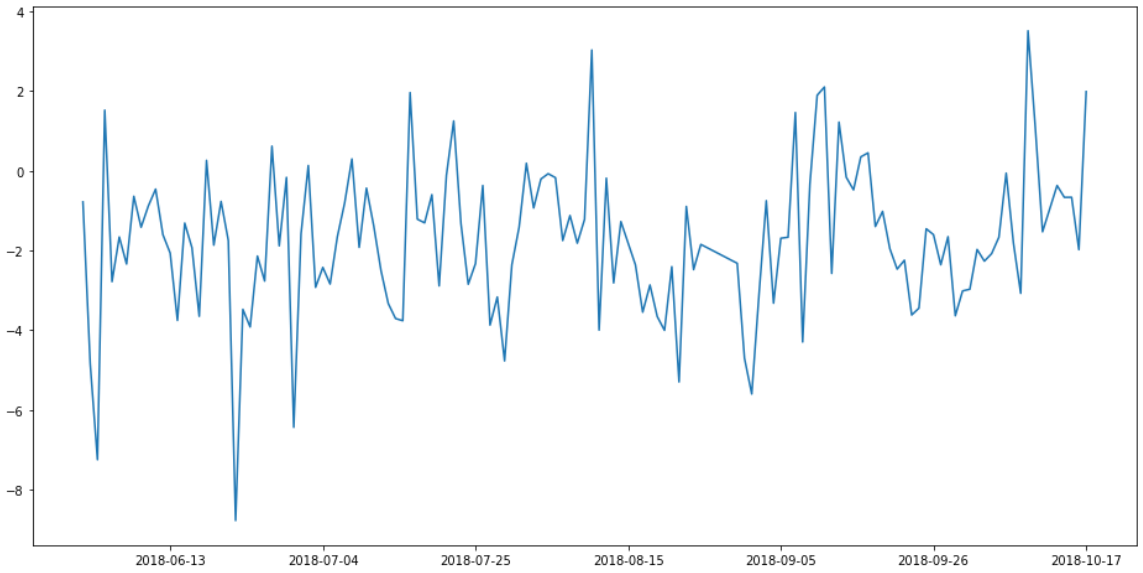
$$PI = 24.0002 + 0.1421 \times GW + 0.0056 \times PA$$

dengan PI adalah kejadian perubahan iklim, GW adalah global warming, dan PA adalah penyebab alami.

Model Runtun Waktu

Dalam pemodelan runtun waktu ini, data yang digunakan adalah data *Average Tone* dari pemberitaan Global Warming dari dataset GDELT. Dataset yang digunakan telah diperiksa tidak memiliki data yang hilang. Data yang tidak lengkap dapat menyebabkan akurasi analisis model ini menjadi kurang baik.

Selanjutnya, dalam memilih metode *time series* (runtun waktu) yang tepat adalah dengan mempertimbangkan pola data, sehingga metode yang paling tepat dengan pola tersebut data diuji. Berikut adalah plot data runtun waktu dari jumlah kejadian Global Warming.



Dapat diperhatikan bahwa data observasi berubah-ubah sekitar tingkatan atau rata-rata yang konstan disebut pola horizontal. Tipe ini pada data runtun waktu disebut stationer dalam rata-rata. Data stasioner didefinisikan sebagai data yang nilai rata-ratanya tidak berubah dari waktu ke waktu atau dapat dikatakan data bersifat stabil.

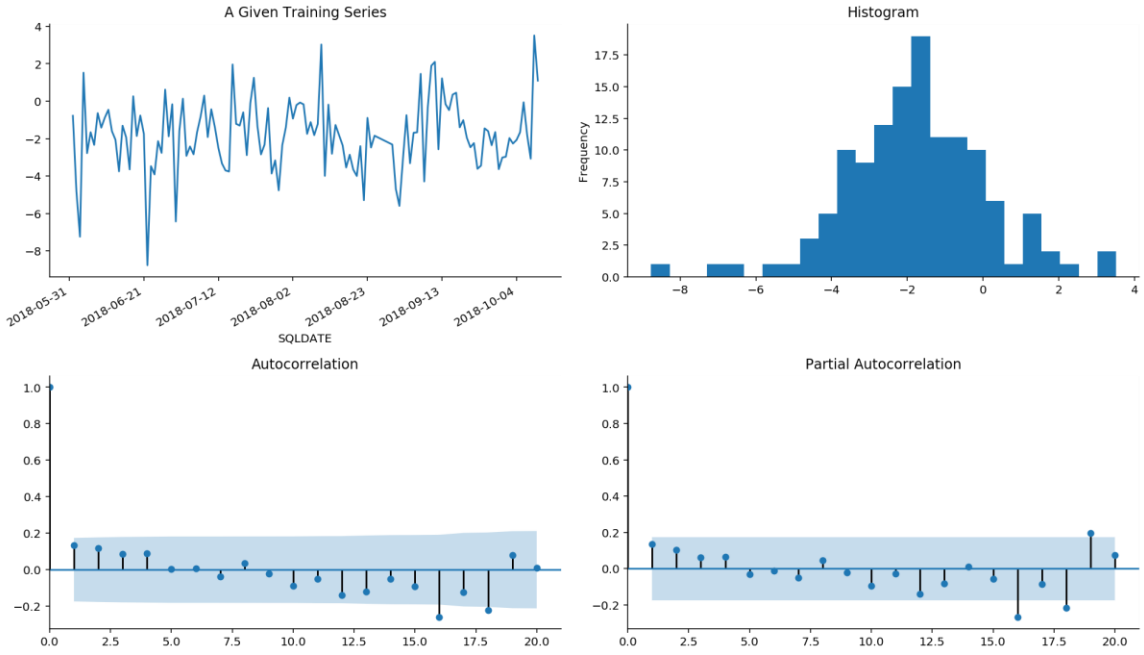
Teknik yang bisa digunakan apabila data stasioner:

1. *Naïve*,
2. *Simple averaging*,
3. *Moving average methods*,
4. *Autoregressive integratrred moving average* (*ARIMA*).

Dalam penelitian ini digunakan metode *ARIMA (Auto Regressive Integrated Moving Average)*.

```
%load_ext autoreload
%autoreload 2
%matplotlib inline
%config InlineBackend.figure_format='retina'
import warnings; warnings.simplefilter('ignore')

from __future__ import absolute_import, division,
print_function

import sys
import os

# TSA from Statsmodels
import statsmodels.api as sm
import statsmodels.formula.api as smf
import statsmodels.tsa.api as smt

pd.set_option('display.float_format', lambda x: '%.5f' % x) #
pandas
np.set_printoptions(precision=5, suppress=True) # numpy

pd.set_option('display.max_columns', 100)
pd.set_option('display.max_rows', 100)
```

Data yang dianalisis selanjutnya di-*split* menjadi data *training* dan data *test*. Selanjutnya, data *training* diplot runtun waktu, histogram, autokorelasi, dan autokorelasi parsialnya.



Hasil dari model ARIMA yang dianalisis dari datanya adalah sebagai berikut.

Dep. Variable:	AvgTone	**No. Observations:**	127
Model:	SARIMAX(2, 0, 0)	**Log Likelihood**	-271.190
Date:	Tue, 27 Nov 2018	**AIC**	548.379
Time:	19:02:53	**BIC**	556.912
Sample:	0	**HQIC**	551.846
	- 127		
Covariance Type:	opg		

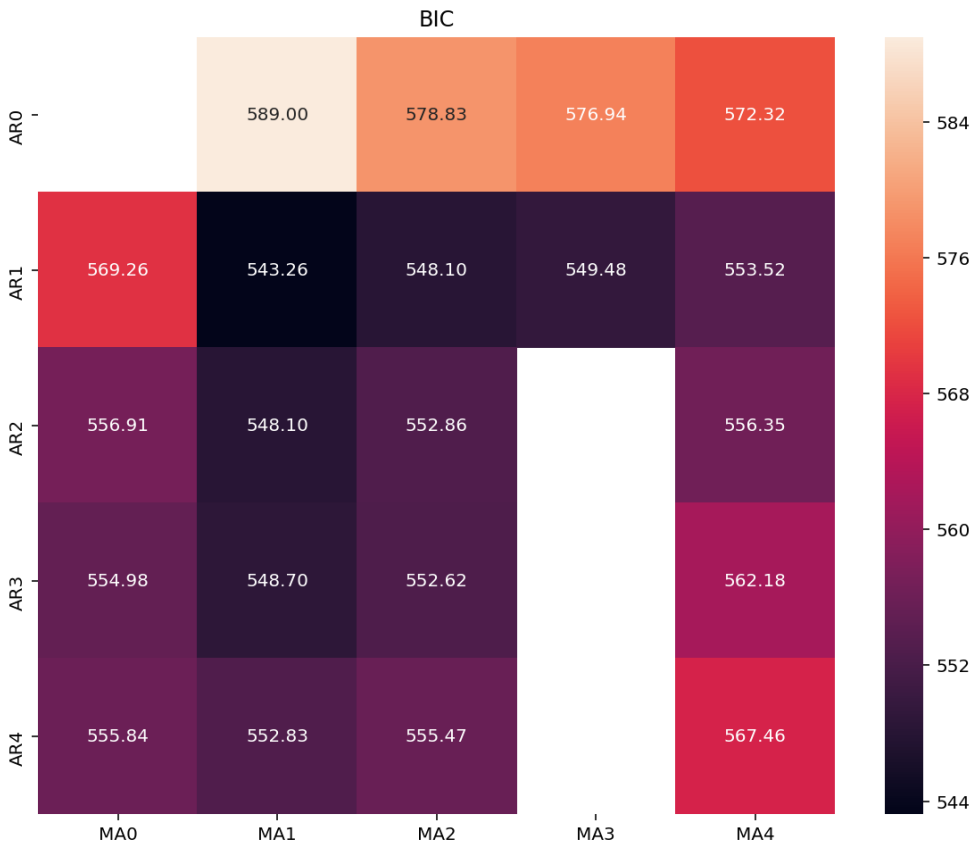
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.3488	0.083	4.183	0.000	0.185	0.512
ar.L2	0.3615	0.099	3.655	0.000	0.168	0.555
sigma2	4.1697	0.376	11.080	0.000	3.432	4.907

Ljung-Box (Q):	47.98	**Jarque-Bera (JB):**	24.91
Prob(Q):	0.18	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.64	**Skew:**	-0.03
Prob(H) (two-sided):	0.15	**Kurtosis:**	5.17

Selanjutnya, dicari lebih dari beberapa model menggunakan petunjuk visual dari plotting sebelumnya sebagai titik awal.

```
import itertools

p_min = 0
d_min = 0
q_min = 0
p_max = 4
d_max = 0
q_max = 4

# Initialize a DataFrame to store the results
results_bic = pd.DataFrame(index=['AR{}'.format(i) for i in range(p_min,p_max+1)],
                           columns=['MA{}'.format(i) for i in range(q_min,q_max+1)])

for p,d,q in itertools.product(range(p_min,p_max+1),
                               range(d_min,d_max+1),
                               range(q_min,q_max+1)):
    if p==0 and d==0 and q==0:
        results_bic.loc['AR{}'.format(p), 'MA{}'.format(q)] = np.nan
        continue

    try:
        model = sm.tsa.SARIMAX(data_train, order=(p, d, q),
                               #enforce_stationarity=False,
                               #enforce_invertibility=False,
                              )
        results = model.fit()
        results_bic.loc['AR{}'.format(p), 'MA{}'.format(q)] = results.bic
    except:
        continue
results_bic = results_bic[results_bic.columns].astype(float)
```

Diperhatikan Bayesian Information Criterion (BIC) sebagai berikut.



Metode pemilihan model alternatif, terbatas hanya mencari parameter AR dan MA ¶

```
train_results = sm.tsa.arma_order_select_ic(data_train,
ic=['aic', 'bic'], trend='nc', max_ar=4, max_ma=4)

print('AIC', train_results.aic_min_order)
print('BIC', train_results.bic_min_order)
```

```
AIC (3, 1)
BIC (1, 1)
```

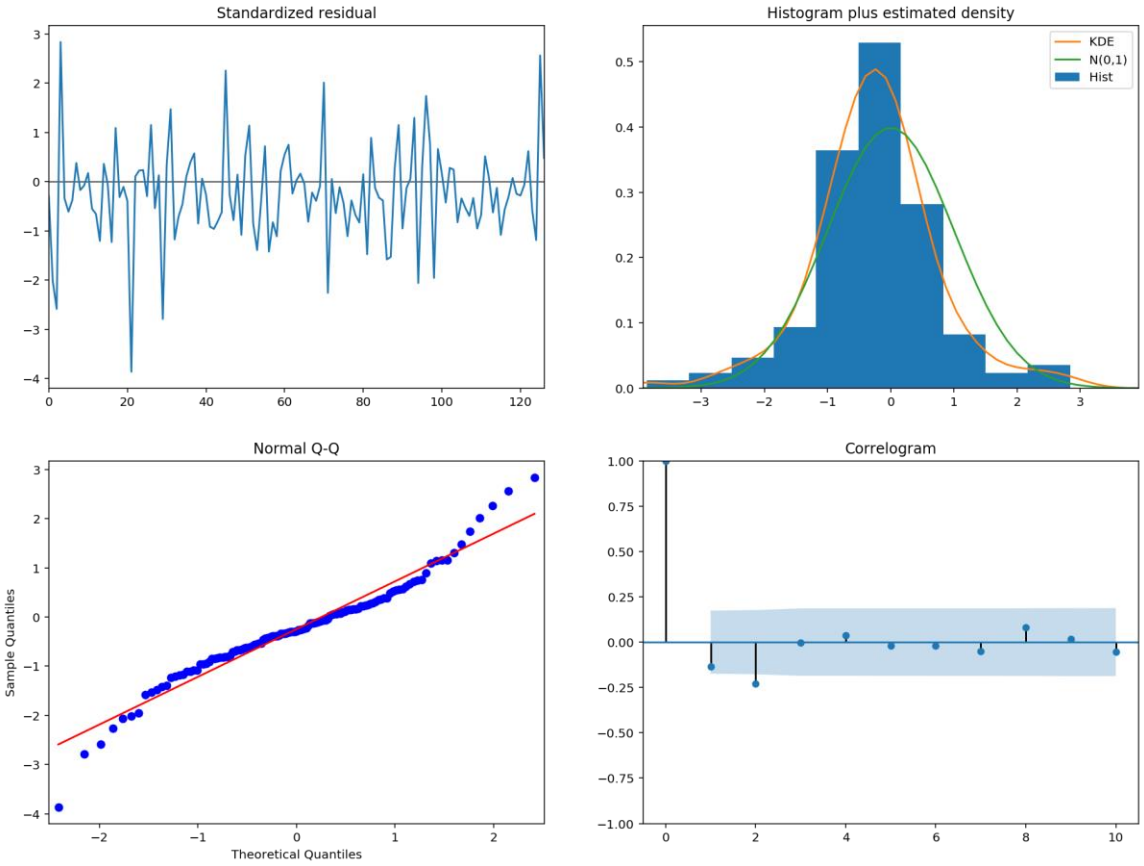
Model Diagnostic Checking

Residual dari model dihitung sebagai perbedaan antara nilai aktual dan nilai yang *fitted*: $e_i = y_i - \hat{y}_i$. Setiap residu adalah komponen observasi yang tidak dapat diprediksi.

Ketika data merupakan *time series*, harus melihat plot ACF residual. Plot ACF residual akan memperlihatkan jika ada autokorelasi di residual (menunjukkan bahwa ada informasi yang belum diperhitungkan dalam model).

Outliers menunjukkan ada sesuatu yang tidak biasa terjadi. Akan sangat berharga untuk menyelidiki pencilan itu untuk melihat apakah ada keadaan atau kejadian yang tidak biasa.

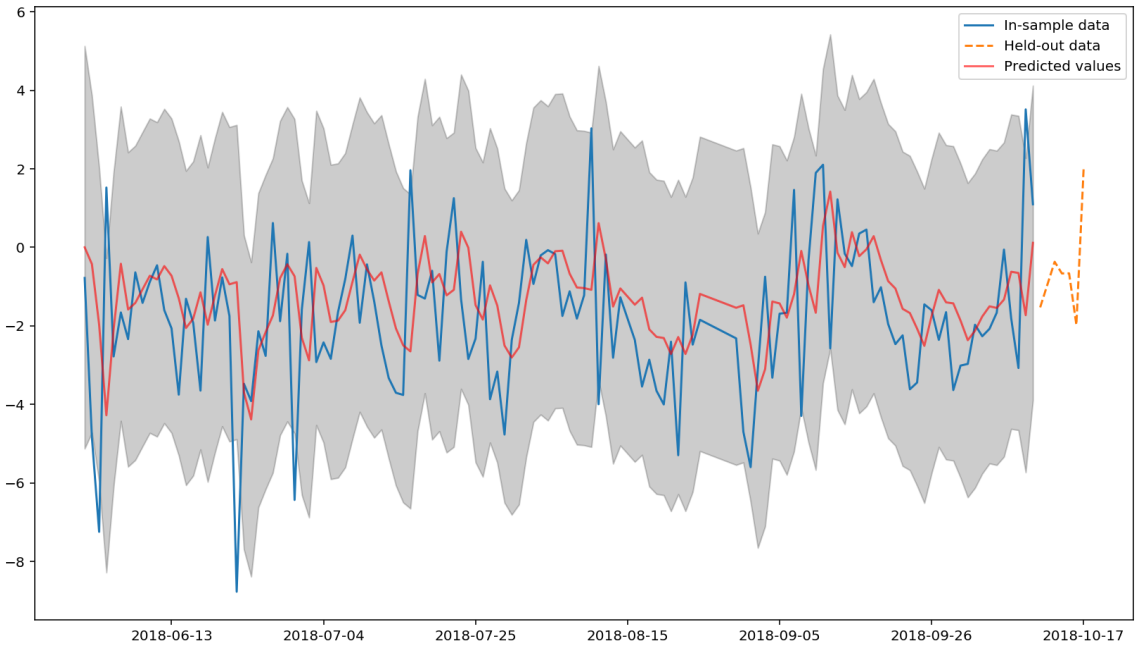
Histogram: baik untuk memeriksa apakah residu terdistribusi secara normal. Seperti yang dijelaskan sebelumnya, ini tidak penting untuk meramalkan, tetapi itu membuat perhitungan interval prediksi jauh lebih mudah.



Beberapa tes (*formal testing*) yang dilakukan menghasilkan keluaran sebagai berikut.

```
Test heteroskedasticity of residuals (breakvar): stat=0.641, p=0.154

Test normality of residuals (jarquebera): stat=24.906, p=0.000

Test serial correlation of residuals (ljungbox): stat=47.984, p=0.181

Durbin-Watson test on residuals: d=2.13
       (NB: 2 means no serial correlation, 0=pos, 4=neg)

Test for all AR roots outside unit circle (>1): True

Test for all MA roots outside unit circle (>1): True
```

Evaluasi performansi model apabila diplot akan menghasilkan grafik seperti berikut.



Dilakukan prediksi untuk data test yang merupakan data Average Tone dari tanggal 11 Oktober 2018 sampai 17 Oktober 2018, dan data yang belum ada yaitu data pada tanggal 18 Oktober 2018. Diketahui, data test aktualnya dari data yang sudah ada adalah sebagai berikut.

```
SQLDATE
2018-10-11   -1.52889
2018-10-13   -0.36819
2018-10-14   -0.66522
2018-10-15   -0.66332
2018-10-16   -1.97722
2018-10-17    1.98598
Name: AvgTone, dtype: float64
```

Diperoleh, data yang diprediksi adalah sebagai berikut.

```
128   0.97286
```

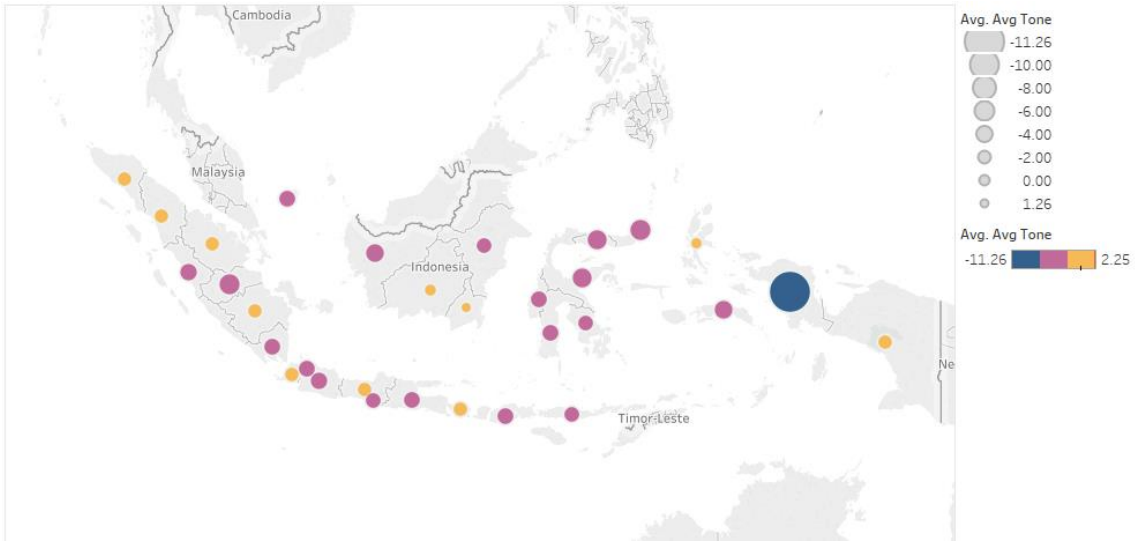
```
129    0.93672
130    0.67844
131    0.57529
132    0.44594
133    0.36353
dtype: float64
```

Selanjutnya, untuk data average tone yang belum ada pada tanggal 18 Oktober 2018 dihasilkan average tone Global Warmingnya adalah 0,28802, yang berarti ada kejadian positif dari global warming (yang dapat berbentuk penanganan global warming atau mitigasi). Dalam tahapan lebih lanjut, prediksi dengan runtun waktu ini dapat digunakan untuk memperkirakan apakah dalam rentang waktu ke depan pemberitaan mengenai global warming ini berangsur membaik atau memburuk.

Model Sebaran dan Visualisasi

Model sebaran yang ditunjukkan dalam subbab berikut memperhatikan tonasi kejadian yang diambil dari angka AVGTone dalam kolom yang termuat dalam dataset GDELT. Tonasi tersebut kemudian dipetakan ke peta Indonesia untuk melihat daerah-daerah terdampak dari label-label kejadian yang telah dikelompokkan. Tonasi dengan angka negatif menunjukkan kejadian negatif, 0 menunjukkan netral, dan positif menunjukkan kejadian positif.

A. Penyakit

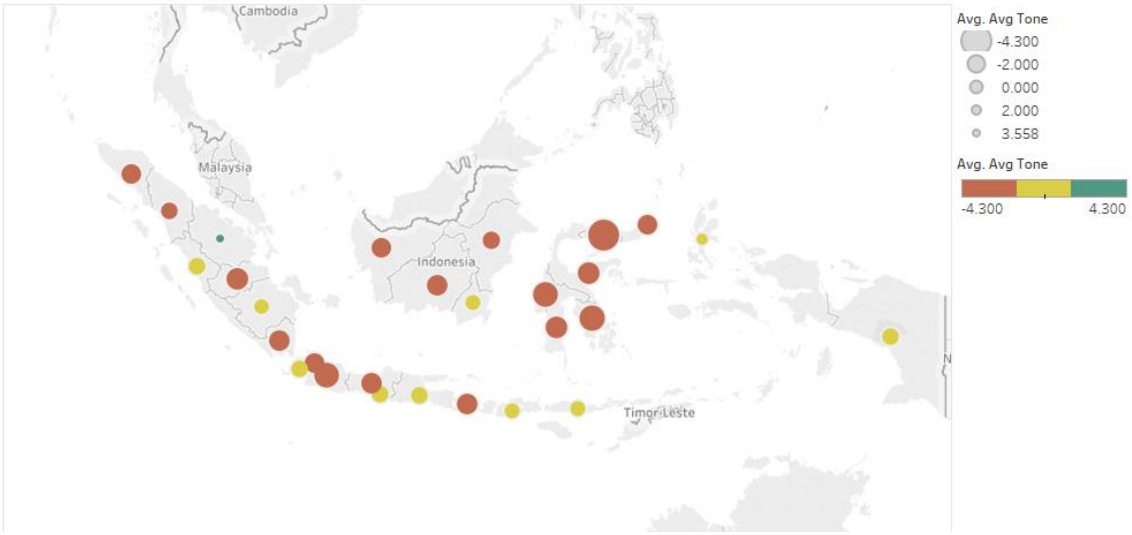


Gambar . Persebaran nilai Avg.Tone terhadap Pemberitaan tentang "Penyakit"

Persebaran rata-rata tonasi terhadap berita/kejadian yang berhubungan dengan "Penyakit" di Indonesia memiliki nilai antara -11,26 sampai 2,25. Nilai tonasi paling negative (-) berada di Papua Barat dengan nilai tonasi -11,26.

Daerah di Papua Barat memiliki dampak/kejadian yang serius terhadap pemberitaan yang berhubungan dengan "Penyakit".

B. Global Warming (Pemanasan Global)

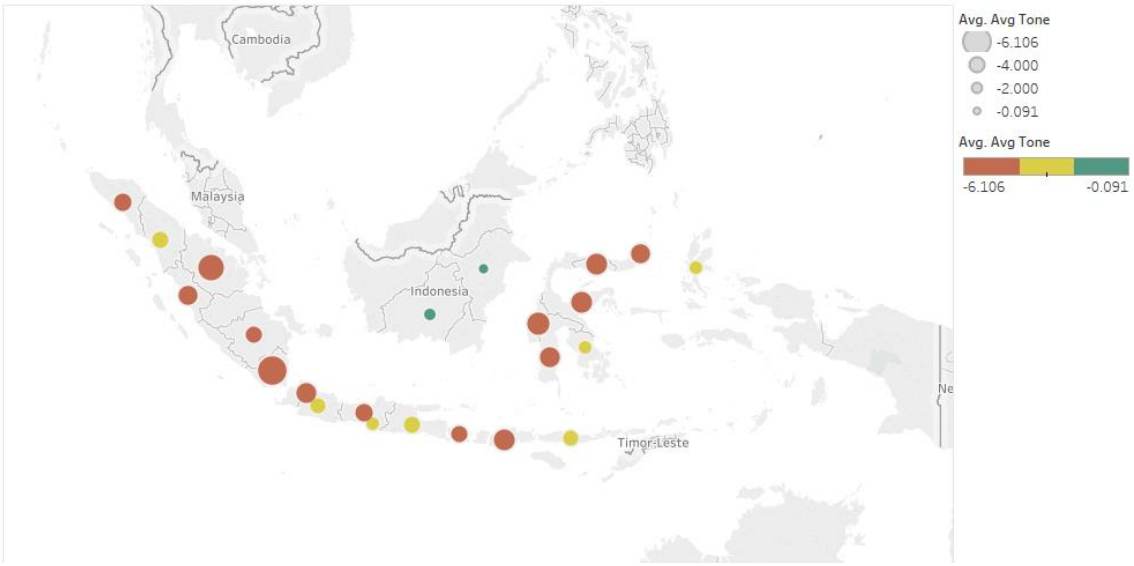


Gambar. Persebaran nilai Avg.Tone terhadap Pemberitaan tentang "Global Warming"

Daerah di Indonesia yang memiliki nilai tonasi paling negative (-) hampir berada di pulau-pulau besar di Indonesia, kecuali Pulau Irian. Sedangkan kota dengan tonasi paling negative berada di Kota Gorontalo dengan nilai notasi -4,30.

Sebagian besar kota-kota besar di Indonesia memiliki dampak/event yang berdampak serius terhadap issue "Global Warming".

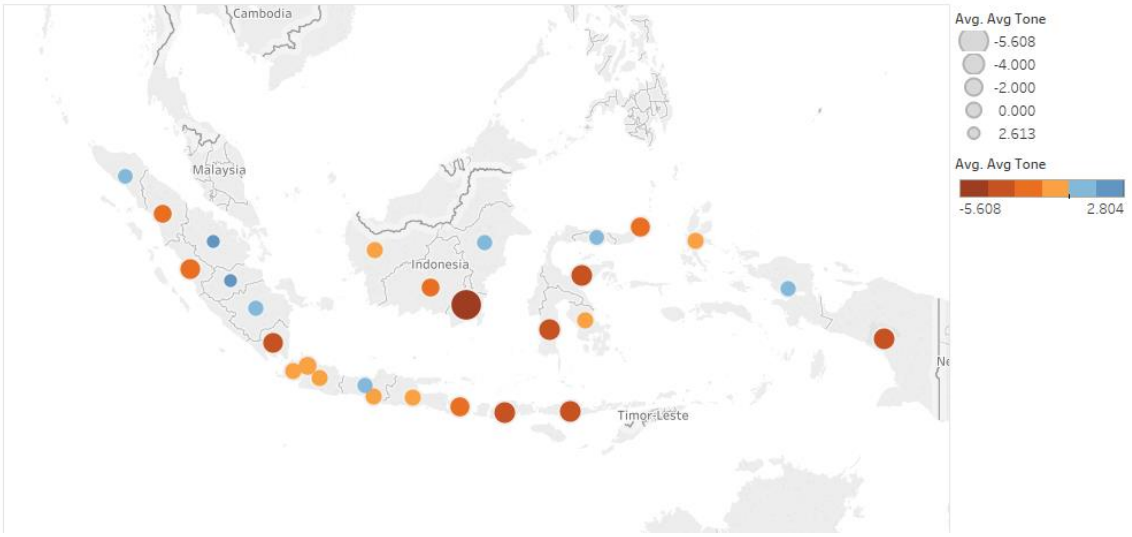
C. Penyebab Alami Perubahan Iklim (Bencana Alam Gunung Meletus)



Gambar. Persebaran Nilai Avg.Tone terhadap Pemberitaan tentang Penyebab Perubahan Iklim yang Berasal dari Alam.

Wilayah di Indonesia yang memiliki dampak serius (tonasi negative) terhadap penyebab perubahan iklim karena factor alam adalah sebagian besar wilayah di Sulawesi, Jawa dan Sumatera. Sedangkan daerah dengan tonasi positif berada di wilayah Pulau Kalimantan. Secara logis, hal ini juga dipengaruhi bahwa tidak ada gunung berapi di Pulau Kalimantan. Hal ini juga terjadi di Pulau Papua yang tidak memiliki indikasi terdapat kejadian di sana.

D. Perubahan Iklim



Gambar. Pesebaran Nilai Tonasi terhadap Pemberitaan tentang "Perubahan Iklim"

Daerah dengan tonasi paling negative terhadap "Perubahan Iklim" adalah Kalimantan Selatan. Sedangkan daerah di Pulau Jawa juga memiliki dampak yang cukup serius terhadap "Perubahan Iklim". Hal ini berbanding terbalik dengan daerah-daerah di Pulau Sumatera yang sebagian besar wilayah memiliki tonasi yang positif.