

**PENERJEMAHAN BAHASA NON FORMAL BAHASA INDONESIA KE  
BAHASA FORMAL BAHASA INDONESIA DENGAN MENGGUNAKAN  
PENDEKATAN *SEMI SUPERVISED TRANSLATION***

**SKRIPSI**

**ACHMAD YUSUF BARMAWI**

**181401119**



**PROGRAM STUDI S-1 ILMU KOMPUTER  
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI  
UNIVERSITAS SUMATERA UTARA  
MEDAN  
2024**

**PENERJEMAHAN BAHASA NON FORMAL BAHASA INDONESIA KE  
BAHASA FORMAL BAHASA INDONESIA DENGAN MENGGUNAKAN  
PENDEKATAN *SEMI SUPERVISED TRANSLATION***

**SKRIPSI**

**Diajukan untuk melengkapi tugas dan memenuhi syarat memperoleh ijazah  
Sarjana Ilmu Komputer**

**ACHMAD YUSUF BARMAWI**

**181401119**



**PROGRAM STUDI S-1 ILMU KOMPUTER  
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI  
UNIVERSITAS SUMATERA UTARA  
MEDAN  
2024**

**PERSETUJUAN**

Judul : PENERJEMAHAN BAHASA NON FORMAL  
BAHASA INDONESIA KE BAHASA FORMAL  
BAHASA INDONESIA DENGAN  
MENGUNAKAN PENDEKATAN *SEMI  
SUPERVISED TRANSLATION*

Kategori : SKRIPSI

Nama : ACHMAD YUSUF BARMAWI

Nomor Induk Mahasiswa : 181401119


Program Studi : SARJANA (S1) ILMU KOMPUTER


Fakultas : ILMU KOMPUTER DAN TEKNOLOGI  
INFORMASI UNIVERSITAS SUMATERA  
UTARA

Komisi Pembimbing :

Pembimbing II

Pembimbing I

  
Sri Melvani Hardi S.Kom., M.Kom  
NIP. 198805012015042006

  
Dr. Amalia ST., M.T  
NIP. 197812212014042001

Diketahui/disetujui oleh

Program Studi S1 Ilmu Komputer

  
Ketua  
  
Dr. Amalia ST., M.T.  
NIP. 197812212014042001

**PERNYATAAN****PENERJEMAHAN BAHASA NON FORMAL BAHASA INDONESIA KE  
BAHASA FORMAL BAHASA INDONESIA DENGAN MENGGUNAKAN  
PENDEKATAN *SEMI SUPERVISED TRANSLATION*****SKRIPSI**

Saya mengakui bahwa skripsi ini adalah hasil karya saya sendiri, kecuali beberapa kutipan dan ringkasan yang masing-masing telah disebutkan sumbernya.

Medan, 5 Januari 2024

A handwritten signature in black ink, appearing to read 'Achmad Yusuf Barmawi', with a horizontal line drawn underneath it.

Achmad Yusuf Barmawi

181401119

## PENGHARGAAN

Penulis ingin menyampaikan puji syukur kepada Allah SWT, karena dengan rahmat dan kuasa-Nya, penulis berhasil menyelesaikan penyusunan skripsi ini sebagai syarat untuk memperoleh gelar sarjana di Program Studi S-1 Ilmu Komputer, Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Sumatera Utara. Dengan rasa syukur yang mendalam, penulis ingin menyampaikan penghargaan dan terima kasih yang tulus kepada semua pihak yang turut membantu dalam proses pembuatan dan penyelesaian skripsi ini, baik melalui doa, bimbingan, kerjasama, dukungan, maupun kata-kata penyemangat. Penulis mengucapkan terima kasih kepada:

1. Bapak Dr. Muryanto Amin, S.Sos, M.Si. selaku Rektor Universitas Sumatera Utara.
2. Ibu Dr. Maya Silvi Lydia, B.Sc., M.Sc. selaku Dekan Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Sumatera Utara.
3. Ibu Dr. Amalia ST., M.T. selaku Ketua Program Studi S1 Ilmu Komputer FASILKOM-TI USU sekaligus Dosen Pembimbing I yang sudah memberikan bimbingan, kritikan, motivasi, dan masukan kepada penulis dalam menyelesaikan tugas akhir.
4. Ibu Sri Melvani Hardi, S.Kom., M.Kom. sebagai Sekretaris Program Studi S1 Ilmu Komputer FASILKOM-TI USU sekaligus Dosen Pembimbing II yang telah memberikan bimbingan, kritikan, motivasi, dan masukan kepada penulis dalam menyelesaikan tugas akhir.
5. Seluruh Bapak/Ibu Dosen Program Studi S-1 Ilmu Komputer yang sudah memberikan waktu dan tenaga untuk mengajar dan membimbing sehingga penulis dapat sampai kepada tahap penyusunan skripsi ini.
6. Seluruh Staf Pegawai FASILKOM-TI USU yang sudah banyak memberikan bantuan kepada penulis selama perkuliahan sampai kepada tahap penyusunan skripsi ini.
7. Kedua orangtua penulis, Syahrul Humaidi dan Sri Medina Sirait yang senantiasa mendoakan, dan mendukung dalam segala hal.

8. Saudara-saudara penulis yang penulis banggakan, Abdurrasyid Aulia Rahman dan Abul Khair Asrul Abdurrahim.
9. Teman-teman Selaras FC Ary Bobby Siregar, Abiyu Dzakwan Khairi, Amhar Syakuro Nasution, Fauzan Zaman, Fiqri Ramadhan, Prima, M. Raja, M. Ridho Rambe, M. Dwiki Prasetyo, Yudha Triya yang memberikan support, motivasi, semangat, dan menjadikan penulis pribadi yang lebih baik.
10. Teman - teman FFL Aditya Wangsa, Afita Puspa Sari, Asri Yulianingrum, Devina, Jessica Graciela, Nikita Masaling, Shelfy Agina G, Rivany, Mita, Rhania, Dini, Shinta, Winda, Zahra.
11. Rekan-rekan seperkuliahannya penulis, stambuk 2018, telah memberi dukungan, semangat, dan menjadikan penulis pribadi yang lebih baik.
12. Seluruh keluarga besar IMILKOM periode 2021/2022 terutama Departemen Komunikasi dan Informasi yang selalu menjadi motivasi.
13. Seluruh rekan - rekan Minat dan Bakat PEMA Fasilkom-TI 2019/2020 yang selalu menjadi motivasi.

Serta kepada seluruh pihak yang sudah membantu penulis, yang belum dapat penulis sebutkan satu-persatu. Semoga Tuhan memberikan keberkahan atas kebaikan dan bantuan yang sudah diberikan kepada penulis. Semoga skripsi yang saya tulis bisa memberikan dampak serta manfaat terhadap diri saya dan orang lain.

Medan, 5 Januari 2024

Achmad Yusuf Barmawi

181401119

## ABSTRAK

Bahasa sebagai alat komunikasi utama manusia, mencerminkan keanekaragaman budaya dan bahasa di Indonesia. Masyarakat sering menggunakan bahasa tidak baku, terutama di media sosial, memunculkan variasi dalam ejaan, slang, dan singkatan. Penelitian bertujuan untuk menerjemahkan bahasa non formal ke bahasa formal Bahasa Indonesia menggunakan pendekatan *Semi Supervised Translation*. Penelitian ini mencermati konteks bahasa non formal pada media sosial, khususnya *platform X* sebelumnya dikenal sebagai Twitter, Facebook, Instagram, dan sebagainya. Penggunaan slang dan singkatan menjadi karakteristik utama dalam bahasa non formal. Oleh karena itu, dibangun suatu sistem model menggunakan arsitektur transformers untuk menerjemahkan bahasa non formal ke bahasa formal. Pengujian dilakukan dengan menganalisis 5 data uji, mengevaluasi akurasi, dan memeriksa efektivitas model penerjemah. Hasilnya menunjukkan bahwa model *Semi Supervised Translation* dapat digunakan untuk membangun sistem penerjemah dengan akurasi mencapai 96%. Meskipun demikian, beberapa kata atau kalimat non formal dan singkatan masih sulit diubah menjadi format formal. Penerjemahan bahasa non formal ke bahasa formal Bahasa Indonesia dapat dilakukan dengan menggunakan *Semi Supervised Translation*.

**Kata Kunci:** *Bahasa non formal, bahasa Indonesia, mesin penerjemahan, semi supervised, natural language processing.*

## ABSTRACT

Language as the primary tool of human communication, reflects the cultural and linguistic diversity in Indonesia. People often use non-standard language, especially on social media, giving rise to variations in spelling, slang, and abbreviations. Research aims to translate non-formal language into formal Indonesian using a Semi-Supervised Translation approach. The study focuses on the context of non-formal language on social media, particularly platforms X previously known as Twitter, Facebook, Instagram, and the like. The use of slang and abbreviations is a prominent characteristic of non-formal language. Therefore, a model system is built using transformer architecture to translate non-formal language into formal language. Testing is conducted by analyzing five test data, evaluating accuracy, and examining the effectiveness of the translation model. The results indicate that the Semi Supervised Translation model can be used to build a translator system with an accuracy of up to 96%. However, some non-formal words or sentences and abbreviations are still challenging to transform into a formal format. In conclusion, this research suggests that the translation of non-formal language into formal Indonesian can be achieved using Semi-Supervised Translation.

**Kata Kunci:** *Non formal language, Indonesian, machine translation, semi supervised, natural language processing.*



## DAFTAR ISI

|                                   |      |
|-----------------------------------|------|
| PERSETUJUAN.....                  | i    |
| PERNYATAAN .....                  | ii   |
| PENGHARGAAN .....                 | iii  |
| ABSTRAK .....                     | v    |
| ABSTRACT .....                    | vi   |
| DAFTAR ISI .....                  | vii  |
| DAFTAR GAMBAR .....               | x    |
| DAFTAR TABEL .....                | xii  |
| DAFTAR LAMPIRAN .....             | xiii |
| BAB I PENDAHULUAN .....           | 1    |
| 1.1. Latar Belakang .....         | 1    |
| 1.2. Rumusan Masalah .....        | 4    |
| 1.3. Batasan Masalah .....        | 5    |
| 1.4. Tujuan Penelitian .....      | 5    |
| 1.5. Manfaat Penelitian .....     | 5    |
| 1.6. Metodologi Penelitian .....  | 6    |
| 1. Studi Pustaka .....            | 6    |
| 2. Analisis dan Perancangan ..... | 6    |
| 3. Implementasi Sistem .....      | 6    |
| 4. Pengujian Sistem .....         | 6    |
| 5. Dokumentasi Sistem .....       | 6    |
| 1.7. Sistematika Penulisan .....  | 7    |
| BAB II LANDASAN TEORI .....       | 8    |

|  |    |
|--|----|
| 2.1. <i>Natural Language Processing</i> .....                  | 8  |
| 2.2. <i>Semi Supervised Learning</i> .....                     | 8  |
| 2.3. Bahasa Indonesia .....                                    | 9  |
| 2.3.1 Bahasa Non Formal .....                                  | 10 |
| 2.4. <i>Machine Translation</i> .....                          | 10 |
| 2.5. <i>Semi-Machine Translation</i> .....                     | 11 |
| 2.6. <i>Neural Machine Translation</i> .....                   | 11 |
| 2.7. <i>Hugging Face Transformer</i> .....                     | 12 |
| 2.8. MBART .....   | 12 |
| 2.9. <i>Bilingual Evaluation Understudy (BLEU Score)</i> ..... | 13 |
| 2.10. <i>Framework Flask</i> .....                             | 14 |
| 2.11. Penelitian yang Relevan .....                            | 14 |
| BAB III ANALISIS DAN PERANCANGAN SISTEM .....                  | 19 |
| 3.1. Analisis Sistem .....                                     | 19 |
| 3.1.1. Analisis Masalah .....                                  | 19 |
| 3.1.2. Analisis Kebutuhan .....                                | 19 |
| 3.1.2.1. Kebutuhan Fungsional .....                            | 19 |
| 3.1.2.2. Kebutuhan Non-Fungsional .....                        | 20 |
| 3.1.3. Arsitektur Umum Sistem .....                            | 20 |
| 3.2. Pemodelan Sistem .....                                    | 22 |
| 3.2.1. <i>Use Case Diagram</i> .....                           | 22 |
| 3.2.2. <i>Activity Diagram</i> .....                           | 23 |
| 3.2.2.1. <i>Activity Diagram</i> Sistem Penerjemah .....       | 24 |
| 3.3. <i>Flowchart</i> .....                                    | 24 |
| 3.3.1. <i>Flowchart</i> Sistem .....                           | 24 |
| 3.4. Perancangan <i>Interface</i> .....                        | 25 |

|   |    |
|---|----|
| 3.4.1. Halaman Home .....                                 | 26 |
| 3.4.2. Halaman Data .....                                 | 26 |
| 3.4.3. Halaman Program .....                              | 27 |
| 3.4.4. Halaman Riwayat .....                              | 29 |
| 3.4.5. Halaman BLEU Score .....                           | 30 |
| BAB IV IMPLEMENTASI DAN PENGUJIAN SISTEM .....            | 32 |
| 4.1. Implementasi .....                                   | 32 |
| 4.1.1. Pengumpulan Data .....                             | 32 |
| 4.1.2. <i>Preprocessing</i> .....                         | 33 |
| 4.1.3. Pembuatan dan <i>training</i> model .....          | 36 |
| 4.1.4. Tampilan Sistem .....                              | 41 |
| 4.2. Pengujian Sistem .....                               | 44 |
| 4.2.1. Pengujian dalam menghitung BLEU <i>Score</i> ..... | 46 |
| BAB V PENUTUP .....                                       | 48 |
| 5.1. KESIMPULAN .....                                     | 48 |
| 5.2. SARAN .....  | 48 |
| DAFTAR PUSTAKA .....                                      | 49 |

## DAFTAR GAMBAR

|   |    |
|---|----|
| Gambar 3.1 Arsitektur Umum Sistem .....                                       | 20 |
| Gambar 3.2 <i>Use Case Diagram</i> .....                                      | 23 |
| Gambar 3.3 <i>Activity Diagram</i> .....                                      | 24 |
| Gambar 3.4 Flowchart Sistem .....   | 25 |
| Gambar 3.5 Halaman <i>Home</i> .....  | 26 |
| Gambar 3.6 Halaman Data .....   | 26 |
| Gambar 3.7 Halaman Program .....  | 27 |
| Gambar 3.8 Halaman program setelah melakukan terjemahan .....                 | 28 |
| Gambar 3.9 Halaman riwayat .....  | 29 |
| Gambar 3.10 Halaman BLEU <i>score</i> .....                                   | 30 |
| Gambar 3.11 Halaman BLEU score setelah dihitung .....                         | 31 |
| <br>  |    |
| Gambar 4.1 Proses pengambilan data .....                                      | 33 |
| Gambar 4.2 Hasil pengambilan data .....                                       | 33 |
| Gambar 4.3 Proses <i>preprocessing</i> .....                                  | 34 |
| Gambar 4.4 Hasil <i>preprocessing</i> .....                                   | 34 |
| Gambar 4.5 Proses menerjemah bahasa non formal .....                          | 35 |
| Gambar 4.6 Hasil terjemahan bahasa non formal .....                           | 35 |
| Gambar 4.7 Proses penghilangan karakter berulang dan <i>export file</i> ..... | 36 |
| Gambar 4.8 <i>Library</i> yang diperlukan .....                               | 36 |
| Gambar 4.9 Proses <i>load dataset</i> .....                                   | 37 |
| Gambar 4.10 Proses <i>split</i> data .....                                    | 37 |
| Gambar 4.11 Proses inialisasi tokenizer .....                                 | 38 |
| Gambar 4.12 Proses tokenisasi data .....                                      | 38 |
| Gambar 4.13 Proses membuat dataset kustom .....                               | 38 |
| Gambar 4.14 Proses inialisasi model dan pembuatan argumen pelatihan .....     | 39 |
| Gambar 4.15 Proses <i>training</i> model .....                                | 39 |
| Gambar 4.16 Hasil <i>training</i> model .....                                 | 40 |
| Gambar 4.17 Hasil <i>training</i> model .....                                 | 40 |
| Gambar 4.18 Proses simpan model .....   | 41 |

|  |    |
|--|----|
| Gambar 4.19 Halaman <i>Home</i> .....                            | 41 |
| Gambar 4.20 Halaman Data .....                                   | 42 |
| Gambar 4.21 Halaman Program .....                                | 42 |
| Gambar 4.22 Halaman program setelah melakukan penerjemahan ..... | 43 |
| Gambar 4.23 Halaman Riwayat .....                                | 43 |
| Gambar 4.24 Halaman BLEU Score .....                             | 44 |

**DAFTAR TABEL**

|   |    |
|---|----|
| Tabel 4. 1 Daftar kalimat yang dimasukan beserta hasil terjemahan ..... | 45 |
| Tabel 4. 2 Daftar kalimat yang dimasukan beserta hasil BLEU .....       | 46 |

## DAFTAR LAMPIRAN

# **BAB I**

## **PENDAHULUAN**

### **1.1. Latar Belakang**

Bahasa merupakan sistem komunikasi yang digunakan oleh manusia untuk mengungkapkan pemikiran, ide, dan informasi kepada orang lain. Fungsi pokok bahasa adalah sebagai alat komunikasi. Manusia memerlukan bahasa untuk berinteraksi dan berkomunikasi dengan sesamanya dalam memenuhi kebutuhan. Bahasa meliputi suatu kata maupun suatu kalimat yang diungkapkan baik secara lisan atau tertulis. Dengan bahasa seseorang dapat mengekspresikan diri, pikiran, perasaan, maupun keinginannya.

Setiap bahasa memiliki ciri nya tersendiri. Mulai dari pembentukan kata, pembentukan kalimat, sampai pengucapannya yang hampir tidak ada yang sama. Indonesia mempunyai banyak ragam budaya dan bahasa yang berbeda didalamnya. Karena Indonesia salah satu negara yang terkenal keanekaragaman bahasa daerah. Masyarakat di Indonesia sering menggunakan bahasa tidak baku di berbagai kesempatan. Pada percakapan, bahasa non formal meliputi penggunaan kata yang berbeda, biasanya menggunakan serapan dari suatu bahasa daerah atau asing. Selain itu, suatu tren dapat mempengaruhi pilihan kata. Hal ini menjadi lebih mencolok ketika masuk ke ranah tertulis, terutama pada media sosial, karena adanya variasi penulisan seperti singkatan dan kemungkinan kesalahan ketik (typo).

Komunikasi secara tertulis yang terutama di media sosial, seringkali digunakan ejaan dan tata bahasa yang tidak resmi dengan tujuan agar komunikasi terasa lebih alami, santai, dan akrab. Namun, hal ini terkadang menjadi hambatan dalam penyebaran informasi karena perbedaan ejaan kata yang ada di setiap media sosial.

Di dalam media sosial seperti X atau dulunya dikenal sebagai twitter, facebook, instagram, dan lainnya biasanya terdapat kata - kata yang diunggah oleh pengguna yang berisi singkatan kata, kata non formal atau biasa disebut dengan kata slang. Kata slang merupakan ragam bahasa non formal yang digunakan oleh



remaja maupun kelompok tertentu yang mungkin dapat dimengerti oleh kaum itu saja. Kata slang ini biasanya sulit dipahami karena setiap orang memiliki perbedaan terhadap pemahaman dari kata slang tersebut. Contoh dari kata slang tersebut ialah seperti “gue” yang memiliki arti “saya”, “gimana” yang memiliki arti “bagaimana” ataupun singkatan seperti “yg” memiliki arti “yang” dan sebagainya.

Meskipun demikian, permasalahan penggunaan bahasa dan ejaan non formal dapat menjadi keuntungan di dunia akademis. Terdapat Beberapa penelitian mengambil permasalahan ini sebagai fokus dalam bidang pemrosesan data teks. Data tersebut diolah menggunakan metode yang dikenal sebagai *Natural Language Processing* (NLP). Mesin penerjemah biasanya digunakan untuk menerjemahkan suatu bahasa ke bahasa yang lainnya yang diinginkan. Contohnya bahasa Inggris ke bahasa Indonesia. Mesin penerjemahan statistik adalah sistem penerjemahan yang umumnya menghasilkan terjemahan berdasarkan model statistik yang menggunakan parameter dari analisis korpus paralel (Pranata, 2016).

Penelitian tentang mesin penerjemah telah banyak dilakukan. Salah satunya adalah “*Phrase Based Statistical Machine Translation Javanese-Indonesian*” (Lestari, Ardiyanti, & Asror, 2021). Bertujuan untuk mengembangkan sistem terjemahan mesin statistik untuk terjemahan Jawa-Indonesia dan untuk memahami dampak dari korpus paralel dan monolingual terhadap kualitas terjemahan. Dilakukannya eksperimen dengan secara bertahap meningkatkan jumlah korpus paralel dan monolingual dalam tujuh konfigurasi berbeda dari sistem terjemahan mesin. Hasilnya menunjukkan bahwa nilai evaluasi terjemahan meningkat seiring dengan peningkatan jumlah korpus paralel dan monolingual menyimpulkan bahwa jumlah data korpus, terutama korpus paralel, memainkan peran penting dalam meningkatkan kualitas terjemahan mesin statistik. Para peneliti merekomendasikan penelitian lebih lanjut untuk mengeksplorasi dampak jenis korpus yang berbeda pada kualitas terjemahan dan untuk mengoptimalkan jumlah data korpus untuk pasangan bahasa tertentu.

Penelitian mesin penerjemah lainnya dilakukan oleh Permata, & Abidin. (2020), dengan judul “*Statistical Machine Translation Pada Bahasa Lampung dialek Api ke Bahasa Indonesia*” didapatkan bahwa penerjemahan bahasa

Lampung ke bahasa Indonesia menggunakan pendekatan *Statistical Machine Translation* (SMT) menunjukkan hasil yang menjanjikan. Meskipun demikian, terdapat beberapa kesalahan yang perlu diperbaiki, seperti kesalahan dalam memahami konteks kalimat, penerjemahan yang tidak beraturan, dan hilangnya kata dalam kalimat. Mesin SMT memiliki kelemahan dalam memahami konteks kalimat dan hanya bekerja berdasarkan kecocokan data trigram, bigram, atau unigram.

Penelitian mesin penerjemah dengan *semi supervised* oleh Wibowo, et. al. (2020) “*Semi-Supervised Low-Resource Style Transfer of Indonesian Informal to Formal Language with Iterative Forward-Translation*”. Dilakukannya perbandingan antara kinerja beberapa mesin penerjemah yaitu dengan menggunakan kamus, mesin penerjemah statistik, transformers dan menggunakan *pretrained* model. Karena kurangnya data, teknik standar yang umumnya digunakan untuk *machine translation*, seperti transformers, tidak memberikan hasil yang memuaskan. Kinerja terbaik dapat dicapai melalui pendekatan statistik (PBSMT) atau menggunakan *pretrained* LM yaitu arsitektur GPT-2 yang juga berbasis transformer. Namun, jika volume data ditingkatkan, teknik statistik menjadi kurang efektif dalam hal kinerja. Selain itu, PBSMT hanya dapat mempelajari dari data yang diberikan, sehingga model ini tidak mampu mengatasi teks informal di luar lingkup pembelajarannya, seperti menghadapi kata-kata baru yang belum pernah dikenali sebelumnya. Sebaliknya, *pretrained language model* telah mempelajari bahasa Indonesia secara general, sehingga telah mempunyai pemahaman tentang kata yang tidak terdapat dalam dataset. Meskipun demikian, masih banyak pendekatan *pre-trained language model* yang belum dijelajahi, seperti menggunakan *multilingual language model*.

Penelitian mesin penerjemahan lainnya dengan menggunakan arsitektur transformers adalah Penelitian Liu, et al (2020) “*Multilingual Denoising Pre-training for Neural Machine Translation*” untuk menunjukkan bahwa *pre-training* multibahasa *denoising* dapat secara signifikan meningkatkan kinerja mesin terjemahan baik yang *supervised* maupun *unsupervised*, baik pada tingkat kalimat maupun dokumen. Penelitian ini juga menganalisis kapan dan bagaimana *pre-training* paling efektif dan dapat dikombinasikan dengan pendekatan lain seperti

*back-translation*. Hasil penelitian menunjukkan kemampuan *transfer learning* dari representasi yang dipelajari dari *pre-training* multibahasa. Ditemukan bahwa *pre-training* multibahasa *denoising* dapat meningkatkan kinerja mesin terjemahan baik pada tingkat kalimat maupun dokumen. Hasilnya juga menunjukkan bahwa model *pre-trained* MBART25 menghasilkan terjemahan yang lebih baik daripada model yang tidak *dipre-training*, terutama pada tingkat dokumen.

Dari beberapa penelitian mesin penerjemahan tersebut dapat disimpulkan bahwa NMT lebih baik daripada SMT dalam melakukan terjemahan dan belum adanya pembangunan suatu model untuk bahasa non formal ke bahasa formal menggunakan *pretrained multilingual language model* yaitu dengan menggunakan *pretrained* MBART.

Pada penelitian ini sistem penerjemahan dilakukan dengan menggunakan pendekatan *semi supervised translation*. Model dibangun dengan menggunakan arsitektur transformers. Transformers merupakan arsitektur model yang dapat digunakan dalam berbagai konteks, termasuk dalam pendekatan *semi-supervised translation*. Transformers telah menjadi dasar bagi banyak model penerjemahan mesin (*machine translation*), termasuk model penerjemah *neural machine translation* (NMT). Pada penelitian ini bahasa non formal bahasa Indonesia merupakan bahasa sumber, dan bahasa formal bahasa Indonesia merupakan bahasa target yang ingin diterjemahkan.

Dengan meningkatnya penggunaan media sosial, banyak informasi dan komunikasi terjadi dalam bentuk bahasa non formal. Namun, ada situasi di mana informasi ini perlu dipindahkan atau diubah menjadi format yang lebih formal, dan sistem penerjemah dapat memfasilitasi proses ini. Penerjemahan dianggap cukup penting karena dengan adanya penerjemahan bahasa, diharapkan dapat membantu mereka yang memerlukan untuk mengetahui maksud dari kata slang yang digunakan.

## 1.2. Rumusan Masalah

Penggunaan bahasa non formal telah menjadi suatu fenomena yang terus berkembang di media sosial dengan seiring berjalannya waktu. Bahasa non formal sering kali memiliki variasi yang dapat menyebabkan ketidakjelasan dan

kesalahpahaman antara penulis dengan para pembaca. Hal ini juga menyebabkan kesulitan dalam komunikasi, terutama komunikasi resmi seperti surat-menyurat ataupun komunikasi bisnis, sehingga diperlukannya suatu sistem atau aplikasi yang dapat konversi kalimat non formal bahasa Indonesia menjadi kalimat formal bahasa Indonesia.

### **1.3. Batasan Masalah**

Berikut ini merupakan batasan masalah yang ada pada penelitian ini:

1. Dataset menggunakan bahasa Indonesia.
2. Bahasa yang diterjemahkan adalah bahasa Indonesia. Tidak dapat menerjemahkan bahasa daerah maupun bahasa asing.
3. Penerjemahan bahasa non formal menjadi bahasa formal menggunakan pendekatan *semi supervised translation*.
4. Bahasa pemrograman yang digunakan adalah bahasa pemrograman Python.

### **1.4. Tujuan Penelitian**

Tujuan yang diharapkan dari penelitian ini adalah:

1. Membangun suatu model yang dapat menerjemahkan bahasa non formal bahasa Indonesia ke bahasa formal bahasa Indonesia.
2. Membangun suatu sistem atau aplikasi yang dapat menerjemahkan bahasa non formal bahasa Indonesia ke bahasa formal bahasa Indonesia.
3. Mengetahui penerapan mesin penerjemah terhadap data bahasa non formal bahasa Indonesia ke bahasa formal bahasa Indonesia.
4. Mengetahui persentase keakuratan dari mesin penerjemah yang dibangun.

### **1.5. Manfaat Penelitian**

Kegunaan yang dapat di manfaatkan dengan adanya penelitian kali ini adalah dapat membantu dalam menerjemahkan bahasa non formal bahasa Indonesia ke dalam bahasa formal bahasa Indonesia untuk kebutuhan organisasi atau individu untuk mematuhi regulasi dan persyaratan suatu hal yang mengharuskan penggunaan bahasa formal dalam beberapa konteks. Penerjemahan menjadi bentuk formal juga berelevansi dengan ketersediaan *pretarined embedding* yang

biasanya tersedia dalam bentuk bahasa formal. Dengan adanya suatu sistem penerjemah bahasa non formal menjadi formal diharapkan dapat meningkatkan *linguistik resources* bagi bahasa Indonesia.

### 1.6. Metodologi Penelitian

Metode penelitian yang diterapkan dalam studi ini adalah sebagai berikut:

#### 1. Studi Pustaka

Pada tahapan ini dilakukan pengumpulan referensi yang diperlukan dalam penelitian. Hal tersebut dilakukan untuk memperoleh data dan informasi yang diperlukan untuk penulisan penelitian. Referensi yang digunakan berupa buku, *e-book*, jurnal, artikel, situs yang berhubungan dengan *Natural Language Processing*, bahasa Indonesia, *machine translation*, dan flask.

#### 2. Analisis dan Perancangan

Merujuk kepada ruang lingkup penelitian, dilakukannya riset ataupun analisis untuk menentukan apa saja yang dibutuhkan dalam penelitian dalam merancang sebuah diagram alir general arsitektur. Dalam tahap ini juga dilakukan pembuatan *usecase diagram*, *activity diagram*, *sequence diagram*, dan perancangan *interface*.

#### 3. Implementasi Sistem

Setelah melakukan perancangan, hasil dari perancangan tersebut di implementasikan menggunakan bahasa pemrograman Python dan menggunakan framework flask.

#### 4. Pengujian Sistem

Di fase ini nantinya dilakukan percobaan dan pemeriksaan pada sistem yang dibuat sesuai, untuk menentukan sistem yang sudah dikerjakan beroperasi sesuai dengan yang diinginkan.

#### 5. Dokumentasi Sistem

Setiap tahapan yang telah dilakukan didokumentasikan agar dapat dijadikan kesimpulan yang akan digunakan oleh penulis dalam bentuk skripsi.

## **1.7. Sistematika Penulisan**

Sistematika penulisan dari skripsi ini terdiri dari lima bab, yakni:

### **BAB I PENDAHULUAN**

Bab ini berfungsi sebagai bagian yang memperkenalkan latar belakang penelitian, menguraikan rumusan masalah, tujuan penelitian, batasan masalah, manfaat penelitian, metode penelitian, dan sistematika penulisan. Ini adalah langkah awal yang penting dalam suatu penelitian ilmiah untuk memberikan pemahaman kepada pembaca mengenai konteks dan kerangka kerja penelitian tersebut.

### **BAB II LANDASAN TEORI**

Pada Landasan teori menjelaskan teori-teori yang digunakan pada penelitian, sebagai dasar penulis untuk melakukan penelitian, di mana teori - teori tersebut didapat dari studi literatur.

### **BAB III ANALISIS DAN PERANCANGAN**

Pada bab ini berisi tentang analisis dan perancangan dari pembuatan sistem penerjemahan bahasa tidak baku bahasa Indonesia ke bahasa baku bahasa Indonesia. Analisis sistem meliputi analisis masalah dan analisis kebutuhan dalam arsitektur umum dan perancangan sistem.

### **BAB IV IMPLEMENTASI DAN PENGUJIAN SISTEM**

Pada bab ini menjelaskan tentang hasil penelitian yang telah dilakukan dan memberikan penjelasan implementasi sistem berdasarkan analisis masalah dan kebutuhan sistem, skema pengujian terhadap sistem yang dikembangkan, serta penjelasan hasil pengujian sistem tersebut.

### **BAB V KESIMPULAN DAN SARAN**

Pada bab ini menjelaskan kesimpulan penelitian yang diperoleh berdasarkan pembahasan pada setiap bab dan saran yang diberikan untuk masukan ataupun rujukan yang diharap dapat dikembangkan penelitian selanjutnya.

## BAB II

### LANDASAN TEORI

#### 2.1. *Natural Language Processing*

Salah satu cabang ilmu *Artificial Intelligence* adalah pengolahan bahasa natural (NLP). Bahasa yang biasa dipakai manusia untuk berbicara satu sama lain disebut bahasa natural. Untuk komputer dapat memahami maksud orang, bahasanya harus diproses dan dipahami terlebih dahulu. Ada banyak aplikasi *natural language processing* (NLP) yang mungkin membuat komputer dapat paham instruksi bahasa yang dimasukkan oleh pengguna seperti chatbot, yang membuat pengguna merasa seperti mereka berbicara dengan komputer, *Stemming*, yang berarti memotong kata dalam menjadi kata dasar setiap kata dalam kalimat, *summarization* yang berarti membuat ringkasan, dan *translation tools*, yang berarti melakukan terjemahan suatu bahasa. Pengolahan bahasa natural (NLP) memungkinkan *transformasi interface* bahasa natural menjadi pengetahuan.

#### 2.2. *Semi Supervised Learning*

*Supervised*, *unsupervised*, dan *semi supervised* adalah jenis pembelajaran yang dikenal dalam pembuatan *Artificial Intelligence* (AI). Ketiganya memiliki kemampuan untuk membuat AI yang akurat. Namun, metodenya cukup berbeda. *Supervised learning* yaitu jenis pembelajaran yang menggunakan satu atau lebih algoritma untuk memahami dan mempelajari fungsi pemetaan dari masukan ke keluaran. dengan tujuan untuk memperkirakan fungsi pemetaan. Oleh karena itu sistem dapat memprediksi outputnya ketika ada input baru. Sistem dilatih untuk mempelajari data dengan label khusus melalui *supervised learning*. Setiap data yang diproses oleh algoritma *supervised learning* harus diberi label. Regresi linier, *forest random*, dan *support vector machine* adalah beberapa contoh algoritma *supervised learning* yang sangat populer.

Berbeda dengan *supervised learning*, *unsupervised learning* hanya melibatkan satu variabel input tanpa adanya variabel output yang ditentukan secara pasti. Tujuannya untuk memodelkan distribusi atau struktur yang memiliki

kemampuan untuk mempelajari data lebih lanjut. *Unsupervised learning* dibagi menjadi dua kelompok, *clustering* dan *association*. *Clustering k-means* dan algoritma apriori adalah beberapa algoritma yang sangat populer.

*Semi-supervised learning* merupakan algoritma yang digunakan untuk melaksanakan pembelajaran data yang tersedia label dan tanpa label. Biasanya, *semi-supervised learning* menggunakan sedikit data, kemudian menggunakan pembelajaran *supervised* dan *unsupervised* untuk membuat data tambahan, dan kemudian membuat model pembelajaran dari tambahan data tersebut (Putra, 2020). Output yang dihasilkan dapat ditingkatkan dengan metode ini. Algoritma pelatihan *semi-supervised* ini termasuk dalam proses pelatihan. Algoritma *semi-supervised* memiliki dua tahap. *Pre-training* yang tidak diawasi adalah tahap pertama, dan *fine-tuning* yang diawasi adalah tahap kedua (Sulistiawan et al., 2020).

Jumlah data input dan output tidak sama untuk semua data. Kumpulan data terkadang memiliki variabel input tetapi hanya sebagian variabel outputnya. Pembelajaran *semi-supervised* dapat digunakan untuk menyelesaikan situasi seperti ini. *Semi-supervised* menggunakan teknik *unsupervised* untuk menemukan atau memahami struktur variabel masukan atau inputan. Kemudian, sistem memanfaatkan algoritma *supervised learning* untuk menghasilkan prediksi optimal dari data yang tidak memiliki label dan memasukkannya kembali ke *supervised learning* sebagai data latih. Kemudian, sistem memakai model tersebut untuk melakukan prediksi baru untuk data masukan yang baru.

### **2.3. Bahasa Indonesia**

Bahasa Indonesia tidak hanya merupakan bahasa persatuan, tetapi juga merupakan bahasa resmi NKRI. Dari perspektif linguistik, bahasa Indonesia adalah variasi dari bahasa melayu. Hingga saat ini, bahasa Indonesia masih hidup dan berkembang dengan penciptaan dan penyerapan kosa kata baru dari bahasa daerah dan bahasa asing.

Jauh sebelum Indonesia dijajah oleh Belanda, masyarakat Indonesia telah memakai bahasa ini. Namun, tidak semua orang memakai metode yang tepat. Salah satunya menggunakan bahasa Indonesia yang tidak sesuai dengan ejaannya,



serta tidak sesuai dengan definisi yang ditemukan dalam KBBI. Karenanya, pemahaman tentang ragam bahasa sangat penting untuk dipelajari secara menyeluruh dalam bahasa Indonesia. Semua orang diharapkan belajar bahasa Indonesia. Dalam konteks bahasa Indonesia, terdapat istilah "ragam bahasa", yang merujuk pada variasi bahasa yang digunakan dengan cara yang berbeda. (Putrayasa, 2018).

Ragam bahasa merupakan perbedaan bahasa berdasarkan cara digunakan, subjek yang menjadi pembicaraan, keterkaitan antara pembicara, dan mitra bicara. Serta sesuai dengan media yang digunakan oleh pembicara (Bachman, 1990). Bahasa juga mengalami perubahan seiring dengan evolusi masyarakat pada zaman sekarang. Perubahan ini terdiri dari penggunaan variasi bahasa sesuai kebutuhan. Dalam kasus ini, banyak variasi tersebut tidak mengurangi efektivitas bahasa sebagai alat komunikasi. Sebagai hasilnya, bahasa memiliki proses untuk menentukan perbedaan tertentu sesuai kebutuhan, yang dikenal sebagai ragam standar (Subianto, 2000).

### **2.3.1 Bahasa Non Formal**

Bahasa non formal mengacu pada penggunaan bahasa yang santai, tidak resmi, atau tidak terikat oleh aturan tata bahasa formal. Bahasa non formal sering digunakan dalam situasi-situasi sehari-hari, percakapan non formal, dan konteks non-akademis. Gaya bahasa ini dapat mencakup penggunaan slang, singkatan, ejaan yang tidak baku, dan ekspresi sehari-hari. Contoh dari kata slang tersebut seperti “gue” yang memiliki arti “saya”, “gimana” yang memiliki arti “bagaimana” ataupun singkatan seperti “yg” memiliki arti “yang” dan sebagainya.

### **2.4. Machine Translation**

Mesin penerjemahan otomatis memungkinkan penerjemahan antara bahasa. Mesin penerjemah bermanfaat karena dapat membantu individu yang berkomunikasi satu sama lain menggunakan bahasa yang berbeda. Metode mesin penerjemah statistik menggunakan analisis korpus teks bilingual atau korpus paralel untuk menghasilkan hasil terjemahan yang didasarkan pada model statistik. Pendekatan ini dikenal sebagai mesin penerjemah statistik (Pratiwi, 2017). Banyak pendekatan

*machine translation* (MT), termasuk *statistical machine translation* (SMT), *rule-based machine translation* (RMT), *phrase-based machine translation* (PBMT), dan *neural machine translation* (NMT), melakukan konversi dari bahasa sumber ke bahasa target atau tujuan.

## **2.5. *Semi-Machine Translation***

*Semi machine translation* mengacu pada pendekatan di mana kedua pihak, yaitu penerjemah manusia dan sistem terjemahan mesin bekerja sama dalam proses penerjemahan. Berbeda dengan terjemahan mesin otomatis, di mana sistem komputer menerjemahkan seluruh teks tanpa intervensi manusia, *semi machine translation* melibatkan kombinasi terjemahan otomatis dan keterlibatan manusia.

Dalam alur *semi machine translation*, sistem terjemahan mesin menghasilkan terjemahan awal, dan kemudian penerjemah manusia meninjau dan menyempurnakan hasilnya. Tujuannya adalah memanfaatkan kelebihan terjemahan mesin dan penerjemah manusia untuk mencapai hasil akhir yang lebih akurat dan berkualitas. Metode ini umumnya diterapkan dalam konteks di mana diperlukan keseimbangan antara efisiensi dan ketepatan linguistik.

## **2.6. *Neural Machine Translation***

NMT merupakan metode inovatif dalam penerjemahan mesin yang memanfaatkan struktur RNNs pada bagian encoder dan decoder-nya. Transformer digunakan dalam model penerjemah NMT, yang merupakan paradigma baru dalam pengembangan sistem penerjemahan. Model NMT menggunakan arsitektur neural network, seperti Transformer, untuk mendekati tugas terjemahan bahasa dengan cara yang berbeda dari model penerjemah statistik tradisional. Model NMT cenderung menghasilkan terjemahan yang lebih akurat dan menangani hubungan yang lebih kompleks antara kata-kata dalam bahasa sumber dan bahasa target. Mesin penerjemah terdapat komponen - komponen didalamnya yaitu bahasa sumber dan bahasa target. Dari beberapa pendekatan, NMT merupakan suatu metodologi pengembangan terbaru pada MT dengan peningkatan yang lebih baik dari pendekatan sebelumnya (Fauziyah, 2022).

## 2.7. *Hugging Face Transformer*

*Hugging Face Transformers* merupakan sebuah *library open source* yang dikembangkan oleh Hugging Face. *Library* ini memungkinkan pengguna untuk dengan mudah menggunakan, mengunduh, dan menjelajahi berbagai model bahasa alamiah (NLP) *state-of-the-art*, termasuk *transformer-based models* seperti BERT, GPT, dan lainnya. Transformers dari Hugging Face menyediakan antarmuka yang ramah pengguna untuk menggunakan model-model tersebut dalam berbagai tugas, seperti pemahaman bahasa alamiah, penerjemahan, generasi teks, dan sebagainya. *Library* ini juga menyertakan model-model *pre-trained* yang dapat digunakan atau disesuaikan untuk tugas-tugas tertentu.

## 2.8. MBART

MBART (*Multilingual BART*) merupakan variasi dari model BART. BART merupakan bagian dari keluarga model transformers. BART (*Bidirectional and Auto-Regressive Transformers*) adalah pendekatan *pretraining* yang mencapai hasil terbaik dalam berbagai tugas pemrosesan bahasa alami, termasuk generasi respons percakapan, ringkasan, abstraktif pertanyaan jawaban, dan terjemahan mesin. Ini mengungguli karya sebelumnya dan model lain seperti Seq2Seq dan BERT. Kinerja BART dikaitkan dengan kemampuannya menghasilkan output yang lancar dan gramatikal, serta kombinasi kuat pemahaman dan generasi bahasa alami. Ini juga menunjukkan peningkatan dalam metrik ringkasan dan analisis kualitatif. Pendekatan BART mengurangi ketidakcocokan antara *pre-training* dan tugas generasi, sehingga efektif untuk tugas diskriminatif maupun generatif (Lewis, 2019). MBART adalah kebalikannya, dirancang untuk tugas multibahasa, yang berarti dapat menangani beberapa bahasa. MBART dapat diterapkan untuk menerjemahkan teks dari satu bahasa ke bahasa lain. Dalam konteks penerjemahan mesin, MBART dapat memahami dan menghasilkan teks dalam beberapa bahasa tanpa harus memiliki model terpisah untuk setiap pasangan bahasa.

Beberapa fitur MBART yaitu *multilingual capability*, *generative text model*, dan *seq2seq architecture*. *Multilingual capability* merupakan kemampuan untuk menangani beberapa bahasa tanpa memerlukan model terpisah untuk setiap

bahasa. *Generative Text Model*, MBART seperti BART umumnya merupakan model yang dapat menghasilkan teks baru berdasarkan pemahaman dari data yang dilatih sebelumnya. MBART menggunakan arsitektur seq2seq (*sequence-to-sequence*) yang umumnya digunakan dalam tugas terjemahan mesin.

Cara kerja MBART yaitu *training* dataset, dan inferensi. MBART dilatih dengan memberikan pasangan teks sumber dan target (dalam bahasa yang berbeda). Model ini belajar untuk memetakan teks sumber ke teks target. Setelah dilatih, MBART dapat diterapkan untuk melakukan terjemahan teks dari satu bahasa ke bahasa yang lain. Selama tahap inferensi, model menghasilkan urutan token yang mewakili terjemahan yang diinginkan.

MBART melakukan *pre-training* model *autoregressive* Seq2Seq lengkap. MBART dilatih satu kali untuk semua bahasa, menyediakan serangkaian parameter yang dapat disesuaikan untuk pasangan bahasa mana pun baik dalam *supervised* dan *unsupervised settings*, tanpa modifikasi atau skema inisialisasi khusus tugas atau bahasa tertentu (Liu et al., 2020).

## 2.9. *Bilingual Evaluation Understudy (BLEU Score)*

Salah satu pengujian yang banyak digunakan pada mesin penterjemah adalah *Bilingual Evaluation Understudy* (BLEU). Matriks BLEU dibuat untuk mengukur seberapa dekat keluaran yang dihasilkan dengan mencocokkan panjang frasa variabel keluaran mesin penterjemah dengan referensi terjemahan (Fauziyah, 2022). BLEU *score* mengukur sejauh mana teks terjemahan mendekati teks referensi atau jawaban manusia yang dianggap sebagai referensi yang benar. Rumus BLEU *score* secara umum adalah sebagai berikut:

$$\text{BLEU} = \text{BP} \times \exp \sum_{n=1}^N \frac{1}{N} \times \log(\text{precision}_n)$$

Keterangan :

1. BP adalah penghitung *brevity penalty* yang dirancang untuk mengurangi skor BLEU jika terjemahan terlalu pendek dibandingkan dengan referensi.

2. N adalah urutan maksimum yang digunakan dalam perhitungan BLEU.
3.  $\text{Precision}_n$  adalah urutan nilai presisi untuk n-gram, dihitung sebagai jumlah n-gram yang cocok antara terjemahan dan referensi dibagi dengan total n-gram dalam terjemahan.

BP dapat dihitung sebagai berikut:

$$\text{BP} = \begin{cases} 1, \\ \exp(1 - \frac{\text{panjang referensi}}{\text{panjang terjemahan}}) \end{cases}$$

Panjang terjemahan mengacu pada jumlah kata dalam terjemahan, dan panjang referensi mengacu pada jumlah kata dalam referensi terjemahan manusia.

#### **2.10. Framework Flask**

Flask merupakan *micro web framework* yang menggunakan bahasa python. Dikembangkan oleh Armin Ronacher, Flask dirilis pertama kali di tahun 2010. Meskipun Flask adalah framework yang relatif baru dibandingkan dengan beberapa framework web lainnya, keunggulannya terletak pada kesederhanaan dan fleksibilitasnya. Menggunakan framework ini dengan tujuan mempercepat pengembangan aplikasi, karena dalam Flask telah tersedia *library* dan kumpulan kode program yang siap digunakan untuk membuat aplikasi web tanpa perlu membuatnya dari dasar atau awal (Ghimire, 2020). Disamping itu, penggunaan Flask juga dapat mengurangi penggunaan sumber daya memori dikarenakan Flask termasuk dalam kategori *micro-framework* (Relan, 2019).

#### **2.11. Penelitian yang Relevan**

Berikut adalah penelitian atau studi sebelumnya yang relevan dengan penelitian yang akan dijalankan, yakni:

1. Pada penelitian Navarro, & Casacuberta (2023) “*Exploring Multilingual Pretrained Machine Translation Models for Interactive Translation*” membandingkan efektivitas model terjemahan mesin multibahasa yang telah

*dipretraining* dengan model yang dapat dilatih dari awal dalam bidang IMT. Kedua model ini mencapai hasil yang serupa, meskipun MBART unggul dalam pasangan bahasa di mana bahasa targetnya adalah bahasa Inggris. Selanjutnya, dengan melakukan *fine-tuning* pada model yang telah dipretraining di domain khusus, pengurangan upaya manusia lebih ditingkatkan, melampaui model. Hal ini mengkonfirmasi bahwa model yang telah dipretraining juga dapat memberikan hasil yang baik dalam bidang ini setelah menyesuaikan model untuk domain tertentu. Dengan melakukan *fine-tuning* daripada melatih model terjemahan dari awal, dapat mengurangi secara signifikan biaya komputasional yang terkait dengan pelatihan. Dapat disimpulkan bahwa MBART dengan *fine-tuning* mencapai hasil yang lebih baik dalam bidang IMT dibandingkan dengan melatih model dari awal.

2. Pada penelitian (Fauziyah, 2022) "Mesin penterjemah Bahasa Indonesia-Bahasa Sunda Menggunakan *Recurrent Neural Networks*". Sistem penerjemahan otomatis dari Bahasa Indonesia ke Bahasa Sunda dikembangkan dengan menggunakan pendekatan *neural machine translation* dan menerapkan struktur *Encoder-Decoder* dengan penyisipan RNN di dalamnya. Kumpulan data yang digunakan terdiri dari 3496 pasangan kalimat paralel dalam Bahasa Indonesia dan Bahasa Sunda. Beberapa pengujian dilakukan, termasuk pengujian terhadap variasi arsitektur RNN, penerapan model NMT dengan dan tanpa attention, uji coba optimalisasi model, dan evaluasi BLEU Score. Hasil pengujian pertama menunjukkan bahwa arsitektur model dengan variasi RNN mencapai nilai optimal menggunakan GRU, dengan tingkat akurasi mencapai 99.17%. Pengujian kedua menunjukkan bahwa model dengan penerapan attention mencapai nilai optimal dengan tingkat akurasi sebesar 99.94%. Pengujian ketiga, yang membandingkan model hasil optimal setelah proses optimisasi, menunjukkan tingkat akurasi sebesar 99.35%. Pengujian terakhir, dengan mengukur BLEU Score, menunjukkan nilai optimal sebesar 92.63%, dengan *brevity penalty* sebesar 0.929.
3. Pada penelitian Wang, et al (2022) dengan judul "*Understanding and Improving Sequence-to-Sequence Pretraining for Neural Machine*

*Translation*” untuk menyelidiki pengaruh dari pretraining Seq2Seq untuk terjemahan mesin neural (NMT) dan untuk mengusulkan strategi yang efektif untuk mengatasi tantangan yang diidentifikasi. Penelitian ini bertujuan memahami dampak dari *pretraining* Seq2Seq untuk NMT dan untuk mengusulkan solusi praktis untuk meningkatkan kinerja model dan ketangguhan. Hasilnya menunjukkan bahwa *pretraining* Seq2Seq dapat menghasilkan terjemahan yang lebih beragam dengan urutan kata yang berbeda, mengurangi kesalahan terjemahan tetapi juga menyebabkan perbedaan domain dan objektif. Selain itu, penelitian ini juga mengusulkan dua strategi, yaitu *pretraining* di dalam domain dan adaptasi input dalam *fine tuning*, yang secara konsisten meningkatkan kinerja terjemahan model pada *pretraining* Seq2Seq.

4. Dalam penelitian Alam, & Suryani (2021) “*Minang and Indonesian Phrase Based Statistical Machine Translation*” Berfokus pada pengembangan mesin penerjemah statistik dari bahasa Minang ke Indonesia, penelitian ini juga mengevaluasi kualitas terjemahan. Sumber data untuk pelatihan dan pengujian mencakup korpus paralel dan monolingual yang diperoleh dari Wikipedia dalam bahasa Minang serta situs berita berbahasa Indonesia. Semua konfigurasi diuji menggunakan 600 baris kalimat sebagai data uji. Evaluasi kualitas terjemahan dilakukan dengan menggunakan metode BLEU *score*. Hasil pengujian enam konfigurasi menunjukkan peningkatan akurasi mesin penerjemah usai menambahkan jumlah korpus monolingual dan paralel. Dalam skenario satu, terdapat peningkatan sebesar 3,6% pada konfigurasi 3 dan 2, sedangkan konfigurasi 2 dan 1 meningkat sebesar 2,59%. Pada skenario kedua, terdapat peningkatan sebesar 0,44% pada konfigurasi 5 dan 4, dan konfigurasi 4 dan 1 meningkat sebesar 0,06%. Hasil tersebut memperlihatkan bahwa pengujian pada skenario satu memiliki konsekuensi yang lebih signifikan dibandingkan dengan pengujian pada skenario kedua dalam konteks kualitas penerjemahan. Kendala utama yang dihadapi adalah kurangnya sumber korpus, yang mempengaruhi kemampuan mesin penerjemah dalam menghasilkan terjemahan yang akurat.

5. Pada penelitian Tang, et al (2021) “*Multilingual Translation from Denoising Pre-Training*”, bertujuan untuk mengeksplorasi potensi gabungan pretraining denoising dengan terjemahan mesin multibahasa dalam satu model dan untuk memperkenalkan konsep *fine-tuning* multibahasa dan mengevaluasi efektivitasnya dalam menciptakan model terjemahan multibahasa. Hasil penelitian menunjukkan bahwa dengan *pretraining* memberikan manfaat yang signifikan dan menyediakan dataset benchmark ML50 untuk pelatihan dan evaluasi standar model terjemahan multibahasa. Selain itu, penelitian ini juga memperkenalkan metode *fine-tuning* multibahasa dari model MBART50, yang menunjukkan peningkatan kinerja yang signifikan dalam terjemahan. Metode ini juga berhasil mengungguli pelatihan bilingual dari awal dan pelatihan multibahasa dari awal, serta memiliki kinerja yang mirip dengan *fine-tuning* bilingual. Studi ini juga menemukan bahwa *fine-tuning* multibahasa lebih baik dalam pengaturan *Many-to-one* namun lebih buruk dalam pengaturan *One-to-Many* dan *Many-to-Many*. Selain itu, penelitian ini juga menunjukkan bahwa *pretraining* kontinu efektif dan dapat memperluas dukungan untuk bahasa tambahan tanpa kehilangan kinerja. Penelitian ini juga memperkenalkan *benchmark* ML50 untuk pelatihan dan evaluasi pada 50 bahasa dan mendiskusikan arah penelitian kedepannya.
6. Pada Penelitian Liu, et al (2020) “*Multilingual Denoising Pre-training for Neural Machine Translation*” untuk menunjukkan bahwa pre-training multibahasa denoising dapat secara signifikan meningkatkan kinerja mesin terjemahan baik yang *supervised* maupun *unsupervised*, baik pada tingkat kalimat maupun dokumen. Penelitian ini juga menganalisis kapan dan bagaimana *pre-training* paling efektif dan dapat dikombinasikan dengan pendekatan lain seperti *back-translation*. Hasil penelitian juga menunjukkan kemampuan transfer *learning* dari representasi yang dipelajari dari *pre-training* multibahasa. Ditemukan bahwa pre-training multibahasa *denoising* dapat meningkatkan kinerja mesin terjemahan baik pada tingkat kalimat maupun dokumen. Hasilnya juga menunjukkan bahwa model *pre-trained* MBART25 menghasilkan terjemahan yang lebih baik daripada model yang tidak *dipre-training*, terutama pada tingkat dokumen.



7. Pada penelitian yang dilakukan oleh Wibowo, et al (2020). “*Semi-Supervised Low-Resource Style Transfer of Indonesian Informal to Formal Language with Iterative Forward-Translation.*” membangun data paralel informal - formal baru di Indonesia dengan memberi anotasi pada *tweet* dari layanan pelanggan domain. Didapatkan bahwa GPT-2 dan PBSMT telah dilatih sebelumnya pendekatan mencapai kinerja terbaik dalam hal BLEU. Pelatihan dengan kumpulan data sintetis tambahan dalam teks bahasa Indonesia informal yang diterjemahkan ke depan meningkatkan performa.
8. Dalam Penelitian yang dilakukan oleh Lewis, et al (2019) “*BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*” penelitian ini menggunakan arsitektur terjemahan mesin *neural* berbasis transformer dan efektif untuk tugas generasi teks, pemahaman, dan terjemahan mesin. BART mengungguli skema *pretraining* lainnya dan mendukung berbagai skema *noising* selama *pretraining*. Ini dapat disesuaikan untuk klasifikasi urutan, klasifikasi token, generasi urutan, dan tugas terjemahan mesin. Model ini terbukti mencapai hasil terbaik dalam berbagai tugas pemrosesan bahasa alami.
9. Pada penelitian Dwiastuti (2019) “*English-Indonesian Neural Machine Translation for Spoken Language Domains*” dalam penelitian ini, tentang *Neural Machine Translation* (NMT) untuk Bahasa Inggris-Indonesia (EN-ID) dan Bahasa Indonesia-Inggris (ID-EN). Berfokus kepada domain bahasa lisan, yaitu bahasa sehari-hari dan bahasa ucapan. Kami membangun sistem NMT menggunakan model Transformer untuk arah terjemahan dan menerapkan adaptasi domain, di mana kami melatih sistem NMT terlatih kami pada data bahasa ucapan (dalam domain). Selain itu, dilakukan evaluasi bagaimana metode adaptasi domain pada sistem EN-ID dapat menghasilkan keluaran terjemahan yang lebih formal.

## **BAB III**

### **ANALISIS DAN PERANCANGAN SISTEM**

#### **3.1. Analisis Sistem**

Cara kerja mendesain serta mengembangkan suatu sistem salah satu langkah pertama yang dilakukan adalah dengan menganalisis sistemnya terlebih dahulu. Analisis sistem bertujuan untuk mengetahui masalah-masalah yang muncul dengan aplikasi agar aplikasi dapat berfungsi dan bekerja baik. Proses analisis sistem terdapat dua tahapan dalam menganalisis, menganalisa masalah dan juga menganalisa apa saja yang dibutuhkan oleh program.

##### **3.1.1. Analisis Masalah**

Dalam konteks ini permasalahannya yang dicakup ialah membuat suatu model dan sistem untuk mengubah atau menerjemahkan bahasa Indonesia yang tidak baku ke bahasa Indonesia baku.

##### **3.1.2. Analisis Kebutuhan**

Kegunaan analisis kali ini dalam perancangan sebuah aplikasi ialah agar dapat mengidentifikasi data yang nantinya diolah sistem. Analisis kebutuhan melibatkan identifikasi mengenai fungsional dan non-fungsional yang nantinya digunakan dalam fase perancangan aplikasi agar dapat berjalan dan dapat memenuhi tujuan dari aplikasi yang dibuat.

###### **3.1.2.1. Kebutuhan Fungsional**

Gabungan berbagai proses yang nantinya dijalankan oleh sistem yang dibangun yang disebut juga sebagai kebutuhan fungsional. Berikut adalah kebutuhan fungsional yang ada pada sistem:

1. Sistem dapat memberikan masukan atau inputan dari user untuk dilakukannya penerjemahan bahasa berupa kata atau kalimat non formal bahasa Indonesia.
2. Sistem dapat menampilkan hasil terjemahan yaitu bahasa formal bahasa indonesia.

3. Sistem dapat menghitung dan menampilkan hasil akurasi dari penerjemahan tersebut dengan BLEU Score.

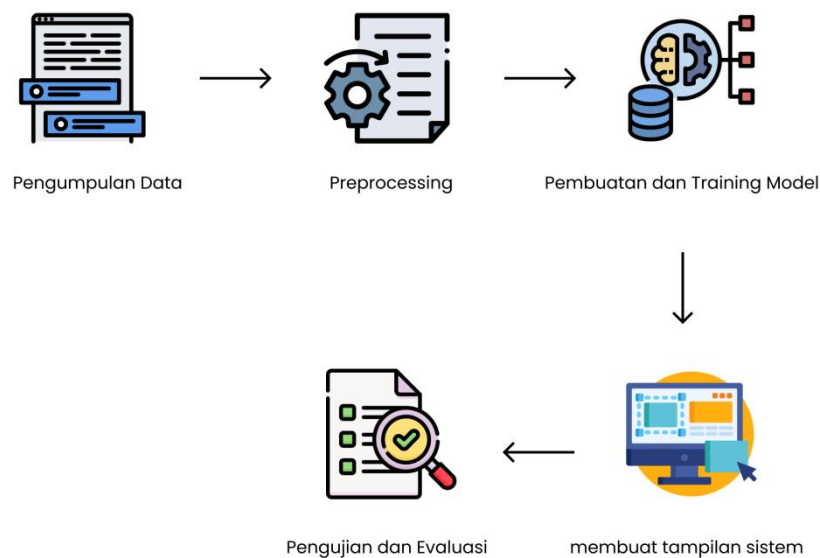
### 3.1.2.2. Kebutuhan Non-Fungsional

Kebutuhan non-fungsional ialah hal yang membahas berupa fitur, karakteristik, makna atau kegunaan yang dapat digunakan oleh aplikasi seperti waktu, batasan pengembangan proses, dan standarisasi sistem yang digunakan sebagai pelengkap. Berikut adalah kebutuhan fungsional yang dibutuhkan dalam sistem ini:

1. Tampilan sistem mudah dipahami agar mudah dimengerti oleh pengguna.
2. Dapat memberikan informasi jika ada kesalahan yang dilakukan pengguna seperti pengguna belum melakukan inputan kata non formal.
3. Perangkat harus terhubung ke internet agar dapat menjalankan sistem.

### 3.1.3. Arsitektur Umum Sistem

Pada tahap ini merupakan tata cara pengerjaan secara dasar dari sistematis langkah-langkah aplikasi yang hendak dibuat pada permasalahan ini, yang bisa terlihat digambar 3.1.



**Gambar 3.1** Arsitektur Umum Sistem

Terdapat beberapa tahapan, yaitu pengumpulan data, membersihkan data, pembuatan dan training model, membuat tampilan sistem serta pengujian. Berikut penjelasan alur proses diagram umum sistem pada gambar di atas:

1. Pengumpulan data bertujuan untuk mengumpulkan dataset yang diperlukan. Data yang diambil adalah bahasa Indonesia informal. Dataset berasal dari media sosial twitter, dikarenakan bahasa Indonesia informal banyak ditemukan di media sosial terutama twitter. Pengerjaan ini dilakukan pengumpulan data *tweet* yang dibutuhkan dalam pengerjaan ini. Pada penelitian ini *tweet* yang diambil adalah dengan kata kunci ‘kuliah luring’ pada media sosial twitter sebanyak 4000 data.
2. Tahapan preprocessing pada penelitian ini untuk membersihkan dataset. Proses pembersihan ini digunakan agar dataset yang digunakan hanya kata - kata yang diperlukan. Preprocessing meliputi proses *case folding* (merubah seluruh data *tweet* menjadi huruf kecil) tahap ini juga yang melibatkan penghilangan karakter tertentu seperti tanda baca. Setelah data dibersihkan kemudian dilakukan penerjemahan data tersebut kedalam bahasa formal. Terjemahan dibantu dengan kamus slang kemudian dilakukan pengoreksian kembali secara manual untuk mendapatkan dan menambah bahasa non formal maupun formal yang tidak ada didalam kamus slang agar terjemahan menjadi lebih baik. Pasangan bahasa sumber dan bahasa target disatukan menjadi korpus untuk model terjemahan.
3. Pembuatan dan *training* model, setelah mendapatkan dataset dan melakukan preprocessing kemudian membuat suatu model untuk penerjemahan bahasa tersebut. Proses ini menggunakan metode pelatihan berbasis Transformer, dan khususnya menggunakan model MBART untuk tugas penerjemahan. Ini merupakan contoh penggunaan *Hugging Face* Transformers untuk tugas penerjemahan mesin dengan model Transformer.
4. Membuat tampilan sistem dengan menggunakan framework flask, terdapat beberapa halaman yaitu halaman *home*, data, program, dan *history* beserta evaluasi. Halaman home merupakan halaman utama pada sistem. Halaman data berisikan contoh data kalimat bahasa non formal dan bahasa formal bahasa Indonesia sebagai referensi untuk inputan pengguna. Halaman

program merupakan halaman untuk menerjemahkan bahasa, didalamnya terdapat tampilan untuk pengguna memasukkan kata atau bahasa non formal yang akan diterjemah ke bahasa formal dan hasil terjemahannya. Halaman riwayat merupakan halaman untuk menyimpan data yang telah diinputkan oleh pengguna, lalu terdapat juga suatu halaman untuk melakukan evaluasi atau pengukuran akurasi BLEU *score* pada data yang telah diinputkan oleh pengguna.

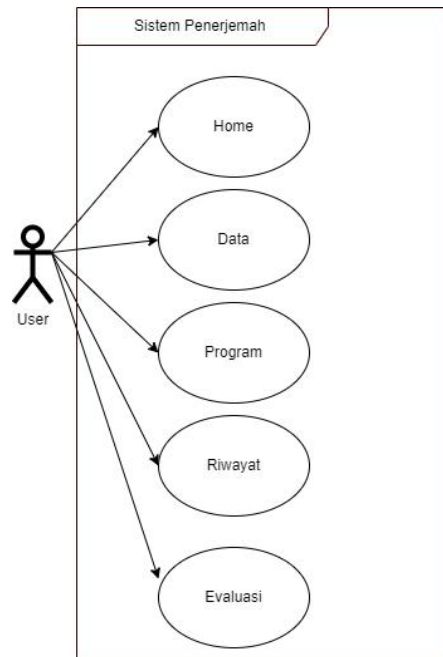
5. Pengujian dan evaluasi dapat dilakukan didalam sistem. Untuk pengujian dilakukan pada halaman program yang berisikan halaman untuk menerjemahkan kata atau kalimat. Evaluasi dilakukan setelah pengguna telah menerjemah kata atau kalimat. Evaluasi dilakukan untuk mengukur keakuratan hasil terjemahan dengan sistem yang digunakan.

### **3.2. Pemodelan Sistem**

Tahap ini ialah perpanjangan tahapan interaksi antara pengguna dengan sistem yang dibangun agar dapat berfungsi secara maksimal. Pada riset kali ini memakai *activity diagram*, *use case diagram* serta *flowchart* sebagai bentuk pemodelan.

#### **3.2.1. Use Case Diagram**

Pemodelan korelasi diantara user dengan sistem yang sedang dibangun merupakan pengertian dari *Use Case*. Pembuatan diagram ini dapat membantu dalam memahami fungsionalitas sistem dan hubungan antar pelaku dan kasus pengguna. Berikut merupakan contoh diagram kasus pengguna pada aplikasi yang dibuat.



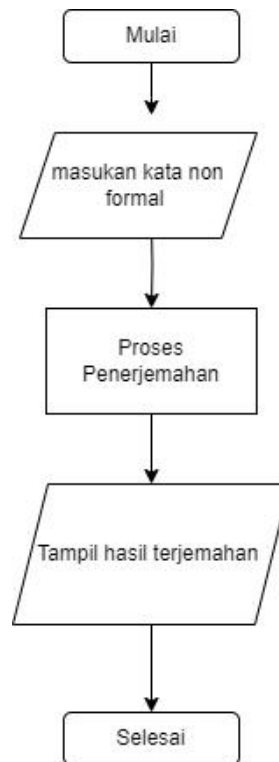
**Gambar 3.2** *Use Case Diagram*

Diagram di atas menunjukkan interaksi dapat dilakukan oleh *user* dengan sistem yang dibangun. Pengguna dapat menggunakan fitur untuk melihat contoh data yang berisikan bahasa non formal pada halaman data serta dapat menggunakan sistem untuk melakukan terjemahan bahasa non formal ke bahasa formal bahasa Indonesia.

### **3.2.2. Activity Diagram**

*Activity Diagram* merupakan penggambaran proses cara kerja yang dibuat sedari awal sampai terakhir sistem. *Activity Diagram* juga dapat digunakan untuk menjelaskan potongan komponen dari *use case diagram*.





**Gambar 3.4** Flowchart Sistem

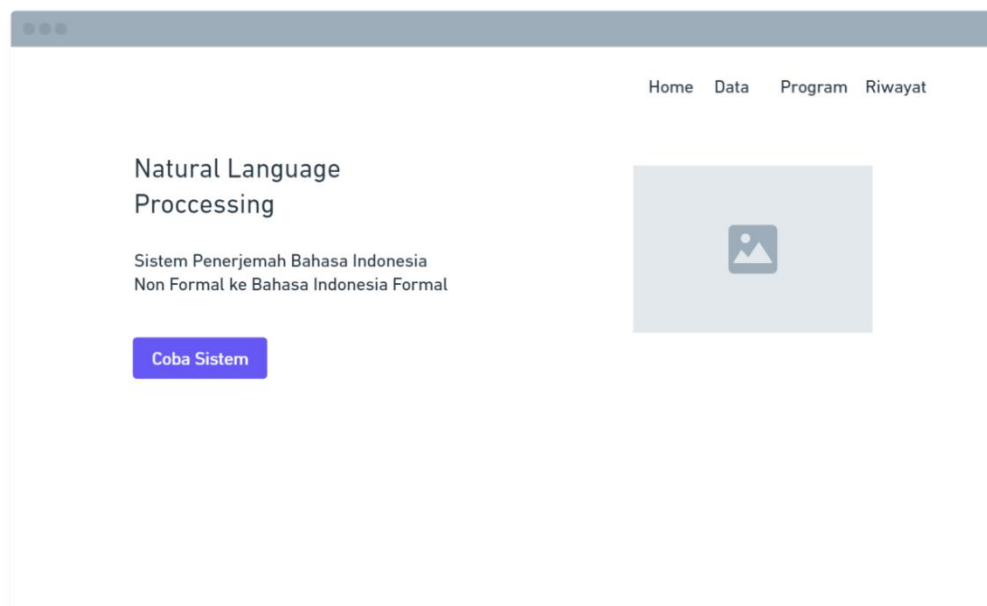
Gambar di atas menjelaskan bagaimana tata cara kerja sistem yang dibuat. Setelah pengguna membuka sistem, pengguna memasukkan kata non formal yang ingin diterjemah, setelah itu sistem menerjemah kata atau kalimat non formal yang dimasukan oleh pengguna dengan model yang dibuat. Lalu setelah proses penerjemahan selesai maka muncul hasil dari terjemahan tersebut.

### **3.4. Perancangan *Interface***

Tahapan dalam pembuatan desain dari tampilan tatap muka daripada suatu sistem dapat dikatakan sebagai tahap perancangan *interface*, Perancangan *Interface* diperlukan agar pada saat pembuatan sistem memiliki patokan pembuatan serta desain yang dibuat.



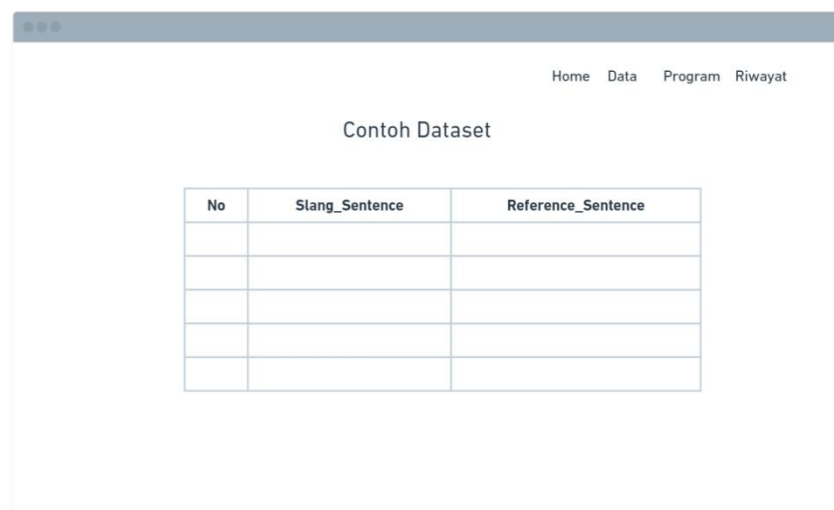
### 3.4.1. Halaman Home



**Gambar 3.5** Halaman *Home*

Halaman ini merupakan antarmuka pertama saat sistem digunakan, pengguna dapat memilih menuju ke halaman data, program, ataupun riwayat.

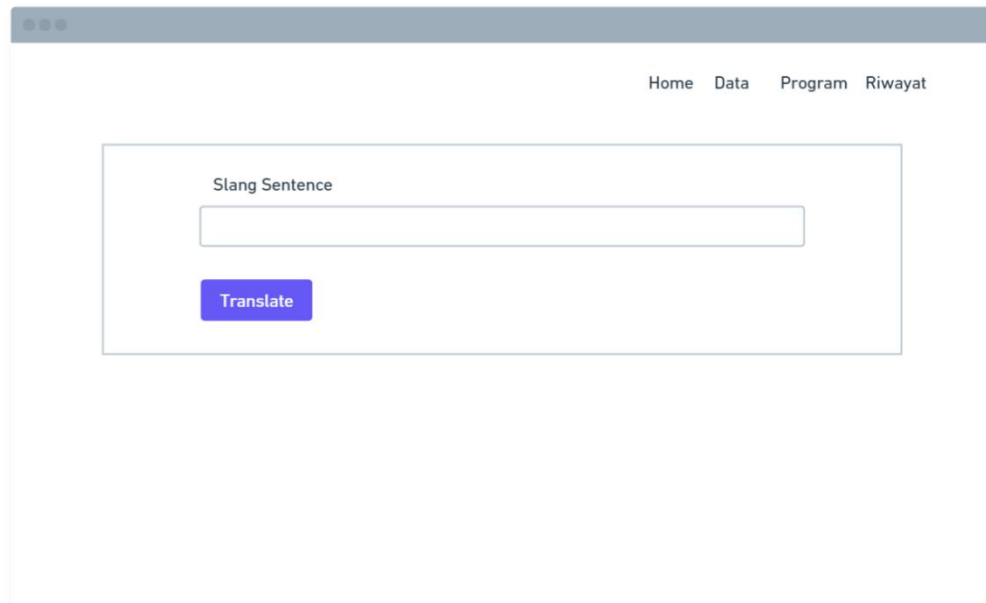
### 3.4.2. Halaman Data



**Gambar 3.6** Halaman Data

Halaman di atas merupakan halaman data, berisikan contoh dataset sebagai referensi masukan untuk pengguna.

### 3.4.3. Halaman Program



The screenshot shows a web application interface. At the top, there is a navigation bar with four links: "Home", "Data", "Program", and "Riwayat". The "Program" link is currently selected. Below the navigation bar, there is a main content area. Inside this area, there is a form titled "Slang Sentence". The form consists of a single-line text input field. Below the input field, there is a blue button with the text "Translate" in white.

**Gambar 3.7** Halaman Program

Halaman ini adalah tampilan halaman program untuk melakukan penerjemahan bahasa non formal ke bahasa formal bahasa Indonesia. Gambar 3.7 merupakan tampilan awal sebelum melakukan penerjemahan atau menekan *button translate*.

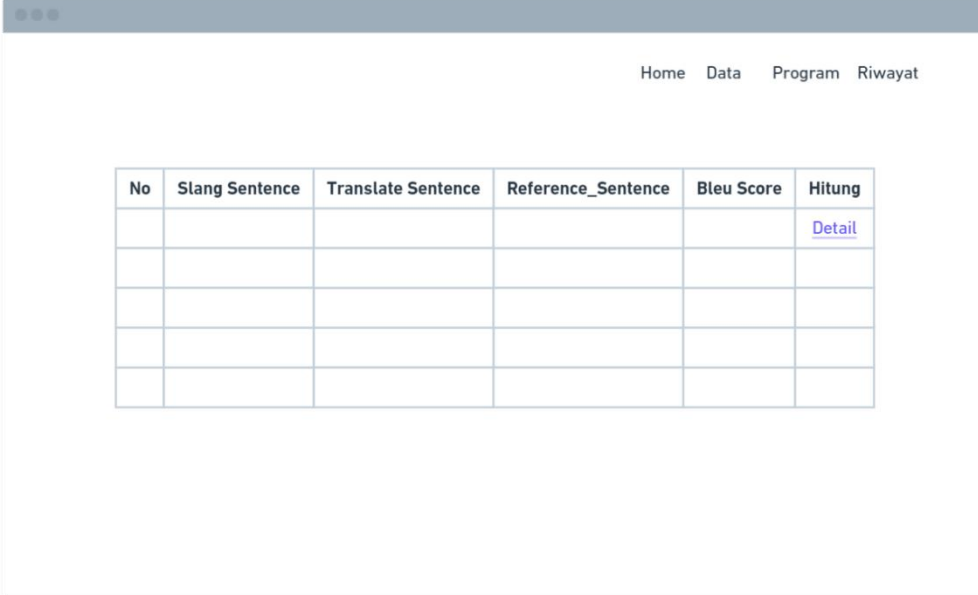
The screenshot shows a web application window with a navigation bar at the top containing links: Home, Data, Program, and Riwayat. The main content area is divided into two sections. The first section, titled 'Slang Sentence', contains a text input field and a blue 'Translate' button. The second section, titled 'Hasil', contains a table with two columns: 'Slang\_Sentence' and 'Hasil Terjemahan'. The table has one empty row for displaying results.

| Slang_Sentence | Hasil Terjemahan |
|----------------|------------------|
|                |                  |

**Gambar 3.8** Halaman program setelah melakukan terjemahan

Gambar di atas merupakan tampilan halaman program setelah melakukan penerjemahan atau setelah menekan *button* translate. Tampilan menghasilkan keluaran yang berisikan kata atau kalimat yang dimasukan oleh pengguna dan hasil terjemahan dari masukan tersebut.

### 3.4.4. Halaman Riwayat

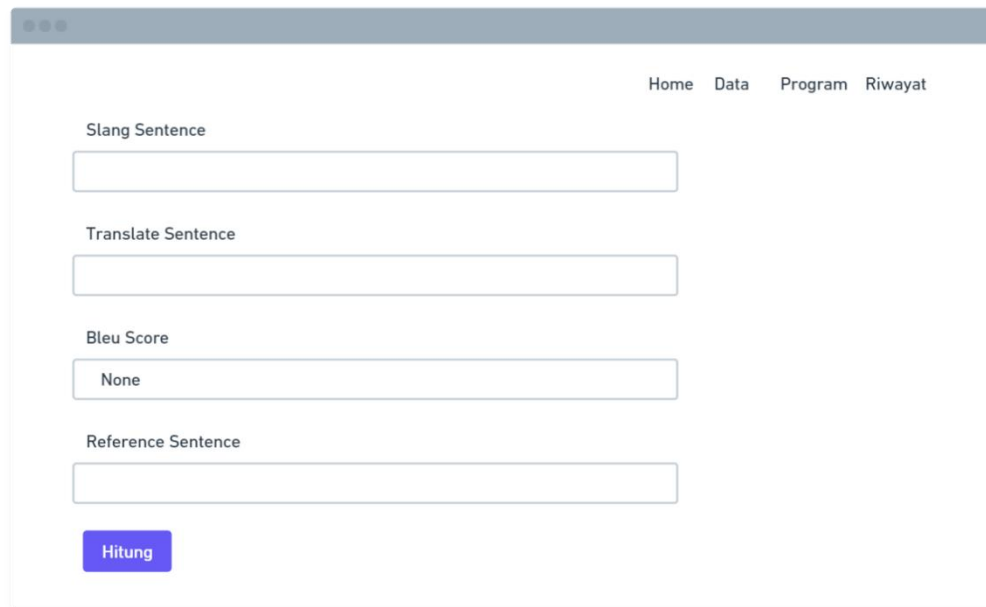


| No | Slang Sentence | Translate Sentence | Reference_Sentence | Bleu Score | Hitung                 |
|----|----------------|--------------------|--------------------|------------|------------------------|
|    |                |                    |                    |            | <a href="#">Detail</a> |
|    |                |                    |                    |            |                        |
|    |                |                    |                    |            |                        |
|    |                |                    |                    |            |                        |
|    |                |                    |                    |            |                        |

**Gambar 3.9** Halaman riwayat

Gambar di atas merupakan tampilan halaman riwayat yang berisikan kata atau kalimat bahasa non formal bahasa Indonesia yang telah dimasukan pengguna dan hasil setelah melakukan penerjemahan atau setelah menekan *button* translate dan hasil akurasi BLEU *score* jika telah melakukan perhitungan akurasi. Terdapat juga suatu link didalam kolom hitung yang menghubungkan ke tampilan halaman untuk menghitung akurasi penerjemahan yang dilakukan sistem.

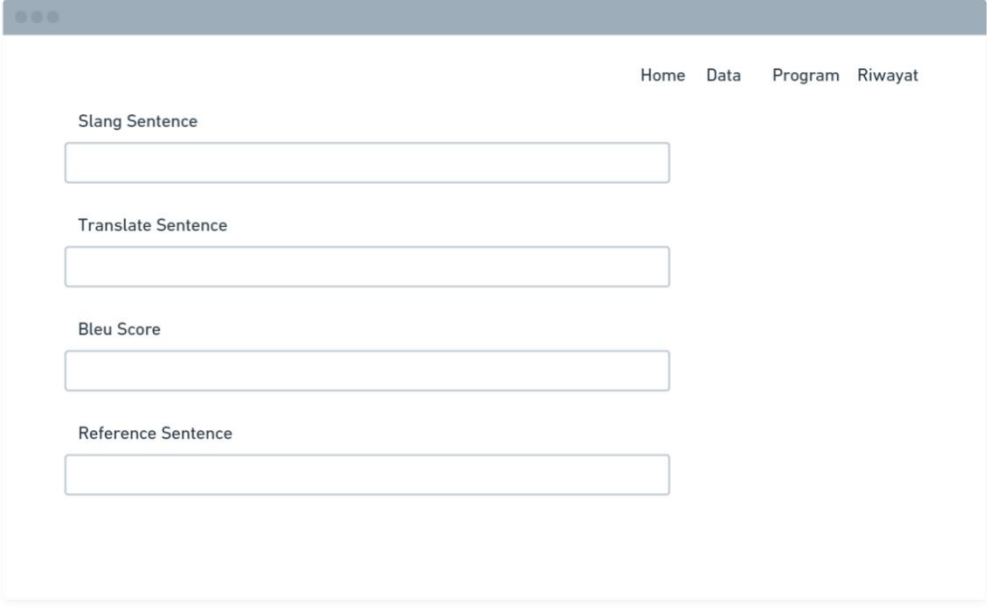
### 3.4.5. Halaman BLEU Score



The image shows a web application interface for calculating BLEU scores. It features a navigation bar with links: Home, Data, Program, and Riwayat. The main form contains four input fields: 'Slang Sentence', 'Translate Sentence', 'Bleu Score' (with a dropdown menu currently showing 'None'), and 'Reference Sentence'. A blue 'Hitung' (Calculate) button is positioned at the bottom left of the form area.

**Gambar 3.10** Halaman BLEU *score*

Gambar di atas merupakan desain daripada halaman BLEU score atau menghitung akurasi dari sistem. Tampilan di atas merupakan tampilan utama BLEU Score sebelum pengguna menghitung akurasi atau sebelum menekan *button* hitung.



The image shows a web application interface with a light blue header bar containing three small circles. Below the header, there is a navigation menu with the links "Home", "Data", "Program", and "Riwayat". The main content area is white and contains four input fields, each with a label above it: "Slang Sentence", "Translate Sentence", "Bleu Score", and "Reference Sentence". Each label is followed by a rectangular input box.

**Gambar 3.11** Halaman BLEU score setelah dihitung

Gambar di atas merupakan desain daripada halaman BLEU *score* setelah menghitung akurasi dari sistem. Hasil dari BLEU *score* tampil pada bagian BLEU *score* dan hasil masuk ke dalam halaman riwayat.

## **BAB IV**

### **IMPLEMENTASI DAN PENGUJIAN SISTEM**

#### **4.1. Implementasi**

Penelitian kali ini, sistem yang dibangun dengan mengimplementasikan metode *semi-supervised translation*. Sistem nantinya dibangun menggunakan bahasa pemrograman python serta mengusung *framework* Flask.

Agar dapat mengimplementasi dan melakukan pengujian terhadap penerjemahan bahasa non formal bahasa Indonesia ke bahasa formal bahasa Indonesia dengan menggunakan pendekatan *semi supervised translation*, perangkat keras dan lunak yang digunakan memiliki spesifikasi sebagai berikut:

1. *Processor* Intel(R) Core(TM) i5-8300H CPU @ 2.30GHz 2.3 GHz.
2. RAM 8GB.
3. Sistem Operasi Windows 11 Home Single Language 64-Bit.
4. Browser Google Chrome.
5. Visual Studio Code.
6. Framework flask.
7. Google Colaboratory.
8. Python 3.10.5.

##### **4.1.1. Pengumpulan Data**

Untuk mendapatkan kalimat atau kata-kata tidak baku pada media sosial twitter maka dilakukan proses crawling data pada media sosial twitter. Pada penelitian ini penulis mengambil *tweet* terlihat pada Gambar 4.1. *Tweet* yang diambil dengan menggunakan kata kunci 'kuliah luring' sebanyak 4000 data *tweet*.

```
# Creating list to append tweet data to
tweets = []

# Using TwitterSearchScrapper to scrape data and append tweets to list
for i,tweet in enumerate(sntwitter.TwitterSearchScrapper('kuliah luring lang:id').get_items()):
    if len(tweets) >= 4000:
        break
    if 'kuliah luring' in tweet.content:
        tweets.append(tweet)

|

tweets_df = pd.DataFrame(tweets)
tweets_df['date'] = tweets_df['date'].apply(lambda a: pd.to_datetime(a).date())
tweets_df
```

**Gambar 4.1** Proses pengambilan data

|    | A  | B   | C          | D            | E             | F        | G          | H          | I          | J         | K        | L       | M     | N  | O  | P                              | Q   | R           | S   | T      | U         | V        |         |
|----|----|---|------------|--------------|---------------|----------|------------|------------|------------|-----------|----------|---------|-------|--|--|--------------------------------|---|-------------|---|--------|-----------|----------|---------|
| 1  |    | url   | date       | content      | ipedia        | id       | user       | replyCount | tweetCount | likeCount | source   | version | lang  | source   | sourceIp   | sourceLabel                    | outlinks                                    | ccountlinks | media                                       | weeted | retweeted | tweetip  | tweetip |
| 2  | 0  | <a href="https://www.ipedia.co.id/">https://www.ipedia.co.id/</a>   | 2023-08-18 | JAPAN SAKAI  | JAPAN SAKAI   | 1,59E+18 | @username  | 0          | 0          | 0         | 1,59E+18 | 1       | <ref= | https://www.ipedia.co.id/                                  | https://www.ipedia.co.id/                                  | Twitter for Android            |   |             | [[previewUrl]: https://pbs.twimg.com/media/ |        |           |          |         |
| 3  | 1  | <a href="https://twitter.com/watakematsu/status/1587646850812919809">https://twitter.com/watakematsu/status/1587646850812919809</a> |            |              |               |          | @username  | 0          | 0          | 0         | 1,59E+18 | 1       | <ref= | https://twitter.com/watakematsu/status/1587646850812919809 | https://twitter.com/watakematsu/status/1587646850812919809 | Twitter for Android            |   |             | [[previewUrl]: https://pbs.twimg.com/media/ |        |           |          |         |
| 4  | 2  | <a href="#">Click to follow the link</a>  |            |              |               |          | @username  | 0          | 0          | 0         | 1,59E+18 | 1       | <ref= | https://twitter.com/watakematsu/status/1587646850812919809 | https://twitter.com/watakematsu/status/1587646850812919809 | Twitter for Android            |   |             | [[previewUrl]: https://pbs.twimg.com/media/ |        |           |          |         |
| 5  | 3  | <a href="https://www.collegienet.co.id/">https://www.collegienet.co.id/</a>   | 2023-08-18 | gynjuvuv     | gynjuvuv      | 1,59E+18 | @username1 | 1          | 1          | 0         | 1,59E+18 | 1       | <ref= | https://www.collegienet.co.id/                             | https://www.collegienet.co.id/                             | Twitter Web App                |   |             |   |        |           | 1,59E+18 |         |
| 6  | 4  | <a href="https://www.gynjuvuv.co.id/">https://www.gynjuvuv.co.id/</a>   | 2023-08-18 | gynjuvuv     | gynjuvuv      | 1,59E+18 | @username1 | 1          | 1          | 0         | 1,59E+18 | 1       | <ref= | https://www.gynjuvuv.co.id/                                | https://www.gynjuvuv.co.id/                                | Twitter for iPhone             |   |             |   |        |           | 1,59E+18 |         |
| 7  | 5  | <a href="https://www.orang.id/">https://www.orang.id/</a>   | 2023-08-18 | orang        | orang         | 1,59E+18 | @username2 | 2          | 0          | 1         | 1,59E+18 | 1       | <ref= | https://www.orang.id/                                      | https://www.orang.id/                                      | Twitter for Android            |   |             |   |        |           |          |         |
| 8  | 6  | <a href="https://www.gynjuvuv.co.id/">https://www.gynjuvuv.co.id/</a>   | 2023-08-18 | gynjuvuv     | gynjuvuv      | 1,59E+18 | @username2 | 0          | 0          | 0         | 1,59E+18 | 1       | <ref= | https://www.gynjuvuv.co.id/                                | https://www.gynjuvuv.co.id/                                | Twitter for Android            |   |             |   |        |           |          |         |
| 9  | 7  | <a href="https://www.collegienet.co.id/">https://www.collegienet.co.id/</a>   | 2023-08-18 | gynjuvuv     | gynjuvuv      | 1,59E+18 | @username2 | 3          | 0          | 1         | 1,59E+18 | 1       | <ref= | https://www.collegienet.co.id/                             | https://www.collegienet.co.id/                             | Pakai masker! Stay safe! - Bot |   |             |   |        |           |          |         |
| 10 | 8  | <a href="https://www.sapa.id/">https://www.sapa.id/</a>   | 2023-08-18 | sapa         | sapa          | 1,59E+18 | @username2 | 0          | 0          | 0         | 1,59E+18 | 1       | <ref= | https://www.sapa.id/                                       | https://www.sapa.id/                                       | Twitter for Android            |   |             |   |        |           |          |         |
| 11 | 9  | <a href="https://www.semenjak.id/">https://www.semenjak.id/</a>   | 2023-08-18 | semenjak     | semenjak      | 1,59E+18 | @username2 | 2          | 0          | 1         | 1,59E+18 | 1       | <ref= | https://www.semenjak.id/                                   | https://www.semenjak.id/                                   | Twitter for Android            |   |             |   |        |           |          |         |
| 12 | 10 | <a href="https://www.bisa.ga.sih.bisa.ga.sih">https://www.bisa.ga.sih.bisa.ga.sih</a>   | 2023-08-18 | bisa         | bisa          | 1,59E+18 | @username2 | 2          | 0          | 0         | 1,59E+18 | 1       | <ref= | https://www.bisa.ga.sih.bisa.ga.sih                        | https://www.bisa.ga.sih.bisa.ga.sih                        | Twitter for Android            |   |             |   |        |           |          |         |
| 13 | 11 | <a href="https://www.kulluh.kulluh.kulluh">https://www.kulluh.kulluh.kulluh</a>   | 2023-08-18 | kulluh       | kulluh        | 1,59E+18 | @username2 | 0          | 0          | 0         | 1,59E+18 | 1       | <ref= | https://www.kulluh.kulluh.kulluh                           | https://www.kulluh.kulluh.kulluh                           | Twitter for iPhone             |   |             |   |        |           |          |         |
| 14 | 12 | <a href="https://www.souyou.kouyou.kouyou">https://www.souyou.kouyou.kouyou</a>   | 2023-08-18 | souyou       | souyou        | 1,58E+18 | @username2 | 1          | 0          | 7         | 1,58E+18 | 1       | <ref= | https://www.souyou.kouyou.kouyou                           | https://www.souyou.kouyou.kouyou                           | Twitter for Android            |   |             | [[previewUrl]: https://pbs.twimg.com/media/ |        |           |          |         |
| 15 | 13 | <a href="https://www.souyou.kouyou.kouyou">https://www.souyou.kouyou.kouyou</a>   | 2023-08-18 | souyou       | souyou        | 1,58E+18 | @username2 | 1          | 0          | 1         | 1,58E+18 | 1       | <ref= | https://www.souyou.kouyou.kouyou                           | https://www.souyou.kouyou.kouyou                           | Twitter for iPhone             |   |             | [[previewUrl]: https://pbs.twimg.com/media/ |        |           |          |         |
| 16 | 14 | <a href="https://www.jukur.jukur.jukur">https://www.jukur.jukur.jukur</a>   | 2023-08-18 | jukur        | jukur         | 1,58E+18 | @username2 | 0          | 0          | 1         | 1,58E+18 | 1       | <ref= | https://www.jukur.jukur.jukur                              | https://www.jukur.jukur.jukur                              | Twitter for iPhone             |   |             |   |        |           |          |         |
| 17 | 15 | <a href="https://www.prodi.sada.prodi.sada">https://www.prodi.sada.prodi.sada</a>   | 2023-08-18 | prodi        | sada          | 1,58E+18 | @username2 | 1          | 0          | 1         | 1,58E+18 | 1       | <ref= | https://www.prodi.sada.prodi.sada                          | https://www.prodi.sada.prodi.sada                          | Twitter for iPhone             |   |             |   |        |           |          |         |
| 18 | 16 | <a href="https://www.twitter.com/tweet/158E+18">https://www.tweet/158E+18</a>   | 2023-08-18 | TWITTER      | TWITTER       | 1,58E+18 | @username2 | 1          | 0          | 0         | 1,58E+18 | 1       | <ref= | https://www.tweet/158E+18                                  | https://www.tweet/158E+18                                  | Twitter for iPhone             |   |             | [[previewUrl]: https://pbs.twimg.com/media/ |        |           |          |         |
| 19 | 17 | <a href="https://www.bisa.ga.kal.bisa.ga.kal">https://www.bisa.ga.kal.bisa.ga.kal</a>   | 2023-08-18 | bisa         |               |          | @username2 | 1          | 0          | 0         | 1,58E+18 | 1       | <ref= | https://www.bisa.ga.kal.bisa.ga.kal                        | https://www.bisa.ga.kal.bisa.ga.kal                        | Twitter for Android            |   |             |   |        |           |          |         |
| 20 | 18 | <a href="https://www.ugm.fess.ugm.fess">https://www.ugm.fess.ugm.fess</a>   | 2023-08-18 | UGM_FESS     | UGM_FESS      | 1,58E+18 | @username2 | 0          | 0          | 8         | 1,58E+18 | 1       | <ref= | https://www.ugm.fess.ugm.fess                              | https://www.ugm.fess.ugm.fess                              |                                |   |             |   |        |           |          |         |
| 21 | 19 | <a href="https://www.gynjuvuv.co.id/">https://www.gynjuvuv.co.id/</a>   | 2023-08-18 | gynjuvuv     | gynjuvuv      | 1,58E+18 | @username2 | 0          | 0          | 1         | 1,58E+18 | 1       | <ref= | https://www.gynjuvuv.co.id/                                | https://www.gynjuvuv.co.id/                                | Twitter Web App                |   |             |   |        |           |          |         |
| 22 | 20 | <a href="https://www.mampus...">https://www.mampus...</a>   | 2023-08-18 | mampus...    | mampus...     | 1,58E+18 | @username2 | 0          | 0          | 0         | 1,58E+18 | 1       | <ref= | https://www.mampus...                                      | https://www.mampus...                                      | Twitter for iPhone             |   |             |   |        |           |          |         |
| 23 | 21 | <a href="https://www.me.as.a.do.mal.a.do">https://www.me.as.a.do.mal.a.do</a>   | 2023-08-18 | me as a do   | mal a do      | 1,58E+18 | @username2 | 0          | 0          | 7         | 1,58E+18 | 1       | <ref= | https://www.me.as.a.do.mal.a.do                            | https://www.me.as.a.do.mal.a.do                            | Twitter for iOS [https://t/... | [[previewUrl]: https://pbs.twimg.com/media/ |             |   |        |           |          |         |
| 24 | 22 | <a href="https://www.mulai.campus.du.mal.campus">https://www.mulai.campus.du.mal.campus</a>   | 2023-08-18 | mulai        | campus du mal | 1,58E+18 | @username2 | 0          | 0          | 0         | 1,58E+18 | 1       | <ref= | https://www.mulai.campus.du.mal.campus                     | https://www.mulai.campus.du.mal.campus                     | Twitter for Android            |   |             |   |        |           |          |         |
| 25 | 23 | <a href="https://www.collegienet.co.id/">https://www.collegienet.co.id/</a>   | 2023-08-18 | @collegienet | @collegienet  | 1,58E+18 | @username2 | 0          | 0          | 0         | 1,58E+18 | 1       | <ref= | https://www.collegienet.co.id/                             | https://www.collegienet.co.id/                             | Twitter for iPhone             |   |             |   |        |           |          |         |
| 26 | 24 | <a href="https://www.krip.sh.smk.krip.sh.smk">https://www.krip.sh.smk.krip.sh.smk</a>   | 2023-08-18 | krip         | sh smk krip   | 1,58E+18 | @username2 | 0          | 0          | 0         | 1,58E+18 | 1       | <ref= | https://www.krip.sh.smk.krip.sh.smk                        | https://www.krip.sh.smk.krip.sh.smk                        | Twitter for Android            |   |             |   |        |           |          |         |
| 27 | 25 | <a href="https://www.first.time.s.first.time">https://www.first.time.s.first.time</a>   | 2023-08-18 | first time s | first time    | 1,58E+18 | @username2 | 0          | 0          | 0         | 1,58E+18 | 1       | <ref= | https://www.first.time.s.first.time                        | https://www.first.time.s.first.time                        | Twitter for Android            |   |             |   |        |           |          |         |
| 28 | 26 | <a href="https://www.baharung.baharung">https://www.baharung.baharung</a>   | 2023-08-18 | baharung     | baharung      | 1,58E+18 | @username2 | 0          | 0          | 0         | 1,58E+18 | 1       | <ref= | https://www.baharung.baharung                              | https://www.baharung.baharung                              | Twitter for Android            |   |             |   |        |           |          |         |
| 29 | 27 | <a href="https://www.kulluh.kulluh.kulluh.kulluh">https://www.kulluh.kulluh.kulluh.kulluh</a>                                       | 2023-08-18 | kulluh       | kulluh kulluh | 1,58E+18 | @username2 | 0          | 0          | 0         | 1,58E+18 | 1       | <ref= | https://www.kulluh.kulluh.kulluh.kulluh                    | https://www.kulluh.kulluh.kulluh.kulluh                    | Twitter for Android            |   |             |   |        |           |          |         |

**Gambar 4.2** Hasil pengambilan data

#### 4.1.2. Preprocessing

Dataset yang sudah didapatkan melalui proses *crawling* selanjutnya dilakukan tahap proses *preprocessing*. Tahap *pre-processing* dalam penelitian ini melibatkan beberapa tahapan yaitu tahap *data cleansing* yang terdiri dari *case folding*, *punctuation removal* , dan *special character removal*.



```

from bs4 import BeautifulSoup

cleanHTML = []
i=0
for column in df.content:
    clean = BeautifulSoup(df.content[i], 'lxml') # menghilangkan text field as '&','&quot','&dl
    cleanHTML.insert(i, clean.get_text())
    i+=1

df["cleanHTML"] = cleanHTML

import re
cleanMention = []
i=0
for column in df.cleanHTML:
    clean = re.sub(r'@[A-Za-z0-9_]+', '', df.cleanHTML[i])
    clean = re.sub('https?://[A-Za-z0-9./]+', '', clean)
    clean = re.sub(r'\\x(.)[2]', "", clean) #menghilangkan kata berlebih
    clean = re.sub(r'^b[\\\'"]|#[A-Za-z0-9]+|Rf|\\n| +|:\\(|:\\)|:v|:v|:':\\(|:':\\(", " ", clean) # menghilangkan karakter lain
    clean = re.sub("\\w+| +", " ", clean)
    clean = clean.lower()
    cleanMention.insert(i, clean)
    i += 1

df["cleanData"] = cleanMention
df

```

**Gambar 4.3** Proses *preprocessing*

|                       | content   | cleanHTML   | cleanData   |
|-----------------------|---|---|---|
| 0                     | [APA SAJA KEUNIKAN DARI UNPAD DAN JATINANGOR?]    | [APA SAJA KEUNIKAN DARI UNPAD DAN JATINANGOR?]    | apa saja keunikan dari unpad dan jatinangor h...  |
| 1                     | Nek ono kuliah luring ngene tapi aku ngantuk b... | Nek ono kuliah luring ngene tapi aku ngantuk b... | nek ono kuliah luring ngene tapi aku ngantuk b... |
| 2                     | banyak kuliah luring ngeluh, banyak kuliah dar... | banyak kuliah luring ngeluh, banyak kuliah dar... | banyak kuliah luring ngeluh banyak kuliah dari... |
| 3                     | @collegemenfess Aku dulunya sih kos khusus cew... | @collegemenfess Aku dulunya sih kos khusus cew... | aku dulunya sih kos khusus cewe tapi pas dah ...  |
| 4                     | @yjnluvv baiklah :( besok ada kuliah luring ga... | @yjnluvv baiklah :( besok ada kuliah luring ga... | baiklah besok ada kuliah luring ga sih            |
| ...                   | ...   | ...   | ...   |
| 3944                  | Kuliah daring kok malah lebih banyak ya tugasn... | Kuliah daring kok malah lebih banyak ya tugasn... | kuliah daring kok malah lebih banyak ya tugasn... |
| 3945                  | Kuliah media daring,\nTidak senikmat kuliah lu... | Kuliah media daring,\nTidak senikmat kuliah lu... | kuliah media daring tidak senikmat kuliah luri... |
| 3946                  | selama daring, waktunya beberes hal-hal yang d... | selama daring, waktunya beberes hal-hal yang d... | selama daring waktunya beberes hal hal yang di... |
| 3947                  | Kalo menurut sy, ketika model kuliah luring di... | Kalo menurut sy, ketika model kuliah luring di... | kalo menurut sy ketika model kuliah luring dig... |
| 3948                  | @KetekYeonjun Percayalah, kuliah daring lebih ... | @KetekYeonjun Percayalah, kuliah daring lebih ... | percayalah kuliah daring lebih menyulitkan dr...  |
| 3949 rows x 3 columns |   |   |   |

**Gambar 4.4** Hasil *preprocessing*

Sebelum dilakukannya pembuatan dan training model, dilakukan terjemahan bahasa non formal ke bahasa formal bahasa Indonesia. Tahapan ini dibantu dengan menggunakan kamus slang untuk mendapatkan beberapa kata slang. Proses terlihat pada gambar 4.5.

```

slang_to_formal = dict(zip(df_slang['slang'], df_slang['formal']))

# Fungsi untuk menerjemahkan kalimat slang ke bahasa baku
def translate_to_baku(kalimat, slang_dict):
    words = kalimat.split()
    translated_words = [slang_dict.get(word, word) for word in words]
    translated_sentence = ' '.join(translated_words)
    return translated_sentence

# Terjemahkan setiap kalimat slang
df['kalimat_baku'] = df['cleanData'].apply(lambda x: translate_to_baku(x, slang_to_formal))

# Tampilkan hasil dalam DataFrame
df_hasil = df[['cleanData', 'kalimat_baku']]
df_hasil = df_hasil.rename(columns={'cleanData': 'slang_sentence', 'kalimat_baku': 'standard_sentence'})
# Print hasil
df_hasil

```

**Gambar 4.5** Proses menerjemah bahasa non formal

|      | slang_sentence                                    | standard_sentence                                 |
|------|---|---|
| 0    | apa saja keunikan dari unpad dan jatinangor h...  | apa saja keunikan dari unpad dan jatinangor ha... |
| 1    | nek ono kuliah luring ngene tapi aku ngantuk b... | nek sono kuliah luring ngene tapi aku ngantuk ... |
| 2    | banyak kuliah luring ngeluh banyak kuliah dari... | banyak kuliah luring ngeluh banyak kuliah dari... |
| 3    | aku dulunya sih kos khusus cewe tapi pas dah ...  | aku dulunya sih kos khusus cewek tapi pas deh ... |
| 4    | baiklah besok ada kuliah luring ga sih            | baiklah besok ada kuliah luring tidak sih         |
| ...  | ...   | ...   |
| 3944 | kuliah daring kok malah lebih banyak ya tugasn... | kuliah daring kok malah lebih banyak ya tugasn... |
| 3945 | kuliah media daring tidak senikmat kuliah luri... | kuliah media daring tidak senikmat kuliah luri... |
| 3946 | selama daring waktunya bebers hal hal yang di...  | selama daring waktunya bebers hal hal yang di...  |
| 3947 | kalo menurut sy ketika model kuliah luring dig... | kalau menurut saya ketika model kuliah luring ... |
| 3948 | percayalah kuliah daring lebih menyulitkan dr...  | percayalah kuliah daring lebih menyulitkan dar... |

3949 rows × 2 columns

**Gambar 4.6** Hasil terjemahan bahasa non formal

Setelah dilakukannya penerjemahan yang dibantu oleh kamus slang, dilakukan penghilangan untuk karakter berlebih dengan menggunakan *regular expression* serta mengeksport file dataset ke dalam bentuk excel. Proses dapat dilihat pada gambar 4.7.

```

# Fungsi menghapus karakter berlebih
def clean_standard_sentence(sentence):
    cleaned_sentence = re.compile(r'(\.|\{1,\}', re.IGNORECASE).sub(r'\1', sentence)
    return cleaned_sentence

# menerapkan fungsi cleaning ke kolom standard_sentence
df_hasil['standard_sentence'] = df_hasil['standard_sentence'].apply(clean_standard_sentence)

# Print clean DataFrame
print(df_hasil)

# menyimpan dataset kedalam bentuk excel
xlsx_file_path = 'file.xlsx'

# Export DataFrame ke file excel
df_hasil.to_excel(xlsx_file_path, index=False)

```

**Gambar 4.7** Proses penghilangan karakter berulang dan *export file*

Setelah dilakukannya *export file*, dilakukan pemeriksaan yang dilakukan secara manual oleh manusia untuk kata dan kalimat yang tidak ada dalam kamus serta pemeriksaan maksud dari suatu kalimat baik dalam kalimat non formal dan kalimat formal.

#### 4.1.3. Pembuatan dan *training* model

Sebelum dilakukannya pembuatan dan training model, dilakukan memasukan modul yang diperlukan untuk pembuatan dan *training* model. Proses dapat dilihat pada gambar 4.8.

```

import torch
from sklearn.model_selection import train_test_split
from transformers import MBart50Tokenizer, MBartForConditionalGeneration, Trainer, TrainingArguments
import pandas as pd

```

**Gambar 4.8** *Library* yang diperlukan

Modul torch, yang merupakan modul utama untuk komputasi tensor di PyTorch. PyTorch digunakan untuk pelatihan dan pengembangan model dalam pembelajaran mesin dan *deep learning*. Fungsi *train\_test\_split* dari modul *model\_selection* dalam *library scikit-learn* (sklearn). Fungsi ini digunakan supaya membagi dataset menjadi subset pelatihan dan validasi. *from transformers import MBart50Tokenizer, MBartForConditionalGeneration, Trainer, TrainingArguments* merupakan beberapa kelas dan fungsi dari pustaka *Hugging Face Transformers*. MBart50Tokenizer merupakan kelas tokenizer untuk model MBart50. MBartForConditionalGeneration merupakan kelas model MBart50 yang

dirancang untuk tugas generasi kondisional, seperti terjemahan bahasa. *Trainer* merupakan kelas yang menyediakan pelatihan yang terintegrasi dengan *library* transformer. *TrainingArguments* merupakan kelas yang menyimpan parameter dan argumen pelatihan untuk *Trainer*.

Kemudian melakukan *load* dataset yang telah diperiksa secara manual oleh manusia dengan membaca file dataset atau corpus dan mengambil kolom '*slang\_sentence*' dan '*standard\_sentence*' yang terdapat didalam file excel. Proses dapat dilihat pada gambar 4.9 dibawah ini.

```
#load data split it into training and validation sets
data = pd.read_excel('file.xlsx')
slang_sentences = data['slang_sentence'].tolist()
standard_sentences = data['standard_sentence'].tolist()
```

**Gambar 4.9** Proses *load* Dataset

Kemudian dataset dibagi menjadi dua data set, satu untuk pelatihan (train) dan satu untuk validasi (val), 90% dataset *training* dan 10% dataset validasi. Dataset *training* dimanfaatkan untuk melatih model. Sedangkan dataset *testing* digunakan untuk mengukur kinerja model selama proses *training*. Gambar 4.10 menunjukkan tahap *split* data.

```
# Split data
slang_train, slang_val, standard_train, standard_val = train_test_split(
    slang_sentences, standard_sentences, test_size=0.1, random_state=42)
```

**Gambar 4.10** Proses *split* data

Setelah melakukan *split* data, dilakukan inisialisasi tokenizer dengan membuat objek tokenizer dari kelas MBart50Tokenizer yang telah di *pretrained* dengan model MBart50. Tokenizer digunakan untuk mengubah teks menjadi representasi token yang dapat dimengerti oleh model. Proses dapat dilihat pada gambar 4.11 dibawah ini.

```
# Initialize the tokenizer
tokenizer = MBart50Tokenizer.from_pretrained("facebook/mbart-large-50")
```

**Gambar 4.11** Proses inialisasi tokenizer

Kemudian melakukan tokenisasi data pelatihan dan data validasi dengan membuat beberapa fungsi untuk pelatihan dan validasi. Proses terlihat pada gambar 4.12.

```
# Tokenize data
train_encodings = tokenizer(slang_train, truncation=True, padding=True, return_tensors="pt")
train_labels = tokenizer(standard_train, truncation=True, padding=True, return_tensors="pt").input_ids
val_encodings = tokenizer(slang_val, truncation=True, padding=True, return_tensors="pt")
val_labels = tokenizer(standard_val, truncation=True, padding=True, return_tensors="pt").input_ids
```

**Gambar 4.12** Proses tokenisasi data

Setelah itu membuat suatu kelas dataset kustom (*TranslationDataset*) yang digunakan selama pelatihan model terjemahan dan inialisasi objek dataset. Membuat dua objek dataset, yaitu *train\_dataset* dan *val\_dataset*, dengan menggunakan kelas *TranslationDataset*. Masing-masing objek ini diinisialisasi dengan pasangan data tokenisasi dan label dari dataset pelatihan dan validasi. Dengan membuat dataset kustom ini, dapat menggunakan objek *train\_dataset* dan *val\_dataset* pada pelatihan model. Proses dapat dilihat pada gambar 4.13 dibawah ini.

```
# Custom dataset
class TranslationDataset(torch.utils.data.Dataset):
    def __init__(self, encodings, labels):
        self.encodings = encodings
        self.labels = labels

    def __getitem__(self, idx):
        item = {key: val[idx] for key, val in self.encodings.items()}
        item["labels"] = self.labels[idx]
        return item

    def __len__(self):
        return len(self.encodings.input_ids)

train_dataset = TranslationDataset(train_encodings, train_labels)
val_dataset = TranslationDataset(val_encodings, val_labels)
```

**Gambar 4.13** Proses Membuat Dataset Kustom



Selanjutnya melakukan inisialisasi model. Menginisialisasi objek model menggunakan kelas MBartForConditionalGeneration dari *library* Transformers. Serta membuat argumen pelatihan yang berisi argumen dan parameter untuk proses pelatihan model. Dengan menggunakan objek TrainingArguments, dapat mengonfigurasi proses pelatihan model sesuai dengan kebutuhan dan ketersediaan sumber daya. Selanjutnya dapat membuat objek Trainer dan menjalankan proses pelatihan dengan menggunakan dataset dan model yang sudah diinisialisasi sebelumnya. Proses dapat dilihat pada gambar 4.14.

```
# Initialize the model
model = MBartForConditionalGeneration.from_pretrained("facebook/mbart-large-50")

# Training arguments
training_args = TrainingArguments(
    output_dir="./translation_model",
    per_device_train_batch_size=8,
    per_device_eval_batch_size=8,
    num_train_epochs=2,
    save_steps=100,
    save_total_limit=2,
    remove_unused_columns=False,
    logging_steps=10,
    logging_dir="./logs",
)
```

**Gambar 4.14** Proses inisialisasi model dan pembuatan argumen pelatihan

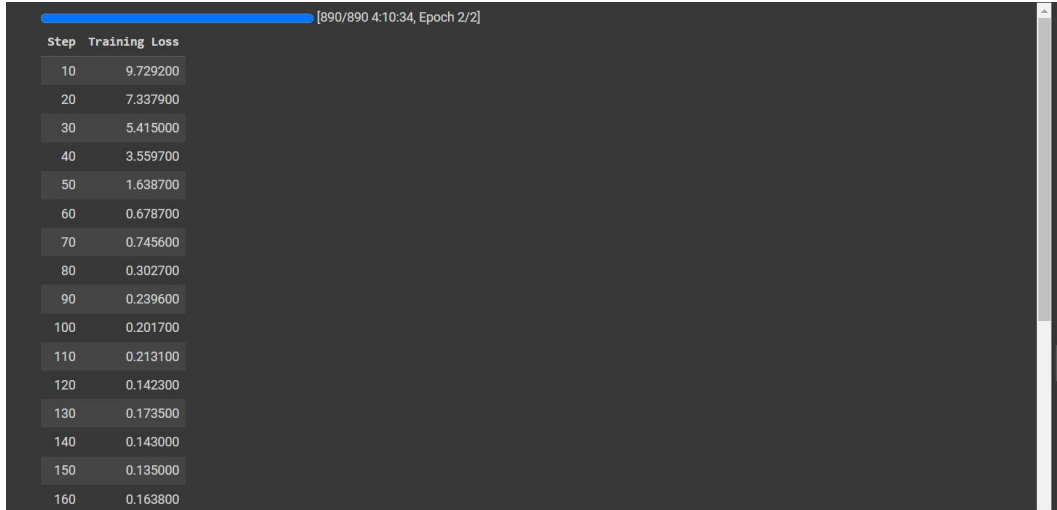
Selanjutnya membuat suatu kelas ‘CustomTrainer’ yang merupakan turunan dari kelas ‘Trainer’ serta mendefinisikan metode baru untuk menghitung loss selama pelatihan. *Loss* diambil dari hasil *output* model. Proses dapat dilihat pada gambar 4.15.

```
# Custom Trainer
class CustomTrainer(Trainer):
    def compute_loss(self, model, inputs):
        labels = inputs["labels"]
        outputs = model(input_ids=inputs["input_ids"], attention_mask=inputs["attention_mask"], labels=labels)
        return outputs.loss

# membuat custom trainer
trainer = CustomTrainer(
    model=model,
    args=training_args,
    train_dataset=train_dataset,
    eval_dataset=val_dataset,
)

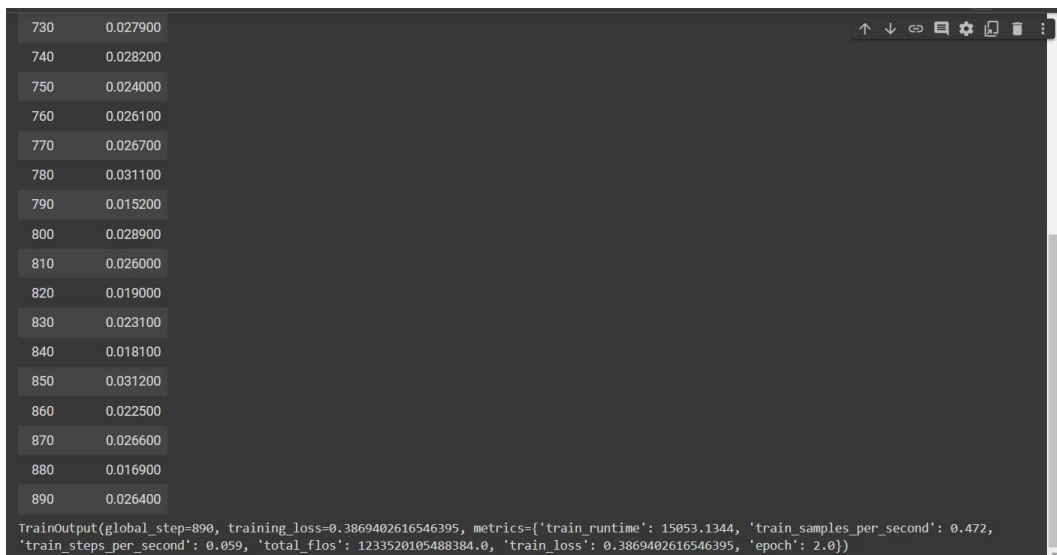
# Train the model
trainer.train()
```

**Gambar 4.15** Proses *training* model



| Step | Training Loss |
|------|---------------|
| 10   | 9.729200      |
| 20   | 7.337900      |
| 30   | 5.415000      |
| 40   | 3.559700      |
| 50   | 1.638700      |
| 60   | 0.678700      |
| 70   | 0.745600      |
| 80   | 0.302700      |
| 90   | 0.239600      |
| 100  | 0.201700      |
| 110  | 0.213100      |
| 120  | 0.142300      |
| 130  | 0.173500      |
| 140  | 0.143000      |
| 150  | 0.135000      |
| 160  | 0.163800      |

**Gambar 4.16** Hasil *training* model



| Step | Training Loss |
|------|---------------|
| 730  | 0.027900      |
| 740  | 0.028200      |
| 750  | 0.024000      |
| 760  | 0.026100      |
| 770  | 0.026700      |
| 780  | 0.031100      |
| 790  | 0.015200      |
| 800  | 0.028900      |
| 810  | 0.026000      |
| 820  | 0.019000      |
| 830  | 0.023100      |
| 840  | 0.018100      |
| 850  | 0.031200      |
| 860  | 0.022500      |
| 870  | 0.026600      |
| 880  | 0.016900      |
| 890  | 0.026400      |

TrainOutput(global\_step=890, training\_loss=0.3869402616546395, metrics={'train\_runtime': 15053.1344, 'train\_samples\_per\_second': 0.472, 'train\_steps\_per\_second': 0.059, 'total\_flos': 1233520105488384.0, 'train\_loss': 0.3869402616546395, 'epoch': 2.0})

**Gambar 4.17** Hasil *training* model

Setelah melakukan *training* model lalu simpan *fine-tuning* model kedalam suatu folder agar dimasukan kedalam tampilan sistem. Proses dapat dilihat pada gambar 4.18.

```
# Save the fine-tuned model
model.save_pretrained("model_datasetLengkap")
tokenizer.save_pretrained("model_datasetLengkap")

('model_datasetLengkap/tokenizer_config.json',
 'model_datasetLengkap/special_tokens_map.json',
 'model_datasetLengkap/sentencepiece.bpe.model',
 'model_datasetLengkap/added_tokens.json')
```

**Gambar 4.18** Proses simpan model

#### 4.1.4. Tampilan Sistem

Berikut adalah tampilan sistem yang terdiri dari halaman *home*, halaman data, halaman program, dan halaman riwayat.



**Gambar 4.19** Halaman *Home*

Tampilan *home* diatas menjadi tampilan awal sistem setelah dibuka. Terdapat button untuk menuju halaman program dan beberapa menu yang menghubungkan pengguna ke halaman data, halaman program, dan halaman riwayat.



| HOME DATA PROGRAM RIWAYAT |  |  |
|---------------------------|--|--|
| Contoh Dataset            |  |  |
| No                        | slang sentence   | standard sentence  |
| 1                         | apa saja keunikan dari unpad dan jatinangor halo sobat warta mayoritas sobat warta pasti sudah menjalankan kuliah luring kan betul atau betul bagaimana rasanya menjalankan keseharian di jatinangor dan di kampus unpad sudah pernah jalan jalan di jatos belu    | apa saja keunikan dari unpad dan jatinangor halo sobat warta mayoritas sobat warta pasti sudah menjalankan kuliah luring kan betul atau betul bagaimana rasanya menjalankan keseharian di jatinangor dan di kampus unpad sudah pernah jalan jalan di jatos belu    |
| 2                         | nek ono kuliah luring ngene tapi aku ngantuk banget materine numpang lewat tok lak ra dosa to ya nasibe kerja kudu pelatihan   | nek sono kuliah luring ngene tapi aku ngantuk banget materine numpang lewat tok lak ra dosa tapi ya nasibe kerja harus pelatihan   |
| 3                         | banyak kuliah luring ngeluh banyak kuliah daring ngeluh jg   | banyak kuliah luring ngeluh banyak kuliah daring ngeluh juga   |
| 4                         | aku dulunya sih kos khusus cewe tapi pas dah mulai kuliah luring dan qodarullah dptnya kost campur padahal maunya kost khusus cewe tapi ga papa sih soalnya disini ad kelebihanannya dan aku aman dan sbg muslimah jg harus pinter2 jaga diri sama lawan jenis non | aku dulunya sih kos khusus cewe tapi pas deh mulai kuliah luring dan qodarullah dapatnya kost campur padahal maunya kost khusus cewe tapi tidak papa sih soalnya disini ada kelebihanannya dan aku aman dan sebagai muslimah juga harus pinter2 jaga diri sama law |
| 5                         | baiklah besok ada kuliah luring ga sih   | baiklah besok ada kuliah luring tidak sih  |
| 6                         | orang yg klo disuru ngisi gform luring daring trus milih luring tp klo ada kuliah luring nyinyir dan klo ada info pergantian jadwal daring seneng bgt tu ada masalah apa sebenere  | orang yang kalau disuru ngisi gform luring daring terus memilih luring tapi kalau ada kuliah luring nyinyir dan kalau ada info pergantian jadwal daring senang sekali itu ada masalah apa sebenere   |
| 7                         | tp harus tidur sapa tau besok ada kabar kuliah luring  | tapi harus tidur siapa tau besok ada kabar kuliah luring   |

**Gambar 4.20** Halaman Data

Pada gambar 4.8 adalah tampilan antarmuka data yang berisikan kalimat non formal dan kalimat formal bahasa Indonesia. Halaman data dapat dikatakan juga sebagai referensi oleh pengguna dalam memasukan kata atau kalimat non formal pada halaman program atau sistem yang dibangun.

| HOME DATA PROGRAM RIWAYAT |
|---------------------------|
|---------------------------|

Slang Sentence

Silahkan masukan kalimat slang

Translate

© 2023 All Rights Reserved By Achmad Yusuf Barmawi

**Gambar 4.21** Halaman Program

Halaman diatas menampilkan tampilan awal sebelum melakukan terjemahan untuk pengguna memasukan bahasa non formal yang ingin diterjemahkan. Terdapat juga button untuk melakukan penerjemahan bahasa non formal ke bahasa formal bahasa Indonesia.

HOME DATA PROGRAM RIWAYAT

Slang Sentence

Silahkan masukan kalimat slang

Translate

**Hasil**

| slang sentence | Hasil Translate      |
|----------------|----------------------|
| gue cape sm lo | saya lelah sama kamu |

© 2023 All Rights Reserved By Achmad Yusuf Barmawi

**Gambar 4.22** Halaman program setelah melakukan penerjemahan

Pada gambar 4.22 menampilkan tampilan ketika pengguna telah memasukan kalimat non formal bahasa indonesia. Sistem menampilkan keluaran seperti kalimat yang dimasukan pengguna beserta hasil terjemahannya.

HOME DATA PROGRAM RIWAYAT

| id | slang sentence                                      | translate sentence                                     | reference sentence                               | bleu score   | Actions                |
|----|---|--|--|--------------|------------------------|
| 1  | knp yh gua begini doang                             | kenapa ya saya seperti ini saja                        | None   | None         | <a href="#">Detail</a> |
| 2  | gue cape sm lo                                      | saya lelah sama kamu                                   | saya lelah sama kamu                             | 1.0000000000 | <a href="#">Detail</a> |
| 3  | bisa ga sih kuliah luring tpi uts uas ny daring trs | bisa tidak kuliah luring tapi uts uas nya daring terus | None   | None         | <a href="#">Detail</a> |
| 4  | capek bgt sm hidup iniii                            | lelah sekali sama hidup ini                            | None   | None         | <a href="#">Detail</a> |
| 6  | gmn cara biar kek kalian gaes                       | bagaimana cara biar seperti kalian teman-teman         | bagaimana cara supaya seperti kalian teman-teman | 0.8429433703 | <a href="#">Detail</a> |

© 2023 All Rights Reserved By Achmad Yusuf Barmawi

**Gambar 4.23** Halaman Riwayat

Tampilan di atas menampilkan riwayat hasil penerjemahan yang telah dilakukan oleh pengguna. Terdapat juga tampilan untuk melakukan perhitungan akurasi terjemahan pada kolom *action*. Ketika sudah melakukan perhitungan akurasi, hasil akurasi ditampilkan juga pada halaman riwayat.

HOME DATA PROGRAM RIWAYAT

Slang Sentence:  
gue cape sm lo

Translate Sentence:  
saya lelah sama kamu

Bleu Score:  
None

Reference Sentence:  
saya lelah sama kamu

Hitung

© 2023 All Rights Reserved By Achmad Yusuf Barmawi

**Gambar 4.24** Halaman BLEU Score

Tampilan di atas menampilkan tampilan untuk menghitung BLEU *score*. Pada proses ini pengguna harus memasukan referensi bahasa untuk menghitung agar mendapatkan akurasi dari sistem yang dibangun.

#### 4.2. Pengujian Sistem

Tahapan selanjutnya dilakukannya pengujian sistem. Sistem yang telah dikembangkan diuji guna mengetahui bahwa aplikasi yang telah dibangun bisa bekerja dengan baik atau sesuai harapan. Pada fase pengujian, bahasa yang dimasukan oleh pengguna adalah bahasa non formal bahasa Indonesia. Kemudian sistem yang dibangun menjalankan penerjemahan berdasarkan *translation model* yang dibuat dengan menggunakan *semi supervised translation*. Setelah melakukan terjemahan, kata atau kalimat yang dimasukan oleh pengguna masuk ke dalam halaman riwayat untuk dapat dilakukannya evaluasi atau pengukuran persentase akurasi dari sistem penerjemahan yang dibangun. Tabel di bawah adalah daftar kalimat non formal yang dimasukan oleh pengguna dan hasil kalimat terjemahannya yang digunakan pada tahap pengujian ini.

**Tabel 4. 1** Daftar kalimat yang dimasukan beserta hasil terjemahan

| No | Slang Sentence                                       | Hasil Terjemahan                                       |
|----|--|--|
| 1. | Knpyh gua begini doang                               | Kenapa ya saya seperti ini saja                        |
| 2. | Gue cape sm lo                                       | Saya lelah sama kamu                                   |
| 3. | Bisa ga sih kuliah luring tapi uts uas ny daring trs | Bisa tidak kuliah luring tapi uts uas nya daring terus |
| 4. | Capek bgt sm hidup iniii                             | Lelah sekali sama hidup ini                            |
| 5. | Gmn cara biar seperti kalian gaes                    | Bagaimana cara biar seperti kalian teman-teman         |

Untuk mendapatkan hasil dibutuhkan beberapa proses agar sistem dapat bekerja yakni:

1. Pengguna memasukkan kata atau kalimat yang ingin diterjemahkan pada halaman yang sudah disediakan.
2. Sistem mengambil data terjemahan pada *translation model* yang dibangun sebagai acuan dalam penerjemahan.
3. Kalimat yang dimasukan pengguna diubah dalam bentuk bahasa formal bahasa Indonesia.
4. Kalimat yang telah dimasukan oleh pengguna masuk ke dalam halaman riwayat beserta hasil terjemahannya.
5. Pada halaman riwayat, pengguna dapat melakukan evaluasi atau mengukur akurasi dari sistem penerjemah yang dibangun. Untuk mengukur akurasi sistem penerjemahannya menggunakan *BLEU score*.
6. Terdapat halaman untuk mengukur akurasi dari sistem penerjemahan. Di dalam halaman tersebut, pengguna memasukkan referensi kalimat formalnya dari kalimat yang dimasukan pengguna yang ingin diterjemah (kalimat non

formal) agar mengetahui perbandingan dan mendapatkan perhitungan akurasi dari sistem yang dibangun.

#### 4.2.1. Pengujian dalam menghitung BLEU Score

Pengujian dalam mengukur atau menghitung akurasi dari sistem penerjemah yang dibangun dilakukan dengan memasukan suatu referensi kalimat untuk hasil terjemahan yang dibuat oleh manusia sebagai perbandingan dengan hasil terjemahan dari sistem penerjemah yang dibangun. BLEU Score dapat memberikan nilai antara 0 dan 1, di mana nilai yang lebih tinggi menunjukkan terjemahan yang lebih baik sesuai dengan referensi kalimat. Berikut adalah daftar hasil dari akurasi.

**Tabel 4. 2** Daftar kalimat yang dimasukan beserta hasil BLEU

| No | Slang Sentence  | Hasil Terjemahan   | Reference Sentence  | BLEU Score |
|----|---|--|---|------------|
| 1. | Knp yh gua<br>begini doang                                    | Kenapa ya saya<br>seperti ini saja                           | Kenapa ya<br>saya seperti ini<br>saja                           | 1.00       |
| 2. | Gue cape sm lo  | Saya lelah sama<br>kamu                                      | Saya lelah<br>sama kamu   | 1.00       |
| 3. | Bisa ga sih<br>kuliah luring<br>tapi uts uas ny<br>daring trs | Bisa tidak kuliah<br>luring tapi uts uas<br>nya daring terus | Bisa tidak<br>kuliah luring<br>tapi uts uas nya<br>daring terus | 1.00       |
| 4. | Capek bgt sm<br>hidup iniii                                   | Lelah sekali sama<br>hidup ini                               | Lelah sekali<br>sama hidup ini                                  | 1.00       |
| 5. | Gmn cara biar<br>seperti kalian<br>gaes                       | Bagaimana cara<br>biar seperti kalian<br>teman-teman         | Bagaimana<br>cara supaya<br>seperti kalian<br>teman-teman       | 0.84       |

Pada data pengujian dapat dikatakan bahwa data pertama mendapatkan hasil yang baik dari hasil perhitungan akurasi mesin penerjemah dikarenakan panjang hasil penerjemahan dan panjang referensi beserta jumlah  $\text{precision}_n$  adalah sama. Pada data pengujian kelima terdapat perbedaan  $\text{precision}_n$  atau jumlah n-gram (kata) yang cocok antara terjemahan dan referensi sehingga mendapatkan hasil 0.84. Dari kelima hasil BLEU *score* dalam data pengujian didapatkan hasil akurasi untuk daftar kalimat dengan nilai rata rata adalah 0,968 atau 96% dengan 5 data pengujian.

## **BAB V**

### **PENUTUP**

#### **5.1. KESIMPULAN**

Setelah melakukan beberapa tahapan pengujian meliputi analisa, perancangan, implementasi, serta pengujian didapatkan kesimpulan mengenai penerjemahan bahasa non formal ke bahasa formal bahasa Indonesia dengan menggunakan *semi supervised translation* sebagai berikut:

1. Model yang dibangun dengan *Semi Supervised Translation* dapat digunakan dalam membangun sistem penerjemah.
2. Penerapan mesin penerjemah terhadap data bahasa Indonesia tidak baku ke bahasa Indonesia baku dapat dilakukan.
3. Masih terdapat beberapa kata atau kalimat non formal dan juga singkatan-singkatan dari kata-kata non formal yang tidak berubah dikarenakan tidak ada dalam korpus.
4. Pengembangan antarmuka mesin penerjemah dapat berjalan dengan baik dalam proses *input* terdapat di aplikasi.
5. Persentase yang didapatkan dalam pengujian sistem untuk menghitung akurasi mesin penerjemahan yang menggunakan 5 data pengujian adalah bernilai 0,968 atau 96%.

#### **5.2. SARAN**

Berikut adalah saran yang dapat dipertimbangkan untuk penelitian selanjutnya.

1. Membuat model dengan menambah data yang lebih banyak agar mendapatkan lebih banyak varian bahasa non formalnya.
2. Lebih memperhatikan kata-kata non formal yang berbentuk singkatan.
3. Sistem dapat menerjemahkan varian bahasa non formal yang menggunakan bahasa daerah maupun asing ke dalam bentuk bahasa formal.

## DAFTAR PUSTAKA

- Alam, M., & Suryani, A. (2021). *Minang and Indonesian Phrase-Based Statistical Machine Translation*. JITE (*Journal of Informatics and Telecommunication Engineering*), 5 (1).
- Bachman. 1990. Ragam bahasa adalah variasi bahasa menurut pemakaian. Jakarta: PT.Gramedia.
- Dwiastuti, M. (2019). *English-Indonesian Neural Machine Translation for Spoken Language Domains*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 309–314, Florence, Italy. *Association for Computational Linguistics*.
- Fauziyah, Y., Ilyas, R., & Kasyidi, F. (2022). Mesin Penerjemah Bahasa Indonesia-Bahasa Sunda Menggunakan *Recurrent Neural Networks*. Jurnal TEKNOINFO, 16 (2). 313-322.
- Ghimire, D. 2020. *Comparative study on Python web frameworks: Flask and Django*. <http://www.theseus.fi/handle/10024/339796>.
- Hidayat, A.(2011). "Aplikasi Penerjemah Dua Arah Bahasa Indonesia – Bahasa Melayu Sambas Berbasis Web dengan Menggunakan Decoder Moses,". Teknik Informatika Universitas Tanjung Pura. Skripsi.
- Lestari, A., Ardiyanti, A., & Asror, I. (2021). Phrase Based Statistical Machine Translation Javanese-Indonesian. *Jurnal Media Informatika Budidarma*, 5(2), 378-386. <https://ejurnal.stmik-budidarma.ac.id/index.php/mib>.
- Lewis et al., (2019). *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. arXiv:1910.13461v1
- Liu, et al. (2020). *Multilingual Denoising Pre-training for Neural Machine Translation*. <https://arxiv.org/abs/2001.08210>



- Navarro, A., & Casacuberta, F. (2023). *Exploring Multilingual Pretrained Machine Translation Models for Interactive Translation*. PRHLT, Universitat Politècnica de Valencia, Spain.
- Pranata, J., & Muljono. (2016). *Mesin Penerjemah Bahasa Indonesia-Bahasa Jawa*. Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro.
- Pratiwi, N., Sujaini, H., Nyoto, R. (2017). Pengembangan Antarmuka Mesin Penerjemah Statistik Multibahasa Berbasis Web. *Jurnal Sistem dan Teknologi Informasi (JUSTIN)*, 5 (1).
- Relan, K. 2019. *Building REST APIs with Flask: Create Python Web Services with MySQL*. DOI:10.1007/978-1-4842-5022-8
- Subarianto. 2000. *Ragam standar*. Yogyakarta: Erlangga.
- Tang et al., (2021). *Multilingual Translation from Denoising Pre-Training*. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 3450–3466
- Wang, et al. (2022). *Understanding and Improving Sequence-to-Sequence Pretraining for Neural Machine Translation*. Department of Computer Science and Engineering, University of Hong Kong. <https://arxiv.org/abs/2203.08442>
- Wibowo et al., (2020). *Semi-Supervised Low-Resource Style Transfer of Indonesian Informal to Formal Language with Iterative Forward-Translation*. <https://doi.org/10.48550/arXiv.2011.03286>