Pengembangan Model Prediksi Hubungan Kutipan Antar-Paper untuk Sistem Rekomendasi Literatur Ilmiah

GammaFest: Data Science Competition

Bryant Farrel Mohammad Raffy Zeidan Evans Kizito

I. PENDAHULUAN

Di tengah kemajuan era digital saat ini, kemampuan mengakses literatur ilmiah yang relevan menjadi faktor penting dalam meningkatkan mutu penelitian. Meskipun perpustakaan digital dan repositori akademik menampung jutaan karya ilmiah dari berbagai bidang, besarnya volume data justru menimbulkan masalah baru: kesulitan peneliti dalam menemukan referensi yang paling sesuai dengan fokus penelitian mereka. Sistem pencarian konvensional seringkali kurang efektif dalam mengidentifikasi relasi kompleks dan dinamis antarmakalah—khususnya dalam memetakan jejaring sitasi dan kontribusi konseptual antarpublikasi.

Masalah utama dalam manajemen literatur akademik adalah kelemahan sistem rekomendasi referensi yang ada. Banyak platform tidak mampu melakukan analisis mendalam terhadap konten artikel atau melacak hubungan sitasi dengan presisi tinggi. Akibatnya, referensi penting yang relevan secara konseptual bisa terlewatkan, sehingga menghambat pengembangan karya akademik yang berlandaskan riset mendalam.

Solusi terhadap tantangan ini adalah penerapan teknik pembelajaran mesin untuk memprediksi hubungan sitasi antarmakalah secara cerdas. Dengan memanfaatkan data yang tersedia, model dapat menilai apakah suatu makalah perlu mengutip makalah lain secara otomatis. Kompetisi ini mendorong peserta untuk merancang model klasifikasi yang andal, dengan evaluasi berbasis *Matthews Correlation Coefficient* (*MCC*), metrik yang ideal untuk klasifikasi biner dengan distribusi label tidak seimbang.

II. Landasan Teori

A. Document Embedding

- BERT Extraction

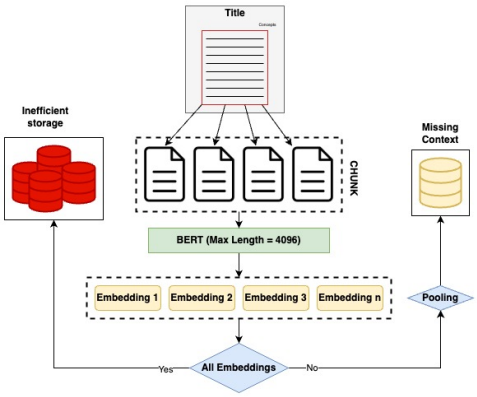


Fig. 1. Document BERT Extraction

Dokumen dibagi menjadi beberapa potongan (chunk), lalu masing-masing diekstrak menggunakan model BERT yang sudah dilatih sebelumnya untuk memperoleh vektor embedding dari setiap potongan tersebut.

- Doc2Vec Extraction

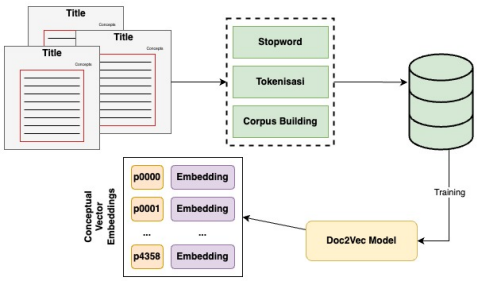


Fig. 2. Document Doc2Vec Extraction

Tidak menggunakan pemrosesan per bagian (chunks), melainkan membangun kosakata atau korpus terlebih dahulu, kemudian melatih model untuk menghasilkan embedding dokumen.

B. Title and Concepts Embedding

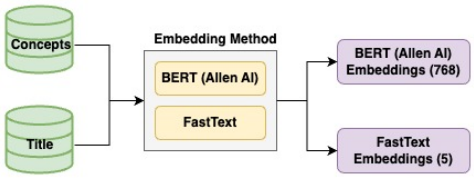


Fig. 3. Text and Concepts Extraction

Mengingat Title dan Concepts memiliki panjang teks yang relatif pendek, metode representasi berbasis FastText dan Allen AI SPECTER dipilih; di mana SPECTER, sebagai model berbasis BERT yang telah dilatih khusus pada korpus akademik, diharapkan mampu menangkap makna semantik yang lebih relevan dalam konteks ilmiah.

C. Distance Similarity

- Cosine Similarity

$$\text{cosine\_similarity} = \frac{A \cdot B}{\|A\|\|B\|} \quad (1)$$

Mengukur sudut antara dua vektor.

- Euclidean Distance

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2)$$

Jarak garis lurus antar dua titik.

- Manhattan Distance

$$d = \sum_{i=1}^{n}|x_i - y_i| \quad (3)$$

Jumlah jarak absolut sepanjang sumbu koordinat.

- Pearson Correlation

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (4)$$

Mengukur korelasi linear antara dua variabel.

D. Vector Calculation

- Mean Combined

$$\text{mean}([v_1, v_2]) \quad (5)$$

Nilai rata-rata dari dua vektor.

- Standar Deviation Diff

$$\text{std}(v_1 - v_2) \quad (6)$$

Variasi perbedaan antar vektor.

- Squared Diff Sumation

$$\sum (v_1 - v_2)^2 \quad (7)$$

”Energi” perbedaan antar elemen.

- Absolute Diff Sumation

$$\sum |v_1 - v_2| \quad (8)$$

Total jarak langsung antar dimensi.

E. Angular and Vector Projection

- **Angle Calculation**

$$\theta = \cos^{-1}(\text{cosine similarity}(v_1, v_2)) \quad (9)$$

Menghitung sudut antara dua vektor berdasarkan nilai cosine similarity-nya. Semakin kecil sudut, semakin mirip arah kedua vektor.

- **Projection Calculation**

$$\text{Proyeksi}_{\vec{b}}\vec{a} = \frac{\vec{a} \cdot \vec{b}}{\|\vec{b}\|} \quad (10)$$

Menghitung panjang proyeksi vektor $\vec{a}$ pada arah vektor $\vec{b}$. Ini menunjukkan seberapa besar komponen $\vec{a}$ searah dengan $\vec{b}$.

III. Proses Analisis

Proses analisis dalam penelitian ini terdiri dari beberapa tahapan sistematis yang dirancang untuk merepresentasikan dokumen akademik secara optimal dan efisien. Berikut adalah tahapan-tahapan yang dilakukan:

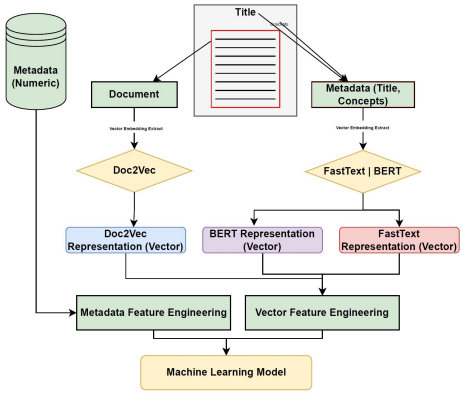


Fig. 4. Research Flow Method

1) **Pemahaman Struktur Data**

Analisis dimulai dengan memahami dua jenis data utama yang tersedia: dokumen penuh (*Full Text*) yang memuat isi lengkap naskah akademik, serta metadata

yang mencakup elemen-elemen ringkas seperti judul (*Title*) dan konsep utama (*Concepts*) dari masing-masing dokumen.

2) **Ekstraksi Embedding Awal**
   Vektor representasi atau embedding diperoleh dari dokumen penuh dan dari metadata. Hal ini dilakukan untuk mengubah informasi teks ke dalam bentuk numerik berdimensi tinggi yang dapat diproses lebih lanjut oleh model pembelajaran mesin.
3) **Evaluasi dan Pemilihan Metode Embedding**
   Dilakukan evaluasi terhadap dua pendekatan embedding, yakni BERT dan Doc2Vec. Meskipun BERT menunjukkan akurasi lebih tinggi dalam menangkap relasi semantik, Doc2Vec dipilih karena efisiensi biaya komputasi serta fleksibilitas dalam penentuan dimensi output. Untuk menjaga efisiensi memori, dimensi embedding Doc2Vec ditetapkan sebesar 90.
4) **Representasi Metadata Pendek**
   Untuk metadata yang terdiri dari teks pendek seperti *Title* dan *Concepts*, digunakan dua pendekatan embedding yaitu FastText dan Allen AI SPECTER. SPECTER dipilih karena merupakan model berbasis BERT yang telah dilatih khusus pada korpus akademik, sedangkan FastText efektif dalam menangkap representasi semantik dari teks pendek. Penggunaan kedua metode ini dianggap tidak membebani secara komputasi karena ukuran input yang kecil.
5) **Rekayasa Fitur Antar Dokumen**
   Dibentuk fitur baru dari kombinasi vektor dokumen dan vektor dokumen referensi. Tujuan dari langkah ini adalah untuk menangkap hubungan semantik antar dokumen yang tidak dapat terwakili hanya oleh embedding individual. Hal ini sangat relevan dalam konteks hubungan sitasi dan referensi dalam literatur ilmiah.
6) **Rekayasa Fitur Temporal dari Metadata**
   Metadata temporal seperti tahun publikasi dinormalisasi, dan selisih tahun antara dokumen dan dokumen referensinya dihitung. Fitur ini ditambahkan untuk mempertimbangkan

aspek kronologis yang dapat memengaruhi pola sitasi atau relevansi antar dokumen.

7) **Eksperimen Model dan Evaluasi Kinerja**
   Model prediktif dikembangkan dengan memanfaatkan kombinasi embedding dan fitur tambahan. Evaluasi dilakukan menggunakan metrik yang sesuai untuk mengukur kualitas dan efektivitas model dalam merepresentasikan dan memprediksi relasi antar dokumen.
8) **Penarikan Kesimpulan**
   Hasil dari seluruh tahapan dianalisis untuk menyimpulkan efektivitas pendekatan yang digunakan, baik dari segi representasi fitur, performa model, maupun efisiensi komputasi yang diperoleh.

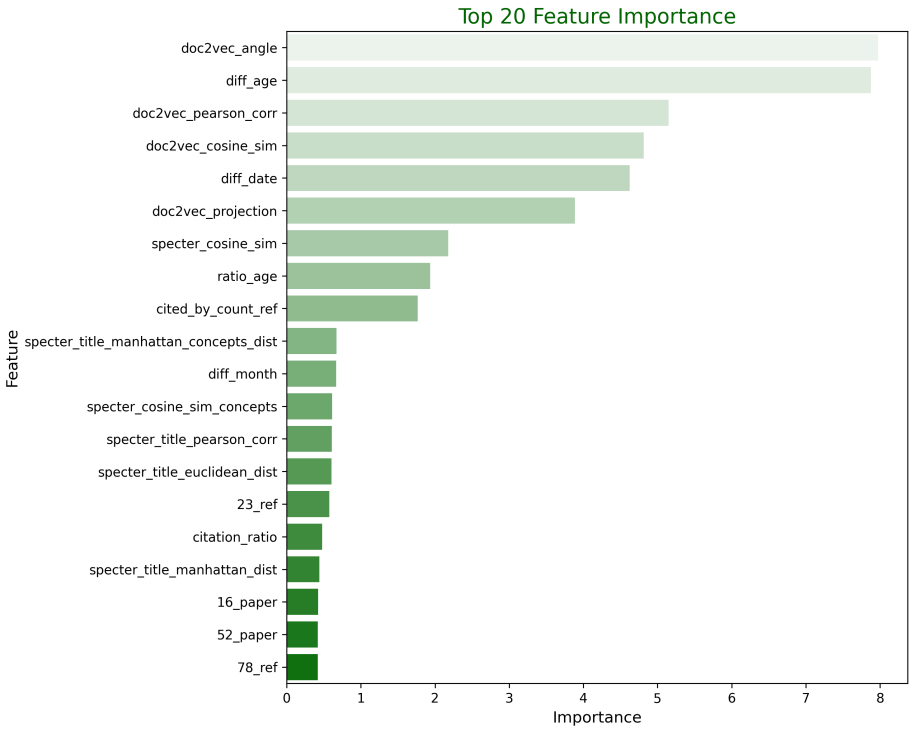
IV. Hasil analisis

Dalam rangka memperoleh hasil yang optimal, sejumlah pendekatan telah diterapkan secara sistematis pada tahap analisis.

TABLE I
PERBANDINGAN PERFORMA MODEL

No.	Model	MCC
1	CatBoost + TF-IDF Document + Metadata Mentah	0.32
2	CatBoost (Tuned) + TF-IDF + TF-IDF Document + Metadata Mentah	0.33
3	LGBM (Tuned) + TF-IDF + Cosine Document + Metadata Mentah	0.41
4	Random Forest (Tuned) + TF-IDF + Cosine Document + Metadata Mentah	0.47
5	CatBoost + Doc2Vec Document + Metadata Mentah	0.50
6	CatBoost + Doc2Vec Document + Fast Text (Concepts + Title) + Metadata Mentah	0.518
7	CatBoost + Doc2Vec Document + Fast Text (Concepts + Title) + FE Metadata	0.520
8	CatBoost + Doc2Vec Document + Fast Text (Concepts + Title) + BERT (Concepts + Title) + FE Metadata + FE Vector Embeddings	0.540
9	CatBoost Tuned + Doc2Vec Document + Fast Text (Concepts + Title) + BERT (Concepts + Title) + FE Metadata + FE Vector Embeddings	0.568

V. KESIMPULAN DAN REKOMENDASI



Representasi vektor embedding dari *Natural Language Processing* (*NLP*) memainkan peran penting dalam merepresentasikan dokumen. Selain itu, informasi tanggal penerbitan turut berkontribusi signifikan dalam mengidentifikasi keterkaitan antar dokumen. Hubungan tersebut dapat dianalisis secara matematis melalui perhitungan jarak atau kesamaan antara dua vektor embedding. Dengan memanfaatkan model machine learning yang sederhana, sistem referensi standar dapat dibangun secara efektif, menghasilkan skor MCC sekitar 0,568 — mendekati 0,57. Pendekatan ini terbukti cukup efisien dari segi biaya, namun tetap mampu memberikan hasil yang memuaskan. Namun, untuk meningkatkan hasil yang lebih optimal, beberapa pendekatan lain dapat dicoba. Salah satunya adalah dengan mengaplikasikan *graph neural network* (*GNN*) yang dapat memanfaatkan hubungan antar dokumen secara lebih komprehensif. Selain itu, penerapan perhitungan vektor yang lebih kompleks, seperti dengan menggunakan teknik *embedding* yang lebih canggih, dapat memberikan representasi yang lebih akurat. Terakhir, perhitungan *similarity* yang lebih *advanced*, misalnya dengan menggunakan model-model *deep learning* atau *hybrid*, dapat memperbaiki analisis kesamaan antar dokumen dan meningkatkan performa model.