KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI
UNIVERSITAS SUMATERA UTARA
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
PROGRAM STUDI S1 TEKNOLOGI INFORMASI
Jalan Alumni No. 3 Gedung C, Kampus USU Padang Bulan, Medan 20155
Telepon/Fax: 061-8210077 | Email: tek.informasi@usu.ac.id | Laman: http://it.usu.ac.id

FORM PENGAJUAN JUDUL

Nama : Johansen Sihombing

NIM : 211402058

Judul diajukan oleh* : ☐ Dosen
☑ Mahasiswa

Bidang Ilmu (tulis dua bidang) :
1. Data Science
2. Intelligent System

Uji Kelayakan Judul** : ◯ Diterima ◯ Ditolak

Hasil Uji Kelayakan Judul :

Calon Dosen Pembimbing I: Dedy Arisandi, S.T., M.Kom.
(Jika judul dari dosen maka dosen tersebut berhak menjadi pembimbing I)

Calon Dosen Pembimbing II: Sarah Purnamawati, S.T., M.Sc.

Paraf Calon Dosen Pembimbing I

Medan, 11 Februari 2025

Ka. Laboratorium Penelitian,

(Fanindia Purnamasari, S.TI., M.IT)

NIP.198908172019032023

** Centang salah satu atau keduanya*

*** Pilih salah satu*

KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI
UNIVERSITAS SUMATERA UTARA
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
PROGRAM STUDI S1 TEKNOLOGI INFORMASI
Jalan Alumni No. 3 Gedung C, Kampus USU Padang Bulan, Medan 20155
Telepon/Fax: 061-8210077 | Email: tek.informasi@usu.ac.id | Laman: http://it.usu.ac.id

RINGKASAN JUDUL YANG DIAJUKAN

*Semua kolom di bawah ini diisi oleh mahasiswa yang sudah mendapat judul

Judul / Topik Skripsi	**Implementasi Model RoBERTa dan SVR untuk Prediksi Kompleksitas Kata Berbahasa Inggris**
Latar Belakang dan Penelitian Terdahulu	**Latar Belakang** Kompleksitas kata berbahasa Inggris sangat mempengaruhi keterbacaan dan pemahaman pembaca pada teks berbahasa Inggris. Mereka mungkin menyerah, salah menafsirkan, atau terus membaca tanpa memahami maknanya. Pembaca yang berkomitmen mungkin meluangkan waktu untuk mencari arti kata tersebut dan memperluas kosakatanya, tetapi bahkan dalam kasus ini mereka harus meninggalkan teks, yang mengganggu konsentrasi mereka. (Matthew et al., 2021) Pendekatan untuk menentukan kompleksitas kata telah berkembang dari anotasi manual hingga metode berbasis machine learning. Anotasi manual bersifat subjektif dan memakan waktu serta biaya, sementara pendekatan berbasis frekuensi kata tidak mempertimbangkan faktor linguistik lainnya. Sebagai contoh, meskipun kata "*ethnic*" muncul 23.752 kali dalam *COCA*, kata ini dikategorikan sebagai C1 (Mahir) dalam *CEFR*, sedangkan "*jazz*" yang hanya muncul 26 kali dikategorikan sebagai A2 (Pemula Lanjutan). Hal ini menunjukkan bahwa frekuensi tidak selalu mencerminkan kompleksitas kata. Metode klasifikasi juga masih terbatas karena hanya membedakan antara kompleks dan tidak kompleks tanpa mampu membandingkan tingkat kompleksitas antar kata kompleks (Zhang et al., 2020; Matthew et al., 2021). Berdasarkan penelitian Kirana dan Basthomi (2020), bahkan mahasiswa jurusan Bahasa Inggris hanya memiliki ukuran kosakata yang jauh di bawah standar minimum untuk komunikasi akademik. Tes Tingkat Kosakata (*Vocabulary Levels Test/VLT*) terhadap 319 mahasiswa di Institut Agama Islam Negeri Ponorogo menunjukkan bahwa rata-rata kosakata mahasiswa hanya 1.366 kata, sedangkan menurut penelitian Paul Nation (2014), pemahaman efektif dalam membaca teks berbahasa Inggris memerlukan sekitar 9.000 kata. Ukuran kosakata berperan penting dalam pemahaman teks, di mana kemunculan kata yang tidak dikenal dalam kalimat dapat menghambat pemahaman pembaca. Oleh karena itu, identifikasi kata-kata yang berpotensi menimbulkan kesulitan perlu dilakukan untuk meningkatkan keterbacaan teks (Matthew et al., 2021). Teknologi *Machine Learning* dan *Deep Learning* dapat membantu dalam memprediksi kompleksitas kata. Penelitian Zhang et al. (2020) dalam *"Automatic Classification and Comparison of Words by Difficulty"* mengusulkan metode klasifikasi dan perbandingan kesulitan kata dengan menganalisis fitur intra-kata, sintaksis, dan semantik. Dalam tugas klasifikasi kesulitan kata, model *Multi-Layer Perceptron* (MLP) mencapai akurasi 42,94%, sedangkan dalam pemeringkatan pasangan kata berdasarkan kesulitan, model *Support Vector Machine* (SVM) memperoleh akurasi 75,59%. Hasil ini menunjukkan bahwa metode yang diusulkan cukup efektif dalam menentukan tingkat kesulitan kata secara otomatis.

Penelitian Bani Yaseen et al. (2021) dalam *"JUST-BLUE at SemEval-2021 Task 1: Predicting Lexical Complexity using BERT and RoBERTa Pre-trained Language Models"* mengembangkan model prediksi kompleksitas kata menggunakan BERT dan RoBERTa. Dengan teknik *ensembling*, model ini menggabungkan prediksi dari input token dan kalimat, menghasilkan skor korelasi Pearson sebesar 0,819 dalam prediksi kompleksitas kata tunggal, yang menunjukkan hubungan kuat antara prediksi model dan data aktual.

Penelitian Pan et al. (2021) dalam *"DeepBlueAI at SemEval-2021 Task 1: Lexical Complexity Prediction with A Deep Ensemble Approach"* mengembangkan model *ensemble* yang menggabungkan BERT, RoBERTa, ALBERT, dan ERNIE untuk memprediksi kompleksitas kata. Dengan teknik *fine-tuning*, *pseudo-labelling*, dan augmentasi data, serta mekanisme *stacking* dengan regresi linier sebagai meta-model, pendekatan ini mencapai skor korelasi Pearson 0,8612 untuk prediksi kompleksitas frasa multi-kata dan 0,7882 untuk kompleksitas kata tunggal.

Penelitian Mosquera (2021) dalam *"Exploring Sentence and Word Features for Lexical Complexity Prediction"* menggunakan rekayasa fitur untuk memprediksi kompleksitas kata. Dengan menerapkan *LightGBM* pada 51 fitur berbasis kata dan kalimat, model ini mencapai skor korelasi Pearson 0,779 untuk kompleksitas kata tunggal, sementara regresi linier memperoleh skor 0,809 untuk frasa multi-kata.

Penelitian Shardlow et al. (2022) dalam *"Predicting Lexical Complexity in English Texts: The Complex 2.0 Dataset"* menggunakan dataset CompLex 2.0 untuk menilai kompleksitas kata secara lebih objektif dan fleksibel dibandingkan pendekatan biner. Dengan regresi linier sebagai model prediksi, evaluasi menggunakan korelasi Pearson menghasilkan skor keseluruhan 0,771, dengan hasil 0,724 pada data *Europarl*, 0,735 pada data *Bible*, dan 0,784 pada data *Biomedical*.

Penelitian Ortiz-Zambrano et al. (2023) dalam *"Combining Transformer Embeddings with Linguistic Features for Complex Word Identification"* menggabungkan *embedding* dari BERT dan XLM-RoBERTa dengan fitur linguistik, seperti morfologi, sintaksis, statistik, dan semantik, untuk memprediksi kompleksitas kata. Model *Support Vector Regressor* (SVR) yang dilatih dengan kombinasi ini mencapai *Mean Absolute Error* (MAE) 0,0688 dan korelasi Pearson 0,8911, menunjukkan bahwa penggabungan *embedding* Transformer dengan fitur linguistik dapat meningkatkan akurasi prediksi kompleksitas kata.

Penelitian Ortiz-Zambrano et al. (2024) dalam *"Deep Encodings vs. Linguistic Features in Lexical Complexity Prediction"* membandingkan efektivitas fitur linguistik dengan representasi vektor dari model Transformer seperti BERT dan XLM-RoBERTa dalam memprediksi kompleksitas kata. Hasil menunjukkan bahwa kombinasi fitur linguistik dengan *deep encodings* meningkatkan akurasi prediksi dan lebih efisien dalam penggunaan sumber daya komputasi dibandingkan Transformer saja. Pada dataset CompLex (Bahasa Inggris), model terbaik adalah *fine-tuned* BERT dengan fitur linguistik, mencapai *Mean Absolute Error* (MAE) 0,0683. Temuan ini menegaskan bahwa meskipun Transformer menjadi standar dalam NLP, fitur linguistik tetap berkontribusi signifikan dalam prediksi kompleksitas kata.

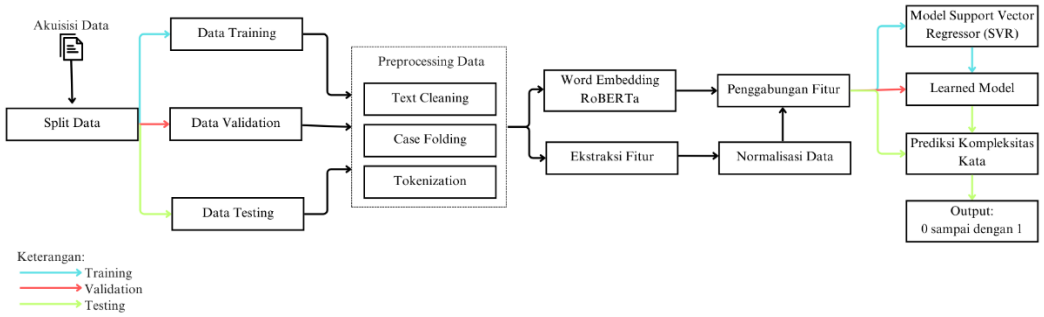
Berdasarkan penelitian yang telah dilakukan, penulis mengusulkan penerapan model *RoBERTa* dan algoritma *Support Vector Regressor* (SVR) untuk memprediksi kompleksitas kata dalam bahasa Inggris. Penulis memberikan penelitian ini judul "Implementasi Model RoBERTa dan SVR untuk Prediksi Kompleksitas Kata Berbahasa Inggris".

Penelitian Terdahulu

No.	Penulis	Judul	Tahun
1.	Shengyao Zhang, Qi Jia, Libin Shen, dan Yinggong Zhao	Automatic Classification and Comparison of Words by Difficulty	2020
2.	Tuqa Bani Yaseen, Qusai Ismail, Sarah Al-Omari, Eslam Al-Sobh, dan Malak Abdullah	JUST-BLUE at SemEval-2021 Task 1: Predicting Lexical Complexity using BERT and RoBERTa Pre-trained Language Models	2021
3.	Chunguang Pan, Bingyan Song, Shengguang Wang, dan Zhipeng Luo	DeepBlueAI at SemEval-2021 Task 1: Lexical Complexity Prediction with A Deep Ensemble Approach	2021
4.	Alejandro Mosquera	Exploring Sentence and Word Features for Lexical Complexity Prediction	2021
5.	Matthew Shardlow, Richard Evans, dan Marcos Zampieri	Predicting Lexical Complexity in English Texts: The Complex 2.0 Dataset	2022

	6.	Jenny A. Ortiz-Zambrano, César Espin-Riofrio, dan Arturo Montejo-Ráez	Combining Transformer Embeddings with Linguistic Features for Complex Word Identification	2023
	7.	Jenny A. Ortiz-Zambrano, César H. Espín-Riofrío, dan Arturo Montejo-Ráez	Deep Encodings vs. Linguistic Features in Lexical Complexity Prediction	2024
Rumusan Masalah	Kompleksitas kata dalam teks berbahasa Inggris memengaruhi pemahaman dan keterbacaan, terutama bagi pembelajar bahasa atau individu dengan keterbatasan linguistik. Prediksi kompleksitas kata yang tersedia masih kurang optimal. Model *Transformer* seperti RoBERTa mampu memahami makna kata dalam berbagai konteks, sementara *Support Vector Regressor* (SVR) efektif dalam memprediksi nilai kontinu. Oleh karena itu, kombinasi RoBERTa dan SVR diharapkan dapat meningkatkan akurasi prediksi kompleksitas kata guna mengidentifikasi kata-kata kompleks dan meningkatkan keterbacaan teks berbahasa Inggris.			

Metodologi



Tahapan Penelitian:

1. **Akuisisi Data**
   Tahap awal penelitian ini adalah pengumpulan data menggunakan *dataset* CompLex 2.0, *dataset* ini mencakup teks dari berbagai sumber serta label berupa nilai kontinu 0 sampai dengan 1 adapun atribut data yang disediakan yaitu, *sentence* (konteks kalimat), *token* (kata target) dan *complexity* (nilai kompleksitas kata).
2. ***Split Data***
   Dataset dibagi menjadi tiga bagian utama, yaitu data pelatihan, validasi, dan pengujian untuk memastikan model dapat belajar secara optimal dan diuji dengan baik. Rasio yang digunakan dalam pembagian data adalah 80% untuk pelatihan, 10% untuk validasi, dan 10% untuk pengujian.
3. ***Preprocessing Data***
   Tahap berikutnya adalah *preprocessing*, yang mencakup beberapa langkah utama, yaitu *text cleaning* dengan menghapus karakter khusus, angka, dan tanda baca yang tidak relevan; *case folding* untuk mengubah seluruh huruf dalam teks menjadi huruf kecil; serta tokenisasi untuk memecah teks menjadi kata-kata/sub-kata (token).
4. **Ekstraksi Fitur**
   Setelah tahap *preprocessing* data, dilakukan ekstraksi fitur linguistik yang mencakup aspek leksikal, morfologi, sintaksis, dan semantik. Fitur leksikal mencakup panjang kata, jumlah suku kata, serta frekuensi kata dalam korpus. Dari aspek sintaksis, dilakukan *POS tagging* untuk menentukan kategori kata dalam suatu kalimat, sementara fitur semantik mencakup jumlah sinonim, hiponim, dan hipernim.
5. ***Word Embedding***
   Selain fitur linguistik, kata akan dikonversi menggunakan *word embedding* dengan RoBERTa untuk menghasilkan representasi kata dalam bentuk vektor sebagai salah satu fitur. *RoBERTa embedding* merupakan pengembangan dari model BERT yang telah dioptimalkan untuk menghasilkan representasi kata yang lebih akurat dalam berbagai konteks.
6. **Normalisasi Data**
   Agar semua fitur memiliki skala yang seragam, dilakukan normalisasi pada data fitur linguistik. Normalisasi ini bertujuan untuk memastikan bahwa perbedaan

skala antara fitur linguistik dan *word embeddings* tidak memengaruhi kinerja model dalam memprediksi kompleksitas kata.

7. **Penggabungan Fitur**
   Setelah fitur linguistik dan *word embedding* diperoleh, tahap berikutnya adalah menggabungkan kedua jenis fitur tersebut. Fitur linguistik yang mencakup aspek leksikal, morfologi, sintaksis, dan semantik akan dikombinasikan dengan representasi vektor dari *word embedding* RoBERTa untuk membentuk satu set fitur yang lebih kaya dan informatif bagi model.
8. **Model**
   Pada tahap ini, model dilatih menggunakan *Support Vector Regressor* (SVR). Setelah proses pelatihan dan validasi, diperoleh *learned model* yang dievaluasi menggunakan *Mean Squared Error* (MSE), *Mean Absolute Error* (MAE), dan Korelasi Pearson. MSE mengukur rata-rata kesalahan kuadrat, MAE menghitung selisih absolut antara prediksi dan nilai sebenarnya, sementara Korelasi Pearson menilai sejauh mana pola prediksi sesuai dengan data aktual.
9. **Output**
   Learned model akan digunakan untuk memprediksi kompleksitas kata dalam bahasa Inggris. Hasil prediksi berupa nilai kontinu dalam rentang 0–1, di mana semakin tinggi nilai output, semakin kompleks kata tersebut.

Referensi

Bani Yaseen, T., Ismail, Q., Al-Omari, S., Al-Sobh, E., & Abdullah, M. (2021). JUST-BLUE at SemEval-2021 Task 1: Predicting lexical complexity using BERT and RoBERTa pre-trained language models. *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, 85. https://aclanthology.org/2021.semeval-1.85.pdf

Kirana, D. P., & Basthomi, Y. (2020). Vocabulary size among different levels of university students. *Universal Journal of Educational Research, 8*(10), 1–6. https://doi.org/10.13189/ujer.2020.081001

Mosquera, A. (2021). Exploring sentence and word features for lexical complexity prediction. *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, 68. https://aclanthology.org/2021.semeval-1.68.pdf

Nation, P. (2014). How much input do you need to learn the most frequent 9,000 words? *Reading in a Foreign Language, 26*(2), 1–16. https://files.eric.ed.gov/fulltext/EJ1044345.pdf

Ortiz-Zambrano, J. A., Espin-Riofrio, C., & Montejo-Ráez, A. (2023). Combining transformer embeddings with linguistic features for complex word identification. *Electronics, 12*(1), 120. https://doi.org/10.3390/electronics12010120

Ortiz-Zambrano, J. A., Espín-Riofrío, C. H., & Montejo-Ráez, A. (2024). Deep encodings vs. linguistic features in lexical complexity prediction. *Neural Computing and Applications.* https://doi.org/10.1007/s00521-024-10662-9

Paetzold, G. H., & Specia, L. (2016). SemEval 2016 Task 11: Complex word identification. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 85. https://aclanthology.org/S16-1085.pdf

Pan, C., Song, B., Wang, S., & Luo, Z. (2021). DeepBlueAI at SemEval-2021 Task 1: Lexical complexity prediction with a deep ensemble approach. *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, 72. https://aclanthology.org/2021.semeval-1.72.pdf

Shardlow, M., Evans, R., Paetzold, G. H., & Zampieri, M. (2021). SemEval-2021 Task 1: Lexical complexity prediction. *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, 1. https://aclanthology.org/2021.semeval-1.1.pdf

Shardlow, M., Evans, R., & Zampieri, M. (2022). Predicting lexical complexity in English texts: The Complex 2.0 dataset. *ResearchGate*. https://www.researchgate.net/publication/359434976_Predicting_lexical_complexity_in_English_texts_the_Complex_20_dataset

Zhang, S., Jia, Q., Shen, L., & Zhao, Y. (2020). Automatic classification and comparison of words by difficulty. In H. Yang, K. Pasupa, A. C.-S. Leung, J. T. Kwok, J. H. Chan, & I. King (Eds.), *Proceedings of the 27th International Conference on Neural Information Processing (ICONIP 2020), Part IV* (pp. 635–642). Springer. https://doi.org/10.1007/978-3-030-63820-7

Medan, 11 Februari 2025
Mahasiswa yang mengajukan,

(Johansen Sihombing)
NIM. 211402058