**Prosiding Seminar Nasional Sains dan Teknologi Seri III**
**Fakultas Sains dan Teknologi, Universitas Terbuka**
**Vol. 2 No. 1 (2025)**
**e-ISSN: 3047-6569**

# PENERAPAN MODEL T5 UNTUK PENERJEMAHAN MESIN ACEH-INDONESIA

**Ivander Perdana Mokhtar*, Aulia Wardani, Hanna Sajidah, M. Faqih, Berliyana Kesuma Hati, Ardika Satria, Christyan Tamaro Nadeak, Ade Lailani**
*Program Studi Sains Data, Institut Teknologi Sumatera, Lampung Selatan*

**Penulis korespondensi: ivander.121450067@student.itera.ac.id*

## ABSTRAK

Bahasa Aceh adalah salah satu bahasa daerah yang memiliki struktur dan kosakata khas, yang sering kali membedakan makna dan konteks dalam proses penerjemah. Penelitian ini bertujuan mengembangkan model terjemahan mesin berbasis *sequence-to-sequence* untuk penerjemahan dari Bahasa Aceh ke bahasa Indonesia dengan memanfaatkan arsitektur T5. Data diproses untuk menghasilkan fitur-fitur yang dapat digunakan dalam model, dengan analisis distribusi panjang kalimat, kata-kata paling umum, serta pembuatan n-grams (bigram dan trigram) untuk mengeksplorasi karakteristik teks lebih lanjut. Distribusi panjang kalimat dalam Bahasa Aceh dan bahasa Indonesia mengindikasikan variasi dalam struktur kalimat antara kedua bahasa. Analisis kata-kata yang paling sering muncul serta visualisasi *word cloud* memberikan gambaran tentang kata-kata dominan dalam kedua bahasa, yang penting untuk memahami konteks linguistik yang ada dalam data pelatihan. Model T5 dilatih dengan Seq2SeqTrainer, lalu dievaluasi menggunakan metrik BLEU dan METEOR. Metrik BLEU mengukur kesesuaian terjemahan yang dihasilkan dengan referensi, sementara METEOR mengevaluasi kualitas terjemahan dengan mempertimbangkan kesamaan makna dan struktur kalimat. Hasil evaluasi menunjukkan bahwa model T5 memperoleh skor BLEU sebesar 10.04 dan skor METEOR sebesar 0.31. Evaluasi METEOR menunjukkan adanya kesulitan model dalam menangkap kesamaan makna dan struktur kalimat, yang mengakibatkan pada perbedaan urutan dan pilihan kata. Model ini dapat menunjukkan potensi dalam menghasilkan terjemahan yang relevan dan mudah dipahami. Hasil penelitian ini dikembangkan menjadi aplikasi berbasis web yang bernama “AcehLingua” untuk penerjemahan kalimat dalam Bahasa Aceh ke bahasa Indonesia.

**Kata kunci:** penerjemahan mesin, model *sequence-to-sequence*, arsitektur T5, metrik evaluasi, pemrosesan teks

## 1 PENDAHULUAN

Indonesia, negara yang kaya akan keanekaragaman budaya dan bahasa, memiliki kekayaan linguistik yang melimpah (Rahman et al, 2022). Di seluruh kepulauan, tercatat lebih dari 700 bahasa daerah yang digunakan, menunjukkan keragaman budaya yang kuat dari setiap kawasan (Handayani, 2019). Meskipun Bahasa Indonesia menjadi bahasa resmi negara dan bahasa pergaulan umum, bahasa daerah, seperti Bahasa Aceh, masih memegang peranan penting dalam kehidupan masyarakat Indonesia, terutama dalam aspek sosial dan budaya (Fajria et al, 2024). Bahasa Aceh, yang dituturkan oleh sekitar tiga juta penduduk, banyak digunakan di Provinsi Aceh dan wilayah sekitarnya (Nisa, 2023). Bahasa Aceh merupakan manifestasi budaya yang sekaligus menjadi sarana komunikasi utama bagi masyarakat Aceh dalam segala aktivitas kesehariannya, meliputi tradisi, kesusastraan, hingga hubungan antarpribadi (Hasibuan et al, 2022).

Namun, keragaman bahasa daerah di Indonesia, termasuk Bahasa Aceh yang digunakan dalam kehidupan keseharian, menghadirkan tantangan tersendiri dalam pengembangan sistem penerjemahan mesin, terutama disebabkan ketersediaan data paralel berkualitas yang terbatas untuk bahasa ini (Anandia et al, 2024). Data paralel merupakan himpunan pasangan naskah dalam dua bahasa yang dimanfaatkan untuk mengasah kecerdasan buatan penerjemah mesin dalam menghasilkan alih bahasa yang cermat dan selaras dengan konteks aslinya. Kurangnya ketersediaan data paralel yang mencukupi menjadi kendala dalam proses pelatihan model terjemahan yang mampu menghasilkan terjemahan akurat dan natural antara Bahasa Aceh dan bahasa Indonesia.

Dalam penerjemahan mesin, Bahasa Aceh menunjukkan ciri-ciri bahasa yang unik yang membedakannya dari bahasa Indonesia. Struktur kalimat, imbuhan, dan kosa kata yang khas perlu diperhatikan secara khusus dalam pengolahan data dan pelatihan model. Sebagai bahasa yang tergolong dalam rumpun Aceh-Chamic, yakni subcabang bahasa Austronesia yang mencakup bahasa Aceh dan bahasa Cham di Vietnam dan Kamboja, Bahasa Aceh memiliki kesamaan linguistik dengan bahasa-bahasa dalam rumpun ini (Mustaqim et al, 2020). Rumpun Aceh-Chamic dikenal memiliki morfologi yang rumit, seperti penggunaan prefiks, infiks, dan sufiks, serta tata bahasa yang kompleks. Kekhasan ini menimbulkan tantangan dalam pengenalan pola bahasa dan penyusunan terjemahan yang tepat. Selain itu, keterbatasan data paralel yang mencakup beragam topik dan ranah memperparah kesulitan dalam pengembangan model penerjemahan yang akurat untuk Bahasa Aceh.

Pengembangan sistem penerjemahan mesin yang memadai untuk Bahasa Aceh krusial dalam memfasilitasi interaksi yang lancar antara masyarakat Aceh dan komunitas global, terutama dalam sektor pariwisata dan penelitian budaya. Meningkatnya geliat pariwisata Aceh menggarisbawahi kebutuhan mendesak antara teknologi penerjemahan mesin yang mampu menunjang komunikasi efektif antara wisatawan domestik dan internasional dengan penduduk setempat (Aji et al, 2022). Berdasarkan data Kementerian Pariwisata dan Ekonomi Kreatif (Kemenparekraf) RI, Provinsi Aceh menempati peringkat lima besar Indeks Pembangunan Kepariwisataan Nasional (IPKN) 2022 pada kategori *travel and tourism demand drivers* dengan skor 3,86 (Patria, 2023). Pencapaian ini mencerminkan daya tarik Aceh bagi wisatawan, baik domestik maupun internasional. Dengan budaya Aceh yang khas, termasuk bahasanya, *machine translation* dapat membantu wisatawan dan peneliti luar yang tertarik memahami budaya Aceh, sehingga tidak hanya mendukung pelestarian Bahasa Aceh tetapi juga memperkuat pengembangan pariwisata dan ekonomi kreatif di wilayah ini.

Penerjemahan mesin merupakan bidang krusial dalam *Natural Language Processing* (NLP) yang bertujuan untuk menerjemahkan teks antar bahasa secara otomatis (Ghirrid et al, 2024). Sejumlah pendekatan berbasis jaringan saraf telah digunakan dalam sistem penerjemahan mesin, termasuk arsitektur berbasis model *Transformers*. Model *Transformers*, yang pertama kali diperkenalkan oleh Vaswani et al. (2017), telah merevolusi pemahaman dan penerapan model NLP untuk berbagai tugas, termasuk terjemahan mesin (Vaswani et al, 2017).

Salah satu arsitektur jaringan saraf berbasis *Transformers* yang banyak digunakan adalah T5 (*Text-to-Text Transfer Transformer*). Arsitektur ini mengonseptualisasikan berbagai tugas pemrosesan bahasa alami, termasuk penerjemahan, sebagai masalah pemetaan teks ke teks lainnya (Maksum

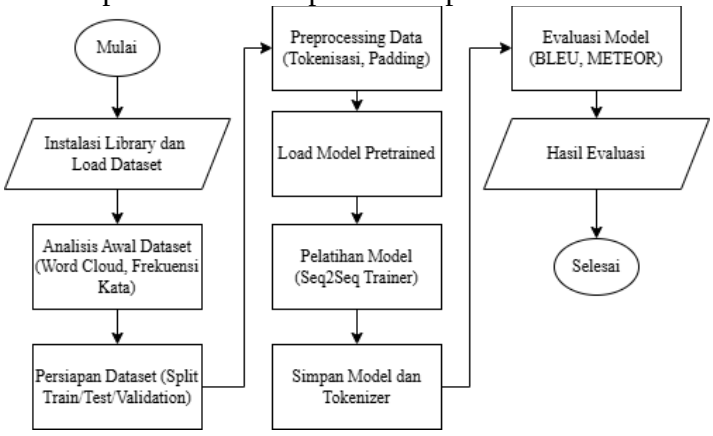
et al, 2024). Penelitian oleh Hafizh (2024) menunjukkan bahwa model *Transformers* memberikan hasil yang unggul lebih unggul dalam mengenali pola data baru pada tugas NLP (Hafizh, 2024). Namun, sumber data paralel berkualitas untuk pasangan bahasa daerah, termasuk Aceh-Indonesia, masih sangat terbatas. Hal ini menggarisbawahi pentingnya eksplorasi lebih lanjut untuk meningkatkan kualitas model penerjemahan dengan data terbatas.

Penelitian ini berfokus pada pengembangan model penerjemahan mesin untuk pasangan Bahasa Aceh-Indonesia dengan memanfaatkan dataset NusaX-MT, salah satu dataset beranotasi untuk bahasa-bahasa daerah di Indonesia. Penelitian ini melakukan serangkaian langkah, termasuk eksplorasi data, analisis distribusi kata, pra pemrosesan data, pelatihan model T5 dengan teknik penalaan halus, serta evaluasi kinerja model melalui metrik standar seperti BLEU dan METEOR. Penelitian ini juga menyelidiki pola linguistik dalam data untuk mengidentifikasi kesulitan dalam penerjemahan dari Bahasa Aceh ke Indonesia.

Penelitian ini bertujuan untuk mengembangkan dan menilai model terjemahan mesin yang akurat untuk bahasa Aceh-Indonesia. Evaluasi akan dilakukan berdasarkan metrik yang relevan untuk mengukur kualitas hasil terjemahan. Penelitian ini juga bertujuan untuk mendorong pengembangan lebih lanjut dalam pemrosesan bahasa alami untuk bahasa-bahasa daerah di Indonesia. Artikel ini disusun dengan struktur sebagai berikut: Bagian 2 menjelaskan tinjauan pustaka dan metode penelitian, Bagian 3 memaparkan hasil dan pembahasan, Bagian 4 menyajikan kesimpulan dan saran untuk penelitian mendatang.

## 2 METODE

Pendekatan penelitian ini dirancang secara sistematis untuk mengembangkan model penerjemahan mesin berbasis arsitektur T5 yang dapat menerjemahkan teks dari Bahasa Aceh ke Bahasa Indonesia dengan hasil yang akurat dan kontekstual. Metode ini terdiri dari beberapa tahapan utama yang dimulai dari pengumpulan data, eksplorasi data, hingga pelatihan dan evaluasi model. Tahapan-tahapan ini dirancang untuk memastikan bahwa setiap proses dilakukan dengan mempertimbangkan karakteristik linguistik kedua bahasa, serta keterbatasan data paralel yang tersedia. Adapun alur dalam penelitian ini dapat dilihat pada Gambar 1.




**Gambar 1.** *Flowchart Alur Penelitian*

### 2.1 Deskripsi Dataset

Penelitian ini mengembangkan model penerjemahan mesin untuk pasangan bahasa Aceh-Indonesia dengan menggunakan dataset NusaX-MT, yang merupakan koleksi data paralel untuk

bahasa-bahasa daerah di Indonesia. Dataset ini memiliki pasangan kalimat dalam bahasa Aceh dan bahasa Indonesia yang telah dianotasi secara manual. Dataset ini terdiri dari 500 baris data dengan lima kolom: "id", "text_1", "text_2", "text_1_lang", dan "text_2_lang". Kolom "text_1" berisi teks dalam bahasa Aceh, sementara "text_2" adalah terjemahannya dalam bahasa Indonesia.

**Tabel 1**. Data Terjemahan Aceh-Indonesia

| id | text_1 | text_2 | lang1 | lang2 |
|---|---|---|---|---|
| 0 | *Neumeuseunang ngon neubayeue bacut-bacut angsuran 0% persen sampoe dua blah buleuen keu nyang bloe tiket kapai teureubang ngon keureutu kredit BNI!* | Nikmati cicilan 0% hingga 12 bulan untuk pemesanan tiket pesawat air asia dengan kartu kredit bni! | ace | ind |
| 1 | *Kueh nyang dihidang peuingat lon masa dilee. Bandum moden kueh jameun, get dari beuntuk atawa rasa. Kuehjih mangat dan yum pih murah.* | Kue-kue yang disajikan bikin saya bernostalgia. Semuanya tipikal kue zaman dulu, baik dari penampilan maupun rasa. Kuenya enak dan harganya juga murah. | ace | ind |
| 2 | *Ibu tom geukereuja bak grab indonesia* | Ibu pernah bekerja di grab indonesia | ace | ind |
| … | … | … | … | … |
| 498 | *Nyan tajaga peunajoh gus, cit lawet nyeo teungoh musem. Nyankeuh geulak nyan nyang peuget seket* | Dijaga ya makannya gus emang lagi musimnya sekarang itu. Iyasi jenuh itu malah bikin sakit | ace | ind |
| 499 | *Trep that inoe hipertensi ta merumpok ngoen pasien BPJS, bayeue hana meupadum pih hawa keu pelayanan eksekutif dan lepah meupep-pep, ek geuritan mantong meunyo meuh'eut nyang eksekutif* | Lama-lama di sini gua bisa hipertensi ngadepin pasien bpjs, bayar tidak seberapa tapi mau pelayanan | ace | ind |

Dataset tersebut terdiri dari tiga bagian utama: 80% *train*, 10% *test*, dan 10% *validation*. Secara matematis, jika dataset terdiri dari N kalimat, maka jumlah data untuk masing-masing bagian adalah sebagai berikut:

$$N_{train} = 0.8 \times N, N_{valid} = 0.1 \times N, N_{test} = 0.1 \times N \quad (1)$$

Keterangan:

- *Training Set* : Digunakan untuk melatih model.
- *Validation Set* : Digunakan untuk melakukan evaluasi selama pelatihan guna mencegah *overfitting*.
- *Test Set* : Digunakan untuk menguji kinerja model yang telah dilatih.

Dalam proses ini, dataset ini diubah menjadi format pandas DataFrame untuk memudahkan analisis dan perosesan lebih lanjut.

## 2.2 Praproses Data

Sebelum melatih model, data mentah perlu diproses terlebih dahulu. Langkah-langkah praproses data meliputi tokenisasi, *padding*, dan pemisahan kalimat menjadi input dan target.

### 2.2.1 Tokenisasi Input dan Target

Tokenisasi adalah proses mengubah kalimat menjadi unit terkecil (token). Dalam hal ini, tokenisasi dilakukan menggunakan T5 *Tokenizer*, yang memisahkan kata-kata dalam kalimat input dan target menjadi token yang dapat diproses oleh model. Misalkan $X_i$ adalah kalimat input dalam bahasa Aceh dan $Y_i$ adalah kalimat target dalam bahasa Indonesia, tokenisasi dilakukan sebagai berikut:

$$Tokenized\ Input\ =\ Tokenizer(X_i) \quad (2)$$

$$Tokenized\ Target\ =\ Tokenizer(Y_i) \quad (3)$$

### 2.2.2 *Padding* dan *Truncation*

Untuk memastikan bahwa semua input dan target memiliki panjang yang seragam, dilakukan *padding* (penambahan tekon kosong) atau *truncation* (pemangkasan) hingga panjang tertentu. Misalnya, panjang maksimum 256 token:

$$Input_{pad}\ =\ Padding\ (Tokenized\ Input, max\_length\ =\ 256) \quad (4)$$

$$Target_{pad}\ =\ Padding\ (Tokenized\ Target, max\_length\ =\ 256) \quad (5)$$

Keterangan:

- *Padding* : Menambahkan token kosong sehingga panjang maksimum yang diinginkan.
- *Truncation* : Memangkas kalimat yang lebih panjang dari panjang maksimum.

### 2.2.3 Pembuatan Input dengan Prefix

Untuk memberikan konteks pada model mengenai tugas yang harus dilakukan (terjemahan), prefix "*Translate Acehnese to Indonesia:*" ditambahkan pada setiap kalimat input:

$$Input_{prefixed}\ =\ "Translate\ Acehnese\ to\ Indonesia:"\ +\ X_i \quad (6)$$

## 2.3 T5 (*Text-to-Text Transfer Transformer*)

Model yang digunakan dalam penelitian ini adalah T5 (*Text-to-Text Transfer Transformer*), yang merupakan model pre-trained dari Google. Model ini dirancang untuk menangani berbagai tugas *Natural Language Processing* (NLP) dalam format teks-ke-teks. Untuk tugas terjemahan ini, model T5-base digunakan. Model T5 bekerja dengan memetakan kalimat input (misalnya X) ke output yang diinginkan, misalnya:

$$\hat{Y}\ =\ T5(X) \quad (7)$$

Keterangan:

- $X$ : Kalimat dalam bahasa Aceh.
- $\hat{Y}$ : Kalimat hasil terjemahan dalam bahasa Indonesia yang dihasilkan oleh model.

## 2.4 Metodologi Pelatihan Model

Proses pelatihan model dilakukan dengan menggunakan *Seq2SeqTrainer* dari pustaka *HuggingFace Transformer*. Pelatihan dilakukan dengan menggunakan *optimizer* AdamW dengan *learning rate* $\alpha = 3 \times 10^{-5}$. Fungsi kerugian yang digunakan adalah *cross-entropy loss*, yang

mengukur perbedaan antara token yang diprediksi oleh model dan token yang benar. Fungsi kerugian ini dapat dinyatakan sebagai:

$$L = -\sum_{i=1}^{n} y_i log(\hat{y}_i) \quad (8)$$

Keterangan:

- $y_i$ : Token target yang benar (referensi)
- $\hat{y}_i$ : Token yang diperoleh oleh model
- $n$ : Jumlah token dalam kalimat

Fungsi kerugian ini mengukur seberapa baik model memprediksi token yang benar pada setiap posisi dalam kalimat.

## 2.5 Evaluasi Model

Setelah model dilatih, kinerjanya dievaluasi menggunakan beberapa metrik evaluasi terjemahan, yang mencakup:

### 2.5.1 BLEU

BLEU (*Bilingual Evaluation Understudy*) mengukur seberapa banyak n-gram dalam kalimat terjemahan yang cocok dengan n-gram pada kalimat terjemahan yang cocok dengan n-gram pada kalimat referensi. Rumus BLEU secara umum adalah:

$$BLEU = min\left(exp(\sum_{n=1}^{N} Weight_n \times Precision_n)\right) \quad (9)$$

Keterangan:

- $Precision_n$ : Akurasi n-gram pada level kalimat.

### 2.5.2 METEOR

METEOR (*Metric for Evaluation of Translation with Explicit ORdering*) adalah metrik yang menggabungkan precision, recall, dan keselarasan sinonim. Rumus METEOR adalah:

$$METEOR = F \times Precision \quad (10)$$

Keterangan:

- $F$ : Faktor rekonsiliasi, yang menghitung penyamaan sinonim dan urutan kata.

## 2.6 Inferensi dan Pengujian

Setelah model dilatih, tahap inferensi dilakukan pada data uji. Misalnya, kalimat dalam Bahasa Aceh $X_{test}$ diterjemahkan ke dalam Bahasa Indonesia oleh model yang telat dilatih. Proses inferensi dapat dinyatakan sebagai:

$$\hat{Y}_{test} = T5(X_{test}) \quad (11)$$

Keterangan:

- Inferensi : Proses menghasilkan *output* dari model berdasarkan input yang baru (data uji).

# 3 HASIL DAN PEMBAHASAN

## 3.1 Analisis Distribusi Kata dan Visualisasi Word Cloud

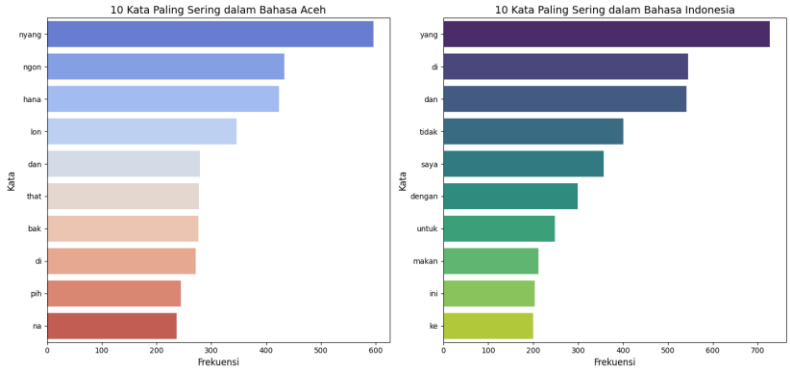
Pengembangan model penerjemahan mesin Aceh-Indonesia berbasis T5 ini menghadirkan sejumlah temuan menarik yang memberikan wawasan baru mengenai tantangan dan peluang dalam mengolah bahasa daerah menggunakan teknologi terkini. Salah satu hasil yang penting adalah analisis distribusi kata dan visualisasi *word cloud* yang memberikan gambaran jelas tentang karakteristik linguistik bahasa Aceh dan Indonesia.


**Gambar 2.** *Word Cloud* Bahasa Aceh

Dalam analisis distribusi kata dan visualisasi *word cloud* bahasa Aceh, terlihat bahwa kata-kata seperti "*nyang*", "*ngon*", dan "*hana*" sering muncul. Hal ini sejalan dengan temuan sebelumnya yang menunjukkan bahwa kata-kata tersebut memiliki peran sentral dalam struktur kalimat dan makna dalam Bahasa Aceh. Kata-kata ini sering digunakan dalam berbagai konteks, yang memperkuat pentingnya pemahaman mendalam tentang distribusi kata untuk pengembangan model penerjemahan yang lebih akurat.


**Gambar 3.** *Word Cloud* Bahasa Indonesia

Sebaliknya, pada *word cloud* bahasa Indonesia yang ditampilkan pada Gambar 3, kata-kata seperti "yang", "dan", dan "tidak" mendominasi, mencerminkan struktur sintaksis dan pola penggunaan kata-kata fungsi yang lebih umum. Hal ini menunjukkan perbedaan yang mencolok dengan bahasa Aceh, di mana kata-kata seperti "*nyang*", "*ngon*", dan "*hana*" lebih sering digunakan dalam konteks yang lebih spesifik dan kontekstual. Perbedaan ini menggarisbawahi tantangan dalam penerjemahan mesin, di mana pengolahan kata-kata fungsi dalam bahasa Indonesia yang lebih banyak digunakan harus dipadukan dengan pemahaman terhadap peran kata-kata inti dalam bahasa Aceh. Pemahaman mendalam terhadap perbedaan distribusi kata ini penting untuk mengembangkan model penerjemahan yang lebih akurat dan efektif dalam mentransfer makna antarbahasa.




**Gambar 4.** Plot Distribusi 10 Kata Paling Sering dalam Bahasa Aceh dan Bahasa Indonesia

Grafik pada Gambar 4 menampilkan distribusi 10 kata yang paling sering muncul dalam Bahasa Aceh dan Bahasa Indonesia berdasarkan analisis dataset yang digunakan, yang semakin memperjelas perbedaan struktural antara kedua bahasa tersebut. Sebagaimana terlihat pada grafik sebelah kiri, dalam Bahasa Aceh, kata "*nyang*" mendominasi, diikuti oleh kata "*ngon*", "*hana*", dan "*lon*". Kata-kata ini berfungsi sebagai elemen inti yang membentuk struktur kalimat dalam bahasa Aceh, dengan penggunaan kata tugas dan kata ganti yang lebih spesifik dalam konteks percakapan sehari-hari.

Di sisi lain, grafik sebelah kanan menunjukkan bahwa dalam Bahasa Indonesia, kata-kata seperti "yang", "di", "dan", dan "tidak" lebih dominan. Hal ini mencerminkan kecenderungan Bahasa Indonesia untuk mengandalkan kata-kata fungsi dalam menyusun kalimat, yang berfungsi untuk menghubungkan elemen-elemen dalam kalimat secara lebih fleksibel. Perbedaan dalam distribusi kata ini menegaskan tantangan yang dihadapi dalam pengembangan model penerjemahan mesin, di mana model tidak hanya harus memahami perbedaan distribusi kata, tetapi juga harus dapat menginterpretasikan makna kata yang sangat bergantung pada konteks dan struktur sintaksis yang khas dari masing-masing bahasa. Analisis ini memberikan wawasan lebih lanjut mengenai kebutuhan akan model penerjemahan yang mampu menangani perbedaan fundamental dalam struktur bahasa, serta pentingnya pemahaman mendalam tentang pola linguistik untuk menghasilkan terjemahan yang akurat dan kontekstual.

### 3.2 Analisis Bigram dan Trigram dalam Teks Bahasa Aceh dan Bahasa Indonesia

Analisis bigram dilakukan untuk mengidentifikasi pola penggunaan kata dalam teks berbahasa Aceh dengan menghitung pasangan kata yang muncul berurutan. Fungsi n-grams dengan parameter n = 2 digunakan untuk membentuk bigram, yang kemudian dihitung frekuensinya.

Hasil analisis pada Tabel 1 menunjukkan beberapa pasangan kata yang sering muncul, seperti "*pajoh bu*" dan "*hana that*" yang masing-masing muncul 37 kali, mengindikasikan penggunaan frasa umum dalam percakapan atau tulisan berbahasa Aceh. Begitu pula bigram seperti "*jak u*", "*makanan nyang*", dan "*jeuet keu*", yang masing-masing muncul dengan frekuensi 34, 33, dan 29 kali. Ini menunjukkan pola penggunaan kata yang konsisten dalam konteks tertentu, seperti percakapan tentang makanan atau tujuan.

**Tabel 2.** Frekuensi Bigram dalam Teks Bahasa Aceh

| Bigram | Frekuensi |
|---|---|
| *pajoh bu* | 37 |
| *hana that* | 37 |
| *jak u* | 34 |
| *makanan nyang* | 33 |
| *jeuet keu* | 29 |
| *nyang ka* | 25 |
| *pih hana* | 25 |
| *na di* | 23 |
| *rumoh makan* | 23 |
| *nyang hana* | 22 |

Selain itu, bigram seperti "*rumoh makan*" dan "*nyang hana*" menunjukkan konteks spesifik, misalnya terkait tempat makan atau ketiadaan sesuatu. Temuan ini memberikan gambaran tentang struktur bahasa Aceh dan kata-kata yang sering digunakan dalam percakapan sehari-hari, serta dapat digunakan sebagai dasar untuk analisis lanjutan dalam pemrosesan bahasa alami atau penelitian linguistik.

**Tabel 3.** Frekuensi Bigram dalam Teks Bahasa Indonesia

| Bigram | Frekuensi |
|---|---|
| makan di | 44 |
| di sini | 39 |
| makanan yang | 33 |
| tidak ada | 31 |
| ke sini | 30 |
| tidak terlalu | 30 |
| enak dan | 29 |
| tempat ini | 28 |
| yang sangat | 27 |
| di tempat | 25 |

Sedangkan, hasil analisis pada Tabel 3 menunjukkan beberapa pasangan kata yang sering muncul, seperti "makan di" dan "di sini", yang masing-masing muncul 44 dan 39 kali. Frasa-frasa ini menunjukkan bahwa kata-kata yang berhubungan dengan aktivitas makan dan lokasi (seperti "makan di" dan "di sini") sering digunakan dalam konteks percakapan atau tulisan. Selain itu, bigram "makanan yang" (33 kali) dan "tidak ada" (31 kali) juga muncul dengan frekuensi tinggi, mengindikasikan pembicaraan yang berkaitan dengan jenis makanan atau ketidakhadiran sesuatu. Beberapa bigram lainnya, seperti "ke sini" (30 kali), "tidak terlalu" (30 kali), dan "enak dan" (29 kali), menunjukkan adanya pola percakapan yang sering mengacu pada tujuan ("ke sini"), penilaian ("tidak terlalu"), dan kualitas ("enak dan"). Bigram seperti "tempat ini" (28 kali) dan "yang sangat" (27 kali) menunjukkan bahwa diskusi sering melibatkan penekanan pada tempat atau tingkat intensitas suatu sifat. Bigram "di tempat" (25 kali) lebih lanjut memperkuat adanya pembicaraan mengenai lokasi atau tempat tertentu.

Secara keseluruhan, analisis bigram ini memberikan wawasan tentang struktur bahasa Indonesia yang digunakan dalam teks, serta pola kata yang sering muncul dalam percakapan atau tulisan. Temuan ini dapat digunakan untuk mendalami lebih lanjut konteks penggunaan bahasa dalam situasi tertentu, serta sebagai dasar untuk aplikasi lebih lanjut dalam pemrosesan bahasa alami atau penelitian linguistik.

Analisis trigram dilakukan untuk menggali pola penggunaan tiga kata yang muncul berurutan dalam teks berbahasa Aceh. Fungsi ngrams dengan parameter n = 3 digunakan untuk membentuk trigram, yang kemudian dihitung frekuensinya. Hasil analisis menunjukkan beberapa trigram yang sering muncul dalam teks yang dianalisis.

**Tabel 4.** Frekuensi Trigram dalam Teks Bahasa Aceh

| Trigram | Frekuensi |
|---|---|
| *pajoh bu leuho* | 11 |
| *Rumoh makan nyoe* | 7 |
| *nyang na di* | 7 |
| *na di jalan* | 7 |
| *rumoh makan nyoe* | 6 |
| *peue lom meunyo* | 5 |
| *pajoh bu bak* | 5 |
| *nyang ka leupah* | 5 |
| *pih hana that* | 5 |
| *mangat dan yum* | 4 |

Analisis trigram dalam bahasa Aceh menunjukkan berbagai kombinasi kata yang sering muncul, memberikan pemahaman yang mendalam tentang pola penggunaan bahasa ini dalam komunikasi lisan maupun tulisan. Salah satu trigram yang paling umum adalah "pajoh bu leuho," yang berarti "makan nasi lauk." Dengan frekuensi kemunculan sebanyak 11 kali, frasa ini mencerminkan popularitasnya dalam menggambarkan aktivitas makan, yang merupakan tema sentral dalam budaya Aceh dan sering digunakan dalam konteks sehari-hari maupun narasi kuliner. Selain itu, trigram "Rumoh makan nyoe," yang berarti "rumah makan ini," dan "nyang na di," yang berarti "yang ada di," masing-masing muncul 7 kali. Kedua frasa ini menegaskan penggunaan kata-kata yang merujuk pada lokasi, menunjukkan kecenderungan bahasa Aceh untuk menyoroti elemen spasial dalam komunikasi.

Trigram lain yang menarik adalah "na di jalan," yang berarti "ada di jalan," juga dengan frekuensi 7 kali, serta "peue lom meunyo," yang berarti "apa lagi jika," muncul sebanyak 5 kali. Penggunaan frasa-frasa ini mencerminkan pola ekspresi kontekstual yang digunakan untuk menggambarkan kondisi atau situasi tertentu dalam percakapan. Frasa "pajoh bu bak," yang berarti "makan nasi dengan," dan "nyang ka leupah," yang berarti "yang sudah lewat," masing-masing juga muncul 5 kali. Ini menunjukkan pentingnya ekspresi yang merujuk pada keadaan atau aktivitas tertentu dalam bahasa Aceh, memberikan gambaran tentang bagaimana masyarakat Aceh mendeskripsikan pengalaman sehari-hari secara spesifik. Trigram "pih hana that," yang berarti "tidak ada itu," muncul sebanyak 5 kali, menegaskan pola penggunaan bahasa yang mencerminkan penolakan atau pernyataan ketiadaan sesuatu, penting untuk memahami elemen negasi dalam struktur bahasa

Aceh. Selain itu, frasa "mangat dan yum," yang berarti "enak dan lezat," muncul 4 kali, menunjukkan bagaimana bahasa Aceh digunakan untuk memberikan penilaian atau deskripsi rasa makanan, topik penting dalam budaya Aceh yang kaya akan tradisi kuliner.

Secara keseluruhan, analisis trigram ini tidak hanya memberikan wawasan tentang pola penggunaan tiga kata berurutan dalam bahasa Aceh tetapi juga memperjelas struktur kalimat, idiom, dan konteks spesifik yang sering muncul dalam komunikasi. Temuan ini sangat relevan untuk pengembangan aplikasi pemrosesan bahasa alami, terutama dalam penerjemahan atau analisis teks otomatis. Dengan memahami pola linguistik ini, peneliti dan pengembang teknologi dapat menciptakan model yang lebih akurat dan kontekstual untuk mengolah bahasa Aceh. Analisis ini juga berkontribusi pada pelestarian bahasa Aceh, membantu mengungkap kompleksitas dan keindahan bahasa daerah yang kaya akan nilai budaya.

**Tabel 5**. Frekuensi Trigram dalam Teks Bahasa Indonesia

| Trigram | Frekuensi |
|---|---|
| makan di sini | 8 |
| makan siang di | 7 |
| untuk makan malam | 7 |
| tidak ada yang | 7 |
| Makanan di sini | 7 |
| makanan dan minuman | 6 |
| ke tempat ini | 5 |
| dengan harga yang | 5 |
| yang berada di | 5 |
| di malam hari | 5 |

Beberapa trigram yang paling sering muncul adalah “makan di sini” (8 kali), yang menunjukkan bahwa frasa ini sering digunakan dalam konteks percakapan atau tulisan yang berkaitan dengan lokasi makan, misalnya restoran atau tempat makan lainnya. Trigram “makan siang di” (7 kali) dan “untuk makan malam” (7 kali) mengindikasikan pembicaraan yang berfokus pada waktu makan tertentu, yaitu makan siang atau makan malam, yang sering menjadi topik dalam teks. Selain itu, trigram “tidak ada yang” (7 kali) menunjukkan pola negatif yang sering digunakan untuk menyatakan ketiadaan atau ketidaktersediaan sesuatu.

Beberapa trigram lainnya seperti “Makanan di sini” (7 kali) dan “makanan dan minuman” (6 kali) mencerminkan fokus pada jenis makanan atau minuman yang tersedia di suatu tempat. Trigram “ke tempat ini” (5 kali) menunjukkan adanya pembicaraan tentang tujuan atau tempat tertentu, sementara “dengan harga yang” (5 kali) menggambarkan pembahasan mengenai harga suatu layanan atau produk. Trigram “yang berada di” (5 kali) mengarah pada informasi lokasi atau tempat, dan “di malam hari” (5 kali) mengindikasikan pembicaraan yang berkaitan dengan waktu, khususnya malam hari.

Secara keseluruhan, analisis trigram ini memberikan wawasan tentang struktur dan pola penggunaan tiga kata berurutan dalam teks berbahasa Indonesia. Temuan ini menunjukkan adanya konsistensi dalam pembicaraan mengenai makanan, tempat, dan waktu tertentu, serta memberikan

gambaran tentang kebiasaan percakapan atau penulisan dalam konteks tersebut. Hasil analisis trigram ini dapat digunakan untuk aplikasi lebih lanjut dalam pemrosesan bahasa alami atau penelitian linguistik untuk memahami konteks percakapan atau tulisan dalam bahasa Indonesia.

### 3.3 Hasil Evaluasi

Pada penelitian ini, evaluasi model terjemahan dilakukan untuk menilai kualitas hasil terjemahan antara teks berbahasa Aceh dan bahasa Indonesia. Model yang digunakan dalam penelitian ini adalah model terjemahan yang telah dilatih sebelumnya dan disimpan dalam direktori yang telah ditentukan. Fungsi "predict_evaluate" digunakan untuk melakukan prediksi pada dataset dan mengevaluasi hasil terjemahan dengan berbagai metrik, seperti BLEU dan METEOR. Dataset yang digunakan mencakup pasangan teks dalam bahasa Aceh (kolom "text_1") dan target teks dalam bahasa Indonesia (kolom "text_2").

**Tabel 6**. Hasil Evaluasi Terjemahan Model dengan Metrik BLEU dan METEOR

| Id | Input | Predict | Target | BLEU | METEOR |
|---|---|---|---|---|---|
| 208 | *Nyum makananjih biasa mantong, yum pih meuha...* | Indah makanannya biasa mahal. Tempat saya bias... | Rasa makanannya biasa saja dan harganya mahal.... | 17.145 | 0.452 |
| 9 | *Meunyo tajak u The Peak memang terasa leupah j...* | Indonesisch: Memang tajak u The Peak memang te... | Untuk menuju ke the peak memang terasa sangat ... | 6.864 | 0.441 |
| 40 | *Bajee gadoh hana tanggong jaweueb.. hotel bint...* | Indonesisch übersetzt: Bajee gadoh hana tanggo... | Baju hilang ga tanggung jawab.. hotel bintang ... | 14.879 | 0.311 |
| 205 | *Jinoe wisatawan nyang galak diving that kecewa...* | Indah wisatawan yang galak diving that kecewa ... | Saat ini para wisatawan pecinta diving sangat ... | 15.540 | 0.394 |
| 402 | *Gojek akan mulai buka layanan bak meupadup boh...* | Indonesisch: Gojek akan mulai buka layanan bak... | Gojek akan mulai membuka layanan mereka di beb... | 17.678 | 0.609 |
| … | … | … | … | … | … |
| 239 | *Atlet muda nanggroe ikut serpong run 2018 untu...* | Indonesisch übersetzt: Atlet muda nanggroe iku... | Atlet muda tanah air ikuti serpong run 2018 un... | 47.855 | 0.733 |
| 232 | *120 peserta nyang luloh akan dibagi lam peut b…* | Indonesisch übersetzen: 120 peserta yang lulu ... | 120 peserta yang lulus ini akan dibagi menjadi... | 10.307 | 0.436 |

| | | | | | |
|---|---|---|---|---|---|
| 187 | *Musem ujeuen lagee nyoe tiep ujeuen pasti hana...* | Indonesisch übersetzt: Musem ujeuen lagi ini t... | Musim hujan begini setiap hujan pasti tidak te... | 2.627 | 0.043 |
| 150 | *Peujioh narkoba nakeuh salah saboh cara nyang ...* | Indonesisch übersetzt: Pejioh narkoba kepada s... | Jauhi narkoba adalah salah satu cara yang bisa... | 22.894 | 0.515 |
| 198 | *Nyang tham GBK keu tempat konser mbong-mbong t...* | Indonesisch übersetzt: Nyang tham GBK ke tempa... | Yang melarang GBK dijadikan venue konser pada ... | 13.250 | 0.331 |

Hasil evaluasi menunjukkan skor BLEU dan METEOR yang mengukur kualitas terjemahan dari model. Misalnya, pada baris pertama, terjemahan dari teks Aceh "*Nyum makananjih biasa mantong, yum pih meuha…*" menghasilkan terjemahan dalam bahasa Indonesia "Indah makanannya biasa mahal. Tempat saya bias…", dengan skor BLEU sebesar 17.15 dan METEOR sebesar 0.45. Skor BLEU ini mencerminkan kualitas kesamaan n-gram antara terjemahan dan referensi, sementara METEOR mengukur kesesuaian dengan memberikan bobot pada sinonim dan struktur kalimat.

Hasil dari model ini menunjukkan variasi skor yang cukup signifikan di berbagai contoh, dengan beberapa terjemahan menunjukkan skor yang lebih tinggi, misalnya pada contoh dengan id 239 yang mencapai skor BLEU sebesar 47.86 dan METEOR sebesar 0.73, yang menunjukkan kualitas terjemahan yang sangat baik.

Secara keseluruhan, hasil evaluasi ini memberikan gambaran mengenai kemampuan model dalam menerjemahkan teks Aceh ke dalam bahasa Indonesia, dengan metrik yang digunakan memberikan wawasan yang lebih komprehensif mengenai kualitas terjemahan secara sintaktis dan semantik. Model ini menunjukkan kinerja yang bervariasi, dengan beberapa terjemahan mencapai skor yang cukup tinggi, namun masih ada ruang untuk perbaikan, terutama pada contoh-contoh yang memiliki skor BLEU dan METEOR lebih rendah. Temuan ini dapat digunakan untuk lebih meningkatkan model terjemahan, baik melalui peningkatan dataset pelatihan maupun pengoptimalan arsitektur model yang lebih sesuai dengan karakteristik bahasa Aceh dan Indonesia.

**Tabel 7**. Hasil Uji Coba

| Index | Predict | Target |
|---|---|---|
| 0 | *Indah makanannya biasa mahal. Tempat saya bias…* | Rasa makanannya biasa saja dan harganya mahal… |

Hasil prediksi model pada contoh pertama menunjukkan perbedaan antara terjemahan yang dihasilkan oleh model dan terjemahan referensi yang diberikan. Teks sumber dalam bahasa Aceh, yang berisi kalimat "*Nyum makananjih biasa mantong, yum pih meuha…*", diterjemahkan oleh model menjadi "Indah makanannya biasa mahal. Tempat saya bias…" dalam bahasa Indonesia. Sementara itu, terjemahan referensi untuk kalimat tersebut adalah "Rasa makanannya biasa saja dan harganya mahal…".

Perbandingan antara terjemahan model dan referensi ini menunjukkan bahwa meskipun terdapat kesamaan dalam konteks, terutama dalam hal deskripsi makanan yang "mahal", terjemahan model tidak sepenuhnya akurat dalam menyampaikan makna yang dimaksud dalam teks sumber. Misalnya, meskipun model berhasil menerjemahkan kata "makanan" dengan tepat, ungkapan "biasa mahal" dalam terjemahan model lebih mengarah pada harga (mahal), sementara referensi menyampaikan makna rasa makanan ("biasa saja"), yang merupakan perbedaan penting dalam konteks.

Hasil ini mencerminkan tantangan yang dihadapi model terjemahan dalam menangkap nuansa dan konteks budaya yang dapat mempengaruhi terjemahan. Hal ini menekankan pentingnya evaluasi kualitas terjemahan yang lebih mendalam, tidak hanya berdasarkan kesamaan kata-kata, tetapi juga berdasarkan kesesuaian semantik antara terjemahan dan makna yang ingin disampaikan dalam bahasa sumber.

## 4 KESIMPULAN

Penelitian ini berhasil mengembangkan model penerjemahan mesin berbasis arsitektur T5 untuk menerjemahkan bahasa Aceh ke bahasa Indonesia. Proses praproses data, seperti analisis distribusi panjang kalimat, kata-kata paling umum, serta pembuatan n-grams, membantu dalam memahami karakteristik linguistik dari bahasa Aceh dan bahasa Indonesia. Hasil analisis menunjukkan adanya variasi struktur kalimat dan dominasi kosakata tertentu yang mempengaruhi kinerja model dalam menerjemahkan. Model dilatih menggunakan pendekatan *sequence-to-sequence* dengan Seq2SeqTrainer, dan evaluasi dilakukan menggunakan metrik BLEU dan METEOR. Hasil evaluasi menunjukkan skor BLEU sebesar 10.04 dan METEOR sebesar 0.31, yang mengindikasikan model memiliki potensi, meskipun masih terdapat tantangan dalam menangkap kesamaan makna dan struktur kalimat antarbahasa. Secara keseluruhan, model T5 menunjukkan kemampuan dasar dalam menerjemahkan bahasa Aceh ke bahasa Indonesia dengan hasil yang dapat dipahami, meskipun akurasi terjemahan masih memerlukan perbaikan. Skor evaluasi yang diperoleh mencerminkan perlunya optimasi lanjutan, seperti peningkatan data pelatihan, penyesuaian parameter model, dan eksplorasi teknik *fine-tuning* yang lebih efektif. Hasil akhir penelitian ini berupa aplikasi web "AcehLingua", yang dapat digunakan sebagai alat bantu penerjemahan untuk bahasa Aceh ke bahasa Indonesia. Aplikasi ini diharapkan menjadi langkah awal dalam melestarikan bahasa Aceh serta meningkatkan aksesibilitas dan pemahaman terhadap bahasa daerah tersebut di era digital.

## UCAPAN TERIMAKASIH

Ucapan terima kasih yang sebesar-besarnya penulis sampaikan kepada pihak-pihak yang telah memberikan dukungan dan kontribusi dalam penyelesaian penelitian ini. Penulis mengucapkan terima kasih kepada institut yang telah memberikan dukungan dana, pihak yang memberikan izin dan akses data untuk penelitian, serta para dosen yang memberikan arahan dan masukan berharga. Tidak lupa, apresiasi mendalam juga ditujukan kepada individu maupun kelompok yang telah membantu dalam proses pengumpulan dan pengolahan data. Penelitian ini, yang berjudul "Pengembangan Model T5 untuk Penerjemahan Mesin Aceh-Indonesia" tidak akan tercapai tanpa dukungan, kerja sama, dan bantuan dari semua pihak yang telah terlibat.

## DAFTAR PUSTAKA

Aji, A. F., Winata, G. I., Koto, F., Cahyawijaya, S., Romadhony, A., Mahendra, R., Kurniawan, K., Moeljadi, D., Prasojo, R. E., Baldwin, T., Lau, J. H., & Ruder, S. (2022). One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 7226–7249. https://doi.org/10.18653/v1/2022.acl-long.500

Anandia, Y., Abila, K., Sinaga, N., & Muhammad. (2024). Pengaruh bahasa daerah terhadap evolusi bahasa Indonesia. Journal of Citizen Research and Development, 1(2), 691–696. https://doi.org/10.57235/jcrd.v1i2.3882

Fajria, N., Sari, S. M., Kasmini, L., & Syarfuni, S. (2024). Peran generasi muda dalam melestarikan tradisi dan bahasa Aceh. TEKAD, 2(1). Retrieved from https://eproceeding.bbg.ac.id/tekad/article/view/290

Ghirrid, A. A., Sari, R. T. K., & Aldisa, R. T. (2024). Algoritma Natural Language Processing untuk aplikasi penerjemah (Indonesia – Jawa) menggunakan metode speech processing. JTIK (Jurnal Teknologi Informasi dan Komunikasi, 8(3), 746–759. https://doi.org/10.35870/jtik.v8i3.2244

Hafizh, R. H. (2024). Pengembangan chatbot berbasis jaringan saraf transformer untuk layanan informasi akademik dan keuangan mahasiswa di Universitas Muhammadiyah Sukabumi. JITET (Jurnal Informatika dan Teknik Elektro Terapan, 12(3), 3412. https://doi.org/10.23960/jitet.v12i3.5002

Handayani, Y. (2019). Ragam bahasa di Indonesia. Bhuana Ilmu Populer.

Hasibuan, A., Siregar, W. V., & Riskina, S. (2022). Sekelumit keberagaman Lhokseumawe dan Aceh Utara. Pelataran Sastra Kaliwungu. ISBN 978-623-5852-08-9.

Maksum, M. S., Arifin, T., Rohidin, R., Prasetya, M. A. B., & Anshori, I. F. (2024). Optimalisasi algoritma terjemahan bahasa dengan model transformer: Pendekatan statistical machine learning. INFOTECHjournal, 10(2), 282–287.

Mustaqim, M., Aziz, Z. A., & Muthalib, K. A. (2020). The use of Acehnese-Chamic derived lexical items used by English department students of UIN Ar-Raniry. English Education Journal, 11(1), 88–100. Retrieved from https://jurnal.usk.ac.id/EEJ/article/view/15279

Nisa, F. (2023, January 20). 12 macam bahasa daerah Aceh yang digunakan penutur lokal di Tanah Rencong. Good News From Indonesia (GNFI). Retrieved from https://www.goodnewsfromindonesia.id

Patria, H. T. (2023, January 20). IPKN Kemenparekraf: Aceh masuk 5 besar sub-indeks travel and tourism demand drivers. Prakarsa Kemenparekraf. Retrieved from https://prakarsa.kemenparekraf.go.id/

Rahman, Y. A., Nurcahyono, O. H., Ummah, S. R., Jondar, A., Supriadi, S., Irsyadi, A. N., Parahita, B. N., Zuhri, S., Koynja, J. J., Sutanto, H., & et al. (2022). Keanekaragaman budaya, bahasa dan kearifan lokal masyarakat Indonesia (Resolusi, pergeseran, pemertahanan, dan konservasi kearifan lokal di Indonesia). Eureka Media Aksara.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, 30 (NIPS 2017). Retrieved from https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf