**Original Article**

https://journal.itera.ac.id/index.php/jsat/

# Generasi Informasi (Teks) dari Kompas yang berkaitan dengan Sains, Teknologi, dan Teknik dengan Large Language Model dan Parallel Airflow Scraping

Miftahul Huda 1 [*a]

[a] miftahul.121450125@student.itera.ac.id

*Corresponding E-mail: journal@itera.ac.id

**Abstract:** *The ability to efficiently extract relevant information from data is increasingly important in the digital age. The purpose of this research is to create a system that can automatically retrieve information from news by using scraping and natural language analysis (NLP) techniques. The main focus is the collection and analysis of news information from the Kompas website, especially those related to the field of science and technology. Kompas news data is collected and presented in tabulation through parallel data scraping exploration with Airflow. To extract relevant information from news texts, natural language processing models such as GPT-2 (Generative Pre-trained Transformer 2) trained with large datasets will be used. Through big data analysis, this research aims to improve data retrieval efficiency, trends, and new insights in science and technology. It is expected that automating the data collection and extraction process will improve accuracy and reduce reliance on manual intervention.*

**Keywords:** *LLM, Kompas, News, Airflow, Scraping*

**Abstrak:** Kemampuan untuk secara efisien mengekstrak informasi relevan dari data semakin penting di era digital. Tujuan dari penelitian ini adalah untuk membuat sistem yang dapat otomatis mengambil informasi dari berita dengan menggunakan *scraping* dan teknik analisis bahasa alami (NLP). Fokus utama adalah pengumpulan dan analisis informasi berita dari situs web Kompas, terutama yang berkaitan dengan bidang sains dan teknologi. Data berita Kompas dikumpulkan dan disajikan dalam tabulasi melalui eksplorasi *scraping* data paralel dengan *Airflow*. Untuk mengekstrak informasi yang relevan dari teks berita, model pemrosesan bahasa natural seperti GPT-2 (Generative Pre-trained Transformer 2) yang dilatih dengan dataset yang besar akan digunakan. Melalui analisis data besar, penelitian ini bertujuan untuk meningkatkan efisiensi pengambilan data, trend, dan wawasan baru dalam sains dan teknologi. Diharapkan bahwa pengotomatisan proses pengumpulan dan ekstraksi data akan meningkatkan akurasi dan mengurangi ketergantungan pada intervensi manual.

**Kata Kunci :** *LLM, Kompas, Berita, Airflow, Scraping*

## Pendahuluan

### Latar Belakang

Dalam konsep pembelajaran mesin saat ini banyak mengembangkan terkait dengan adanya pemrosesan tugas yang mengeksekusi data latih menggunakan kumpulan data besar, model yang berkapasitas tinggi [1]. Pemrosesan data yang berkapasitas tinggi banyak dimanfaatkan dalam

pengambilan suatu informasi sebagai suatu *insight* yang bisa diambil dan dikembangkan menjadi suatu ide yang dapat diterapkan dimana saja untuk era digital ini. Informasi menjadi aset yang sangat berharga. Oleh karena itu kemampuan dalam ekstrak data memahami data dan menghasilkan suatu informasi yang relevan dari data yang tersedia secara efisien menjadi sangat penting [2]. Salah satu pendekatan yang menjadikan penggunaan model pemrosesan bahasa besar seperti GPT-2 (*Generative Pre-trained Transformer 2*) sangat efektif adalah kemampuannya untuk menangkap pola dan konteks bahasa secara mendalam. Menurut laporan Balai Diklat Keuangan Pontianak model ini dilatih menggunakan sejumlah besar data teks dari berbagai sumber terdata dalam versi kedua model GPT yang telah diresmikan pada tanggal 5 November 2019. Model GPT-2 memiliki 1,5 miliar parameter dan dilatih menggunakan dataset 8.000.000 halaman website [3][4].

Informasi saat ini tersebar luas dalam era digital yang memungkinkan mengumpulkan data dalam berbagai format dapat dilakukan kapan saja, namun proses pengumpulan data yang dilakukan secara manual dapat menjadi tidak efisien karena memakan waktu yang terlalu banyak. Teknik *scraping* bisa menjadi salah satu kunci penting dalam memungkinkan pengumpulan dan ekstraksi data dalam jumlah besar dari berbagai sumber online dengan cepat dan terstruktur[5].

Dari berbagai banyaknya sumber informasi *online* yang ada, Kompas merupakan salah satu jawaban dari sumber berita dan informasi di Indonesia yang berdiri sejak tahun 1965. Dengan jangkauan nasional yang luas, Kompas menjadi salah satu sumber referensi utama bagi masyarakat Indonesia dalam berbagai bidang, termasuk sains, teknologi, dan teknik. Mendasari tujuan penelitian ini, dilakukan eksplorasi penggunaan teknik *scraping* data paralel dengan Airflow untuk mengumpulkan data yang relevan dari situs berita online Kompas .

Dalam penelitian ini menggunakan data yang didapatkan dari hasil *scraping* laman berita kompas disajikan dalam bentuk tabulasi data. *Web Scraping* adalah proses pengambilan dokumen semi-terstruktur dari internet, biasanya berupa halaman *web* dalam bahasa *markup* seperti HTML atau XHTML. Proses ini melibatkan analisis dokumen tersebut untuk mengekstrak data tertentu dari halaman tersebut guna digunakan untuk kepentingan lain. Meskipun sering disebut sebagai *screen scraping*, *Web Scraping* tidak dapat dianggap sebagai bagian dari data mining karena data mining bertujuan untuk memahami pola semantik atau tren dari sejumlah besar data yang telah diperoleh. Aplikasi *Web Scraping*, yang juga dikenal sebagai *intelligent, automated, atau autonomous agents*, hanya fokus pada cara memperoleh data melalui pengambilan dan ekstraksi data dengan ukuran data yang bervariasi[6].

Penelitian ini berfokus pada data yang didapatkan pada website kompas untuk topik ilmu pengetahuan. Berdasarkan penelitian terdahulu dalam publikasi yang disajikan oleh Alec Radford dkk, didapatkan hasil bahwa ketika diumpankan dengan sebuah dokumen dan pertanyaan, model bahasa besar dapat menghasilkan jawaban yang mencapai skor F1 55 pada dataset CoQA (*Question Answering*) - setara atau lebih baik dari 3 dari 4 sistem baseline tanpa menggunakan 127.000+ contoh pelatihan[7]. Ini yang menjadikan dasar penelitian ini untuk mengembangkan lebih lanjut dalam implementasi *Natural Language Processing* dengan menggunakan data hasil *scraping* website kompas.

**Rumusan Masalah**

1. Bagaimana cara mengembangkan sistem yang dapat mengambil informasi yang relevan dari berita dengan cepat dan efisien.
2. Bagaimana cara mengurangi waktu yang dibutuhkan dan meningkatkan akurasi dibandingkan dengan metode manual.

**Tujuan Penelitian**

1. Mengembangkan sistem otomatisasi pengambilan informasi dari berita.
2. Mengurangi waktu pengambilan informasi dengan mengotomatisasi proses pengumpulan data.

**Batasan Masalah**

Dalam Penelitian ini mempunyai Batasan Masalah :

1. Informasi Yang diambil dan diberikan hanya berdasarkan informasi mengenai Sains dan Teknologi.
2. Informasi Yang diambil dan diberikan hanya bersumber dari satu sumber, yaitu Kompas.

**Manfaat Penelitian**

1. Peningkatan efisiensi pengambilan informasi dengan mengolah volume data besar dengan cepat dan efisien menggunakan teknologi *Big Data*.
2. Mengurangi kebutuhan intervensi manual dan biaya operasional melalui pemrosesan data yang efisien.
3. Menggunakan analisis data besar untuk menemukan pola, trend, dan wawasan dalam bidang sains dan teknologi.

4. Menerapkan dan mengembangkan teknik *Big Data* dalam kasus nyata, meningkatkan pemahaman dan ekstraksi informasi dari teks berita.

## Tinjauan Pustaka

### Analisis Big Data

*Big Data* adalah istilah yang menggambarkan volume data yang besar, baik terstruktur dan tidak terstruktur yang dijadikan sebagai informasi untuk pengambilan keputusan. *Big Data* dapat dianalisis untuk wawasan yang mengarah pada keputusan yang lebih baik dan bergerak bisnis strategis [8]. *Big data* dapat diartikan dalam 3v yaitu *volume* (data set disimpan dalam jumlah besar), *velocity* (kebutuhan untuk mengakses data sangat besar), dan *variety* (format data yang sangat bervariasi saat ini)[9].

Analisis data adalah proses meneliti data untuk mengetahui pola tersembunyi, korelasi yang belum diketahui, dan informasi berguna lainnya. Sedangkan pengertian big data analytics adalah proses meneliti, mengolah data set besar (*Big Data*) untuk mengetahui pola tersembunyi, korelasi yang tidak diketahui, tren pasar, preferensi pelanggan dan informasi bisnis berguna lainnya. *Analisis big data* di era teknologi saat ini memungkinkan semua sektor mendapat hasil olah data secara cepat bahkan real time[10].

### Apache Spark

*Apache Spark* adalah sebuah *framework* untuk melakukan proses analisis *big data* yang *open source*, cepat, dan disimpan dalam memori. *Apache Spark* menyediakan *API* yang bersifat *high-level* untuk melakukan proses pengolahan data dalam bahasa *Java, Scala, Python*, dan *R*[11]. 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 application, iterative algorithms, interactive queries* dan *streaming*[12].

### Web Scraping

*Web Scraping* adalah sebuah program atau kode otomatis yang dapat mengunjungi dan mencari informasi yang sesuai dengan kata kunci dalam sebuah website. *Web scraping* ini berkelana di *web* dan berfungsi untuk mengumpulkan segala informasi tentang suatu halaman *web* dan mengindeksnya ke dalam suatu *database*. Informasi tentang halaman *web* tersebut didapat dari kata-kata yang terdapat di dalam halaman *web* tersebut. Kata-kata tersebut kemudian diindeks menjadi sebuah daftar kata-kata yang biasanya dipakai sebagai kata kunci untuk menemukan halaman *web*. Proses pengumpulan informasi tentang halaman *web* dari situs web ini disebut *Web scraping*[13].

### Large Language Models (LLMs)

*Large Language Models* (LLMs) adalah jenis model kecerdasan buatan yang dirancang untuk memahami dan menghasilkan teks dalam bahasa alami dengan tingkat kemampuan yang semakin menyerupai manusia[14]. LLMs seperti GPT (*Generative Pre-trained Transformer*) dikembangkan oleh OpenAI dan BERT (*Bidirectional Encoder Representations from Transformers*) oleh *Google* merupakan contoh terkenal dari jenis model ini. LLMs memiliki kemampuan untuk menganalisis konteks teks yang kompleks, mengenali pola bahasa, dan menghasilkan teks yang berkualitas. Mereka dilatih menggunakan metode *deep learning* pada dataset teks yang sangat besar, seperti arsip web atau koleksi buku, untuk mempelajari pola bahasa yang umumnya digunakan dalam komunikasi manusia.

Kegunaan utama LLMs terletak pada berbagai aplikasi di bidang pemrosesan bahasa alami (NLP). Mereka digunakan untuk melakukan tugas-tugas seperti penerjemahan bahasa, ringkasan teks, generasi teks, dan pemahaman bahasa alami. Misalnya, LLMs dapat digunakan untuk menerjemahkan teks dari satu bahasa ke bahasa lain, menghasilkan ringkasan singkat dari teks panjang, atau menjawab pertanyaan berdasarkan informasi yang terdapat dalam teks.

LLMs telah menjadi fokus utama dalam penelitian dan perkembangan di bidang NLP sejak pertengahan 2010-an. Perkembangan utama dalam arsitektur LLMs dimulai dengan penemuan transformer, sebuah arsitektur jaringan saraf yang sangat efisien untuk memproses urutan data, yang memungkinkan pelatihan model bahasa yang lebih besar dan lebih canggih. Sejak itu, telah terjadi serangkaian peningkatan dalam ukuran dan kompleksitas LLMs, yang telah menghasilkan kemajuan besar dalam kemampuan komputasi bahasa alami.

### Nilai Informasi dalam Era Digital

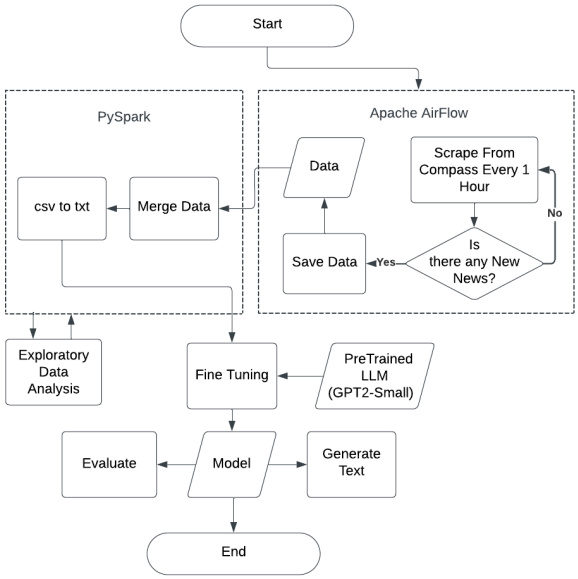
Nilai informasi dalam era digital sangat penting dan melibatkan aksesibilitas, pengelolaan, dan pemanfaatan informasi. Era digital telah mengubah cara kita mengakses, menyimpan, dan berbagi informasi dengan cepat dan efisien melalui internet. Nilai informasi tidak hanya terletak pada

kuantitasnya, tetapi juga pada kualitas, relevansi, dan kemampuan untuk mengubah informasi menjadi pengetahuan yang berguna. Meskipun memberikan manfaat besar dalam pengelolaan data dan analisis informasi, era digital juga membawa tantangan baru, seperti penilaian kebenaran informasi dan perlindungan privasi. Oleh karena itu, penting untuk menggunakan informasi secara bijaksana dan kritis dalam era digital ini[15].

### Airflow

Apache Airflow adalah platform manajemen alur kerja open-source yang memungkinkan pengguna untuk mengotomatisasi, menjadwalkan, dan memantau alur kerja yang kompleks dalam lingkungan komputasi terdistribusi. Dengan Airflow, pengguna dapat mendefinisikan alur kerja sebagai Directed Acyclic Graphs (DAGs) menggunakan kode Python, yang memungkinkan pembangunan alur kerja yang fleksibel dan dinamis. Airflow juga menawarkan fitur penjadwalan yang kuat, yang memungkinkan penjadwalan tugas dengan menggunakan cron expressions atau interval waktu tertentu. Platform ini dilengkapi dengan antarmuka pengguna berbasis web yang memungkinkan pengguna untuk memantau status tugas, melihat log, dan mendapatkan notifikasi jika terjadi kegagalan. Selain itu, Airflow dapat berintegrasi dengan berbagai sistem big data dan alat ETL seperti Hadoop, Spark, Hive, dan Presto, sehingga memudahkan integrasi dengan ekosistem data yang besar. Dengan fitur-fitur ini, Airflow menjadi pilihan populer bagi perusahaan teknologi dan tim data untuk otomatisasi proses ETL, pipeline data, pengolahan batch, dan orkestrasi berbagai proses dalam lingkungan data yang kompleks.

## Metode




**Gambar 1.** Flowchart Metode

### *Scraping* Data

Data didapat melalui hasil *scraping* dari *website* kompas.com dengan kata kunci yang berkaitan dengan sains, teknologi, dan teknik. Proses *scraping* menggunakan komputasi paralel secara asinkron dengan $n_{worker} = n_{core}$ yang di mana semakin banyak jumlah core dalam CPU/GPU/TPU semakin cepat proses *scraping*. Proses *scraping* dijadwalkan dengan *Apache Airflow* yang terhubung dengan server. *Apache Airflow* adalah *platform* dari *Apache* untuk menulis, menjadwalkan, dan memantau alur kerja secara terprogram[17].


**Gambar 2.** Gambaran Umum Proses *Airflow Scraping*

Penjadwalan proses *scraping* dilakukan setiap 1 jam, jika terdapat berita baru (update pada kompas) data tersebut

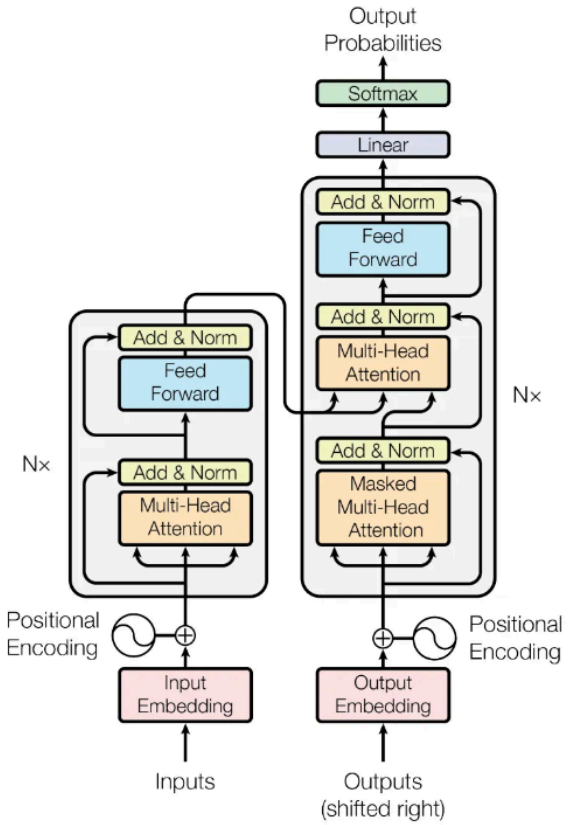
akan ditambahkan ke dalam data yang sudah ada. Pada penelitian ini menggunakan server lokal sebagai uji coba.

**Preprocessing Data dan Eksplorasi Data Analisis**

Pada proses ini menggunakan *PySpark* untuk melakukan preprocessing data. *PySpark* adalah API *Apache Spark* untuk *python*, memungkinkan melakukan pemrosesan data skala besar di dalam lingkungan *python*[16]. Data hasil *scraping* berbentuk csv terpisah setiap file, oleh karena itu data digabung menjadi satu csv. Kemudian data dieksplorasi untuk melihat gambaran data dan dijadikan file txt untuk input *Large Language Model*.

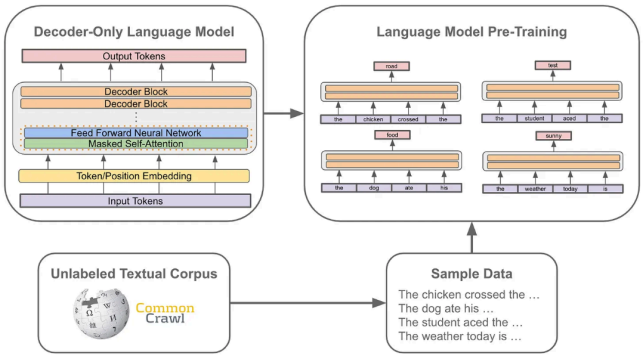
***Fine Tuning Large Language Model***

*Large Language Model* (LLM) khususnya *Generative Pre-Trained Transformer* (GPT) mempunyai arsitektur *transformer* besar yang telah dilatih sebelumnya untuk memprediksi token berikutnya berdasarkan token input[18].




**Gambar 3.** Arsitektur *Transformer*

*Encoder* mengubah input menjadi representasi yang memiliki informasi menggunakan *multi-head attention* dan jaringan *feed-forward*. Kemudian decoder menghasilkan teks output token demi token, dengan memastikan hanya konteks sebelumnya yang digunakan untuk setiap prediksi, sambil memperhatikan representasi dari *encoder*. *Output* dihasilkan melalui proses linier dan *softmax* untuk memberikan probabilitas setiap token.




**Gambar 4.** Proses Pelatihan *Generative Pre-Trained Transformer*

Pada proses pre-training di mana model belajar dari banyak data yang tidak berlabel untuk memahami struktur dan pola bahasa tujuannya untuk memprediksi token berikutnya. Model seperti GPT hanya menggunakan bagian dekoder dari arsitektur *transformer*, yang dirancang untuk tugas-tugas generatif seperti melanjutkan teks. *Language Model* GPT-2 melalui pendekatan *unsupervised learning* dan probabilitas kondisional untuk memprediksi token berikutnya. Misal terdapat serangkaian kalimat dalam data $(x_1, x_2, \ldots, x_n)$ masing-masing terdiri dari serangkaian token $(s_1, s_2, \ldots, s_n)$ maka $p(x)$ adalah serangkaian token yang dihasilkan mengikuti probabilitas kondisional[7].

$$p(x) = \prod_{i=1}^{n} p(s_n|s_1, s_2, \ldots, s_{n-1}) \quad \textbf{(1)}$$

Pendekatan ini secara dinamis dengan jumlah $k$ token yang akan dihasilkan $p(s_{n-k}, \ldots, s_n|s_1, \ldots, s_{n-k-1})$. Dengan begini cocok untuk banyak token yang akan di generate. *Language Model* bertujuan untuk memaksimalkan peluang tersebut yang berarti meminimumkan *negative log-likelihood* berdasarkan hal tersebut fungsi kerugian *(loss function)* dapat didefinisikan sebagai:

$$L(S, \Theta) = \sum_{i=1}^{N} log(P(s_i|s_{i-k}, \ldots, s_{i-1}, \Theta)) \quad \textbf{(2)}$$

Dengan $\Theta$ adalah parameter-parameter dalam model yang terus diperbarui selama pembelajaran oleh model *(backpropagation)*. Dalam *transformer backpropagation* diterapkan pada *feedforward neural network* dan *multi-head attention.* Misal terdapat $\Theta$ matriks bobot dalam

jaringan pada *feedforward* dan bobot *query* pada *multi-head attention:*

$$m_t^{(\Theta)} = \beta_1 m_{t-1}^{(\Theta)} + (1 - \beta_1)\nabla L(\Theta) \quad \textbf{(3)}$$

$$v_t^{(\Theta)} = \beta_2 v_{t-1}^{(\Theta)} + (1 - \beta_2)(\nabla L(\Theta))^2 \quad \textbf{(4)}$$

$$\Theta_{new} = \Theta - \eta \frac{(m_t^{(\Theta)})/(1-\beta_1^{\,t})}{\sqrt{(v_t^{(\Theta)})/(1-\beta_2^{\,t})}+\epsilon} \quad \textbf{(5)}$$

Di mana $m$ dan $v$ adalah momen pertama dan kedua dengan koefisien $\beta_1$dan $\beta_2$ lalu konstanta kecil pencegah pembagian dengan nol $\epsilon$ dan laju pembelajaran *(learning rate)* $\eta$.

**Evaluasi dan Generasi Teks**

Evaluasi model difokuskan dengan penurunan *loss* dan nilai gradien *loss* pada setiap *epochs*. Model setelah *fine tuning* akan diuji dengan menggenerate beberapa *prompt* yang sesuai dengan tema dalam data (konten berita).

## Hasil dan Pembahasan

**Scraping**

Dalam penelitian ini data diambil menggunakan metode *Scraping web* yang bersumber dari *web* Kompas.com, dengan bantuan dari library yang digunakan sebagai contoh ada *library multiprocessing* yang berfungsi untuk menentukan *thread* optimal untuk *scraping paralel* berdasarkan cpu yang dimiliki oleh *developer*, lalu ada *library request* yang digunakan untuk mengirim HTTP *request* atau bisa dibilang *library* ini membantu kita untuk mengambil konten halaman di kompas, adapun bs4 atau biasa disebut dengan *BeautifulSoup* yang berfungsi untuk mengekstrak elemen-elemen yang dimiliki oleh *web* kompas nantinya, lalu terdapat *library json* yang digunakan untuk menyimpan semua format ataupun hasil *scraping* ke dalam format *json*. *library-library* tersebut membantu penelitian ini untuk mencari konteks berita yang dibutuhkan, dalam penelitian ini dilakukan pencarian atau *scraping* ke dalam *web* kompas dengan kata kunci *AI*, komputer, sains, teknik, matematika, statistika, pemrograman, data, fisika, biologi, algoritma.

**Dataset**

Dataset yang peneliti gunakan adalah data yang dihasilkan dari proses *scraping* terhadap *web* kompas, data memiliki struktur seperti beberapa sampel berikut :

| Tanggal | Judul | Konten | Link |
|---|---|---|---|
| 2024-05-09 14:03:00 | Kontroversi Reka..... | LONDON, KOMPAS.com | http://www.kompas... |
| 2024-05-06 10:30:00 | X Twitter Rilis F..... | KOMPAS.com- Media..... | http://tekno.komp... |
| 2024-05-06 15:01:00 | Mengenal Gemini A... | KOMPAS.com – Artif..... | http://tekno.komp... |
| 2024-05-11 10:05:00 | Siapkan Generasi..... | KOMPAS.com -Sekol..... | http://www.kompas... |
| 2024-05-06 12:41:00 | [HOAKS] Foto Rest..... | Hoaks!Berdasarkan..... | http://www.kompas... |
| 2024-05-10 15:35:00 | [HOAKS] Foto Riha..... | Hoaks!Berdasarkan..... | http://www.kompas... |
| 2024-05-07 17:58:00 | Hoaks, Spongebob..... | KOMPAS.com- Sebuah... | http://www.kompas... |
| 2024-05-11 11:11:00 | INFOGRAFIK: Konten. | KOMPAS.com- Di me..... | http://www.kompas... |
| 2024-05-08 15:22:00 | Novel Dilan 1983:..... | JAKARTA, KOMPAS.co.... | http://tekno.komp... |
| 2024-05-08 09:10:00 | Chip Apple M4 Mel.... | KOMPAS.com- Di me..... | http://www.kompas... |

**Tabel 1**. Informasi Dataset

pada gambar diatas adalah gambar visualisasi dari dataset yang dihasilkan dan digunakan sebagai sumber informasi untuk program *generate*, data tersebut memiliki atribut tanggal, judul yang menginterpretasikan bahwa judul dari artikel atau konten yang diterbitkan oleh *web* kompas, dan konten sendiri adalah kandungan yang berisi informasi-informasi yang dimiliki oleh konten itu sendiri, seperti isi berita,penulis dan sebagainya yang dimiliki oleh konten tersebut, sedangkan untuk atribut *link* adalah atribut yang diambil sebagai sumber dimana konten itu didapatkan, data ini sendiri mempunyai jumlah berita 28839 dan jumlah kata dalam berita sebanyak 8913668, dan untuk berita terlama nya sendiri berasal dari tanggal 2013-05-03 dan berita terbaru di tanggal 2024-05-12.

**Preprocessing Data**

Dalam penelitian ini *preprocessing* dilakukan sebelum dataset terbentuk, hal ini dikarenakan proses *scraping* yang membuat penelitian ini harus melakukan *preprocessing* sebelum dataset terbentuk karena tujuan dari *scraping* sendiri adalah untuk membuat sebuah dataset, dalam *preprocessing* ini telah dilakukan seperti penghapusan karakter-karakter yang tidak penting.

**Visualisasi Word Cloud**

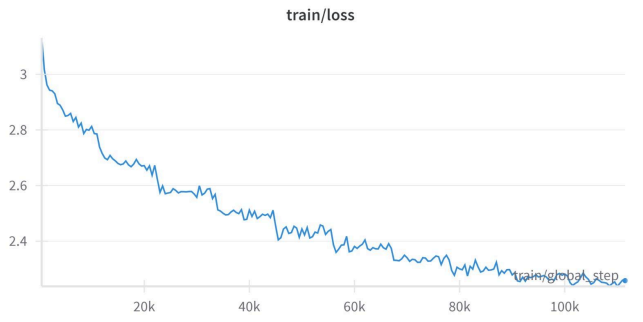
Dari hasil dataset penelitian ini bisa dilihat informasi apa saja yang banyak terkandung dalam konten kompas yang sudah didapatkan, dari hasil visualisasi *word cloud* didapatkan bahwa kata atau kalimat-kalimat seperti Aplikasi *Whatsapp*, Baca Juga, Kamu Sudah, Langsung, install aplikasi

adalah kalimat atau kata yang memiliki frekuensi paling banyak dimiliki, dari hasil visualisasi ini bisa dikatakan bahwa konten yang diambil masih dijaga untuk sumber dan keakuratannya tanpa menghilangkan sedikit informasi. berikut gambar hasil dari visualisasi word cloud.


**Gambar 5.** *Word Cloud*

## Training Argument

*Training Argument* adalah tahap dimana informasi yang dimiliki dari dataset di lakukan pelatihan menggunakan bantuan dari *DaracollactorFor Language Modelling* yang berguna untuk mengatur cara pengumpulan data berdasarkan tokenizer yang digunakan dan menggunakan *Masked Language Modelling*(MLM) tetapi *Autoregressive Language Modeling* seperti yang dilakukan pada GPT-2. *Training* argumen dalam penelitian ini dilakukan dengan menggunakan 10 *epoch* dan 10 *batch size*, yang berarti setiap *batch* akan terdiri dari 10 sampel. berikut adalah *linechart* dari monitoring *train/loss* yang dihasilkan dari proses *train*:




**Gambar 6.** *Linechart monitoring train/loss*

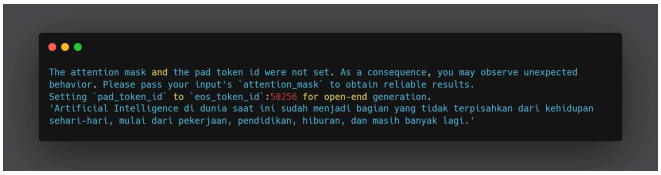
dari *line chart* tersebut dapat dilihat bahwa garis mengalami penurunan, yang mengindikasikan bahwa setiap kali model melakukan pelatihan model semakin akurat dalam mempelajari data pelatihan, pada awal pelatihan kita bisa melihat bahwa penurunan begitu tajam yang mengindikasikan bahwa model cepat belajar terhadap pola dasar data, dan seiring berjalannya waktu pola penurunan menjadi sedikit lambat yang mengindikasikan bahwa model mendekati konvergensi.

## Modelling

Dalam *modelling* di penelitian ini digunakan untuk menghasilkan teks respon berdasarkan model GPT2 yang telah dilatih, dalam *modeling* ini peneliti membuat sebuah fungsi yang bernama *generate response* dengan panjang maksimal kata dari *response* itu sendiri adalah 40 kata dan untuk hasil keluaran sendiri di kontrol serendah mungkin agar bisa mendapat jawaban yang konsisten atau deterministik. dari setiap output teks yang akan keluar akan truncate output untuk memangkas atau mengakhiri kalimat berdasarkan tanda baca yang sudah ditentukan seperti “.” sebagai contoh berikut adalah contoh penggunaan dan hasilnya dari *Generate Box*.


**Gambar 7.** *Inputan key*

jika dilakukan sebuah input terhadap fungsi yang dikhususkan untuk menerima dan memberikan informasi berdasarkan pertanyaan yg diberikan, maka hasilnya akan memberikan jawaban yang masih termasuk ke dalam konten pertanyaan tersebut. seperti gambar diatas jika kita memberikan *generate* kalimat “Artificial Intelligence di dunia” maka keluaran yang akan diberikan oleh model ini adalah sebagai berikut.




**Gambar 8.** *Output generate*

Bisa dilihat pada gambar tersebut terdapat informasi yang menjawab pertanyaan atau kalimat yang kita generate dalam fungsi sebelumnya.

**Kesimpulan**

Penelitian ini berhasil mengembangkan sistem yang secara otomatis dapat mengambil informasi relevan dari berita dengan menggunakan teknik pengumpulan data dan pengolahan bahasa alami (NLP). Sistem ini menggunakan model bahasa besar seperti GPT-2 yang telah dilatih dengan kumpulan data teks besar untuk memahami konteks dan pola bahasa dalam teks berita. Dengan menggunakan metode pembelajaran mesin dan pembelajaran mendalam, sistem ini dapat mengekstrak informasi penting dari teks berita dengan akurasi lebih tinggi dibandingkan metode manual.

Proses pengambilan data berita dari situs Kompas dilakukan secara paralel menggunakan *Apache Airflow* untuk meningkatkan efisiensi pengumpulan data. Data yang dikumpulkan kemudian diproses dan digunakan untuk melatih model NLP seperti GPT-2. Setelah dilatih dengan data berita yang relevan, model tersebut dapat menghasilkan ringkasan, jawaban atas pertanyaan, atau informasi lain yang diperlukan berdasarkan konteks teks berita. Sistem ini memberikan manfaat dengan meningkatkan efisiensi pengambilan informasi dari data dalam jumlah besar, mengurangi kebutuhan akan intervensi manual dan menemukan pola, trend, dan wawasan baru dalam bidang keilmuan sains dan teknologi.

Penggunaan teknologi *Big Data* dan NLP memungkinkan pengelolaan dan analisis informasi yang lebih cepat, akurat, dan terstruktur. Secara keseluruhan, penelitian ini menunjukkan potensi besar untuk mengotomatisasi proses pengambilan informasi dari sumber data online seperti berita, dengan memanfaatkan kemajuan terkini di bidang NLP dan pembelajaran mesin. Namun, pengembangan lebih lanjut masih diperlukan untuk mengeksplorasi kumpulan data yang lebih beragam, meningkatkan performa model, dan mengatasi aspek lain seperti bias dan keamanan data.

## References

[1] Krizhevsky, Alex, et al. “ImageNet classification with deep convolutional neural networks.” *ImageNet Classification with Deep Convolutional Neural Networks*, vol. 60, no. 6, 2017, pp. 84–90.

[2] Yudhanto, Yudho, and Abdul Aziz. *Pengantar Teknologi Internet of Things IoT)*. UNSPress, 2019. Accessed 23 May 2024.

[3] Dwi Hatmanto, Endro. “Revolusi Pembelajaran Bahasa Inggris: Menggali Potensi Transformatif Chat GPT.” *Inovasi dan Daya Saing dalam Linguistik dan Pembelajaran Bahasa Asing Masa Kini*, 2023, p. 102.

[4] “Artikel :: CHAT GENERATIVE PRE-TRAINED TRANSFORMER PELUANG, TANTANGAN, ATAU ANCAMAN DUNIA PENDIDIKAN?” *BPPK Kemenkeu*, 28 August 2023,

[5] Firdaus, Ali, and Wahyu Istalama Firdaus. *Text Mining Dan Pola Algoritma Dalam Penyelesaian Masalah Informasi : (Sebuah Ulasan)*, vol. 13, no. 1, 2021, pp. 66 - 78.

[6] Abdillah, L. A., Josi, A., & Suryayusra. (2023). “Penerapan Teknik Web Scraping pada Mesin Pencari Artikel Ilmiah”. Jurnal Informatika, 12(2), 123-134. Universitas Bina Darma.

[7] Radford, Alec, et al. “language models are unsupervised multitask learners.” *Language Models are Unsupervised Multitask Learners*, 2018.

[8] Hapsari, N. F. A. (2020). “Big Data dan Pemanfaatannya di Perpustakaan”. *Jurnal Ilmu Perpustakaan*, 2(1), 24-32. Universitas Muhammadiyah Mataram.

[9] Adriyana, L., & Darumoyo, K. (2020). “Big Data Analytics in Public Sector (Library and Archive)”. *ACARYA PUSTAKA*, 7(2),2442-4366.https://ejournal.undiksha.ac.id/index.php/AP

[10] Zhai, C., & Massung, S. (2016). “Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining”. Morgan & Claypool Publishers.

[11] Jeremia, D., Palit, H. N., & Gunawan, A. (2020). “Platform Big Data Analytic Berbasis Apache Spark Bagi Pemula Dalam Menyusun Data Analysis Workflow”. Universitas Kristen Petra. *Program Studi Informatika Fakultas Teknologi Industri,* Surabaya.

[12] Oliviandi, S., Osmond, A. B., & Latuconsina, R. (2018). “Implementasi Apache Spark pada Big Data Berbasis Hadoop Distributed File System”. *e-Proceeding of Engineering, 5(1)*, 1005. ISSN: 2355-9365.

[13] Fadhlullah, N., Setiawansyah, & Surahman, A. (2022).

"Penerapan Teknologi Web Scraping sebagai Pengumpulan Data COVID-19 di Provinsi Lampung". *Jurnal Informatika dan Rekayasa Perangkat Lunak (JATIKA), 3(1),* 25-30. http://jim.teknokrat.ac.id/index.php/informatika

[14] Kalyan, K. S. (2024). A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, 6. Diakses dari www.elsevier.com/locate/nlp

[15] Danuri, M. (2019). "Perkembangan dan Transformasi Teknologi Digital". *INFOKAM, 15(II)*, 116. Manajemen Informatika, AMIK Jakarta Teknologi Cipta Semarang.

[16] "PySpark Overview — PySpark master documentation." *Apache Spark*, 24 February 2024, https://spark.apache.org/docs/latest/api/python/index.html. Accessed 23 May 2024.

[17] *Apache Airflow*, https://airflow.apache.org/. Accessed 23 May 2024.

[18] Ph. D. Cameron R. Wolfe, "Language models: GPT and GPT-2," Language Models: GPT and GPT-2 - by Cameron R. Wolfe, Ph.D., https://cameronrwolfe.substack.com/p/language-models-gpt-and-gpt-2 (accessed May 23, 2024).

## Lampiran

Berikut kami lampirkan code yang kami kerjakan :

https://github.com/Mfys212/Generasi-Information-Teks-from-Kompas/tree/main