**MODEL PEMBELAJARAN DAN LAPORAN AKHIR**

**PROJECT-BASED LEARNING**

**MATA KULIAH BIG DATA**

**KELAS B**



**“ANALISIS DATA VIDEO TRENDING YOUTUBE MENGGUNAKAN PYSPARK DAN VISUALISASI MENGGUNAKAN HVPLOT DAN PANEL ”**

**DISUSUN OLEH KELOMPOK “I” :**

1. ACH. ARIF SETIAWAN ( 21083010014 ) - KETUA
2. HAJJAR AYU CAHYANI K. ( 21083010044 ) - ANGGOTA
3. ELFIRA RATNA SYAHARANI ( 21083010056 ) - ANGGOTA
4. MUTIARA IRMADHANI ( 21083010079 ) - ANGGOTA
5. ABDUL MUKTI ( 21083010098 ) - ANGGOTA

**DOSEN PENGAMPU:**

TRESNA MAULANA FAHRUDIN, S.ST., MT ( 199305012022031007 )

KARTIKA MAULIDA HINDRAYANI, S.KOM, M.KOM ( 199209092022032009 )

PROGRAM STUDI SAINS DATA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS PEMBANGUNAN NASIONAL “VETERAN” JAWA TIMUR

2023

# KATA PENGANTAR

Segala puja dan puji syukur kehadiran Tuhan Yang Maha Esa., yang telah melimpahkan rahmat dan hidayah-Nya kepada kita semua, sehingga penyusun dapat membuat laporan akhir project-based learning Mata kuliah Big Data. Walaupun demikian, penyusun berusaha dengan maksimal mungkin demi kesempurnaan penyusunan laporan ini baik dari hasil kegiatan dalam perkuliahan. Saran dan kritik yang sifatnya membangun begitu diharapkan oleh penyusun demi kesempurnaan dalam penulisan laporan berikutnya

Dalam kesempatan ini, penulis mengucapkan banyak terima kasih kepada semua pihak yang telah membantu dalam penyusunan Laporan Akhir Mata Kuliah Big Data ini, di antaranya:

1. Bapak Tresna Maulana Fahrudin, S.S.T., MT., selaku Dosen Pengampu Mata kuliah Algoritma dan pemrograman lanjut
2. Ibu Kartika Maulida Hindrayani, S.KOM., M.KOM., selaku Dosen Pengampu Mata kuliah Algoritma dan pemrograman lanjut.
3. Rekan se-kelompok.

Akhir kata, penyusun berharap laporan ini dapat bermanfaat bagi pembaca. Saya ucapkan terima kasih banyak kepada semua pihak yang telah membantu, semoga Tuhan Yang Maha Esa membalas semua kebaikan kalian. Amin.

|  |
| --- |
| Surabaya, 1 Juni 2023 |
|  |
| Penyusun |

# DAFTAR ISI

[**KATA PENGANTAR 2**](#_v8l8rdo0jmvx)

[**DAFTAR ISI 3**](#_30j0zll)

[**DAFTAR GAMBAR 4**](#_1fob9te)

[**1. BAB I: PENDAHULUAN 1**](#_2et92p0)

[1.1 Latar Belakang 1](#_tyjcwt)

[1.2 Permasalahan 2](#_3dy6vkm)

[1.3 Tujuan 2](#_1t3h5sf)

[1.4 Manfaat 2](#_4d34og8)

[**2. BAB II: TINJAUAN PUSTAKA 3**](#_2s8eyo1)

[2.1 Teori Penunjang 3](#_17dp8vu)

[2.1.1 Konsep Big Data 3](#_18phbwxa7w62)

[2.1.2 Youtube 3](#_r1sfjtvjvhjm)

[2.1.3 PySpark 3](#_r61ul9q650ut)

[2.1.4 HvPlot 4](#_zfkz1vxrcgt9)

[2.2 Penelitian Terkait 4](#_3rdcrjn)

[2.2.1 “Analisa Pola Pengguna Youtube Trending Menggunakan Algoritma Partitioning Around Medoids dan Frequent Pattern Growth” oleh S Seimahuira (2022) 4](#_p4h3xz9wep0e)

[2.2.2 “Implementasi Sentimen Analisis Komentar Channel Video Pelayanan Pemerintah Di Youtube Menggunakan Algoritma Naïve Bayes” oleh Pramana Yoga Saputra, Dian Hanifudin Subhi dan Fahmi Zain Afif Winatama (2019). 5](#_18zczqltt0tt)

[2.2.3 “Analysis of Sentiments on Indonesian YouTube Video Comments: Case Study of The Indonesian Government's Plan to Move the Capital City” oleh R. A. Maisal, A. N. Hidayanto, N. F. Ayuning Budi, Z. Abidin and A. Purbasari (2019). 5](#_my3bnx1i01sv)

[**3. BAB III: METODOLOGI PENELITIAN 6**](#_26in1rg)

[3.1 Input 6](#_8t6pyhm36y6o)

[3.2 Proses Analisis 6](#_ep1i15p1o7ki)

[3.3 Visualisasi 7](#_qisd2clmi0bz)

[3.4 Output 7](#_f9mysy7kjzgn)

[**4. BAB IV: HASIL DAN PEMBAHASAN 8**](#_i6z5np560fno)

[4.1 Pengumpulan Data 8](#_n3lno3e4hpjx)

[4.2 Pembersihan Data 8](#_dguj12jczbja)

[4.3Visualisasi Data 10](#_v4qwb83iyqxi)

[4.3.1 Jumlah Video Trending Dari Waktu Ke Waktu Berdasarkan Kategori 12](#_vl9p4pkqrdsi)

[4.3.2 Jumlah Penonton Video Trending Dari Waktu Ke Waktu Berdasarkan Kategori 13](#_o72yt01ddfwr)

[4.3.3 Membuat Dashboard 15](#_i6izcsrlfn5y)

[4.4 Hasil Dan Analisis Data 16](#_ov7st6ir37fj)

[4.4.1 Jumlah Video Trending Dari Waktu Ke Waktu Berdasarkan Kategori 16](#_94ppqyvplxdb)

[4.4.2 Jumlah Penonton Video Trending Dari Waktu Ke Waktu Berdasarkan Kategori 19](#_mx438azglwf5)

[**5. BAB V: KESIMPULAN 23**](#_2jxsxqh)

[5.1 Kesimpulan 23](#_qnxpjdooa0p8)

[**6. DAFTAR PUSTAKA 24**](#_z337ya)

[**7. LAMPIRAN 25**](#_3j2qqm3)

# DAFTAR GAMBAR

[Gambar 1.](#_lnxbz9) Contoh Desain sistem yang diusulkan dalam proyek2

[Gambar 2. Diagram Alur](#_1ksv4uv) 6

[Gambar 3. Jumlah Kategori Video Trending](#_1ksv4uv) 7

[Gambar 4. Jumlah Views Kategori Video Trending](#_1ksv4uv) 13

[Gambar 5. Dashboard Analisis](#_1ksv4uv) 16

[Gambar 6. Jumlah Video Trending Awal](#_1ksv4uv) 16

[Gambar 7. Jumlah Video Trending Seminggu](#_1ksv4uv) 17

[Gambar 8. Jumlah Video Trending Sebulan](#_1ksv4uv) 17

[Gambar 9. Jumlah Video Trending Tahun Baru](#_1ksv4uv) 18

[Gambar 10. Jumlah Video Trending Akhir](#_1ksv4uv) 18

[Gambar 11. Line Plot Kategori Music](#_1ksv4uv) 19

[Gambar 12. Line Plot Kategori Entertainment](#_1ksv4uv) 19

[Gambar 13. Line Plot Kategori Gaming](#_1ksv4uv) 20

[Gambar 14. Line Plot Kategori Travel & Events](#_1ksv4uv) 20

[Gambar 15. Line Plot Kategori Nonprofit & Activism](#_1ksv4uv) 21

[Gambar 16. Line Plot Kategori Shows](#_1ksv4uv) 22

# BAB I: PENDAHULUAN

## 1.1 Latar Belakang

Popularitas platform YouTube di kalangan masyarakat Indonesia saat ini tidak bisa dipungkiri. Melansir katadata, jumlah pengguna aktifnya di Indonesia mencapai 129 juta. Seiring semakin populernya platform ini, menjadi YouTuber atau pembuat konten di YouTube saat ini semakin dipandang masyarakat sebagai satu aktivitas yang dapat menghasilkan uang, apalagi ketika video yang diunggah berhasil masuk pada halaman trending YouTube [1]. YouTube (situs web berbagi video terkenal di dunia) menyimpan daftar video trending teratas di platform. Menurut majalah Variety, “Untuk menentukan video trending teratas tahun ini, YouTube menggunakan kombinasi faktor termasuk mengukur interaksi pengguna (jumlah penayangan, pembagian, komentar, dan suka). YouTube menyediakan mekanisme interaksi sosial untuk menilai pendapat pengguna dan pandangan tentang video dengan cara voting, rating, favorit, berbagi (share) dan komentar negatif, dll. Informasi ini berguna dalam mempelajari perilaku pengguna dan masyarakat. Kemampuan untuk mengekstrak pendapat dari baris-baris teks yang ada pada YouTube dapat menjadi sangat berguna, dan ini adalah area studi yang banyak dikaji, tidak diragukan karena kemungkinan nilai komersialnya [2].

Big data menjadi tren dalam dunia teknologi informasi saat ini. Big data merupakan sumber data yang memiliki volume yang besar, variasi yang banyak, dan aliran data yang sangat cepat [3]. Big data dapat digunakan untuk menggambarkan video yang sedang tren terjadi saat ini. Contoh big data yang dapat digunakan untuk menayangkan video tren saat ini adalah data dari platform YouTube. Data tersebut mampu melacak aktivitas masyarakat yang sedang membuat konten. Pengolahan big data bukanlah suatu hal yang mudah. Pengolahan big data tidak dapat disamakan dengan pengolahan data dengan ukuran yang relatif kecil. Single computer akan terhambat kinerjanya atau juga tidak akan dapat mengolah data jika ukurannya melebihi kapasitas memori pada komputer tersebut. Oleh karena itu diperlukanlah suatu tool atau kerangka kerja yang dapat membantu proses pengolahan terhadap big data [4].

Dalam penelitian ini, dilakukan analisis data video YouTube yang sedang tren dengan menggunakan PySpark dengan menganalisis jumlah video yang sedang tren berdasarkan kategori dan jumlah penonton video yang sedang tren dari waktu ke waktu berdasarkan kategori. Selanjutnya memvisualisasikan hasil analisis data dalam bentuk grafik menggunakan hvplot. Penelitian ini bertujuan untuk memberikan wawasan tentang pola dan tren dalam video YouTube yang sedang tren berdasarkan kategori dan jumlah penonton dari waktu ke waktu.

## 

## 1.2 Permasalahan

1. Bagaimana cara mengoptimalkan penggunaan PySpark untuk menganalisis data video YouTube?
2. Bagaimana cara memvisualisasikan hasil analisis data pada dashboard secara efektif?

## 1.3 Tujuan

1. Menjelaskan cara untuk mengoptimalkan penggunaan PySpark dalam menganalisis data video YouTube.
2. Menjelaskan cara memvisualisasikan hasil analisis data pada dashboard secara efektif.

## 1.4 Manfaat

Manfaat bagi penyusun:

1. Mendapatkan pengetahuan dan keterampilan baru dalam mengoptimalkan penggunaan PySpark untuk menganalisis data video YouTube dan cara memvisualisasikan hasil analisis data pada dashboard secara efektif.
2. Membantu penyusun dalam mengembangkan kemampuan analitis dan pemecahan masalah.

Manfaat bagi pembaca:

1. Mendapatkan informasi dan wawasan baru tentang cara mengoptimalkan penggunaan PySpark untuk menganalisis data video YouTube dan cara memvisualisasikan hasil analisis data pada dashboard secara efektif.
2. Membantu pembaca dalam meningkatkan pemahaman mereka tentang topik tersebut dan menerapkannya dalam pekerjaan atau kegiatan mereka.

# BAB II: TINJAUAN PUSTAKA

## 2.1 Teori Penunjang

### 2.1.1 Konsep Big Data

Konsep big data mencakup kumpulan data dengan ukuran di luar kemampuan alat perangkat lunak yang biasa digunakan untuk menangkap, menyimpan, mengelola, dan memproses data (Kataria & Mittal, 2014). Konsep big data didefinisikan melalui model 3V oleh Gartner (n.d.) yang menjelaskan bahwa big data merupakan aset informasi yang memiliki volume, kecepatan, dan variasi yang tinggi, yang menuntut bentuk pengolahan informasi yang hemat biaya dan inovatif untuk meningkatkan wawasan dan pengambilan keputusan. Adapun 3V tersebut, diantaranya [5]

1. Volume mengacu pada kuantitas data yang dihasilkan dan dikumpulkan dari berbagai sumber.
2. Velocity mengacu pada laju di mana data dihasilkan dan perlu diproses.
3. Variety mengacu pada berbagai jenis data yang ada, baik terstruktur maupun tidak terstruktur.

### 2.1.2 Youtube

YouTube merupakan salah satu perusahaan milik Google. YouTube diciptakan oleh 3 orang mantan karyawan PayPal (website online komersial), Chad Hurley, Steve Chen, dan Jawed Karim pada Februari 2005. Sejak awal diluncurkan, YouTube langsung mendapat sambutan baik di masyarakat. Youtube adalah video online dan yang utama dari kegunaan situs ini ialah sebagai media untuk mencari, melihat dan berbagi video yang asli ke dan dari segala penjuru dunia melalui suatu web (Budiargo, 2015; 47). Diluncurkan pada bulan Mei 2005, YouTube telah memudahkan miliaran orang untuk menemukan, menonton, dan membagikan beragam video. YouTube menyediakan forum bagi orang-orang untuk saling berhubungan, memberikan informasi, dan menginspirasi orang lain di seluruh dunia, serta bertindak sebagai Platform distribusi bagi pembuat konten asli dan pengiklan, baik yang besar maupun kecil. YouTube mudah dipergunakan, tidak memerlukan biaya tinggi, dan dapat diakses dimanapun, tentunya dengan gadget yang kompatibel. Hal itu membuat pembuat video amatir dapat dengan bebas mengunggah konten-konten video mereka untuk dipublikasikan [6].

### 2.1.3 PySpark

Apache Spark merupakan teknologi komputasi cluster yang cepat, yang dirancang untuk perhitungan cepat. Hal ini didasarkan pada Hadoop MapReduce dan memperluas model dari MapReduce untuk efisiensi lebih banyak jenis perhitungan, yang mencakup query interaktif dan stream processing. Fitur utama Apache Spark adalah komputasi cluster di memory yang meningkatkan kecepatan pemrosesan aplikasi. Apache Spark dirancang untuk mencakup berbagai macam beban kerja seperti batch aplication, iterative algorithms, interactive queres dan streaming Apache Spark bekerja dengan cara menyimpan semua proses iterasi ke dalam memory, bukan ke dalam disk seperti halnya MapReduce. Apache Spark mempunyai kecepatan 100 kali lebih cepat di dalam memory di bandingkan MapReduce. PySpark dapat membantu dalam menggunakan Resilient Distributed Datasets(RDD) dalam Bahasa pemrograman Apache Spark dan Python dengan memanfaatkan library Py4j. Library Py4j merupakan library yang telah terinterigasi dalam PySpark [7].

### 2.1.4 HvPlot

Hvplot adalah sebuah pustaka Python yang digunakan untuk memvisualisasikan data secara interaktif dengan menggunakan library Bokeh. Hvplot menyediakan antarmuka yang mudah digunakan untuk membuat grafik tingkat tinggi dengan sintaks yang mirip dengan library plotting populer seperti Matplotlib dan Pandas. Hvplot memanfaatkan kekuatan Bokeh untuk membuat visualisasi interaktif yang dapat digunakan dalam berbagai aplikasi. Bokeh adalah sebuah pustaka Python yang dirancang khusus untuk menghasilkan visualisasi interaktif yang menarik dengan dukungan untuk tampilan interaktif dan responsif di browser web [8].

## 2.2 Penelitian Terkait

### 2.2.1 “Analisa Pola Pengguna Youtube Trending Menggunakan Algoritma Partitioning Around Medoids dan Frequent Pattern Growth” oleh S Seimahuira (2022)

Penelitian ini bertujuan untuk membantu memberikan informasi bahwa adanya faktor yang mempengaruhi video di youtube menjadi trending diantaranya comment, comment count, dislike dan like pada video. Penggunaan algoritma Partitioning Around Medoids (PAM) bersamaan dengan algoritma FP-Growth diterapkan pada sampel video trending yang digunakan dalam penelitian diperoleh pola masyarakat dalam menggunakan Youtube bahwa lebih banyak penonton Youtube tidak memberikan respon terhadap video yang ditonton baik berupa like dislike maupun comment. Sehingga jumlah viewers berbanding jauh dengan jumlah respon atau interaksi dari penonton [9].

### 2.2.2 “Implementasi Sentimen Analisis Komentar Channel Video Pelayanan Pemerintah Di Youtube Menggunakan Algoritma Naïve Bayes” oleh Pramana Yoga Saputra, Dian Hanifudin Subhi dan Fahmi Zain Afif Winatama (2019).

Penelitian ini bertujuan untuk melakukan tolok ukur dalam menilai tingkat kepuasan masyarakat dengan Analisa sentimen menggunakan klasifikasi data Youtube pada komentar video yang dipublish pemerintah tentang kinerjanya kepada masyarakat. Data tersebut nantinya akan diklasifikasikan menjadi klasifikasi positif, negatif dan netral. Komentar tersebut kemudian diproses menggunakan metode Naïve Bayes Classifier. Berdasarkan proses pengujian dan analisis yang telah dilakukan, dapat diambil kesimpulan. Hasil akurasi klasifikasi algoritma Naive Bayes Classifier pada channel Kemen PUPR menghasilkan nilai 69.23% persen dan pada channel Kemenkeu menghasilkan nilai 64.10% [10].

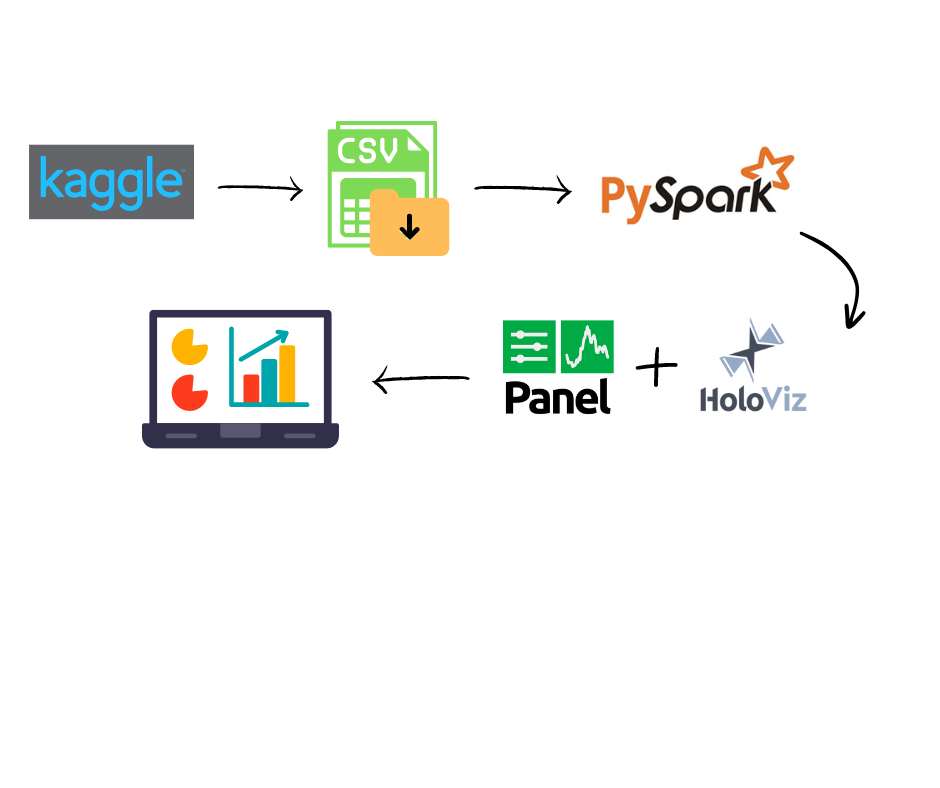
### 2.2.3 “Analysis of Sentiments on Indonesian YouTube Video Comments: Case Study of The Indonesian Government's Plan to Move the Capital City” oleh R. A. Maisal, A. N. Hidayanto, N. F. Ayuning Budi, Z. Abidin and A. Purbasari (2019).

Penelitian ini bertujuan untuk menganalisis sentimen dari komentar video YouTube yang berkaitan dengan rencana pemerintah Indonesia untuk memindahkan ibu kota. Metode pengumpulan data yang digunakan adalah pengambilan komentar dari video YouTube yang relevan dengan topik tersebut. Dari hasil penelitian, terdapat variasi sentimen yang signifikan dalam komentar video YouTube terkait rencana pemindahan ibu kota. Beberapa komentar menunjukkan dukungan dan optimisme terhadap rencana tersebut, sementara yang lain mengungkapkan kekhawatiran, kritik, atau ketidaksetujuan. Analisis sentimen ini memberikan wawasan penting mengenai pandangan masyarakat terhadap rencana pemindahan ibu kota [11].

# 

# BAB III: METODOLOGI PENELITIAN

pada Bab III ini akan dijelaskan mengenai rancangan desain sistem dari proyek yang diusulkan mulai dari input-proses-output. Rancangan desain sistem yang digunakan direpresentasikan dalam notasi *flowchart* maupun diagram secara umum seperti di bawah.



**Gambar 1.** Contoh Desain sistem yang diusulkan dalam proyek

## 3.1 Input

Pada bagian ini, dilakukan pengambilan data awal yang digunakan sebagai dasar untuk analisis dan visualisasi. Sumber data yang digunakan pada penelitian ini berasal dari Kaggle. Dataset ini berisi data tentang video-video YouTube yang sedang populer dalam beberapa bulan terakhir. Data tersebut mencakup wilayah Amerika Serikat, Britania Raya, Jerman, Kanada, Prancis, Rusia, Meksiko, Korea Selatan, Jepang, dan India. Setiap wilayah memiliki file data terpisah. Informasi yang disediakan meliputi judul video, judul saluran (channel), waktu publikasi, tag, jumlah penonton, jumlah suka dan tidak suka, deskripsi, jumlah komentar, dan kategori video. Kategori video dapat ditemukan dalam file JSON yang terkait dengan setiap wilayah. Namun, dalam penelitian ini, digunakan dataset dari negara Amerika Serikat.

## 3.2 Proses Analisis

Data yang telah diunduh kemudian diproses menggunakan Pyspark, sebuah framework pemrosesan data yang kuat dan efisien. Pyspark membantu dalam melakukan manipulasi, pemfilteran, pemrosesan, dan analisis data yang besar dengan cepat. Dalam tahap ini, data diolah untuk menggali wawasan atau informasi yang berguna. Namun sebelum dilakukan proses analisis terhadap data, akan dilakukan tahap pra-proses terlebih dahulu dimana data akan dibersihkan dari missing value yang ada, menyesuaikan format dan tipe data, dan lain sebagainya.

## 3.3 Visualisasi

Setelah data diolah, langkah selanjutnya adalah memvisualisasikan hasil analisis. Dalam hal ini, digunakan library hvplot, yang memungkinkan pembuatan visualisasi interaktif dengan mudah. Hvplot memungkinkan pengguna untuk membuat berbagai jenis visualisasi, seperti grafik garis, grafik batang, plot scatter, dan banyak lagi.

## 3.4 Output

Setelah desain rancangan selesai, outputnya adalah dashboard visualisasi yang mencakup visualisasi interaktif dari data yang telah diolah dan dianalisis. Pengguna dapat menggunakan dashboard ini untuk memperoleh wawasan dan pemahaman yang lebih baik tentang data yang diproses. Untuk menyajikan visualisasi secara terintegrasi, dapat menggunakan library Panel. Panel memungkinkan pengguna untuk membuat dashboard interaktif dengan mudah. Dengan menggunakan Panel, visualisasi yang telah dibuat dengan hvplot dapat digabungkan dalam satu tampilan yang terorganisir dengan baik. Dalam dashboard visualisasi terdapat berbagai elemen interaktif, seperti filter, slider, tombol, atau widget lainnya untuk meningkatkan dan memudahkan dalam mengeksplorasi data.

Alur dari tahapan yang dilakukan pada penelitian ini juga dapat dilihat pada diagram alur berikut.



**Gambar 2.** Diagram Alur

# BAB IV: HASIL DAN PEMBAHASAN

## 4.1 Pengumpulan Data

Pengumpulan data merupakan proses awal pengumpulan data atau informasi dari berbagai sumber untuk digunakan dalam analisis, penelitian, pengambilan keputusan, atau keperluan lainnya. Tujuan dari pengumpulan data sendiri untuk memperoleh informasi yang akurat, komprehensif, dan representatif yang dapat digunakan untuk analisis dan pemahaman yang lebih baik tentang suatu fenomena atau masalah yang sedang diteliti.

Data yang digunakan diambil dari dataset Trending YouTube Video Statistics untuk wilayah negara Amerika Serikat yang bersumber pada kaggle.com. Kumpulan data ini mencakup beberapa bulan data tentang video YouTube yang sedang tren setiap harinya, dengan hingga 200 video tren terdaftar setiap harinya. Dataset tersebut dipisah menjadi dua format file yang berbeda yaitu .json berisi daftar id kategori video dan nama kategori video sedangkan untuk format file .csv berisi berbagai fitur atau informasi tentang video tersebut seperti judul video, judul saluran, waktu publikasi, tag, jumlah tayangan, jumlah suka dan tidak suka, deskripsi, dan jumlah komentar. Dataset ini diambil mulai dari tanggal 14 November 2017 sampai dengan 14 Juni 2018 dan berisi 48137 record.

## 4.2 Pembersihan Data

Proses pembersihan data dilakukan agar data tersebut menjadi lebih bersih, terstruktur, dan berkualitas. Untuk mencapai hal tersebut, proses pembersihan data dilakukan menggunakan PySpark, sebuah framework pengolahan data yang kuat dan efisien. Dengan menggunakan PySpark, data dapat diolah dengan efektifitas tinggi serta memungkinkan penanganan volume data yang besar secara cepat dan efisien. untuk menginstall PySpark jalankan perintah dibawah ini pada command line atau terminal.

|  |
| --- |
| #install pyspark  ! pip install pyspark |

Setelah menginstall library PySpark, langkah selanjutnya adalah mengimport modul yang diperlukan dan membuat sesi (session). Dalam tahap ini, modul-modul yang relevan seperti SparkSession dan modul lainnya diimpor ke dalam skrip Python kita. Kemudian, dengan menggunakan SparkSession, sesi Spark dibuat. Sesi ini berperan penting dalam menginisialisasi dan mengatur lingkungan kerja PySpark. Dengan membuat sesi, kita dapat mengkonfigurasi pengaturan seperti nama aplikasi, konfigurasi cluster, dan pengaturan lainnya. Selain itu, sesi juga memungkinkan kita untuk berinteraksi dengan cluster Spark, menjalankan operasi pemrosesan data, dan mengakses berbagai fitur dan utilitas yang disediakan oleh PySpark.

|  |
| --- |
| #mengimport modul yang dibutuhkan  from pyspark.sql import SparkSession  from pyspark.sql.types import \*  from pyspark.sql.functions import \*  import json  #membuat session  appName = "Analisis Trending Youtube"  spark = SparkSession \  .builder \  .appName(appName) \  .config("spark.some.config.option", "some-value") \  .getOrCreate() |

Setelah membuat sesi, tahap pembersihan data. Tahap ini melibatkan penggunaan berbagai teknik dan metode untuk membersihkan dan mempersiapkan data agar siap untuk analisis lebih lanjut. Dalam proses ini, data yang tidak relevan, duplikat, atau tidak lengkap akan dihapus atau diolah sehingga data yang tersisa menjadi lebih bersih, terstruktur, dan siap untuk diolah.

|  |
| --- |
| #membaca data dari file ke DataFrame dengan skema yang diinfer  trend\_csv = spark.read.csv("C:/Users/LENOVO/Desktop/Praktik Coding/Big Data\_TA/USvideos.csv", inferSchema=True, header=True)  # Konversi kolom 'trending\_date' ke tipe tanggal  trend\_csv = trend\_csv.withColumn('trending\_date', to\_date(trend\_csv['trending\_date'], 'yy.dd.MM'))  # Konversi kolom 'publish\_time' ke tipe timestamp  trend\_csv = trend\_csv.withColumn('publish\_time', to\_timestamp(trend\_csv['publish\_time'], 'yyyy-MM-dd\'T\'HH:mm:ss.SSS\'Z\''))  # Baca file JSON  multiline\_df = spark.read.option("multiline", "true").json("C:/Users/LENOVO/Desktop/Praktik Coding/Big Data\_TA/US\_category\_id.json")  # Cetak skema JSON  multiline\_df.printSchema()  # Memecah nilai array menjadi baris terpisah  exploded\_df = multiline\_df.select(explode("items").alias("item"))  # Mengakses kolom yang tepat dalam struktur JSON  category\_df = exploded\_df.select("item.id", "item.snippet.title")  category\_df = category\_df.withColumnRenamed("title", "category\_title")  # Menggabungkan DataFrame berdasarkan kolom tertentu  joined\_df = trend\_csv.join(category\_df,trend\_csv.category\_id == category\_df.id,"inner")  # Memeriksa nilai NaN (null) pada DataFrame  nan\_counts = joined\_df.select([sum(col(c).isNull().cast("int")).alias(c) for c in joined\_df.columns])  # Menggantikan nilai '[none]' dengan NaN pada kolom 'tags'  joined\_df = joined\_df.withColumn('tags', when(col('tags') == '[none]', None).otherwise(col('tags')))  # Memeriksa nilai NaN (null) pada DataFrame  nan\_counts = joined\_df.select([sum(col(c).isNull().cast("int")).alias(c) for c in joined\_df.columns])  # Daftar kolom yang ingin diubah tipe datanya menjadi integer  int\_columns = ["views", "likes", "dislikes", "comment\_count"]  # Mengubah tipe data kolom menjadi integer menggunakan loop  for col\_name in int\_columns:  joined\_df = joined\_df.withColumn(col\_name, col(col\_name).cast("int")) |

Data yang telah melalui proses pembersihan, langkah selanjutnya adalah melakukan visualisasi data menggunakan grafik. Dengan visualisasi data menggunakan grafik, kita dapat memperoleh pemahaman yang lebih baik tentang pola, tren, dan hubungan antara variabel dalam data.

## 4.3Visualisasi Data

Untuk membuat dashboard visualisasi data yang interaktif, pastikan Anda telah menginstall library panel dan hvplot. Library panel menyediakan berbagai komponen interaktif untuk membangun dasbor yang dapat dioperasikan oleh pengguna. Sementara itu, hvplot adalah sebuah library yang berintegrasi dengan panel dan digunakan untuk membuat visualisasi data yang interaktif secara cepat dan mudah.

|  |
| --- |
| #install panel dan hvplot  ! pip install panel hvplot |

Untuk meningkatkan performa dashboard, disarankan untuk menyimpan data yang digunakan dalam cache agar akses ke data tersebut lebih cepat dan efisien. Dengan menyimpan data dalam cache, dashboard dapat menghindari operasi yang mahal secara komputasi, seperti pengambilan atau pemrosesan ulang data, setiap kali interaksi pengguna terjadi. Sebaliknya, data yang telah disimpan dalam cache dapat dengan cepat diambil dan digunakan untuk memperbarui visualisasi atau responsivitas dashboard. Dengan menggunakan cache, waktu respons dashboard dapat dikurangi secara signifikan, memberikan pengalaman yang lebih lancar dan responsif kepada pengguna.

|  |
| --- |
| import panel as pn  # Konversi DataFrame  youtube\_df = joined\_df.toPandas()  # menyimpan data untuk meningkatkan kinerja dasbor  if 'data' not in pn.state.cache.keys():  df = youtube\_df  pn.state.cache['data'] = df.copy()  else:  df = pn.state.cache['data'] |

Selanjutnya membuat Data Interaktif untuk Pipeline dengan melakukan beberapa transformasi pada DataFrame youtube\_df, seperti mengubah tipe data kolom tertentu dan membuat DataFrame interaktif dengan bantuan pustaka seperti Pandas, Numpy, Panel dan holoviews.

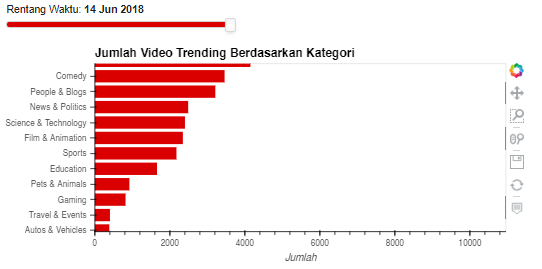
|  |
| --- |
| import pandas as pd  import numpy as np  import panel as pn  import holoviews as hv  from holoviews import opts  hv.extension('bokeh')  pn.extension('tabulator')  # %%  # konversi kolom 'trending\_date' menjadi datetime  youtube\_df['trending\_date'] = pd.to\_datetime(youtube\_df['trending\_date'])  youtube\_df['publish\_time'] = pd.to\_datetime(youtube\_df['publish\_time'])  youtube\_df['views'] = youtube\_df['views'].astype(int)  youtube\_df['likes'] = youtube\_df['likes'].astype(int)  youtube\_df['dislikes'] = youtube\_df['dislikes'].astype(int)  youtube\_df['comment\_count'] = youtube\_df['comment\_count'].astype(int)  # Membuat DataFrame Pipeline Interactive  idf = youtube\_df.interactive()  idf.dtypes |

### 4.3.1 Jumlah Video Trending Dari Waktu Ke Waktu Berdasarkan Kategori

Selanjutnya membuat widget DateSlider dan plot bar horizontal yang menunjukkan jumlah video trending dari waktu ke waktu berdasarkan kategori, dengan rentang waktu yang dapat dipilih oleh pengguna melalui widget DateSlider.

|  |
| --- |
| import datetime  # Define Panel widgets  min\_tahun\_upload = youtube\_df['trending\_date'].min().date()  max\_tahun\_upload = youtube\_df['trending\_date'].max().date()  range\_year\_slider = pn.widgets.DateSlider(  name='Rentang Waktu',  start=min\_tahun\_upload,  end=max\_tahun\_upload,  value=(max\_tahun\_upload),  bar\_color='#D95252'  )  cat\_count\_pipeline = (  idf[(idf.trending\_date.dt.date <= range\_year\_slider)]  .groupby(['category\_title']).size()  .to\_frame(name='jumlah\_data')  .reset\_index()  .sort\_values(by='jumlah\_data')  .reset\_index(drop=True))  plot\_count\_cat = cat\_count\_pipeline.hvplot.barh(y='jumlah\_data', x='category\_title', color='#D95252', line\_width=0, title="Jumlah Video Trending Dari Waktu Ke Waktu Berdasarkan Kategori", xlabel = '', ylabel = 'Jumlah')  plot\_count\_cat |

**Output:**



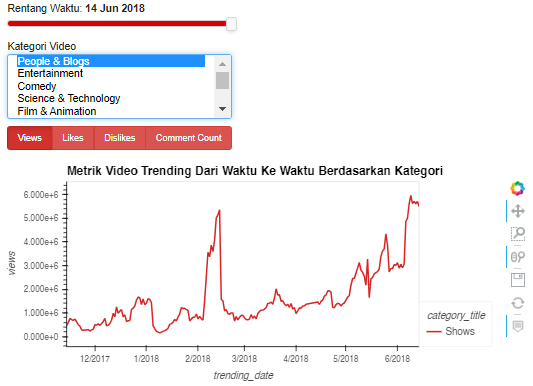
**Gambar 3.** Jumlah Kategori Video Trending

### 4.3.2 Jumlah Penonton Video Trending Dari Waktu Ke Waktu Berdasarkan Kategori

Selanjutnya membuat widget DateSlider dan line plot yang menunjukkan jumlah video trending dari waktu ke waktu berdasarkan kategori, dengan rentang waktu yang dapat dipilih oleh pengguna melalui widget DateSlider dan kategori melalui MultiSelect.

|  |
| --- |
| kategori = list(youtube\_df['category\_title'].unique())  multi\_choice = pn.widgets.MultiSelect(name='Kategori Video', value=['People & Blogs'], size=5,  options=kategori)  # Radio buttons  yaxis\_yutub = pn.widgets.RadioButtonGroup(  name='Y axis',  options= {'Views':'views', 'Likes':'likes', 'Dislikes':'dislikes', 'Comment Count':'comment\_count'},  button\_type='danger'  )  cat\_metric\_pipeline = (  idf[  (idf.trending\_date.dt.date <= range\_year\_slider) &  (idf.category\_title.isin(multi\_choice))  ]  .groupby(['category\_title', 'trending\_date'])[yaxis\_yutub].mean().round(0)  .to\_frame()  .reset\_index()  .sort\_values(by='trending\_date')  .reset\_index(drop=True)  )  time\_line\_plot = cat\_metric\_pipeline.hvplot(x = 'trending\_date', by='category\_title', y=yaxis\_yutub, line\_width=2, title='Metrik Video Trending Dari Waktu Ke Waktu Berdasarkan Kategori')  time\_line\_plot |

**Output:**



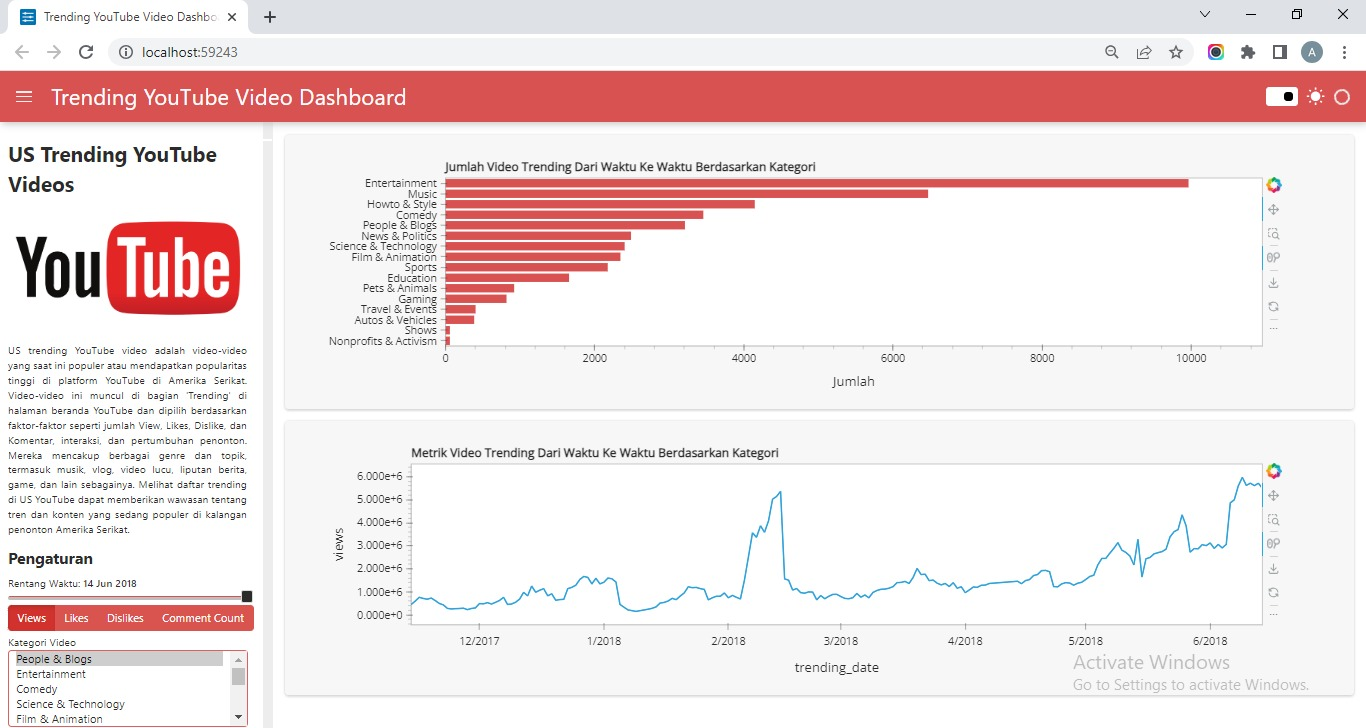
**Gambar 4.** Jumlah Views Kategori Video Trending

### 4.3.3 Membuat Dashboard

Tahapan terakhir dalam visualisasi data yaitu membuat sebuah dashboard yang menampilkan judul, gambar, penjelasan, plot visualisasi, dan pengaturan yang dapat digunakan untuk mengendalikan dashboard tersebut. Semua hasil dari kodingan sebelumnya dijadikan satu dan diimplementasikan dalam bentuk dashboard

|  |
| --- |
| import panel as pn  # membuat tata letak template dashboard  template = pn.template.FastListTemplate(  title='Trending YouTube Video Dashboard',  sidebar=[pn.pane.Markdown("# US Trending YouTube Videos"),  pn.pane.PNG("C:/Users/LENOVO/Desktop/Praktik Coding/Big Data\_TA/yutub.png", sizing\_mode="scale\_both"),  pn.pane.Markdown("<p align='justify'>US trending YouTube video adalah video-video yang saat ini populer atau mendapatkan popularitas tinggi di platform YouTube di Amerika Serikat. Video-video ini muncul di bagian 'Trending' di halaman beranda YouTube dan dipilih berdasarkan faktor-faktor seperti jumlah View, Likes, Dislike, dan Komentar, interaksi, dan pertumbuhan penonton. Mereka mencakup berbagai genre dan topik, termasuk musik, vlog, video lucu, liputan berita, game, dan lain sebagainya. Melihat daftar trending di US YouTube dapat memberikan wawasan tentang tren dan konten yang sedang populer di kalangan penonton Amerika Serikat."),  pn.pane.Markdown("## Pengaturan"),  range\_year\_slider, yaxis\_yutub, multi\_choice],  main=[pn.Row(pn.Column(plot\_count\_cat.panel(width=1200), margin=(0,25))),  pn.Row(pn.Column(time\_line\_plot.panel(width=1200), margin=(0,25)))],  accent\_base\_color="#D95252",  header\_background="#D95252",  )  template.show()  #template.servable(); |

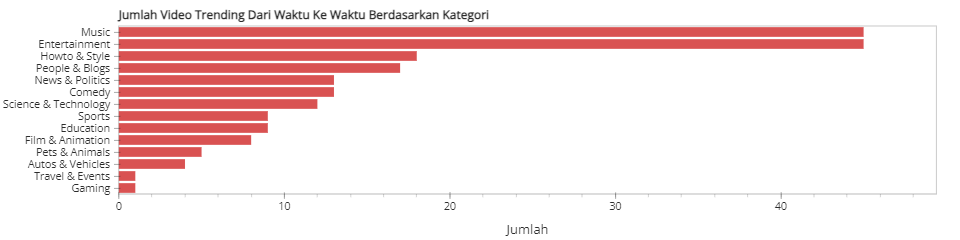
**Output:**



**Gambar 5.** Dashboard Analisis

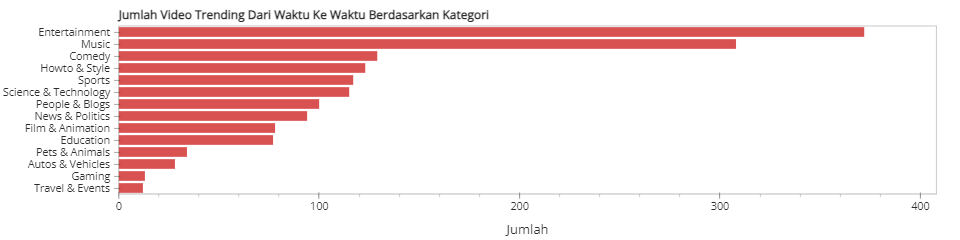
## 4.4 Hasil Dan Analisis Data

### 4.4.1 Jumlah Video Trending Dari Waktu Ke Waktu Berdasarkan Kategori



**Gambar 6.** Jumlah Video Trending Awal

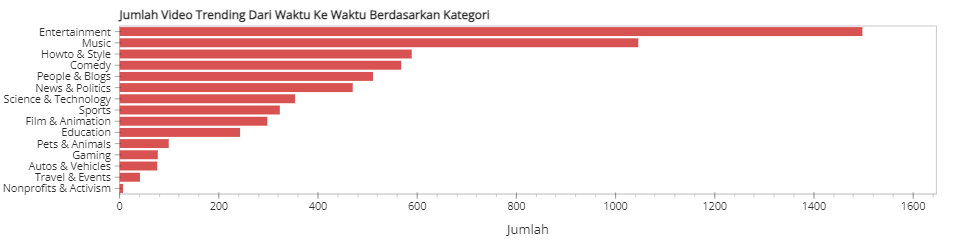
Pada tanggal awal dataset yaitu 14 November 2017, kategori video “Music” dan “Entertainment” memiliki jumlah video trending yang paling tinggi pada tanggal tersebut. Hal ini menunjukkan bahwa kategori video “Music” dan “Entertainment” memiliki daya tarik yang kuat di antara pengguna YouTube. Sedangkan kategori video “Gaming” dan “Travel & Event” menunjukkan jumlah video trending yang rendah pada tanggal tersebut. Hal ini bisa menunjukkan bahwa video-video dalam kategori tersebut mungkin memiliki target audiens yang lebih terbatas atau mungkin tidak memiliki popularitas sebesar kategori lainnya pada tanggal tersebut.



**Gambar 7.** Jumlah Video Trending Seminggu

Setelah satu minggu dari tanggal awal dataset, tepatnya pada tanggal 21 November 2017, terlihat bahwa kategori video "Entertainment" tetap mempertahankan posisinya sebagai kategori dengan jumlah video trending yang paling tinggi pada tanggal tersebut. Hal ini menunjukkan bahwa konten hiburan terus menarik perhatian pengguna YouTube. Namun, jumlah video trending dalam kategori "Music" mulai mengalami penurunan. Hal ini mungkin mengindikasikan pergeseran minat pengguna yang sedang terjadi dalam hal preferensi konten. Faktor-faktor seperti perubahan tren musik atau munculnya konten hiburan lainnya dapat mempengaruhi popularitas video-video dalam kategori ini.

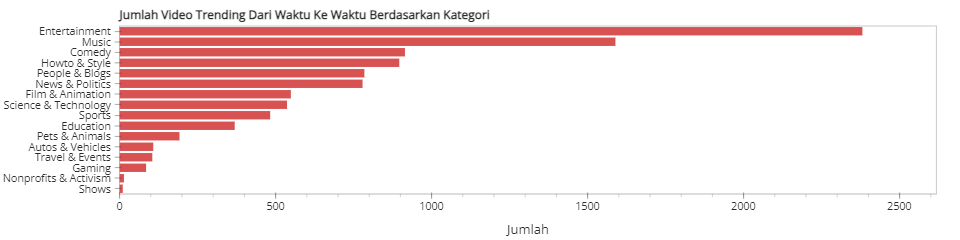
Kategori video “Gaming” dan “Travel & Event” menunjukkan jumlah video trending yang tetap rendah pada tanggal tersebut. Hal ini bisa menunjukkan bahwa video-video dalam kategori tersebut mungkin memiliki target audiens yang lebih terbatas atau mungkin tidak memiliki popularitas sebesar kategori lainnya pada tanggal tersebut.



**Gambar 8.** Jumlah Video Trending Sebulan

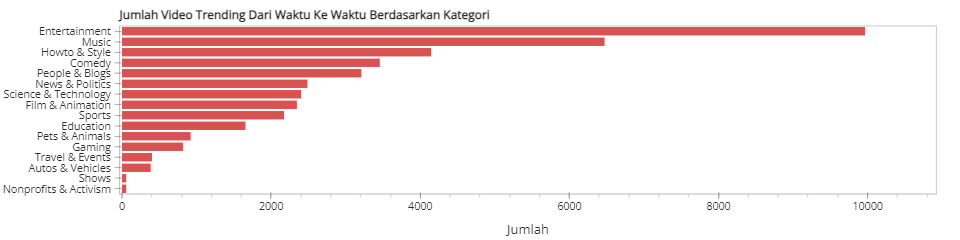
Setelah satu bulan dari tanggal awal dataset, tepatnya pada tanggal 14 November 2017, terlihat bahwa kategori video "Entertainment" tetap mempertahankan posisinya sebagai kategori dengan jumlah video trending yang paling tinggi pada tanggal tersebut. Hal ini menunjukkan bahwa konten hiburan terus menarik perhatian pengguna YouTube. Dan jumlah video trending dalam kategori "Music" terus mengalami penurunan. Hal ini mungkin mengindikasikan pergeseran minat pengguna yang sedang terjadi dalam hal preferensi konten. Faktor-faktor seperti perubahan tren musik atau munculnya konten hiburan lainnya dapat mempengaruhi popularitas video-video dalam kategori ini.

Selain itu muncul kategori video baru yaitu "Nonprofits & Activism" yang menunjukkan jumlah video trending yang paling rendah pada tanggal tersebut, menggantikan posisi kategori video "Gaming" dan "Travel & Event" yang sebulan sebelumnya memiliki jumlah video trending yang rendah. Munculnya kategori video baru yaitu "Nonprofits & Activism" yang menjadi trending mungkin disebabkan oleh peristiwa atau kampanye penting yang sedang berlangsung pada periode tersebut yang memicu peningkatan minat dan perhatian terhadap konten-konten yang terkait.



**Gambar 9.** Jumlah Video Trending Tahun Baru

Pada saat awal tahun baru, tepatnya pada tanggal 1 Januari 2018, terlihat bahwa kategori video "Entertainment" tetap mempertahankan posisinya sebagai kategori dengan jumlah video trending yang paling tinggi pada tanggal tersebut. Hal ini menunjukkan bahwa konten hiburan terus menarik perhatian pengguna YouTube. Selain itu muncul kategori video baru yaitu "Shows" yang menunjukkan jumlah video trending yang paling rendah pada tanggal tersebut, menggantikan posisi kategori video "Nonprofits & Activism" yang sebelumnya memiliki jumlah video trending yang rendah. Munculnya kategori video baru yaitu "Shows" yang menjadi trending mungkin disebabkan oleh adanya peningkatan minat pengguna terhadap konten-konten yang berkaitan dengan acara televisi, pertunjukan, atau serial. Pada periode tersebut, mungkin terdapat acara atau serial yang sedang populer atau menarik perhatian banyak orang, sehingga konten-konten dalam kategori "Shows" menjadi tren.



**Gambar 10.** Jumlah Video Trending Akhir

Pada tanggal akhir dataset yaitu 14 Juni 2018, terlihat bahwa kategori video "Entertainment" tetap mempertahankan posisinya sebagai kategori dengan jumlah video trending yang paling tinggi pada tanggal tersebut. Hal ini menunjukkan bahwa konten hiburan terus menarik perhatian pengguna YouTube. Selain itu jumlah video trending kategori “Shows” meningkat Hal ini menunjukkan bahwa konten-konten dalam kategori "Shows" menjadi semakin populer dan menarik minat pengguna YouTube. Kemungkinan adanya acara televisi atau serial yang sedang tren pada periode tersebut bisa menjadi salah satu faktor yang mendukung peningkatan jumlah video trending dalam kategori ini.

### 4.4.2 Jumlah Penonton Video Trending Dari Waktu Ke Waktu Berdasarkan Kategori

#### A. Kategori Video “Music”



**Gambar 11.** Line Plot Kategori Music

Dari grafik tersebut, terlihat bahwa terjadi kenaikan jumlah penonton (views) yang signifikan mulai dari tanggal 9 Maret 2018. Jumlah penonton terus meningkat hingga mencapai puncaknya pada tanggal 24 Maret 2018 sebelum mengalami penurunan yang signifikan. Setelah melakukan analisis lebih lanjut dengan melihat dataset yang digunakan, dapat disimpulkan bahwa pada rentang waktu tersebut, terdapat tiga musisi yang merilis lagu baru yang cukup populer. Lagu-lagu tersebut antara lain "Maluma - El Préstamo (Official Video)", "Taylor Swift - Delicate", dan "Camila Cabello - Never Be the Same". Kesuksesan penampilan mereka dalam periode tersebut dapat menjadi faktor peningkatan jumlah penonton yang signifikan pada tanggal-tanggal tersebut.

#### B. Kategori Video “Entertainment”



**Gambar 12.** Line Plot Kategori Entertainment

Dari grafik tersebut, terlihat bahwa terjadi kenaikan jumlah penonton (views) yang signifikan mulai dari tanggal 29 November 2017. Jumlah penonton terus meningkat hingga mencapai puncaknya pada tanggal 13 Desember 2017 sebelum mengalami penurunan yang signifikan. Setelah melakukan analisis lebih lanjut dengan melihat dataset yang digunakan, dapat disimpulkan bahwa pada rentang waktu tersebut terdapat beberapa video yang menjadi penyebab kenaikan tersebut.

Salah satu video yang menjadi faktor kenaikan adalah "YouTube Rewind: The Shape of 2017 | #YouTubeRewind". Video ini berisi rangkuman kejadian-kejadian viral yang terjadi selama tahun 2017 dan mendapatkan perhatian yang signifikan dari pengguna YouTube. Selain itu, terdapat juga video promosi dari studio film yang turut berkontribusi dalam kenaikan jumlah penonton. Video promosi tersebut antara lain "Marvel Studios' Avengers: Infinity War Official Trailer" dan "Jurassic World: Fallen Kingdom - Official Trailer [HD]". Kedua video ini mendapatkan popularitas yang tinggi dan menarik minat banyak orang untuk menonton trailer film tersebut.

Dengan demikian, video-video tersebut merupakan faktor utama dalam kenaikan jumlah penonton yang signifikan pada rentang waktu tersebut, baik melalui kejadian viral maupun melalui video promosi film yang populer.

#### C. Kategori Video “Gaming”

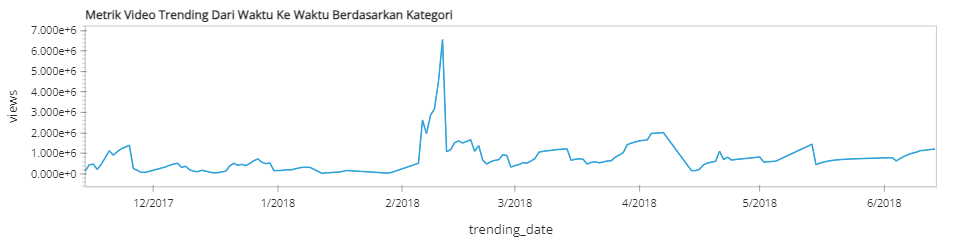


**Gambar 13.** Line Plot Kategori Gaming

Dari grafik tersebut, terlihat bahwa terjadi kenaikan jumlah penonton (views) yang signifikan mulai dari tanggal 10 Mei 2018. Jumlah penonton terus meningkat hingga mencapai puncaknya pada tanggal 15 Mei 2018 sebelum mengalami penurunan yang signifikan. Setelah melakukan analisis lebih lanjut dengan melihat dataset yang digunakan, dapat disimpulkan bahwa pada rentang waktu tersebut, terdapat dua game yang menjadi penyebab kenaikan tersebut. Salah satu game yang mempromosikan fitur baru mereka adalah "Clash Royale" dengan video berjudul "Clash Royale: Meet the Rascals! (New Card!)". Video ini berhasil menarik perhatian banyak penonton dan mengundang minat para pemain game untuk mengetahui lebih lanjut tentang fitur baru yang diperkenalkan.

Selain itu, game "Fortnite" juga berkontribusi dalam kenaikan jumlah penonton pada rentang waktu tersebut. Video promosi berjudul "SEASON 4 | ANNOUNCE TRAILER" berhasil menarik minat banyak penonton dengan mengungkapkan informasi tentang musim baru dalam permainan tersebut. Dengan demikian, dapat disimpulkan bahwa video-video promosi dari game "Clash Royale" dan "Fortnite" merupakan faktor utama dalam kenaikan jumlah penonton yang signifikan pada rentang waktu tersebut.

#### D. Kategori Video “Travel & Events”



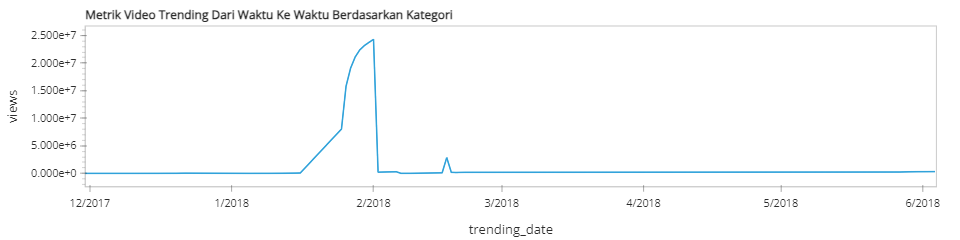
**Gambar 14.** Line Plot Kategori Travel & Events

Dari grafik tersebut, terlihat bahwa terjadi kenaikan jumlah penonton (views) yang signifikan mulai dari tanggal 29 Januari 2018. Jumlah penonton terus meningkat hingga mencapai puncaknya pada tanggal 12 Februari 2018 sebelum mengalami penurunan yang signifikan. Setelah melakukan analisis lebih lanjut dengan melihat dataset yang digunakan, dapat disimpulkan bahwa pada rentang waktu tersebut, terdapat beberapa video promosi yang menjadi faktor utama dalam kenaikan tersebut.

Salah satu video promosi yang mencolok adalah "Turkish Airlines - 5 Senses with Dr. Oz". Video ini adalah promosi dari Turkish Airlines yang menampilkan Dr. Oz dan menggambarkan pengalaman perjalanan dengan mengaktifkan kelima indra manusia. Selain itu, terdapat juga video promosi dari Tourism Australia dengan judul "Dundee Super Bowl Ad 2018 w/ Chris Hemsworth and Danny McBride". Video ini memanfaatkan momen Super Bowl untuk mempromosikan pariwisata Australia dengan melibatkan aktor terkenal seperti Chris Hemsworth dan Danny McBride. Tidak hanya itu, video "INSANE! SpaceX Falcon Heavy Side Boosters Landing Simultaneously at Kennedy Space Center" juga berkontribusi dalam kenaikan jumlah penonton pada rentang waktu tersebut. Video ini menampilkan pendaratan simultan dua booster SpaceX Falcon Heavy di Kennedy Space Center yang mengundang minat penggemar eksplorasi antariksa.

Dengan demikian, dapat disimpulkan bahwa video promosi acara seperti "Turkish Airlines - 5 Senses with Dr. Oz", "Tourism Australia Dundee Super Bowl Ad 2018 w/ Chris Hemsworth and Danny McBride", dan "INSANE! SpaceX Falcon Heavy Side Boosters Landing Simultaneously at Kennedy Space Center" merupakan faktor utama dalam kenaikan jumlah penonton yang signifikan pada rentang waktu tersebut.

#### E. Kategori Video “Nonprofits & Activism”

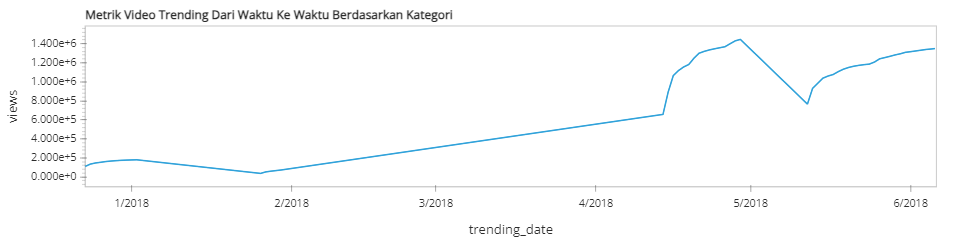


**Gambar 15.** Line Plot Kategori Nonprofit & Activism

Dari grafik tersebut, terlihat bahwa terjadi kenaikan jumlah penonton (views) yang signifikan mulai dari tanggal 16 Januari 2018. Jumlah penonton terus meningkat hingga mencapai puncaknya pada tanggal 2 Februari 2018 sebelum mengalami penurunan yang signifikan. Setelah melakukan analisis lebih lanjut dengan melihat dataset yang digunakan, dapat disimpulkan bahwa pada rentang waktu tersebut, video yang berjudul "Suicide: Be Here Tomorrow." menjadi faktor utama dalam kenaikan jumlah penonton. Video tersebut berfokus pada isu serius mengenai bunuh diri, dengan tujuan meningkatkan kesadaran dan menyediakan sumber daya serta bantuan bagi mereka yang mungkin sedang mengalami masalah mental. Dalam periode tersebut, video ini mendapatkan perhatian yang signifikan dari penonton yang peduli dengan isu kesehatan mental.

Kenaikan jumlah penonton yang signifikan menunjukkan bahwa isu kesehatan mental menjadi perhatian yang penting bagi banyak orang pada rentang waktu tersebut. Video ini dapat dianggap sebagai upaya untuk mempromosikan kesadaran dan mengurangi stigma terkait dengan masalah mental. Dengan demikian, dapat disimpulkan bahwa video "Suicide: Be Here Tomorrow." merupakan faktor utama dalam kenaikan jumlah penonton yang signifikan pada rentang waktu tersebut, karena berhasil menarik perhatian penonton yang peduli dengan isu kesehatan mental.

#### F. Kategori Video “Shows”



**Gambar 16.** Line Plot Kategori Shows

Dari grafik tersebut, terlihat bahwa terjadi kenaikan jumlah penonton (views) yang signifikan mulai dari tanggal 16 Januari 2018. Jumlah penonton terus meningkat hingga mencapai puncaknya pada tanggal 2 Februari 2018 sebelum mengalami penurunan yang signifikan. Setelah melakukan analisis lebih lanjut dengan melihat dataset yang digunakan, dapat disimpulkan bahwa pada rentang waktu tersebut, video yang berjudul "Game of Zones - S5:E1: 'A Golden Summer'" menjadi faktor utama dalam kenaikan jumlah penonton. Video tersebut merupakan episode pertama dari musim kelima serial animasi "Game of Zones", yang merupakan parodi dari serial televisi "Game of Thrones". Video ini berhasil menarik minat penggemar kedua seri tersebut, dengan menggabungkan elemen komedi dan referensi yang cerdas.

Kenaikan jumlah penonton yang signifikan menunjukkan popularitas seri "Game of Zones" dan antusiasme penggemar dalam menyaksikan episode baru. Video ini berhasil membangun loyalitas penggemar yang ingin mengikuti perkembangan cerita yang unik dan lucu. Dengan demikian, dapat disimpulkan bahwa video "Game of Zones - S5:E1: 'A Golden Summer'" merupakan faktor utama dalam kenaikan jumlah penonton yang signifikan pada rentang waktu tersebut, karena berhasil menarik perhatian penggemar dengan konten yang menghibur dan menggabungkan dua serial populer.

# 

# BAB V: KESIMPULAN

## 5.1 Kesimpulan

Dalam penelitian ini, kami menggunakan PySpark untuk melakukan pembersihan data dan teknik analisis time series untuk memahami pola dan tren dalam data video yang trending di YouTube. Dengan memanfaatkan tanggal trending sebagai variabel waktu, kami dapat melihat bagaimana jumlah penonton (views) berubah seiring waktu dan mengidentifikasi periode kenaikan atau penurunan yang signifikan. berdasarkan hasil penelitian dan analisis yang telah dilakukan, kami dapat menyimpulkan hal-hal sebagai berikut:

1. Pyspark cocok digunakan untuk melakukan pembersihan data pada data yang berukuran besar hal ini ditunjukkan dengan rata-rata waktu yang dibutuhkan untuk menjalankan program lebih cepat.

2. Kategori video "Entertainment" memiliki jumlah video trending yang paling banyak di antara kategori video lainnya. Hal ini menunjukkan bahwa konten hiburan memiliki daya tarik yang tinggi bagi penonton dan seringkali mendapatkan perhatian yang besar.

3. Jumlah penonton pada kategori video "Entertainment" dan "Show" menunjukkan tren yang positif, yang mengindikasikan bahwa minat dan minat penonton terhadap konten hiburan terus meningkat dari waktu ke waktu. Namun, untuk kategori "Music", "Game", "Travel & Events", dan "Nonprofits & Activism", jumlah penonton cenderung menunjukkan tren yang datar. Hal ini menggambarkan bahwa minat penonton terhadap konten-konten tersebut tidak mengalami pertumbuhan yang signifikan seiring berjalannya waktu.

# 

# DAFTAR PUSTAKA

|  |  |
| --- | --- |
| [1] | Laraswati, B. D. (2022, July 14). *Analisis Trending YouTube Indonesia dengan Data*. Algoritma Data Science School. https://blog.algorit.ma/analisis-trending-youtube/ |
| [2] | EW Wirga. (2017). “Analisis Konten Pada Media Sosial Video Youtube Untuk Mendukung Strategi Kampanye Politik”. Universitas Gunadarma. Jawa Barat. |
| [3] | J. Hurwitz, A. Nugent, F. Halper, and M. Kaufman, Big Data for Dummies. New Jersey: John Wiley & Sons, Inc. |
| [4] | C. A. M. Toledo, C. Degener, L. Vinhal, G. Coelho, W. Meira, C. Codeco, and M. Teixeira, “Dengue prediction by the web: Tweets are a useful tool for estimating and forecasting dengue at country and city level,” PLOS, vol. 11, no. 7, pp. 1–13,. |
| [5] | NUR, S. K. (2020). Pemanfaatan Big Data Pada Konsep Smart City: Kajian Pustaka. *Jurnal INSTEK (Informatika Sains dan Teknologi)*, *5*(1), 27-36. |
| [6] | A. Ryanto, “Analisis kinerja framework big data pada cluster tervirtualisasi: Hadoop mapreduce dan apache spark,” Makassar. |
| [7] | David, E. R., Sondakh, M., & Harilama, S. (2017). Pengaruh Konten Vlog Dalam Youtube Terhadap Pembentukan Sikap Mahasiswa Ilmu Komunikasi Fakultas Ilmu Sosial Dan Politik Universitas Sam Ratulangi. *Acta Diurna*, *6*(1), 93363. https://www.neliti.com/publications/93363/pengaruh-konten-vlog-dalam-youtube-terhadap-pembentukan-sikap-mahasiswa-ilmu-kom |
| [8] | Oliviandi, S., Osmond, A., & Latuconsina, R. (2018). *IMPLEMENTASI APACHE SPARK PADA BIG DATA BERBASIS HADOOP DISTRIBUTED FILE SYSTEM* |
| [9] | Seimahuira, S. (2022). Analisa Pola Pengguna Youtube Tranding Menggunakan Algoritma Partitioning Around Medoids (PAM) dan FP-Growth. *Jurnal Penelitian Inovatif*, *2*(1), 59–66. https://doi.org/10.54082/jupin.38 |
| [10] | Saputra, P. Y., Subhi, D. H., & Winatama, F. Z. A. (2019). IMPLEMENTASI SENTIMEN ANALISIS KOMENTAR CHANNEL VIDEO PELAYANAN PEMERINTAH DI YOUTUBE MENGGUNAKAN ALGORITMA NAÏVE BAYES. *Jurnal Informatika Polinema*, *5*(4), 209–213. https://doi.org/10.33795/jip.v5i4.259 |
| [11] | R. A. Maisal, A. N. Hidayanto, N. F. Ayuning Budi, Z. Abidin and A. Purbasari. (2019). Analysis of Sentiments on Indonesian YouTube Video Comments: Case Study of The Indonesian Government's Plan to Move the Capital City. |

# LAMPIRAN

Pada bagian lampiran tercantum tambahan lampiran penunjang yang membantu dalam proses pengerjaan proyek.

Dataset: [Trending YouTube Video Statistics | Kaggle](https://www.kaggle.com/datasets/datasnaek/youtube-new)