

AUTOMATIC PODCAST SUMMARIZATION DENGAN METODE WAV2VEC DAN BART

Zhilaan Abdurrasyid Rusmawan 1^a, Dheaz Kelvin Harahap 2^b, Azhar Dzakwan
Azizi 3^c, Reynaldo Arya Budi Trisna 4^d

^a162012233009 Teknik Robotika dan Kecerdasan Buatan, Fakultas Teknologi Maju dan Multidisiplin, Universitas Airlangga,
Surabaya

^b162012233011 Teknik Robotika dan Kecerdasan Buatan, Fakultas Teknologi Maju dan Multidisiplin, Universitas Airlangga,
Surabaya

^c162012233035 Teknik Robotika dan Kecerdasan Buatan, Fakultas Teknologi Maju dan Multidisiplin, Universitas Airlangga,
Surabaya

^d162012233087 Teknik Robotika dan Kecerdasan Buatan, Fakultas Teknologi Maju dan Multidisiplin, Universitas Airlangga,
Surabaya

Abstrak

Saat ini konten video merupakan sebuah trend yang sangat banyak diminati oleh masyarakat dari berbagai kalangan, kemudahan akses menonton menjadikan video yang beredar sangat beragam, hal tersebut dapat membuat user kesulitan mencari konten mana yang benar-benar sesuai dengan yang dicari, oleh karena itu penelitian rangkuman video podcast ini dibuat. Penelitian ini berfokus dalam bidang pemrosesan bahasa alami, khususnya pada pengembangan metode automatic podcast summarization menggunakan teknologi Wav2Vec dan BART. Penelitian ini mengatasi tantangan dalam mengolah data audio podcast menjadi ringkasan informatif, dengan mempertimbangkan keberagaman bahasa dan kebakuan bahasa, khususnya Bahasa Indonesia. Penelitian ini melibatkan analisis dan perbandingan efektivitas kedua metode ini dalam konteks podcast summarization, mengeksplorasi potensi penggunaan Wav2Vec untuk memproses sinyal suara dan BART untuk rangkuman teks. Tujuan akhirnya adalah untuk mengembangkan sistem yang lebih efisien dan akurat dalam rangkuman podcast, meningkatkan aksesibilitas dan kebermanfaatan konten podcast. Evaluasi dilakukan menggunakan metrik WER dan ROUGE untuk mengukur akurasi transkripsi dan kualitas rangkuman dan kualitas rangkuman. Hasilnya diharapkan dapat meningkatkan aksesibilitas dan kebermanfaat podcast melalui rangkuman yang efisien.

Kata Kunci : Pemrosesan Bahasa Alami, *Podcast Summarization*, Wav2vec, Model BART, Word Error Rate (WER), ROUGE, Akurasi.

1. Pendahuluan

Dalam era informasi digital yang terus berkembang, *podcast* telah menjadi salah satu media yang semakin populer untuk menyampaikan berbagai informasi, hiburan, dan wawasan kepada pendengarnya. Sebuah studi telah menunjukkan peningkatan yang signifikan dalam penggunaan dan popularitas *podcast*. Studi ini menemukan bahwa podcast dapat meningkatkan kemampuan belajar dan mendengar, dengan mencatat secara khusus dampak positif pada skor tes siswa (Nurwahidah et al., 2023). *Podcast summarization* adalah metode untuk membuat ringkasan dan mengidentifikasi informasi serta poin penting yang terkandung dalam sebuah *podcast* (Khaiqiang et al., 2022)

Kompleksitas pengolahan data podcast audio menjadi ringkasan yang informatif dan relevan merupakan tantangan utama di era informasi digital (Yubiantara et al., 2020). Ini sangat penting dalam

konteks industri 4.0, di mana penyebaran informasi melalui podcast semakin populer (Akhyar et al., 2022). Meskipun penggunaan podcast untuk pembelajaran dan pengumpulan informasi sudah luas, hambatan bahasa masih bisa menjadi tantangan (Norhayati et al., 2020). Namun, peran *podcast* dalam menyebarkan informasi, khususnya dalam program pemerintah, telah diakui dan dimanfaatkan (Santhia et al., 2022). Oleh karena itu, meskipun kompleksitas pengolahan data podcast audio menjadi ringkasan yang informatif merupakan tantangan, potensi podcast sebagai media untuk menyebarkan informasi sangat signifikan.

Dalam hal ini, Wav2Vec, yang dirancang khusus untuk memproses sinyal suara, menjadi opsi alternatif yang menarik. Kendala dan tantangan yang terkait dengan penggunaan metode ini dalam rangkuman *podcast* seperti keberagaman bahasa yang dipakai, kebakuan sebuah bahasa, terutama Bahasa Indonesia juga perlu diteliti lebih lanjut untuk memastikan efektivitas dan ketercapaian tujuan tersebut. Gondi (2022) dan Lim et al (2021) fokus pada model Wav2Vec dalam pengenalan suara. Gondi (2022) mengevaluasi kinerjanya pada perangkat tepi yang berdaya rendah dan sumber daya rendah, sementara Lim et al (2021) menyesuaikan penyematan untuk teks ke suara, keduanya dengan hasil yang positif. Secara kolektif, studi-studi ini menyarankan bahwa model Wav2Vec 2.0 dapat menjadi model yang baik untuk konversi audio ke teks. Selain itu, salah satu masalah utama yang diangkat dalam penelitian ini adalah kompleksitas dalam mengolah data audio podcast menjadi ringkasan yang informatif dan sesuai dengan maksud dari dibuatnya video itu sendiri. Model BART mampu menghasilkan ringkasan teks yang linguistik dan akurat, meskipun teks input mungkin berisi noise, kesalahan, atau hilang. BART memiliki keunggulan untuk menggenerasi teks yang benar, membuatnya fleksibel dan dapat disesuaikan untuk berbagai tugas teks generasi bawah, seperti machine translation, penangan pertanyaan, atau klasifikasi teks (Lewis et al., 2019)

Studi "*WaveNet: A Generative Model for Raw Audio*" memperkenalkan WaveNet, sebuah jaringan neural dalam yang menghasilkan gelombang audio mentah. Model ini bersifat probabilistik dan *autoregressive*, dengan distribusi prediktif untuk setiap sampel audio dikondisikan pada semua sampel sebelumnya. WaveNet dapat menangkap karakteristik berbagai pembicara dengan baik. Penelitian ini relevan dengan topik yang akan diteliti karena mengusulkan model generatif canggih untuk audio yang dapat menjadi dasar bagi sistem rangkuman podcast yang lebih alami dan efektif. (Oord et al., 2016)

Namun, metode Wav2Vec dan BART lebih unggul dari metode dalam WaveNet dalam beberapa hal. Pertama, Wav2Vec secara khusus dirancang untuk merepresentasikan suara mentah secara efektif, menangkap suara yang halus dalam audio yang mungkin terlewat oleh model yang lebih umum seperti WaveNet. Ini sangat berguna dalam konteks rangkuman *podcast*, di mana pemahaman nuansa dalam ucapan penting. Kedua, BART, sebagai model pemrosesan bahasa alami, sangat mahir dalam memahami dan menghasilkan teks yang koheren dan kontekstual. Ini berarti bahwa ketika digunakan bersama dengan Wav2Vec, sistem tersebut tidak hanya dapat memahami nuansa audio tetapi juga menghasilkan rangkuman yang lebih akurat dan mudah dibaca, memberikan pendekatan yang lebih holistik dan terintegrasi untuk merangkum podcast. (Oord et al., 2016)

Studi lain berjudul "*Abstractive Summarization of Podcast Transcripts with BART using Semantic Self-segmentation*" dari *University of Bologna* berfokus pada rangkuman *podcast* dengan menggunakan kombinasi segmentasi semantik dan BART untuk rangkuman abstraktif, mengatasi tugas kompleks untuk merangkum konten yang diucapkan secara efektif, yang sering kali panjang, tidak lancar, dan bervariasi. Ini menunjukkan peningkatan yang signifikan dibandingkan dengan *baseline*, dengan metrik

menunjukkan peningkatan hingga 30%. Metodologi melibatkan pra-pemrosesan deskripsi podcast yang menyeluruh, segmentasi semantik dari transkrip, dan penggunaan arsitektur BART. Kekuatan studi ini terletak pada pendekatannya yang komprehensif, memanfaatkan teknik NLP canggih untuk meningkatkan kualitas rangkuman secara signifikan. Namun, ketergantungan pada BART, meskipun efektif, juga membawa kompleksitas pelatihan dan mungkin membatasi adaptabilitas. Lebih lanjut, studi ini mengakui tantangan untuk secara akurat mencerminkan konten nuansa dari *podcast* (Boezio et al., 2022)

Dalam rangka mencapai tujuan penelitian ini, langkah-langkah analitis dan perbandingan akan diambil untuk mengevaluasi efektivitas masing-masing metode. Dengan mengidentifikasi kelebihan dan kekurangan dari BART dan Wav2Vec dalam konteks *podcast summarization*, penelitian ini diharapkan dapat memberikan wawasan yang mendalam untuk pengembangan sistem yang lebih efisien dan akurat. Selain itu, penelitian ini juga diarahkan untuk mengatasi tantangan khusus yang mungkin muncul dalam pengolahan konten audio, seperti pemrosesan sinyal suara dan ekstraksi informasi yang lebih akurat. Kesimpulan dari penelitian ini diharapkan dapat memberikan kontribusi yang berarti dalam pengembangan teknologi podcast summarization untuk meningkatkan aksesibilitas dan kebermanfaatan konten podcast secara keseluruhan.

2. Landasan Teori

2.1 Automatic Speech Recognition

Automatic Speech Recognition (ASR) adalah bidang penelitian dalam pengolahan bahasa alami yang bertujuan untuk mengubah sinyal suara atau ucapan manusia menjadi teks secara otomatis. ASR memiliki berbagai aplikasi, termasuk sistem transkripsi audio, asisten virtual berbasis suara, dan aplikasi lainnya yang melibatkan interaksi manusia dengan mesin melalui ucapan. Karena keinginan untuk mengotomatisasi tugas-tugas sederhana yang memerlukan interaksi antar manusia dan mesin, minat terhadap teknologi *Automatic Speech Recognition* (ASR) semakin meningkat (Fleck et al., 2022). ASR dapat didefinisikan sebagai proses penghasilan transkripsi ucapan, yang dikenal sebagai urutan kata, dengan fokus pada bentuk gelombang ucapan (Yu et al., 2016). Secara aktual, pengenalan ucapan sulit karena keragaman dalam sinyal suara (Yu et al., 2016). Saat ini, ASR secara luas diterapkan dalam berbagai fungsi, seperti laporan cuaca, penanganan panggilan otomatis, kutipan saham, dan sistem konsultasi. Penelitian-penelitian dalam bidang Automatic Speech Recognition (ASR) dipengaruhi oleh beberapa faktor kunci. Pertama, jumlah pembicara memiliki peran penting dalam pelatihan sistem, di mana dibutuhkan ucapan dari sejumlah besar pengguna untuk memastikan keberagaman dan representasi yang memadai. Selanjutnya, sifat dari ucapan juga memainkan peran krusial, di mana suara pengguna lebih mudah dikenali dalam sistem pengenalan yang terisolasi dengan mengucapkan kata-kata satu per satu dengan jeda di antaranya. Ukuran kosa kata juga menjadi faktor determinan, karena sistem pengenalan ucapan bervariasi tergantung pada jumlah kata yang dapat mereka kenali. Terakhir, lebar gelombang spektral juga mempengaruhi performa sistem ASR yang dilatih; jika lebar gelombang berkurang, kinerjanya akan memburuk, dan sebaliknya. Dengan memahami faktor-faktor ini, penelitian dan pengembangan di bidang ASR dapat lebih efektif mengatasi tantangan dalam pengenalan ucapan otomatis (Fleck et al., 2022).

2.1 Wav2Vec

Prinsip kerja dari wav2vec2 adalah membagi prosedur pelatihan menjadi dua bagian terpisah: Pertama, sebuah model ucapan diproduksi dengan jumlah data audio mentah yang sangat besar dalam proses pembelajaran mesin self-supervised. Seperti yang ditunjukkan oleh tim Meta (Baevski et al., 2020), model yang telah dilatih sebelumnya ini tidak harus dilatih dengan bahasa yang sama dengan model fine-tuned berikutnya, tetapi juga dapat dilatih menggunakan korpora multi-bahasa. Model yang telah dilatih sebelumnya kemudian digunakan untuk *fine-tuning* dengan model ucapan yang memiliki jumlah data pelatihan yang relatif sedikit yang telah dilabeli (Conneau et al., 2020). Metode ini menggunakan konsep kontrastif learning, di mana model dilatih untuk membedakan antara bagian-bagian yang relevan dari sinyal suara dan bagian yang tidak relevan. Representasi vektor yang dihasilkan oleh Wav2Vec dapat digunakan untuk tugas-tugas ASR, memungkinkan pengenalan ucapan yang lebih baik dan akurat (Baevski et al., 2020).

Arsitektur dasar dari Wav2vec 2.0 terdiri dari tiga jaringan: pemberi fitur (feature encoder), transformer kontekstual, dan modul kuantisasi (Conneau et al., 2020). Pemberi fitur, yang terdiri dari jaringan saraf konvolusional (CNN) berlapis ganda, mengodekan audio mentah X dan menghasilkan representasi ucapan laten Z . Transformer kontekstual, yang merupakan rangkaian encoder transformer, mempelajari representasi konteks C dengan mengambil representasi ucapan laten sebagai input. Modul kuantisasi digunakan untuk memetakan representasi laten ke dalam ruang terdisritisasi Q , memilih entri buku kode diskrit secara sepenuhnya dapat dibedakan. Dalam pra-pelatihan, sebagian representasi laten diacak sebelum dimasukkan ke dalam transformer kontekstual. Model dilatih dengan menyelesaikan tugas kontrastif dengan representasi tersembunyi yang diacak, membedakan vektor laten terkuantisasi sejati dari vektor laten diskrit yang telah diambil acak dari langkah waktu tersembunyi lainnya. Selama pra-pelatihan, model mempelajari representasi terkontekstualisasi hanya dari data audio ucapan yang tidak berlabel (Riviere et al., 2020)

2.2 Text Summarization

Text summarization adalah bidang dalam pemrosesan bahasa alami yang bertujuan untuk membuat ringkasan atau rangkuman dari dokumen teks dengan mempertahankan informasi kunci dan esensial. Tugas ini dapat diterapkan pada berbagai jenis dokumen, termasuk artikel berita, makalah riset, atau konten web (Widyassari et al., 2022). Pendekatan utama dalam *text summarization* terbagi menjadi dua yaitu ekstraktif dan abstraktif. Pendekatan ekstraktif mencoba untuk mengekstrak kalimat-kalimat atau frasa-frasa yang paling penting dari dokumen sumber, sedangkan pendekatan abstraktif menciptakan ringkasan baru dengan menggunakan bahasa yang berbeda dari dokumen sumber, menjelaskan inti informasi dengan cara yang lebih umum atau ringkas. Pada metode ekstraksi, cara ini menggunakan teknik-teknik seperti analisis frekuensi kata, penghitungan bobot kata (TF-IDF), dan model pembelajaran mesin sederhana untuk menentukan kalimat-kalimat yang paling penting dalam dokumen. Algoritma seperti Textrank dan Graph-based ranking sering digunakan untuk mengekstrak kalimat-kalimat kunci. Sedangkan dengan metode abstraksi, metode bekerja dengan melibatkan pembangkitan kalimat baru yang mencerminkan inti informasi dari dokumen sumber dengan menggunakan teknik-teknik seperti model bahasa berbasis transformer, seperti GPT (Generative Pre-trained Transformer) dan BART (Bidirectional and Auto-Regressive Transformers) (Yadav et al., 2020).

2.3 Bidirectional and Auto-Regressive Transformer (BART)

BART adalah model *sequence-to-sequence* yang telah dilatih sebelumnya dan menunjukkan hasil yang menjanjikan dalam tugas ringkasan teks. ra-pelatihan terdiri dari dua tahap: (1) teks dikorupsi menggunakan fungsi *noise*, dan (2) model *sequence-to-sequence* dipelajari untuk merekonstruksi teks asli. Arsitektur terjemahan mesin berbasis Transformer BART dapat dilihat sebagai generalisasi dari BERT (karena adanya *bidirectional encoder*), GPT (dengan decoder dari kiri ke kanan), dan banyak pendekatan *pre-training* kontemporer lainnya (Yadav et al., 2020). Selain keunggulannya dalam tugas pemahaman, efektifitas BART meningkat dengan fine-tuning untuk generasi teks. BART menghasilkan hasil terbaik baru pada berbagai tugas ringkasan percakapan, QnA, dan tugas ringkasan, sejajar dengan kinerja RoBERTa dengan sumber daya pelatihan yang sebanding pada GLUE dan SQuA (Payong, 2023).

Arsitektur BART mengikuti desain umum *sequence-to-sequence* Transformer (Raval, 2023), dengan enam lapisan dalam encoder dan decoder untuk model dasar, dan dua belas lapisan masing-masing untuk model besar. Perbedaan dengan arsitektur BERT terletak pada tidak adanya lapisan decoder tambahan yang melakukan cross-attention pada lapisan tersembunyi akhir encoder (seperti dalam model *sequence-to-sequence* Transformer); dan tidak adanya jaringan feed-forward tambahan yang digunakan sebelum prediksi kata. Penelitian terdahulu mendiskusikan beberapa transformasi mereka gunakan melibatkan penyamaran token, penghapusan token, pengisian token dalam teks, permutasi kalimat, dan rotasi dokumen (Payong, 2023).

3. Sumber Data dan Metodologi

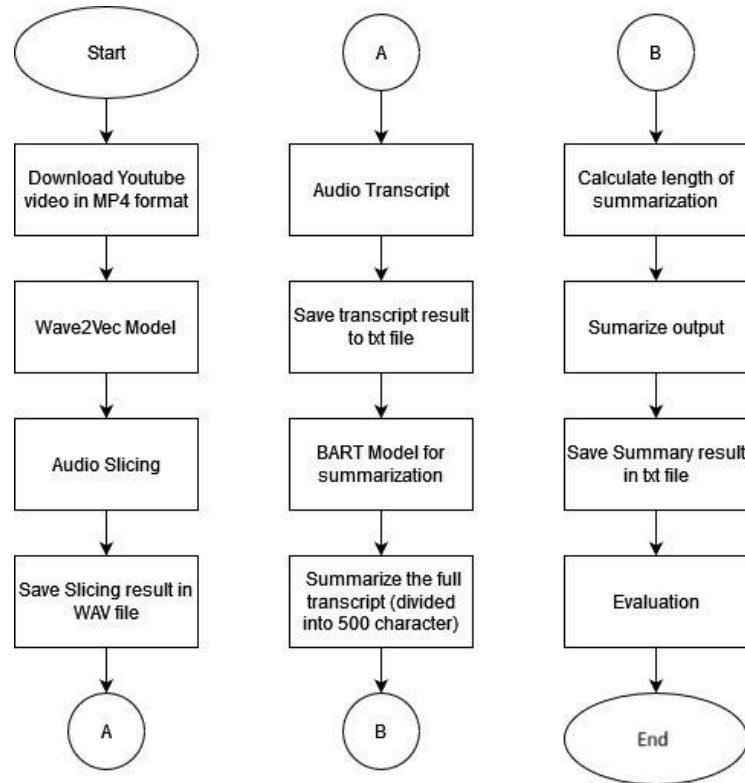
3.1 Sumber Data

Data yang digunakan pada penelitian ini merupakan sebuah video bersumber dari kanal YouTube.com dengan detail sebagai berikut :

Judul Video	: Pidato Presiden Joko Widodo Pada Pembukaan KTT ke-3 BRF, Beijing, 18 Oktober 2023
Tipe File	: MP4
Author	: Sekretariat Presiden
Tanggal Upload	: 18 Oktober 2023
Durasi	: 3 Menit 54 Detik
Link Video	: https://www.youtube.com/watch?v=2j8CqOCmEDM

Data MP4 tersebut nantinya akan diolah menjadi data audio dengan format .wav menggunakan library dari python bernama pytube yang dirancang khusus untuk interaksi dengan YouTube. Lalu, dipotong per-30 detik audio, sehingga menghasilkan 8 parsial video berdurasi 30 detik.

3.2 Metodologi



Penelitian diawali dengan pengumpulan data podcast dari platform satu sumber sebagai percobaan, Data podcast ini melibatkan transkripsi audio yang akan menjadi dasar untuk proses *podcast summarization*. Setelah pengumpulan data, dilakukan tahap pra-pemrosesan untuk membersihkan dan merapikan transkripsi, serta memecahnya menjadi segmen-segmen yang lebih kecil untuk memudahkan proses selanjutnya.

Langkah berikutnya adalah mengimplementasikan model Wav2Vec untuk ekstraksi fitur suara dari transkripsi podcast. Model ini dilatih menggunakan *pre-trained* model dari sumber [10] untuk menyesuaikan kebutuhan data berbasis bahasa Indonesia, selanjutnya hasil dari proses ini nantinya akan digunakan sebagai masukan untuk model BART.

Selanjutnya, model BART diaplikasikan untuk merangkum teks dari transkripsi podcast yang telah diolah oleh Wav2Vec. Fine-tuning dilakukan menggunakan data podcast yang telah dilabeli, diambil dari *pre-trained* model yang disediakan oleh sumber [11] untuk menyesuaikan kebutuhan dan corpus berbasis bahasa Indonesia yang memungkinkan BART untuk memahami konten podcast dengan lebih baik dan menghasilkan ringkasan yang informatif. Proses ini dilakukan untuk mencapai tingkat akurasi dan keberlanjutan optimal dalam menangkap esensi dari podcast.

Evaluasi kinerja model dilakukan dengan menggunakan matrik evaluasi ringkasan seperti ROUGE.

Metrik-metrik ini membantu mengukur sejauh mana ringkasan yang dihasilkan mencerminkan informasi kunci dari transkripsi podcast asli. Perbandingan hasil summary antara hasil keluaran oleh BART dengan keluaran menggunakan pipeline API dari google juga dilakukan untuk mengetahui Hasil eksperimennya. Hasil transkrip dievaluasi dan dibandingkan dengan hasil transkrip yang asli.

4. Analisis dan Pembahasan

4.1.1 Konversi Audio

Setelah file audio berhasil diunduh dalam format MP4, langkah selanjutnya adalah mengubah format file tersebut menjadi WAV. Konversi ini dilakukan menggunakan alat ffmpeg, yang merupakan perangkat lunak open-source sangat populer untuk memproses audio dan video. Alat ini memungkinkan konversi format file dengan cara yang efisien dan efektif.

Format WAV dipilih karena kecocokannya dengan kebutuhan transkripsi audio. Berbeda dengan format MP4 yang merupakan format kontainer dan biasanya mengandung kompresi data, format WAV menyajikan data audio dalam bentuk yang tidak terkompresi dan lebih murni, yang ideal untuk analisis dan pengolahan audio lebih lanjut.

Perintah ffmpeg yang digunakan dalam script melakukan konversi dengan mengatur codec audio (-acodec) menjadi pcm_s16le, yang menentukan format penyimpanan data audio di file WAV, dan menetapkan sample rate (-ar) menjadi 16000 Hz. Sample rate ini penting karena menentukan seberapa sering sampel audio diambil per detik dan mempengaruhi kualitas serta ukuran file hasil konversi.

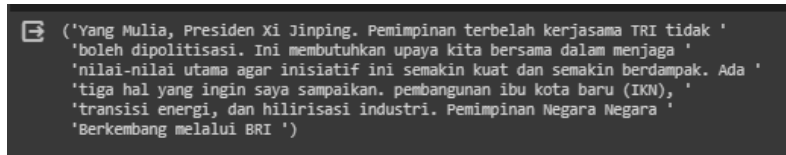
4.1.2 Transkripsi Audio menggunakan BERT (Wav2Vec) Model

Wav2Vec merupakan terobosan dalam bidang pemrosesan bahasa alami dan pengenalan suara. Dikembangkan dengan fokus pada pengolahan audio, model ini efektif dalam menangkap ciri-ciri unik dari gelombang suara dan mengubahnya menjadi teks. Berbeda dengan model tradisional yang bergantung pada data teks berlabel yang memakan waktu dan sumber daya untuk dibuat, Wav2Vec dirancang untuk mempelajari representasi audio yang berguna dari data suara yang tidak berlabel. Ini memungkinkan Wav2Vec untuk mengenali pola dalam data audio dengan lebih efisien dan menghasilkan transkripsi yang lebih akurat.

Model Wav2Vec menggunakan pendekatan self-supervised learning, di mana sebagian besar pembelajaran dilakukan pada data yang tidak berlabel, dengan hanya sebagian kecil dari data berlabel yang digunakan untuk menyetel akhir model. Pendekatan ini membuat Wav2Vec tidak hanya canggih dalam hal teknologi tetapi juga efisien dalam hal waktu dan biaya pelatihan. Kemampuannya dalam mengenali berbagai aksen dan dialek menjadikannya alat yang sangat berguna dalam aplikasi pengenalan suara, terutama dalam konteks bahasa dan variasi ucapan.

Pada tahap berikut adalah mengonversi konten audio tersebut menjadi teks. Ini dilakukan menggunakan perpustakaan Python huggingsound, yang merupakan bagian dari ekosistem Hugging Face dan menyediakan model-model untuk pengenalan suara. Dalam tahap ini, digunakan model "cahya/whisper-medium-id" dari huggingsound, yang dirancang khusus

untuk mengenali dan transkripsi bahasa Indonesia. Proses ini diawali dengan memuat file audio WAV yang telah dikonversi. Menggunakan librosa, sebuah perpustakaan pemrosesan audio di Python, script ini membagi audio menjadi blok-blok atau potongan berdurasi 60 detik. Pembagian ini penting untuk memastikan bahwa setiap segmen audio dapat diproses secara efisien oleh model transkripsi, mengingat keterbatasan kapasitas memori dan prosesor. Hasil akhir dari proses ini adalah sebuah file teks yang berisi transkripsi lengkap dari audio yang diunduh seperti gambar 4.1 berikut :

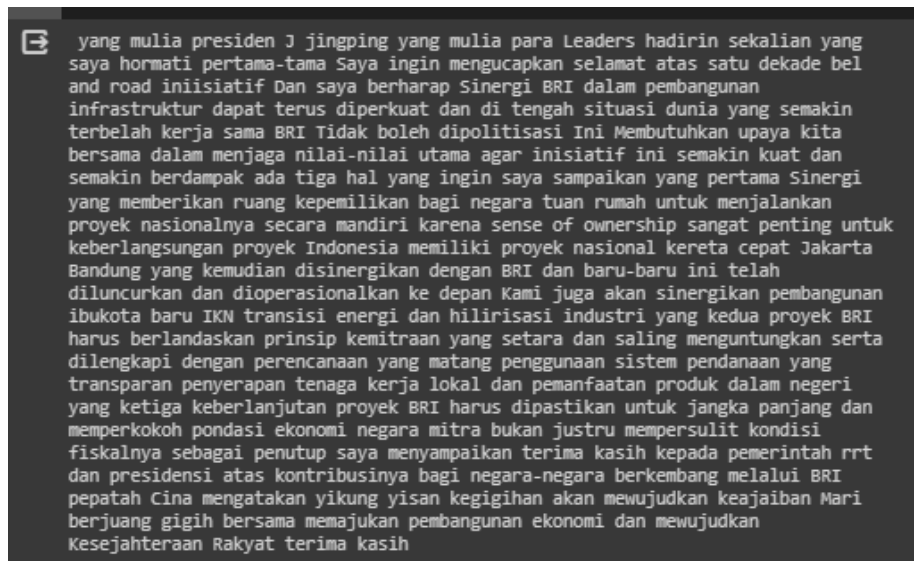


```
( 'Yang Mulia, Presiden Xi Jinping. Pimpinan terbelah kerjasama TRI tidak '
'boleh dipolitisasi. Ini membutuhkan upaya kita bersama dalam menjaga '
'nilai-nilai utama agar inisiatif ini semakin kuat dan semakin berdampak. Ada '
'tiga hal yang ingin saya sampaikan, pembangunan ibu kota baru (IKN), '
'transisi energi, dan hilirisasi industri. Pimpinan Negara Negara '
'Berkembang melalui BRI ' )
```

Gambar 4.1 Hasil Transkripsi Audio dengan Wav2Vec (BERT)

4.1.3 Transkripsi Audio menggunakan Google Cloud API

Proses berlanjut dengan metode Google Cloud Speech-to-Text API. Metode ini memungkinkan transkripsi audio yang lebih akurat dan efisien untuk bahasa yang beragam, termasuk Bahasa Indonesia. Google Cloud API menganalisis audio WAV dan menghasilkan transkripsi teks yang lebih akurat, memanfaatkan kecanggihan cloud computing. Dengan integrasi Google Cloud API, proses transkripsi diperkuat dengan teknologi pengenalan suara yang lebih canggih, memberikan hasil transkripsi yang lebih baik.



```
yang mulia presiden J jingping yang mulia para Leaders hadirin sekalian yang
saya hormati pertama-tama Saya ingin mengucapkan selamat atas satu dekade bel
and road inisiatif Dan saya berharap Sinergi BRI dalam pembangunan
infrastruktur dapat terus diperkuat dan di tengah situasi dunia yang semakin
terbelah kerja sama BRI Tidak boleh dipolitisasi Ini Membutuhkan upaya kita
bersama dalam menjaga nilai-nilai utama agar inisiatif ini semakin kuat dan
semakin berdampak ada tiga hal yang ingin saya sampaikan yang pertama Sinergi
yang memberikan ruang kepemilikan bagi negara tuan rumah untuk menjalankan
proyek nasionalnya secara mandiri karena sense of ownership sangat penting untuk
keberlangsungan proyek Indonesia memiliki proyek nasional kereta cepat Jakarta
Bandung yang kemudian disinergikan dengan BRI dan baru-baru ini telah
diluncurkan dan dioperasionalkan ke depan Kami juga akan sinergikan pembangunan
ibukota baru IKN transisi energi dan hilirisasi industri yang kedua proyek BRI
harus berlandaskan prinsip kemitraan yang setara dan saling menguntungkan serta
dilengkapi dengan perencanaan yang matang penggunaan sistem pendanaan yang
transparan penyerapan tenaga kerja lokal dan pemanfaatan produk dalam negeri
yang ketiga keberlanjutan proyek BRI harus dipastikan untuk jangka panjang dan
memperkokoh pondasi ekonomi negara mitra bukan justru mempersulit kondisi
fiskalnya sebagai penutup saya menyampaikan terima kasih kepada pemerintah rrt
dan presidensi atas kontribusinya bagi negara-negara berkembang melalui BRI
pepatah Cina mengatakan yikung yisan kegigihan akan mewujudkan keajaiban Mari
berjuang gigih bersama memajukan pembangunan ekonomi dan mewujudkan
Kesejahteraan Rakyat terima kasih
```

Gambar 4.2 Hasil Transkripsi Audio dengan Google Cloud API

4.1.4 Summarization menggunakan BART Model

Model BART (*Bidirectional and Auto-Regressive Transformers*) telah menjadi populer dalam tugas rangkuman teks di bidang pemrosesan bahasa alami (NLP). Kelebihan utama BART

terletak pada kemampuannya untuk menggabungkan pemahaman kontekstual yang mendalam dengan kemampuan generatif. Ini memungkinkan BART untuk memahami nuansa teks asli secara efektif dan menghasilkan rangkuman yang koheren dan kontekstual. BART juga unggul dalam berbagai tugas NLP lainnya, termasuk terjemahan dan generasi teks, berkat pelatihannya yang melibatkan teks yang sengaja 'dirusak' atau dimanipulasi. Kemampuan ini membuatnya tangguh dalam memperbaiki kesalahan dan mengisi kekosongan informasi.

Namun, BART juga menghadapi beberapa *challenge*. Sebagai model yang besar dan kompleks, BART memerlukan sumber daya komputasi yang signifikan, yang bisa menjadi penghalang bagi pengguna dengan keterbatasan sumber daya. Selain itu, BART sangat bergantung pada kualitas data pelatihan; rangkuman yang dihasilkan mungkin terpengaruh jika data pelatihan tidak representatif atau bias. Model ini juga rentan terhadap overfitting, terutama jika tidak dilatih dengan dataset yang cukup beragam. Terakhir, BART mungkin memerlukan fine-tuning yang cermat untuk tugas rangkuman tertentu, membutuhkan keahlian dan pemahaman mendalam tentang model dan data.

Pada tahap summarization ini, adalah proses merangkum teks transkripsi yang telah didapat dari audio YouTube. Proses ini menggunakan model BART dari *library transformers*, yang khusus dirancang untuk tugas rangkuman teks. Model BART, singkatan dari Bidirectional and Auto-Regressive Transformers, merupakan model canggih yang dapat mengolah dan menyederhanakan teks panjang menjadi versi yang lebih ringkas tanpa kehilangan esensi penting.

Proses rangkuman dimulai dengan membaca teks transkripsi dari file yang telah disimpan sebelumnya. Mengingat model BART memiliki batasan jumlah token yang dapat diolah dalam satu sesi, skrip ini membagi teks transkripsi menjadi segmen-segmen yang lebih kecil menggunakan fungsi `chunk_text`. Fungsi ini memecah teks berdasarkan jumlah karakter, memastikan setiap segmen tidak melebihi batas token yang ditentukan (1024 token).

Setelah teks dibagi menjadi segmen-segmen yang lebih kecil, setiap segmen dirangkum secara terpisah. Model BART memproses setiap segmen, memahami konteks dan isi utamanya, dan menghasilkan versi yang lebih singkat dan padat. Rangkuman ini tetap mempertahankan informasi kunci dari teks asli.

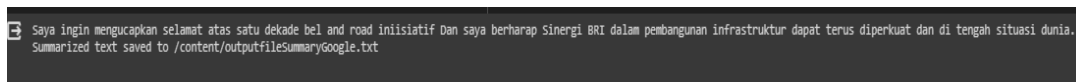
Selanjutnya, semua rangkuman segmen tersebut digabungkan untuk membentuk rangkuman akhir dari seluruh teks transkripsi. Hasilnya adalah teks yang jauh lebih ringkas dibandingkan transkripsi asli tetapi masih memuat poin-poin penting dari keseluruhan konten. Seperti hasil yang dapat dilihat pada gambar 4.2 berikut.

```
Pemimpin terbelah kerjasama TRI tidak boleh dipolitisasi. Ini membutuhkan upaya kita bersama dalam menjaga nilai-nilai utama. Ada tiga hal yang ingin saya sampaikan.

Summarized text saved to /content/summarized_transcript.txt
```

Gambar 4.2 Hasil Summarization Transcript Text BERT Model

Pada Proses Selanjutnya dilakukan perbandingan text rangkuman yang dihasilkan sebelumnya pada Gambar 4.2 dengan hasil text transcript yang menggunakan Google Cloud API Speech To Text. Hasil dari text transcript dengan metode tersebut dapat dilihat pada gambar 4.3 berikut.



Gambar 4.3 Hasil Summarization Transcript Text Google Cloud API Text to Speech

4.1.5 Evaluasi *Transcription Text*

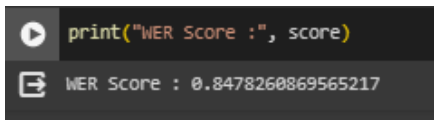
Evaluasi akurasi transkripsi bertujuan untuk menentukan seberapa akurat proses transkripsi audio ke teks telah berlangsung. Akurasi ini diukur menggunakan metrik Word Error Rate (WER).

WER adalah metrik umum untuk menilai kinerja sistem pengenalan ucapan. WER mengukur persentase kata yang salah dalam transkripsi dibandingkan dengan transkripsi referensi (dianggap benar). Ini dihitung sebagai jumlah kesalahan pengeditan (penyisipan, penghapusan, dan penggantian) dibagi dengan jumlah kata dalam transkripsi referensi.

Pada tahapan ini menggunakan WER untuk membandingkan hasil transkripsi yang dihasilkan oleh model pengenalan ucapan dengan transkripsi referensi (misalnya, transkripsi manual atau transkripsi yang dihasilkan oleh sistem lain). Skor WER memberikan gambaran tentang seberapa sering kesalahan terjadi dalam transkripsi dan menjadi indikator kualitas transkripsi. Pada kasus kami, WER diukur dari perbedaan antara teks transkripsi yang dihasilkan dengan transkripsi yang diambil dari Google Cloud API Text to Speech (referensi) dengan metode BERT yang kami gunakan (sistem). Hasil dari Perbandingan text yang dihasilkan mungkin dapat dilihat dari tabel 1. berikut.

Tabel 1. Perbandingan hasil Text dan Score WER antara Referensi dan Sistem

Text Transcript Wav2Vec (Sistem)	Text Transcript Google Cloud API (REF)
Yang Mulia, Presiden Xi Jinping. Pemimpinan terbelah kerjasama TRI tidak boleh dipolitisasi. Ini membutuhkan upaya kita bersama dalam menjaga nilai-nilai utama agar inisiatif ini semakin kuat dan semakin berdampak. Ada tiga hal yang ingin saya sampaikan. pembangunan ibu kota baru (IKN), transisi energi, dan hilirisasi industri. Pemimpinan Negara Negara Berkembang	yang mulia presiden J jingping yang mulia para Leaders hadirin sekalian yang saya hormati pertama-tama Saya ingin mengucapkan selamat atas satu dekade bel and road inisiatif Dan saya berharap Sinergi BRI dalam pembangunan infrastruktur dapat terus diperkuat dan di tengah situasi dunia yang semakin terbelah kerja sama BRI Tidak boleh dipolitisasi Ini Membutuhkan upaya

melalui BRI	<p>kita bersama dalam menjaga nilai-nilai utama agar inisiatif ini semakin kuat dan semakin berdampak ada tiga hal yang ingin saya sampaikan yang pertama Sinergi yang memberikan ruang kepemilikan bagi negara tuan rumah untuk menjalankan proyek nasionalnya secara mandiri karena sense of ownership sangat penting untuk keberlangsungan proyek Indonesia memiliki proyek nasional kereta cepat Jakarta Bandung yang kemudian disinergikan dengan BRI dan baru-baru ini telah diluncurkan dan dioperasionalkan ke depan Kami juga akan sinergikan pembangunan ibukota baru IKN transisi energi dan hilirisasi industri yang kedua proyek BRI harus berlandaskan prinsip kemitraan yang setara dan saling menguntungkan serta dilengkapi dengan perencanaan yang matang penggunaan sistem pendanaan yang transparan penyerapan tenaga kerja lokal dan pemanfaatan produk dalam negeri yang ketiga keberlanjutan proyek BRI harus dipastikan untuk jangka panjang dan memperkokoh pondasi ekonomi negara mitra bukan justru mempersulit kondisi fiskalnya sebagai penutup saya menyampaikan terima kasih kepada pemerintah rrt dan presidensi atas kontribusinya bagi negara-negara berkembang melalui BRI pepatah Cina mengatakan yikung yisan kegigihan akan mewujudkan keajaiban Mari berjuang gigih bersama memajukan pembangunan ekonomi dan mewujudkan Kesejahteraan Rakyat terima kasih</p>
<p>WER SCORE :</p> <div data-bbox="562 1606 995 1725">  <pre>print("WER Score :", score) WER Score : 0.8478260869565217</pre> </div>	

Word Error Rate (WER) adalah metrik yang digunakan untuk mengevaluasi akurasi transkripsi dalam sistem pengenalan suara. WER menghitung proporsi kesalahan yang dibuat oleh sistem saat mengubah ucapan menjadi teks. Ini dilakukan dengan membandingkan transkripsi yang dihasilkan sistem dengan transkripsi referensi yang dianggap benar. Pada hal ini metode BERT yang digunakan dibandingkan dengan referensi transkripsi yang benar yaitu Google Cloud API *Speech To Text*, dengan nilai Word Error Rate 0.847 menunjukkan bahwa ada tingkat kesalahan yang cukup tinggi dalam transkripsi. Ini berarti bahwa hampir 85% dari kata-kata dalam transkripsi mengalami kesalahan.

4.1.6 Evaluasi *Summarization Text* Model BART

Setelah proses rangkuman teks selesai, langkah selanjutnya adalah mengevaluasi kualitas rangkuman tersebut. Evaluasi ini dilakukan menggunakan metrik ROUGE (Recall-Oriented Understudy for Gisting Evaluation). ROUGE adalah standar dalam evaluasi rangkuman teks otomatis dan memberikan skor berdasarkan seberapa baik rangkuman mencerminkan isi dan inti dari teks asli.

Metrik ROUGE digunakan untuk mengukur kesamaan antara rangkuman teks dan teks aslinya dengan membandingkan elemen-elemen seperti unigram (ROUGE-1), bigram (ROUGE-2), dan urutan kata terpanjang yang cocok (ROUGE-L).

Menghitung skor ROUGE antara teks transkripsi asli dan rangkuman yang dihasilkan oleh model BART. Skor ini memberikan indikasi numerik tentang kualitas rangkuman, meliputi aspek recall (seberapa banyak informasi penting dari teks asli yang tercakup dalam rangkuman) dan precision (seberapa relevan informasi dalam rangkuman terhadap teks asli). Hasil dapat dilihat dari tabel 2. berikut.

Tabel 2. Evaluasi Summary dengan ROUGE

	rouge-1	rouge-2	rouge-L
recall	1.000000	0.954545	1.000000
precision	0.479167	0.411765	0.479167
f1-score	0.647887	0.575342	0.647887

ROUGE-1: Recall (r) 1.000000: Ini menunjukkan bahwa 100% dari unigram (kata individu) dalam teks referensi terdapat dalam teks yang dirangkum. Precision (p) 0.479167: Menunjukkan bahwa 47.91% dari unigram dalam teks yang dirangkum juga terdapat dalam teks referensi. Skor F1 (f) 0.693878: Skor F1 adalah rata-rata dari recall dan precision, menunjukkan keseimbangan yang baik antara kedua aspek tersebut.

ROUGE-2: Recall (r) 0.954545: Menunjukkan bahwa 95.45% dari bigram (pasangan kata) dalam teks referensi ada dalam teks yang dirangkum. Precision (p) 0.411765: 41.17% dari bigram dalam teks yang dirangkum ditemukan dalam teks referensi. Skor F1 (f) 0.575342: Skor F1 yang lebih rendah untuk ROUGE-2 dibandingkan dengan ROUGE-1 menunjukkan bahwa kecocokan bigram kurang baik dibandingkan dengan kecocokan unigram.

ROUGE-L (*Longest Common Subsequence*): Recall dan Precision (r dan p) 1.000000 dan 0.479167: Menunjukkan tingkat kecocokan yang tinggi untuk urutan kata terpanjang yang sama antara teks yang dirangkum dan referensi. Skor F1 (f) 0.647887. Hal ini mengindikasikan keseimbangan yang baik antara recall dan precision untuk urutan kata terpanjang yang sama.

5. Kesimpulan

Penelitian kami tentang Pemrosesan Bahasa Alami, khususnya pengembangan metode untuk merangkum podcast secara otomatis menggunakan teknologi Wav2Vec dan BART. Penelitian ini bertujuan mengatasi tantangan dalam mengolah data audio podcast menjadi ringkasan yang informatif, dengan fokus pada keberagaman bahasa dan kebakuan Bahasa Indonesia. Metode ini melibatkan analisis efektivitas kedua teknologi tersebut dalam rangkuman podcast, menggunakan metrik evaluasi seperti WER dan ROUGE untuk mengukur akurasi transkripsi dan kualitas rangkuman. Hasil penelitian diharapkan meningkatkan aksesibilitas dan kebermanfaatan konten podcast melalui rangkuman yang efisien.

Future work kami adalah Emotion and Tone Recognition bertujuan membuat komputer bisa mengerti perasaan kita lebih baik lagi. Komputer nanti tidak hanya tahu saat kita sedih atau senang, tapi juga bisa mengerti perasaan rumit lainnya dan situasi di sekitarnya. Untuk Emotion and Tone Recognition, akan menggunakan teknologi canggih dan belajar lebih dalam tentang cara manusia menunjukkan perasaan.

Di masa yang akan datang, sistem ini diharapkan bisa bekerja lebih cepat dan tepat. Kami ingin membuat agar komputer bisa memperhatikan hal-hal seperti ekspresi wajah kita, gerakan tubuh, bahkan detak jantung untuk mengerti perasaan kita dengan lebih lengkap. Mereka juga akan memastikan bahwa semua informasi pribadi kita aman. Dengan semua peningkatan ini, diharapkan komputer bisa lebih baik lagi dalam membantu kita, seperti dalam belajar, pengobatan, atau bahkan dalam berbelanja.

Daftar Pustaka

- [1] Fleck, M. & Göderle, W., 2022, *wav2vec and its current potential to Automatic Speech Recognition in German for the usage in Digital History*, arXiv:2303.06026.
- [2] Yu, Dong, & Li Deng. 2016, *Automatic Speech Recognition*. Springer London limited.
- [3] Baevski, A., Henry, Z., Mohamed, A., & Michael, A., 2020 *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*, arXiv:2006.11477.
- [4] Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M., 2020, *Unsupervised Cross-Lingual Representation Learning For Speech Recognition.*, arXiv:2006.13979.
- [5] Riviere, M., Joulin, A., Mazare, P.-E., and Dupoux, 2020, E. *Un-Supervised Pretraining Transfers Well Across Languages*. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
- [6] Widyassari, A., Rustad, S., Shidik, G., Noersasongko, E., Syukur, A., Affandy, Setiadi, D., 2022, *Review of automatic text summarization techniques & methods*, Journal of King Saud University - Computer and Information Sciences, Volume 34, Issue 4.
- [7] Yadav, D., Desai, J., Yadav, A., 2020, *Automatic Text Summarization Methods: A Comprehensive Review*.
- [8] Payong, A., 2023, *BART Model for Text Summarization*, paperspace.com/bart-model-for-text-summarization-part1/ (diakses: 16 Desember, 2023).
- (Raval, 2023) Raval, P., 2023, *Transformers BART Model Explained for Text Summarization*, www.projectpro.io/article/transformers-bart-model-explained/553, (diakses: 16 Desember, 2023)
- [10] Nurwahdaniah, Ari Prasetyaningrum, & Fathurrohman. (2023). Using Podcasts to Enhance Students' Listening Ability . *Jurnal LENTERA: Jurnal Studi Pendidikan*, 5(2), 79-90. <https://doi.org/10.51518/lentera.v5i2.134>
- [11] S. Kaiqiang , L. Chen, W. Xiaoyang, Y. Dong and L. Fei , "Towards Abstractive Grounded Summarization of Podcast Transcripts," *Association for Computational Linguistics*, vol. 1, p. 4407, 2022.
- [12] Yubiantara, M. I., & Retnasary, M. (2020). "Podcast: Media Baru Pemenuhan Kebutuhan Informasi di Era Disruptif." , 2(1). Universitas Adhirajasa Reswara Sanjaya.
- [13] Akhyar, H., Mariyani, D., Rahayu, S., & Ali, M. (2022). DISEMINASI PENERAPAN TEKNOLOGI AUDIO ON DEMAND MELALUI PODCAST SEBAGAI MEDIA INFORMASI DI ERA INDUSTRI 4.0. *Jurnal Abdi Insani*, 9(3), 800-809. <https://doi.org/10.29303/abdiinsani.v9i3.633>
- [14] Norhayati, N., & Jayanti, S. (2020). Pemanfaatan Teknologi untuk Mendukung Kegiatan Belajar Secara Mandiri (Studi Kasus: Penggunaan Podcast oleh Mahasiswa di Kota Palangkaraya). *Jurnal Humaniora Teknologi*, 6(1), 29–36. <https://doi.org/10.34128/jht.v6i1.73>
- [15] Santhia, B. A., & Soedarsono, D. K. (2022). Peran Podcast Sebagai Media Penyebaran Informasi Program Kerja Dinas Komunikasi Dan Informatika Kota Bandung. *Medialog: Jurnal Ilmu Komunikasi*, 5(2). <https://doi.org/10.35326/medialog.v5i2.1840>
- [16] Gondi, S. (2022). Wav2Vec2.0 on the Edge: Performance Evaluation. *ArXiv*, abs/2202.05993.
- [17] Y. Lim, N. Kim, S. Yun, S. Kim and S. -I. Lee, "A Preliminary Study on Wav2Vec 2.0 Embeddings for Text-to-Speech," 2021 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Korea, Republic of, 2021, pp. 343-347, doi: 10.1109/ICTC52510.2021.9621175.

- [18] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation Translation and Comprehension. Retrieved from arXiv:1910.13461v1.
- [19] Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. Retrieved from arXiv:1609.03499v2
- [20] Boezio, G., Montali, S., & Murro, G. (2022). Abstractive Summarization of Podcast Transcripts with BART using Semantic Self-segmentation. Natural Language Processing. Alma Mater Studiorum - University of Bologna