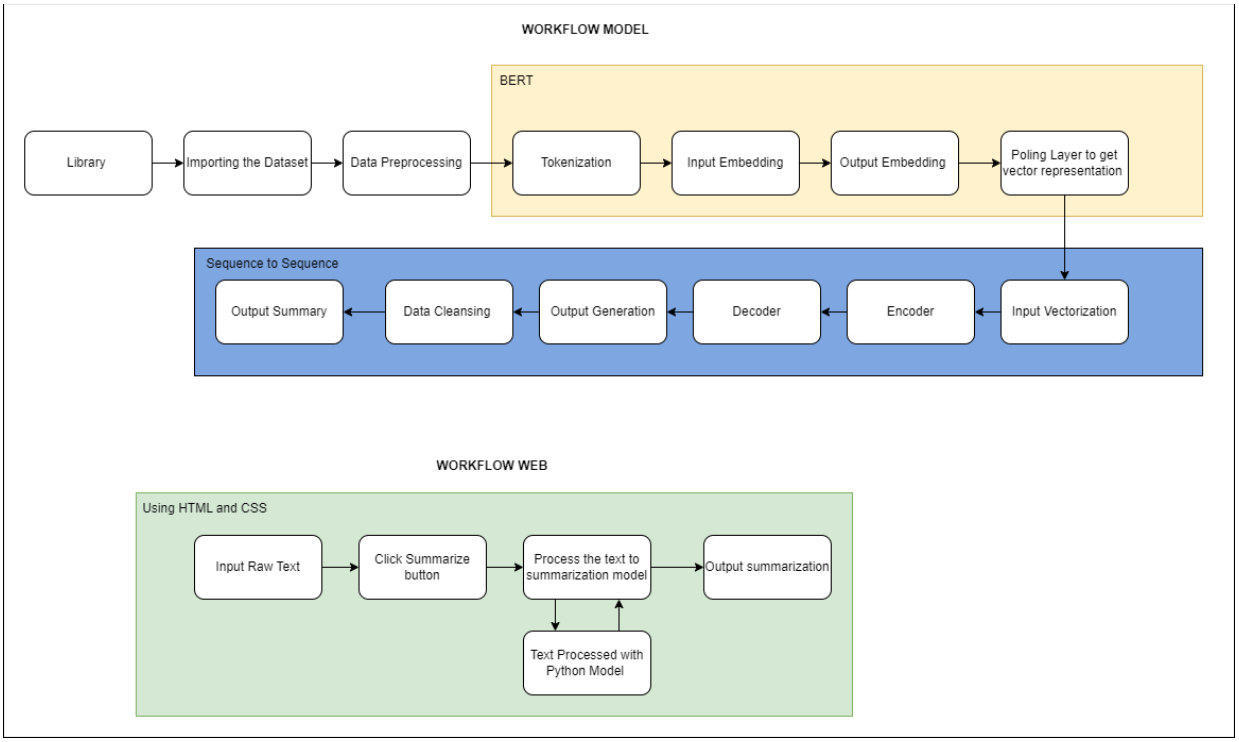
Nama Anggota :

- KANIA GALIH WIDOWATI – 2502047070
- KEZIA FOEJIONO – 2540131014
- NICHOLAS – 2540133291

Abstractive Text Summarization using BERT for Feature Extraction and Seq2Seq Model for Summary Generation

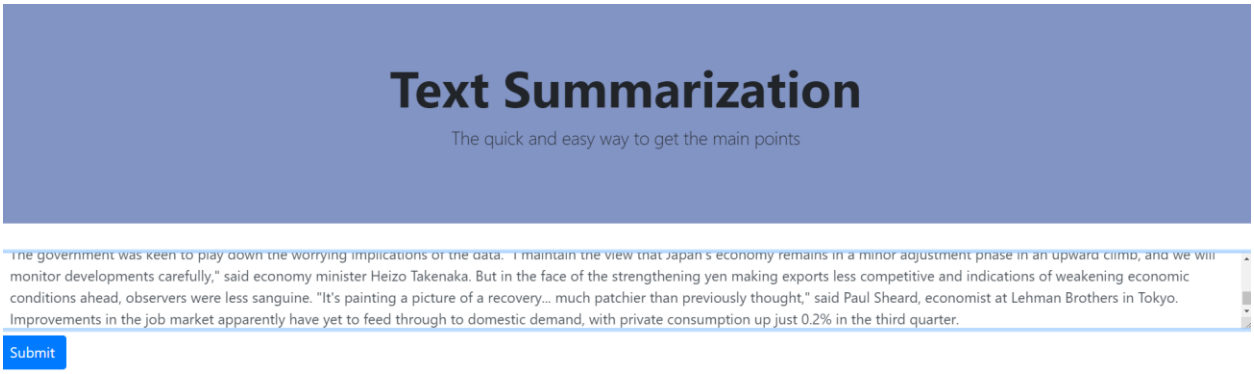
1. Workflow

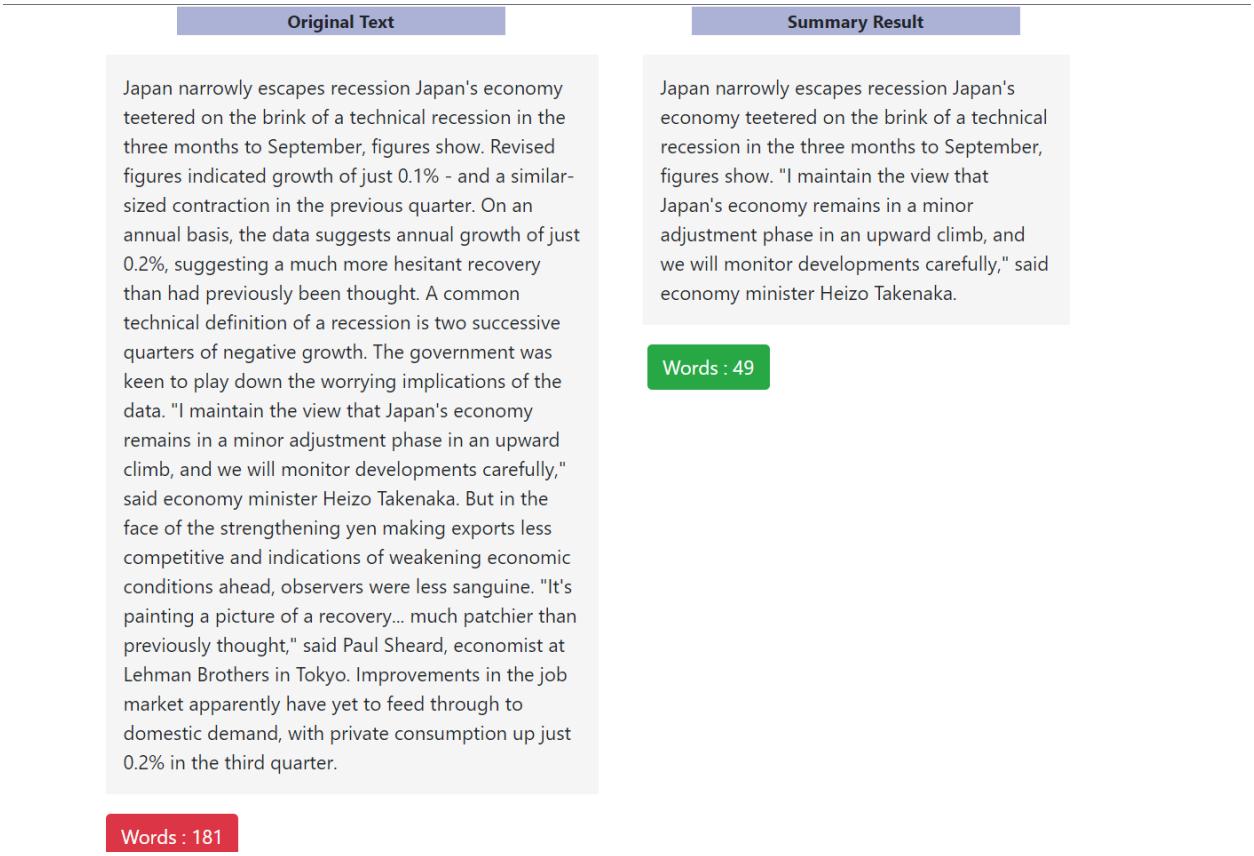


Pada model yang kami buat terbagi menggunakan method BERT dan Sequence to Sequence. Langkah awal yaitu menuliskan library yang dibutuhkan kemudian import dataset. Dataset yang kami gunakan untuk training adalah CNN Daily Mail dataset, kemudian Ketika testing menggunakan BBC Dataset dimana kami melakukan declare variable dengan nama text dan berisikan beberapa paragraph yang nantinya akan dilakukan tokenization dan menentukan stopword. Selain itu juga dilakukan perhitungan word count untuk mengetahui seberapa banyak kata tersebut muncul dan akan digunakan untuk menghitung probability untuk setiap kata. Pada flowchart diatas terdapat input embedding dan output embedding dan menghasilkan vector representation yang akan digunakan pada tahapan mneggunakan method sequence to sequence. Pada

sequence to sequence vector yang didapatkan dari hasil perhitungan menggunakan method BERT, akan dilakukan encoder dan decoder dan menghasilkan output berupa summary. Namun hasil ini masih belum final dikarenakan perlu dilakukan cleansing seperti menghilangkan tanda baca (\ “ ~). Setelah dilakukan data cleaning maka akan dihasilkan output summary.

Pada bagian website hal yang perlu dilakukan adalah memasukkan text yang akan dilakukan summarize pada bagian input text, kemudian user melanjutkan dengan menekan tombol “summarize”. Pada tahapan ini text akan diproses dengan python model yang sudah dibuat pada bagian atas, kemudian akan memberikan output berupa hasil summarize yang akan dimunculkan pada bagian Summary result.





2. Dataset

Dataset yang kami gunakan adalah **CNN Daily Mail dataset**. Kami langsung memanggil pretrained model dengan method seq2seq namun tokenizer dilakukan menggunakan method BERT. Untuk menjalankan model ini diperlukan import library transformer untuk AutoTokenizer dan AutoModelForSeq2SeqLM. CNN Daily Mail dataset diambil melalui Hugging Face Model Hub sehingga ketika dilakukan pemanggilan maka akan secara otomatis melakukan download. Kemudian pada bagian training dan test kami menggunakan **BBC Dataset**, dimana data ini memiliki beberapa topic yaitu Business, Politic, Entertainment, Sport dan Tech. Dataset BBC ini berbentuk file.txt dan setiap kategori memiliki kurang lebih 500 file. Sehingga kami hanya menggunakan satu file untuk melakukan testing kemudian juga memasukkan file dari hasil summary manusia kemudian membandingkan hasil summary dari model dan hasil summary dari manusia, yang terdapat pada dataset BBC. Pada model yang kami buat test1_.txt merupakan file yang kami ambil dari BBC dataset untuk melakukan testing, dan key1.txt merupakan hasil summary dari manusia. Berikut merupakan hasil summary dari model yang kami buat.

```
2   summary = " ".join(summaries)
3   pprint(summary)
4
```

```
('quarterly profits jumped 76 % to $ 1. 13bn from $ 639m year - earlier. '
 'timewarner said fourth quarter sales rose 2 % to 11. 1bn from $ 10. 9bn. the '
 "firm is now one of google's biggest investors in google. aol lost 464, 000 "
 'subscribers in the fourth quarter profits were lower than in the previous '
 'three quarters. its own internet business, aol, had mixed fortunes. it hopes '
 'to offer the online service free to timewarners and timewarner internet '
 'customers. the company said aols underlying profits rose 8 % on the back of '
 "stronger advertising revenues. time warner's fourth quarter profits were "
 'slightly better than analysts expectations. the us securities exchange '
 'commission ( sec ) is close to concluding. timewarner also has to restate '
 "2000 and 2003 results following a probe by the sec. time warners'fourth "
 'quarter profit was 27 % to $ 284m. " our financial performance was strong, '
 'meeting or exceeding all of our full - year objectives, " chairman and chief '
 'executive richard parsons says. timewarner is to restate its accounts as '
 'part of a probe into aol by us market regrehending aol in us market. for '
 '2005, $ 3. 36bn was up 27 % from its 2003 performance. the company has '
 'already offered to pay $ 300m to settle charges. it will now book the sale '
 'of its stake in aol europe as a loss on the value of that stake. the deal is '
 'under review by the sec and will now be reviewed by the the sec.')
```

Berikut hasil summary dari file txt yang terdapat pada BBC Dataset :

```
1   pprint(data1)
```

```
('TimeWarner said fourth quarter sales rose 2% to $11.1bn from $10.9bn.For the '
 'full-year, TimeWarner posted a profit of $3.36bn, up 27% from its 2003 '
 'performance, while revenues grew 6.4% to $42.09bn.Quarterly profits at US '
 'media giant TimeWarner jumped 76% to $1.13bn (£600m) for the three months to '
 "December, from $639m year-earlier.However, the company said AOL's underlying "
 'profit before exceptional items rose 8% on the back of stronger internet '
 'advertising revenues.Its profits were buoyed by one-off gains which offset a '
 'profit dip at Warner Bros, and less users for AOL.For 2005, TimeWarner is '
 'projecting operating earnings growth of around 5%, and also expects higher '
 'revenue and wider profit margins.It lost 464,000 subscribers in the fourth '
 'quarter profits were lower than in the preceding three quarters.Time '
 "Warner's fourth quarter profits were slightly better than analysts' "
 'expectations.')
```

3. Model Abstractive Summarization using BERT and Seq2Seq

Dalam membangun model Abstractive Summarization menggunakan gabungan BERT dan Sequence-to-Sequence, dibutuhkan beberapa modul dan class. Kami menggunakan bahasa pemrograman Python dalam membangun model Abstractive Summarization ini. Modul yang kami gunakan adalah torch dan transformers. Modul torch digunakan sebagai framework deep learning untuk membuat, melatih, dan mengevaluasi model summarization. Kemudian, modul transformers digunakan untuk mengolah bahasa alami dengan cara mempelajari ketergantungan jarak jauh dalam urutan kata menggunakan neural network. Dari modul transformers, kami menggunakan dua class untuk membuat model Abstractive Summarization. Kedua class tersebut adalah AutoTokenizer dan AutoModelForSeq2SeqLM. Class AutoTokenizer berfungsi untuk menyediakan tokenisasi dari teks untuk digunakan dengan transformers. Class AutoTokenizer secara otomatis dapat menghasilkan token pada teks verdasarkan model transformator yang telah dilatih sebelumnya. Kemudian, class AutoModelForSeq2SeqLM berfungsi untuk menyediakan cara untuk menggunakan transformer untuk tugas Sequence-toSequence. AutoModelForSeq2SeqLM akan digunakan untuk menghasilkan ringkasan teks berdasarkan model transformer yang telah dilatih sebelumnya.

Untuk menghasilkan Abstractive Summarization, model akan melakukan beberapa proses sebagai berikut. Pertama model memerlukan beberapa modul seperti torch, transformers, pprint, dan rouge. Selanjutnya, dilakukan pemrosesan tokenizer menggunakan BERT dan model summarization menggunakan Sequence-to-Sequence. Tokenizer akan digunakan untuk mengubah teks menjadi foemaat token yang mampu dipahami oleh model. Kemudian model akan menggunakan token-token untuk belajar dan menghasilkan ringkasan teks. Selanjutnya, disediakan satu file yang berisi teks yang akan dibuat ringkasannya. Model akan membaca teks dalam file dan melakukan beberapa preprocessing data. Untuk menghasilkan ringkasan yang lebih baik, pisahkan data teks menjadi bagian-bagian yang terdiri dari maksimal 512 karakter. Lalu setiap bagian akan diringkas secara terpisah. Bagian teks tersebut akan di encoding menggunakan tokenizer, lalu bagian yang telah di encoding dimasukkan ke model summarization. Selanjutnya model akan menghasilkan teks ringkasan. Hasil ringkasan akan didecoding oleh tokenizer sehingga menghasilkan bagian ringkasan semi-final. Bagian-bagian ringkasan semi-final tersebut akan digabungkan menjadi satu ringkasan final. Dan terakhir, untuk menganalisis kinerja model, kami

menggunakan Rouge Metric untuk membandingkan ringkasan yang dihasilkan oleh model dengan ringkasan yang dihasilkan oleh manusia.