Implementation and Evaluation of Lexical Complexity Prediction Models from SemEval-2021 Task 1

Anggota Kelompok Ahmad:
Tegar Prasetyo (23/520277/PA/22364) Ahmad Sudais (23/520192/PA/22352)
Kevin Andreas Sitanggang (23/512227/PA/21874) Muhamad Hafiz Saputra (23/518511/PA/22252)
Vincentius Davin Febrillianagata (23/520016/PA/22330)

1. Shared Task yang Dipilih

a. Nama Shared Task: Lexical Complexity Prediction
b. Sumber: https://sites.google.com/view/lcpsharedtask2021

2. Definisi Permasalahan

Input dari permasalahan ini berupa sebuah berkas yang berisi beberapa kolom, yaitu id (identitas setiap entri), corpus (sumber data kata), sentence (kalimat tempat kata muncul), token (kata tunggal yang diukur tingkat kompleksitasnya), dan complexity (nilai tingkat kesulitan kata dalam skala 0–1). Output dari program ini adalah prediksi nilai kompleksitas dari kata atau frasa dalam konteks kalimat. Shared task ini dirancang untuk mengukur kemampuan model dalam memprediksi tingkat kompleksitas leksikal suatu kata berdasarkan konteksnya dalam kalimat. Tujuannya adalah untuk mengembangkan sistem yang dapat secara otomatis menilai sejauh mana suatu kata dianggap sulit dipahami oleh pembaca umum. Untuk menilai kinerja sistem, Evaluasi akan dilakukan menggunakan metrik Pearson correlation, MAE, Spearman's Rho (), dan R^2, dengan Pearson sebagai metrik utama.

3. Dataset

Dataset yang digunakan berasal dari SemEval-2021 Task 1. Dataset ini terdiri atas tiga berkas, yaitu train, trial, dan test. Setiap berkas memiliki beberapa kolom: id (identitas setiap entri), corpus (sumber data kata), sentence (kalimat tempat kata muncul), token (kata tunggal yang diukur tingkat kompleksitasnya), dan complexity (nilai tingkat kesulitan).

a. Nama Dataset:

SemEval-2021 Task 1: Lexical Complexity Prediction

b. Sumber Dataset:

`https://sites.google.com/view/lcpsharedtask2021/data`

c. Ukuran Data:

- Training: 9.279 samples (1.517 Multiword Expressions | 7.662 kata tunggal)
- Test: 1.101 (184 Multiword Expressions | 917 kata tunggal)

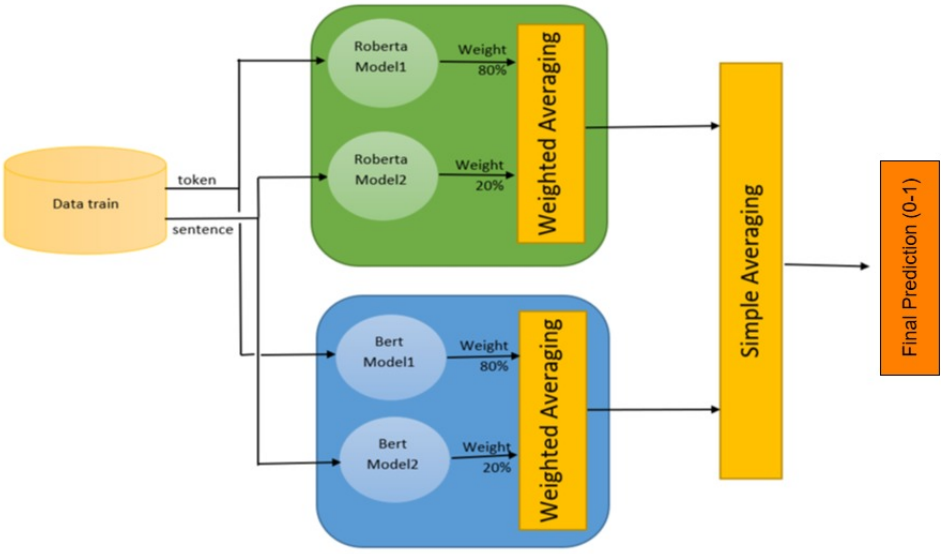
d. Rencana Pembagian Data:

- Training: Diambil sebanyak 80% dari setiap jenis data di data training, yaitu 7.344 sampel
- Validation: Diambil sebanyak 20% dari setiap jenis data di data training, yaitu 1.835 sampel
- Test: Digunakan sesuai ukuran data tes yang tersedia, yaitu 1.101 sampel

4. Pendekatan yang Direncanakan

Arsitektur yang digunakan mengacu pada paper "JUST-BLUE at SemEval-2021 Task 1: Predicting Lexical Complexity using BERT and RoBERTa Pre-trained Language Models" (Bani Yaseen et al., 2021). Model ini dipilih karena relatif mudah diimplementasikan, memiliki dokumentasi parameter yang lengkap, serta telah terbukti efektif dalam tugas Lexical Complexity Prediction (LCP). Arsitektur JUST-BLUE memanfaatkan dua pre-trained language model, yaitu BERT dan RoBERTa, yang diintegrasikan melalui pendekatan ensembling. Implementasi dilakukan menggunakan pustaka Simple Transformers untuk mempermudah proses pelatihan dan pengujian. Pada tahap pelatihan, model menerima masukan berupa pasangan token (kata yang dinilai) dan label complexity (nilai kompleksitas). Selain itu, kompleksitas konteks kalimat juga dipertimbangkan agar prediksi lebih akurat. Hasil dari masing-masing

model kemudian digabungkan menggunakan metode weighted averaging untuk menghasilkan nilai akhir kompleksitas yang lebih stabil dan representatif.



Sumber: Bani Yaseen et al. (2021)

5. Linimasa dan Pembagian Pekerjaan

Rencana pengerjaan proyek akan mengikuti linimasa yang diuraikan pada Tabel 1.

Tabel 1: Rencana Pengerjaan Proyek

Task	**Week 10 (6/11/2025)**	**Week 11 (13/11/2025)**	**Week 12 (20/11/2025)**	**Week 13 (27/11/2025)**	**Week 14 (4/12/2025)**
Proposal Projek					
Mengumpulkan Data(Davin)					
Data Preprocessing (Hafiz)					
Training dan Evaluasi Model(Tegar)					
Uji Model dan Optima-si(Sudais)					
Menulis Laporan Akhir(Kevin)					
Presentasi Hasil Akhir(Kevin)					

6. Daftar Pustaka

Bani Yaseen, T., Ismail, Q., Al-Omari, S., Al-Sobh, E. and Abdullah, M., 2021. JUST-BLUE at SemEval-2021 Task 1: Predicting Lexical Complexity using BERT and RoBERTa Pre-trained Language Models. In: *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Bangkok, Thailand (online), 5–6 August 2021. Association for Computational Linguistics, pp. 661–666. Available at: <`https://aclanthology.org/2021.semeval-1.85`> [Accessed 13 November 2025].