Anastasia Indri T.K., Bernardia Vitri A., Yannissa M.R.
Group 2

BIG DATA CLASS C
MILESTONE 1

5 November 2018

Studying Mass Media Influence on Tourist Visits in Indonesia with Global Database of Events, Language, and Tone

Anastasia Indri T.K., Bernardia Vitri A., Yannissa M.R.
Group 2

Deskripsi Proyek

Kami akan melakukan analisis terhadap hubungan berita tentang pariwisata di Indonesia di luar negeri dengan tingkat kunjungan wisatawan mancanegara ke Indonesia. Kami menggunakan data Gdelt sebagai acuan untuk menentukan sentiment berita dan asal berita. Sentimen berita ditentukan dengan meng-cluster AvgTone menjadi positif, neutral, dan negatif. Asal berita ditentukan melalui domain URL berita. Hasil dari peng-clusteran akan digunakan sebagai dataset final Gdelt yang akan digabung dengan dataset kunjungan wisatawan mancanegara BPS. Dataset yang akhir ini yang akan digunakan untuk menjawab pertanyaan penelitian.

Deskripsi Dataset

1. Judul : Event Turisme Indonesia 2017
2. Sumber : GDELT (https://www.gdeltproject.org/)
   Tanggal : 1 January 2017 – 31 December 2017
3. Informasi Relevan :
   Data diambil dari GDELT Project dengan mengambil tema turisme di Indonesia, sebagai salah satu negara dengan potensi turisme yang tinggi. Dataset ini memiliki atribut Month, AvgTone, SOURCEURL, dan country source
4. Jumlah sampel : 1350
5. Jumlah atribut : 4
6. Informasi atribut :
   1. Month
   2. AvgTone in between -100 until +100. AvgTone adalah rerata tone semua dokumen even. Nilainya berkisar dari -100 (sangat negatif) hingga +100 (sangat positif). Nilai umum berkisar antara -10 dan +10, dengan 0 mengindikasikan kenetralan. Nilai AvgTone dapat digunakan sebagai metode penyaringan konteks event dan dampak dari event tersebut.
   3. SOURCEURL
   4. country source adalah identitas negara berdasarkan domain
7. Missing Attribute Values : None

Data Exploration

Untuk menentukan apakah tren di media massa dapat mempengaruhi tingkat kunjungan wisatawan ke tujuan ini, data kunjungan wisatawan diperlukan. Data pendukung yang akan digunakan adalah Data Kunjungan Wisata 2017 per provinsi yang diperoleh dari Biro Pusat Statistik Indonesia.

Anastasia Indri T.K., Bernardia Vitri A., Yannissa M.R.
Group 2

BIG DATA CLASS C MILESTONE 2

15 November 2018

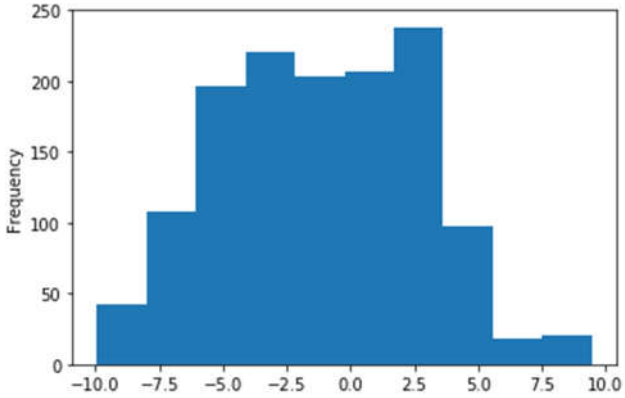
Eksplorasi Statistik Dataset

Dengan AvgTone, dapat dilihat bahwa angka negatif seperti -9.93 terhubung dengan event-event yang buruk seperti “Serangan Bunuh Diri”, dan angka positif terhubung dengan events yang positif seperti “Kemenangan Penghargaan Pemasaran.”

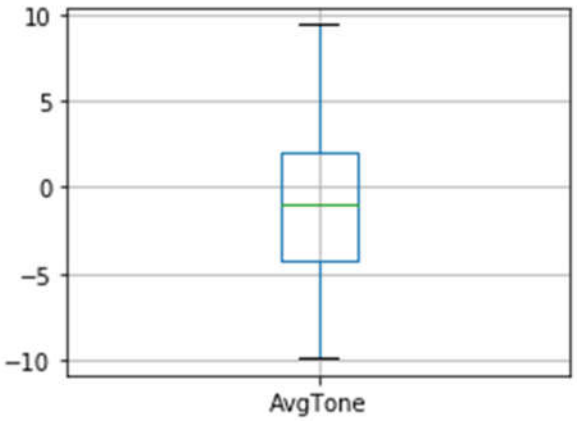
Event Turisme Indonesia 2017

	count	mean	std	min	25%	50%	75%	max
AvgTone	1350	-1.14	3.87	-9.93	-4.23	-1.02	1.99	9.45

Kita dapat melihat bahwa kebanyakan ialah event negatif pada berita turisme. Dimana dapat dilihat bahwa mean dan median ialah berita dengan AvgTone negatif. Dibawah ini ialah histogram persebaran AvgTone dan boxplotnya.



Gambar 1. Histogram Frekuensi AvgTone



Gambar 2. AvgTone Boxplot

Anastasia Indri T.K., Bernardia Vitri A., Yannissa M.R.
Group 2

Research Question

1. Apakah terdapat hubungan berita Pariwisata Indonesia di luar negeri dan jumlah kunjungan wisatawan mancanegara ke Indonesia?
2. Apakah berita secara global dapat menjadi alat promosi pariwisata Indonesia?

Anastasia Indri T.K., Bernardia Vitri A., Yannissa M.R.
Group 2

BIG DATA CLASS C MILESTONE 3

20 November 2018

Model Machine Learning

Model machine learning yang paling relevan ialah KMeans.

Dasar Metode Pemilihan Model

Metode K-Means dipilih berdasarkan Flowchart Map Machine Learning dari https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Berikut urutan pertanyaan dari Flowchart Map Machine Learning sehingga didapat metode K-means

1. >50 samples = Ya (Data berjumlah 1350)

2. Predicting a category = Ya (Sentimen berita dari AVGTone)

3. Do you have labeled data? = Tidak

4. Number of categories known = Ya (AVGTone)

5. <10 = Ya

Maka metode yang dapat digunakan sesuai Map adalah K-means. Jika tidak bekerja sesuai Map juga dapat dilanjutkan menggunakan Spectral Clustering atau GMM.

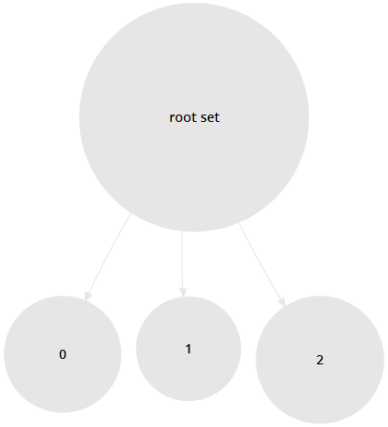
Anastasia Indri T.K., Bernardia Vitri A., Yannissa M.R.
Group 2

BIG DATA CLASS C MILESTONE 4

27 November 2018

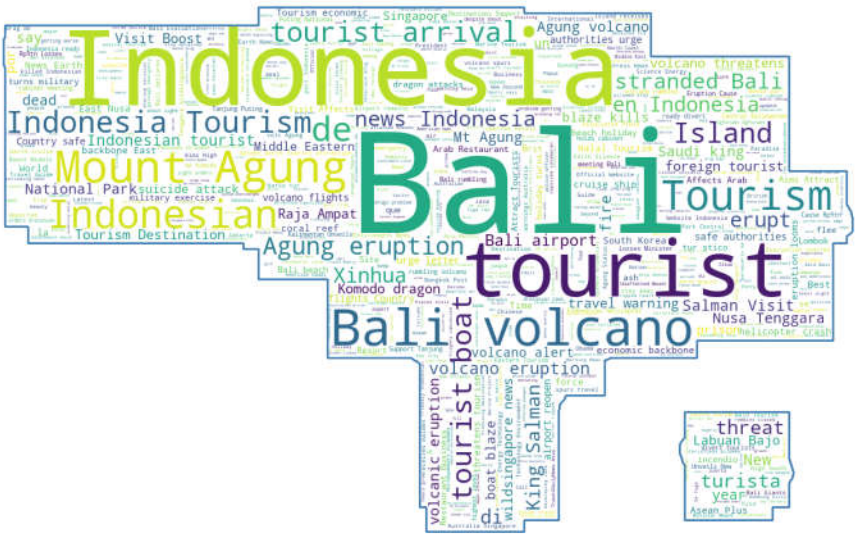
Hasil K-Means

Didapatkan hasil terbaik K-Means menggunakan k=3 sehingga didapat 3 kluster dengan nilai DBi yaitu -0.559. Lalu cluster dilabeli dengan sentimen event yaitu positif, negatif dan netral.



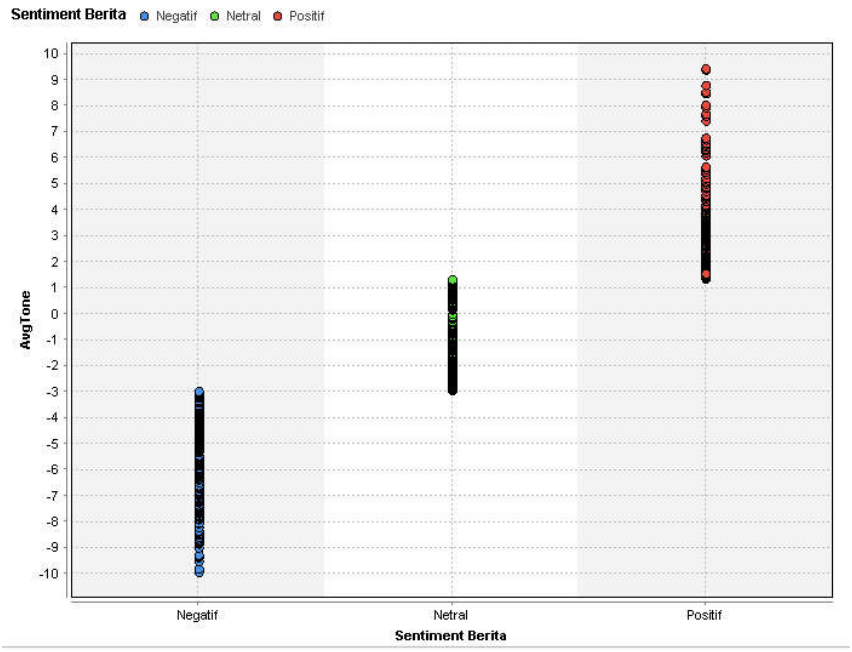
Gambar 4. Visualisasi K-Means

Data Visualisations



Gambar 11. Visualisasi Top Words Judul Berita

Dapat dilihat pada Gambar 11 di atas bahwa kata-kata utama dalam judul berita yang didapatkan dari URL dataset antara lain Indonesia, Bali, Mount Agung, tourist, dan volcano. Visualisasi ini menggunakan fitur wordcloud dengan Python. Lalu Kami menggunakan Rapid Miner dalam melakukan proses perhitungan dan visualisasi metode K-Means. Scatter plot untuk tiap cluster sentiment berita dapat dilihat pada plot dibawah ini.

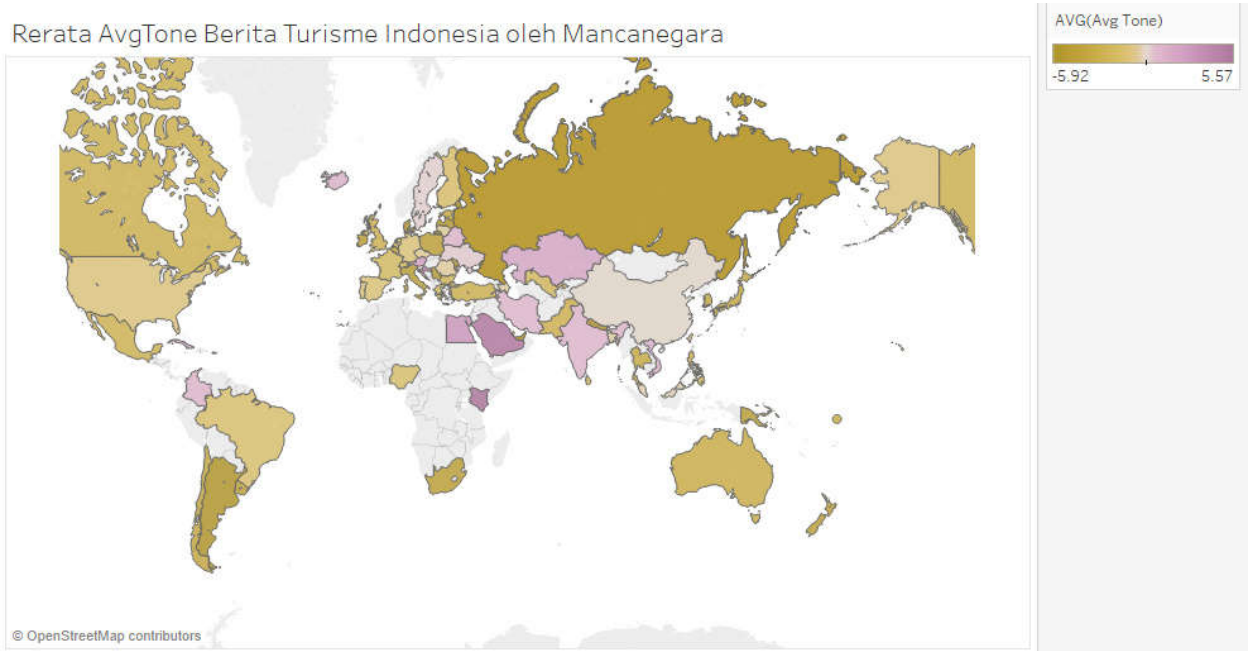


Gambar 5. Persebaran Sentiment Berita setiap cluster

Selanjutnya kami menggunakan Tableau untuk memvisualisasikan jumlah berita dan sentiment berita per negara dan bulan. Berikut peta persebaran jumlah sumber berita dari mancanegara tentang turisme Indonesia. Dari peta ini diketahui bahwa negara yang paling banyak menulis tentang turisme Indonesia ialah Amerika Serikat.

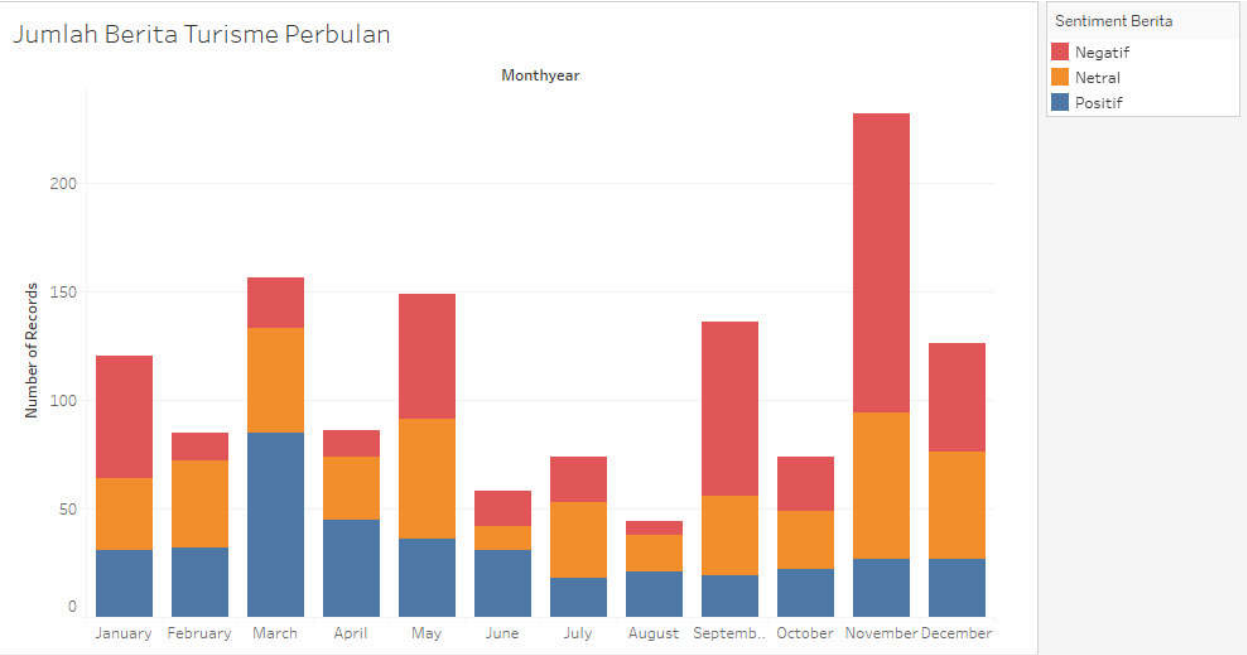
Gambar 6. Sumber Berita Mancanegara tentang Turisme Indonesia

Berikut peta persebaran berita dengan warna yang menjadi indikator rerata AvgTone berita turisme Indonesia yang dihasilkan oleh negara tersebut. Dari peta ini diketahui bahwa negara Saudi Arabia menjadi negara sumber berita turisme Indonesia yang rata-rata beritanya paling positif. Sedangkan Rusia sebaliknya. Negara-negara lain dengan rerata berita positif tentang turisme Indonesia antara lain Mesir, Ukraina, Islandia, dan Swedia.



Gambar 7. Rerata AvgTone Berita Turisme Indonesia oleh Mancanegara

Pada Gambar 10 di bawah ini dapat dilihat grafik jumlah berita turisme setiap bulannya yang diterbitkan oleh negara selain Indonesia. Dari grafik ini diketahui bahwa jumlah berita tentang turisme Indonesia oleh negara lain paling banyak diterbitkan pada bulan November, sedangkan paling sedikit diterbitkan pada bulan Agustus. Pada bulan November pula didapati berita dengan sentiment negatif terbanyak. Hal ini dikarenakan pada bulan November tersebut ada erupsi Gunung Agung di Bali yang menyumbang banyak berita dengan AvgTone yang negatif. Sedangkan pada bulan Maret didapati berita dengan sentiment positif terbanyak. Hal ini dikarenakan pada bulan Maret terdapat kunjungan Raja Salman dan pangeran-pangeran Arab Saudi ke Bali. Hal ini senada dengan Gambar 7 di atas yang menyatakan bahwa nilai AvgTone berita dari negara Arab Saudi menjadi yang paling tinggi.



Gambar 10. Jumlah Berita Turisme Perbulan

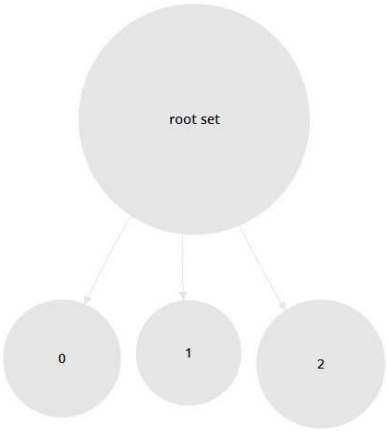
Anastasia Indri T.K., Bernardia Vitri A., Yannissa M.R.
Group 2

BIG DATA CLASS C MILESTONE 5

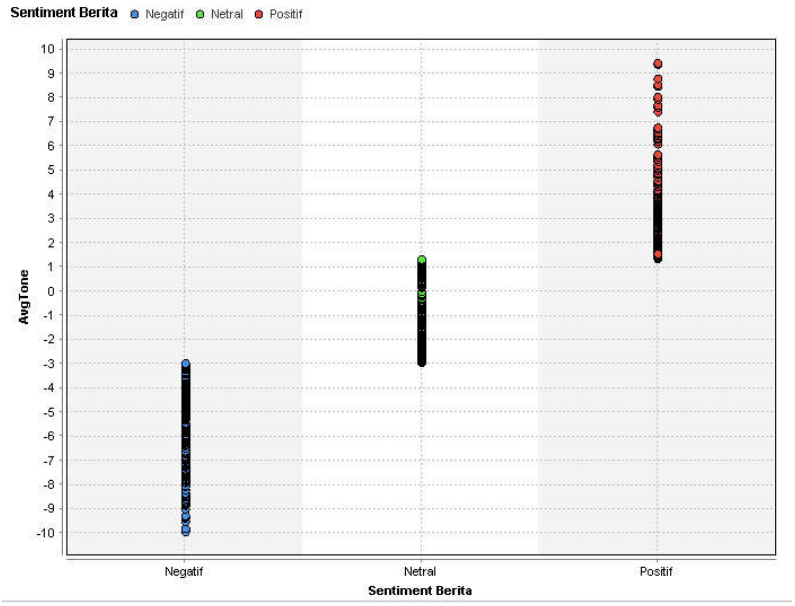
6 Desember 2018

Results and Discussion

Didapatkan hasil terbaik K-Means menggunakan k=3 sehingga didapat 3 kluster dengan nilai DBi yaitu -0.559. Lalu cluster dilabeli dengan sentimen event yaitu positif, negatif dan netral.sesuai nilai AVGTone pada rata - rata Cluster.

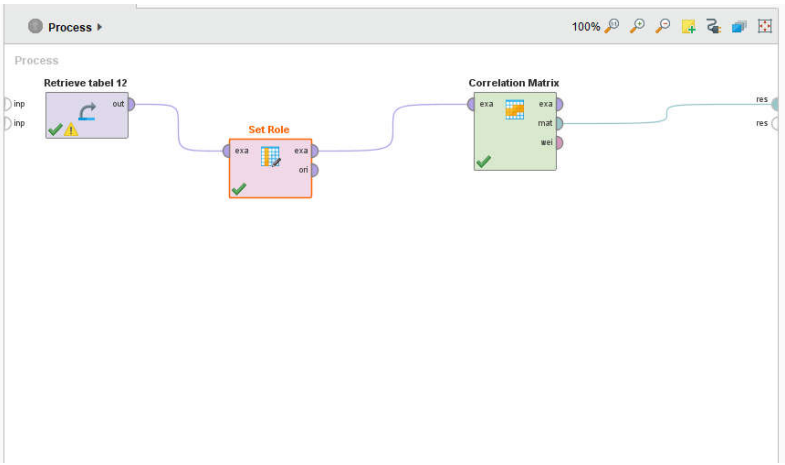


Gambar 12. Visualisasi K-Means



Gambar 13. Persebaran AvgTone setiap cluster

Korelasi dilakukan untuk menguji hubungan antara Jumlah Berita, AvgTone, dan Jumlah Kunjungan Wisatawan Mancanegara. Kami menggunakan tools Rapidminer dalam pengujian ini. Berikut adalah peta kerja di rapidminer:



Gambar 14. Peta Kerja Rapidminer untuk Correlation Matrix

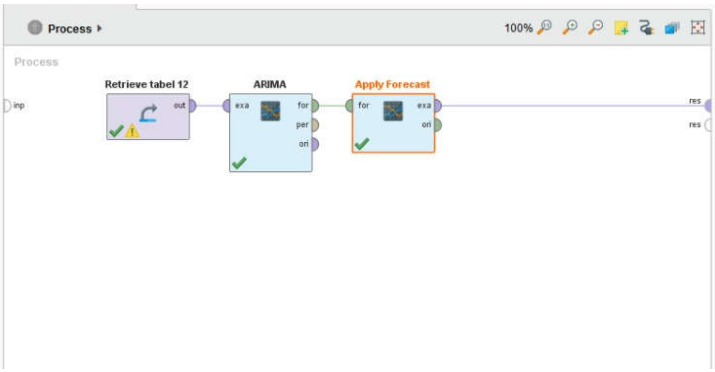
Berikut hasilnya:

Attribut...	Jumlah...	AvgTone	TotalWi...
JumlahB...	1	-0.637	-0.503
AvgTone	-0.637	1	0.082
TotalWis...	-0.503	0.082	1

Gambar 15. Correlation Matrix

Berdasarkan hasil dari perhitungan korelasi, didapatkan bahwa terdapat korelasi kuat negatif pada jumlah berita dengan AvgTone dan Total Wisatawan Mancanegara dengan Jumlah Berita. Hubungan AvgTone dengan Total wisatawan memiliki korelasi positif. Semakin tinggi AvgTone, maka semakin tinggi jumlah kunjungan wisatawan.

Masih menggunakan Rapidminer, kami membuat model untuk prediksi tingkat kunjungan dengan melihat AvgTone pada berita di bulan tersebut. Kami menggunakan model ARIMA dengan parameter default. Berikut adalah peta kerja di Rapidminer:



Gambar 16. Peta Kerja Rapidminer untuk ARIMA

Prediksi tingkat kunjungan bulan pertama hingga bulan kelima tahun 2018 dapat dilihat pada Gambar 17 dibawah ini pada baris ketiga-belas dan seterusnya.

ExampleSet (17 examples, 2 special attributes, 2 regular attributes)

Row No.	Bulan	forecast of ...	TotalWisata...	TotalWisata...
1	1	?	1085629	1085629
2	2	?	1006406	1006406
3	3	?	1039224	1039224
4	4	?	1151653	1151653
5	5	?	1127140	1127140
6	6	?	1117121	1117121
7	7	?	1351325	1351325
8	8	?	1374634	1374634
9	9	?	1232213	1232213
10	10	?	1145037	1145037
11	11	?	1040941	1040941
12	12	?	1126645	1126645
13	13	1183134.932	?	1183134.932
14	14	1157724.519	?	1157724.519
15	15	1151701.685	?	1151701.685
16	16	1150274.140	?	1150274.140
17	17	1149935.780	?	1149935.780

Gambar 17. Hasil Pemodelan ARIMA

Conclusions

Berdasarkan hasil penelitian, berikut adalah jawaban atas pertanyaan penelitian:

1. Berita memiliki hubungan terhadap tingkat kunjungan wisatawan mancanegara ke Indonesia. Penelitian menunjukkan bahwa kunjungan dipengaruhi secara positif oleh sentiment berita.
2. Berita secara global dapat dijadikan sebagai alat promosi pariwisata terutama berita positif.