UAS Eksplorasi dan Visualisasi Data

Kelas B

No	Nama Lengkap	NPM	Kontribusi
1	Vanny Khairunnisaa	2206051506	Melakukan proses *coding* dan merancang ide-ide analisis data
2	Makayla Adzra Kumaladewi	2206053884	Membuat laporan *preprocessing* data
3	Khalila Izzatunnisa	2206051544	Membuat laporan *preprocessing* data

I. Pendahuluan

Sea Games 2023 ke-23 resmi diadakan di Phnom Penh, Kamboja pada tanggal 5 - 17 Mei 2023. Melalui sumber dari situs web berita detik.com dengan kata kunci “Sea Games”, kami memperoleh data yang kemudian akan dianalisa kemungkinan kata kunci lain yang relevan dengan kata kunci “Sea Games”. Kami juga melakukan penyaringan data berdasarkan waktu dalam rentang 5 April 2023 - 17 Juni 2023 untuk melihat tren kata kunci “Sea Games” yang dibagi berdasarkan lini masa pra-Sea Games 2023, Sea Games 2023, dan pasca-Sea Games 2023.

a) Ukuran dan Tipe Data

	Headline	Date	Link	Content
0	Lolos 8 Besar Indonesia Open, Leo/Daniel Belum...	15 Jun 2023 20:52	https://sport.detik.com/raket/d-6775241/lolos-...	Leo Rolly Carnando/Daniel Marthin sukses menju...
1	Andai Ramadhan Sananta Main saat Indonesia Vs ...	15 Jun 2023 10:30	https://www.detik.com/sulsel/sepakbola/d-67736...	Timnas Indonesia kesulitan mencetak gol saat m...
2	Indonesia Sulit Gol Lawan Palestina, Netizen: ...	15 Jun 2023 00:00	https://sport.detik.com/sepakbola/liga-indones...	Timnas Indonesia kesulitan membuat gol saat me...
3	Lawan lebih Tenang, Chico Gagal Menang	14 Jun 2023 12:59	https://sport.detik.com/raket/d-6771988/lawan-...	Chico Aura Dwi Wardoyo tersingkir di babak per...
4	Lawan lebih Tenang, Chico Gagal Menang	14 Jun 2023 12:39	https://sport.detik.com/sepakbola/liga-indones...	Chico Aura Dwi Wardoyo tersingkir di babak per...
...	...	...	...	...
392	Indonesia Vs Argentina Dikomentari Nyinyir Net...	28 Mei 2023 04:12	https://www.detik.com/sumbagsel/sepakbola/d-67...	Rencana Fifa Matchday Indonesia Vs Argentina t...
393	Kapan dan di Mana FIFA Matchday Indonesia Vs A...	25 Mei 2023 00:49	https://www.detik.com/sumbagsel/sepakbola/d-67...	Ketum PSSI Erick Thohir menjanjikan laga ciami...
394	Cerita Astrid Rahmad, Atlet Hoki Peraih Emas S...	24 Mei 2023 21:03	https://www.detik.com/sumbagsel/berita/d-67372...	Astrid Rahmad, salah satu atlet cabang olahrag...
395	Melihat JSC, Venue Olahraga yang Kini Jadi Tem...	23 Mei 2023 15:28	https://www.detik.com/sumbagsel/wisata/d-67342...	Bagi masyarakat Sumatera Selatan (Sumsel) past...
396	7 Atlet Peraih Medali SEA Games 2023 Terima Ta...	23 Mei 2023 14:48	https://www.detik.com/sumbagsel/berita/d-67343...	Gubernur Jambi Al Haris menyerahkan tali asih ...

397 rows × 4 columns

Penyaringan data berita dengan kata kunci “Sea Games” berdasarkan rentang waktu 5 April 2023 - 17 Juni 2023 menghasilkan data dengan ukuran 397 baris x 4 kolom dengan tipe data sebagai berikut :

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 397 entries, 0 to 396
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   Headline  397 non-null    object
 1   Date      397 non-null    object
 2   Link      397 non-null    object
 3   Content   397 non-null    object
dtypes: object(4)
memory usage: 12.5+ KB
```

b) Variabel Data

- Headline : deskripsi singkat berupa judul yang merepresentasikan keseluruhan konten atau isi berita.
- Date : tanggal dan waktu (dalam jam dan menit) pada saat berita dipublikasikan.
- Link : tautan yang mengarah ke halaman berita.
- Content : keseluruhan dari isi berita.

  Selanjutnya dalam proses pengolahan data, variabel “Link” akan dihilangkan sehingga hanya dipilih variabel “Headline”, “Date”, dan “Content” untuk dianalisis lebih lanjut.

II. Pengolahan Data

a) *Scraping* Data

Scraping data adalah proses mengambil atau mengumpulkan data dari *website*. Kami menggunakan Python library, yaitu BeautifulSoup. Pada proses ini, kami mengekstrak data dari element yang dipilih, yaitu link, headline, date, dan content. Setelah itu, kami menjalankan skrip atau kode yang sebelumnya didefinisikan, lalu menyimpan data yang dihasilkan ke dalam file dengan judul "main.csv".

b) Import Module dan Data

Tahap pertama yang dilakukan dalam proses pengolahan data adalah mengimpor module dan data. Module yang akan digunakan adalah warnings untuk mengabaikan peringatan dan pandas untuk memproses data (seperti read csv). Kemudian dengan menggunakan module pandas, akan diimport data dari format csv menjadi bentuk tabel data frame "df" berukuran 397 baris x 4 kolom.

c) *Preprocessing* Data

- Pemilihan Variabel

Terdapat 4 kolom yang merepresentasikan jumlah variabel dari data yang dihasilkan dari proses impor data pada tahap sebelumnya. Keempat variabel itu adalah variabel "Headline", "Date", "Link", dan "Content". Variabel yang kami pilih untuk dianalisis lebih lanjut pada tahap berikutnya adalah variabel "Headline", "Date", dan "Content" sehingga variabel "Link" akan dihapus dari tabel.

- Import Module NLTK

Module yang selanjutnya akan diimport untuk secara spesifik digunakan untuk tahap *preprocessing* data adalah nltk (Natural Language Toolkit) yaitu sebuah library yang digunakan untuk pengolahan data bentuk teks dalam bahasa manusia.

- Cek Duplikasi Judul

Pada tahap ini, kami akan melakukan pengecekan terhadap adanya kemungkinan duplikasi judul yaitu pada variabel "Headline". Hasilnya, kami menemukan duplikasi sebanyak 292 pada variabel tersebut yang kemudian akan dihilangkan sehingga menghasilkan ukuran data frame yang baru yaitu 105 baris x 3 kolom.

- Penghapusan Kalimat yang Tidak Sesuai dengan Konten

Pada data yang ada, kami mendapatkan beberapa kalimat di dalam tanda "( )" yang tidak mengandung unsur apapun dan dominan berisikan tanggal atau unsur lain, apabila tidak dihapus sebelum penghapusan tanda baca dikhawatirkan hanya akan menumpuk dan tidak berguna dalam proses selanjutnya. Selain itu, kami juga menemukan beberapa

kalimat di ujung konten berita yang berisikan link untuk berita selanjutnya. Kalimat tersebut kami pilih dengan bantuan keyword dan dihapus.

- Analisis Sentimen

Analisis sentimen merupakan proses yang dilakukan untuk menganalisis sebuah teks apakah nada pesan tersebut positif, netral, atau negatif. Modul yang akan diimpor untuk menjalankan proses ini adalah 'SentimentIntensityAnalyzer' dari 'nltk.sentiment'. Semakin tinggi nilai dari sentimen tersebut, maka akan semakin positif isi dari berita dan berlaku sebaliknya. Nilai dari sentimen memiliki rentang nilai -1 sampai dengan 1

- Normalisasi Teks

Tahap terakhir dalam preprocessing data adalah menormalisasi teks dengan mengubah semua kata yang semula terdapat huruf besar menjadi huruf kecil, serta menghilangkan tanda baca.

- Tokenisasi dan *Stopwords Removal*

Stopword adalah kata umum yang biasanya muncul dalam jumlah besar dan dianggap tidak memiliki makna sehingga akan dihilangkan. Kata ini akan dihilangkan dengan lebih dahulu melakukan proses tokenisasi (memecah kalimat menjadi kepingan kata) untuk memudahkan proses penghapusan stopwords. Sebelumnya, kami sudah mengimpor modul nltk (Natural Language Toolkit) yang kemudian akan diimpor 'stopwords' dari 'nltk.corpus' dan 'word_tokenize' dari 'nltk.tokenize' yang akan digunakan dalam proses tokenisasi dan penghapusan stopwords. Setelah menghilangkan stopwords, proses pengolahan data menjadi lebih efektif dan diharapkan juga menjadi lebih akurat. Proses tokenisasi juga dilakukan dalam penghapusan tanda baca.

- Penyimpanan hasil pre-processing

Kami menyimpan seluruh data hasil pre-processing dalam drive https://drive.google.com/drive/folders/1845UqhF78FtPGgK3PH90B3Tgitrb5fQj?usp=drive_link dengan rincian:

a) main_preprocessed : hasil *preprocessing* seluruhnya
b) main_v : hasil pre-processing untuk visualisasi Voyant.Tools yang berisikan kolom konten dan headline
c) main_t : hasil *preprocessing* untuk visualisasi Time Series yang berisikan kolom tanggal
d) main_s : hasil *preprocessing* untuk visualisasi sentimen yang berisikan kolom tanggal dan sentimen

III. Analisis dan Visualisasi Data

Selanjutnya, untuk mempermudah proses visualisasi, data yang sebelumnya sudah diolah akan disimpan dalam file dengan judul "main_preprocessed.csv". File tersebut kemudian diunggah dan dipecah ke dalam tiga data frame yang berbeda, yaitu 'df.v' untuk visualisasi Voyant Tools, 'df.t' untuk visualisasi Time Series, dan 'df.s' untuk visualisasi Analisis Sentimen.

a) Voyant Tools

Dengan menggunakan Voyant Tools (https://voyant-tools.org/) , diperoleh ringkasan dari main_v dan visualisasi sebagai berikut:

- Ringkasan (*summary*)

Vocabulary Density: 0.150

Readability Index: 17.977

Average Words Per Sentence: 36567.0

Most frequent words in the corpus:

- indonesia (476); games (363); sea (318); 2023 (249); atlet (248)

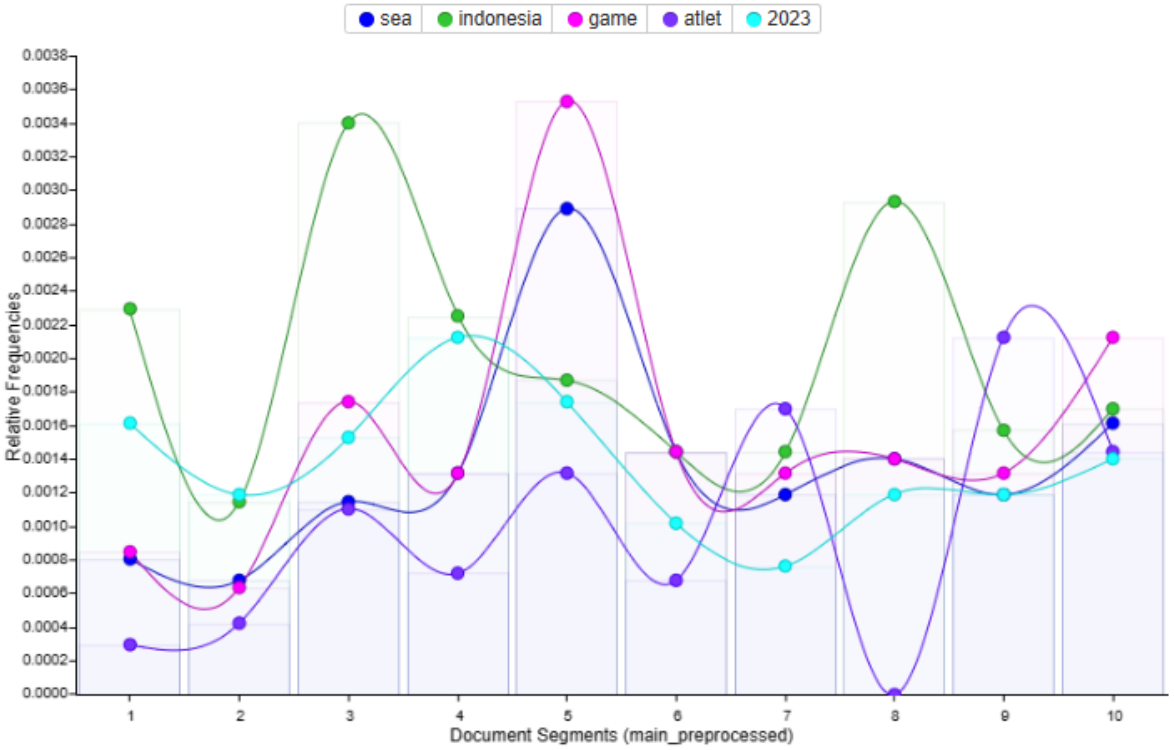
Didapatkan informasi bahwa pada data main_v didapatkan 5 jumlah kata terbanyak yaitu "indonesia" sebanyak 476 kata, "games" sebanyak 363 kata, "sea" sebanyak 318 kata, "2023" sebanyak 249 kata, dan atlet sebanyak "248" kata.

- Cirrus

Visualisasi menggunakan cirrus merupakan visualisasi yang digunakan untuk menampilkan kata-kata yang ada pada main_v.csv dalam bentuk awan, dimana

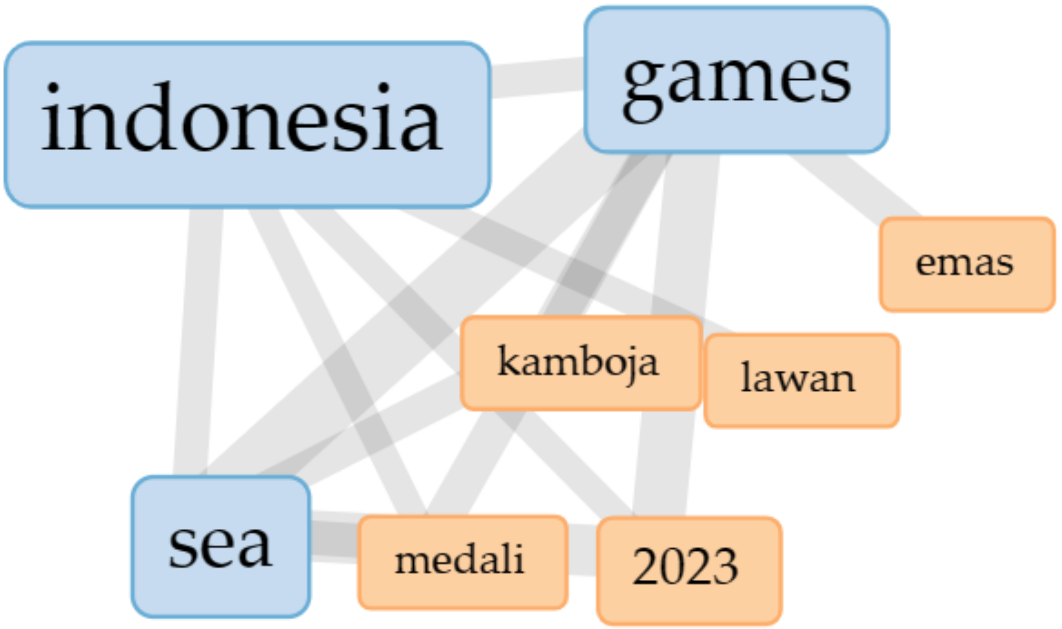
semakin besar kata yang muncul menandakan semakin banyak frekuensi munculnya kata tersebut. Warna pada visualisasi ini tidak memiliki fungsi apapun selain estetika. Untuk melihat jumlah frekuensi kata secara interaktif, dapat dilihat pada google colab milik kami.

- Trends



Dengan visualisasi Trends, dapat dilihat berapa perubahan frekuensi kemunculan kata kata untuk setiap kata-kata kunci terkait “Sea Games”. Kata-kata kunci yang memiliki frekuensi kemunculan tertinggi adalah “Game” dan “Indonesia”.

- Links



Dengan visualisasi Links, dapat dilihat bahwa terdapat kata-kata kunci lain yang memiliki keterhubungan dengan kata kunci “Sea Games”. Semakin besar kata menandakan semakin banyak frekuensi dari penggunaan kata tersebut, pada visualisasi ini didapatkan bahwa “indonesia”, “sea”, dan “games” merupakan kata dengan urutan terbanyak yang disusul dengan kata-kata berikutnya yang berwarna oranye. Pada visualisasi ini juga memberikan visual atas adanya hubungan antar kata satu sama lain.

b) Time Series

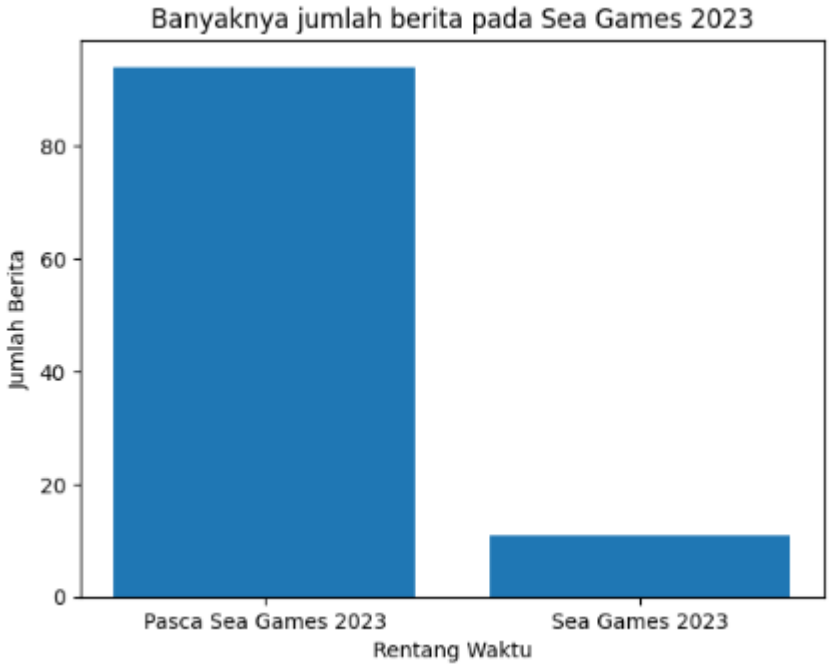
Kami mengelompokkan banyak data yang ada pada main_t dengan rincian waktu yaitu pra-Sea Games 2023 (5 April 2023 s.d. 4 Mei 2023); Sea Games 2023 (5 Mei 2023 s.d. 17 Mei 2023); Pasca Sea Games 2023 (18 Mei 2023 s.d. 17 Juni 2023). Berdasarkan lini waktu tersebut, diperoleh visualisasi menggunakan bar chart dengan informasi sebagai berikut:

- Informasi data

```
Pasca Sea Games 2023    94
Sea Games 2023          11
Name: Kelompok, dtype: int64
```

Didapatkan sebanyak 94 data pada kurun waktu Pasca Sea Games 2023 2023 dan 11 data pada kurun waktu Sea Games 2023.

- Frekuensi Berita yang Dipublikasikan

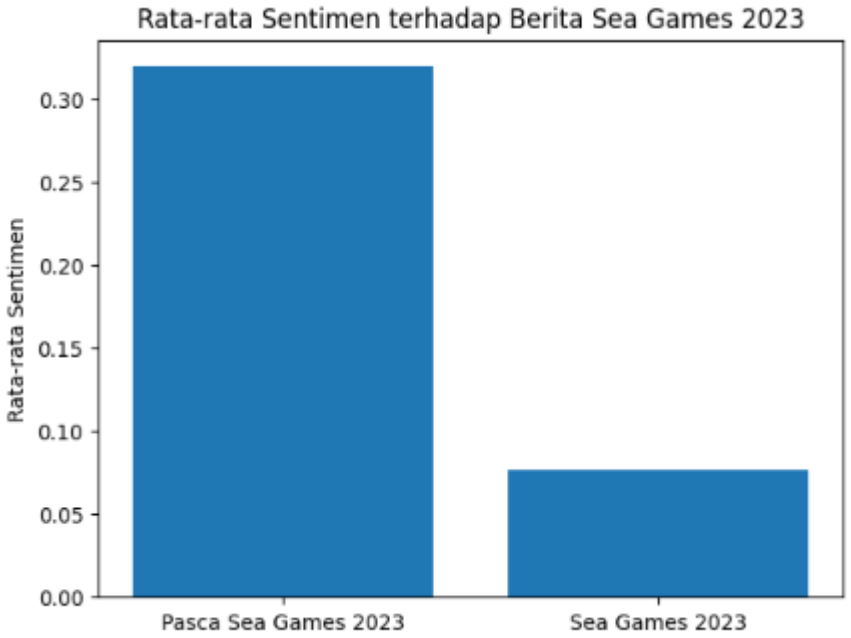


Informasi yang didapatkan dari visualisasi tersebut adalah berita mengenai Sea Games 2023 yang dipublikasikan melalui halaman web berita detik.com lebih banyak saat pasca acara dibandingkan saat acara berlangsung. Banyaknya frekuensi publikasi berita untuk Pra-Sea Games 2023 tidak muncul di dalam visualisasi karena tidak adanya berita yang dipublikasikan dalam rentang waktu yang kami tentukan.

c) Rata-rata Sentimen

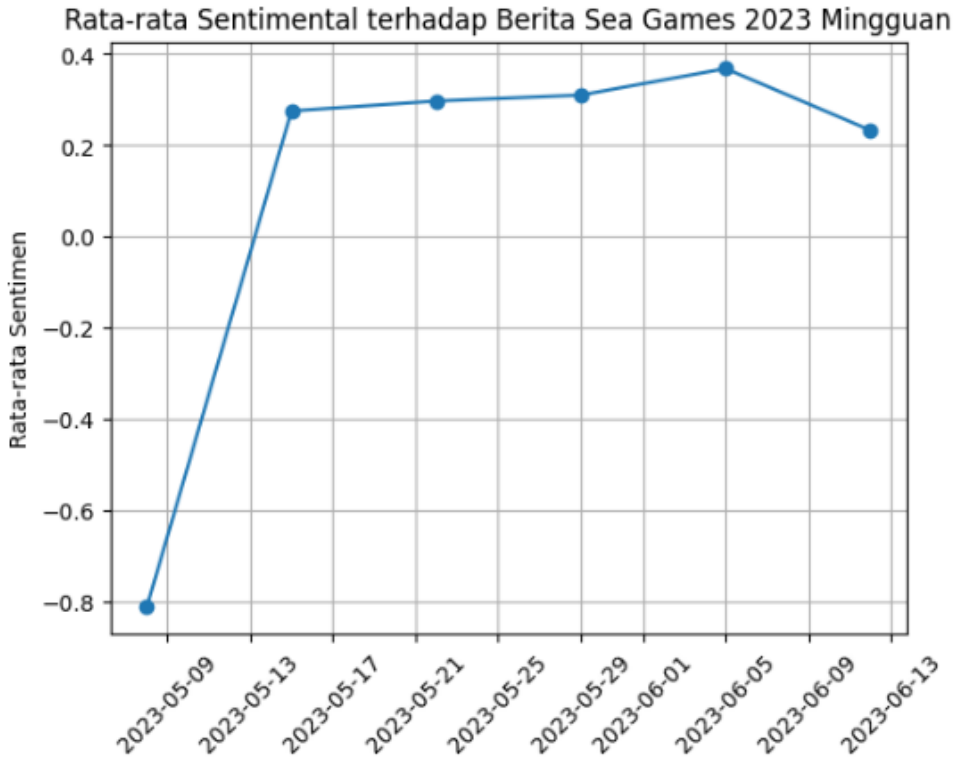
Dengan pengelompokkan yang sama dengan proses pengelompokkan waktu Time Series, dilakukan penghitungan rata-rata pada nilai sentimen pada rentang waktu tersebut dan dihasilkan visualisasi sebagai berikut:

- Visualisasi Bar chart



Terlihat skor rata-rata sentimen Pasca Sea Games 2023 lebih besar dibandingkan dengan saat pelaksanaan Sea Games 2023, dengan keduanya mengarah kepada sentimen netral menuju positif.

- Visualisasi Line chart



Kami juga memberikan visualisasi rata-rata sentimen yang kami kelompokkan perminggunya menggunakan line chart, terlihat perubahan sentimen yang cukup signifikan dimulai dari minggu pertama yang bernilai negatif (-0.8) hingga minggu kedua yang sudah bernilai positif, dimana pelaksanaan Sea Games 2023 telah berlangsung sekitar 4 hari dengan nilai sentimen yang lebih baik. Setelah minggu kedua, dapat dilihat perubahan sentimen relatif stabil jika dibandingkan dengan perubahan pada minggu pertama - minggu kedua.

IV. Penutup

Berdasarkan keseluruhan proses pengolahan data hingga visualisasi data, informasi yang diperoleh berdasarkan pengambilan data di halaman web berita detik.com mengenai Sea Games 2023 adalah adanya kata-kata kunci dengan frekuensi penggunaan kata tertinggi yaitu “Game”, “Indonesia”, dan “Sea”. Kami juga

menyimpulkan bahwa berita mengenai Sea Games 2023 lebih banyak dipublikasikan saat pasca acara dibandingkan dengan saat acara berlangsung, yaitu nilai sentimen terhadap berita tersebut diatas 0.05 (positif). Sedangkan, berita mengenai Sea Games 2023 tidak kami temukan saat masa pra-Sea Games 2023. Berita Sea Games 2023 yang muncul saat acara berlangsung mengalami perubahan signifikan pada nilai sentimen (nada emosional) dimana pada minggu pertama memiliki kecenderungan nilai nada emosional yang negatif, akan tetapi setelah melewati empat hari, nilai nada emosional menjadi positif.

Tautan Google Drive berisi file sumber data, laporan Google Colaboratory, hasil preprocessing data, dan video presentasi : UAS EVDA- EVDA B - Google Drive
Video Presentasi:
https://drive.google.com/drive/folders/1-Mwet_qo93L7FuBiHDVFiY0xoJJfcCjS?usp=sharing