UNIVERSITAS MERCU BUANA	UNIVERSITAS MERCU BUANA FAKULTAS ILMU KOMPUTER PROGRAM STUDI SISTEM INFORMASI	**Q**

No.Dokumen	120.423.4.010.00	Distribusi		
Tgl. Efektif	Juni 2008	Kaprodi	POP	Dosen

UJIAN AKHIR SEMESTER GANJIL TAHUN AKADEMIK 2024/2025

Mata Kuliah/SKS : Simulasi Data

Hari/Tanggal : Senin/28 Juli 2025 Waktu : 13:15 – 15:45

Sifat Ujian : Buka Buku Ruang : M-511

Dosen : Muhammad Ali Akbar, M.Kom

---

1. Buatlah Hipotesa Awal dari tujuan kompetisi tersebut: (20 Poin)

CPMK: 3,4 CPL: S8, P1,P2, K1, U9,U10,U11

Dalam upaya mengidentifikasi dan mengklasifikasikan kutipan data dalam literatur ilmiah menjadi dua jenis utama, yaitu primer (data yang dihasilkan dalam makalah) dan sekunder (data yang diperoleh dari sumber lain), dirumuskan hipotesis awal sebagai berikut:

- Hipotesis Nol (H_0):
  Penggunaan model berbasis BERT yang dikombinasikan dengan metode clustering K-Means tidak memberikan perbedaan performa yang signifikan dalam mengklasifikasikan kutipan data dibandingkan dengan pendekatan berbasis TF-IDF dan Random Forest.
- Hipotesis Alternatif (H_1):
  Model berbasis BERT dan K-Means mampu memberikan performa klasifikasi kutipan data yang lebih baik secara signifikan dibandingkan pendekatan lain seperti TF-IDF dan Random Forest.

Namun, berdasarkan hasil simulasi yang telah dilakukan, hipotesis alternatif tersebut tidak terbukti. Justru pendekatan TF-IDF untuk representasi teks dan Random Forest sebagai algoritma klasifikasi memberikan hasil yang lebih memuaskan. Pendekatan ini lebih efisien secara komputasi serta menghasilkan skor evaluasi yang tinggi, khususnya pada metrik F1 Score. Hal ini menunjukkan bahwa metode yang lebih sederhana pun dapat memberikan performa yang kompetitif dalam tugas klasifikasi kutipan data.

2. Buatlah Exploratory Data Analysis dari permasalahan di atas (20 Poin)

CPMK: 1,3,4 CPL: S8, P1,P2, K1, U9,U10,U11

Sebagai langkah awal untuk memahami kompleksitas kutipan data dalam literatur ilmiah, dilakukan eksplorasi menyeluruh terhadap struktur, distribusi, serta karakteristik teks kutipan data yang tersedia. EDA ini bertujuan mengidentifikasi pola-pola tersembunyi, potensi bias dalam data, serta indikasi awal mengenai fitur yang relevan untuk model klasifikasi.

A. Pemeriksaan Struktur Dataset

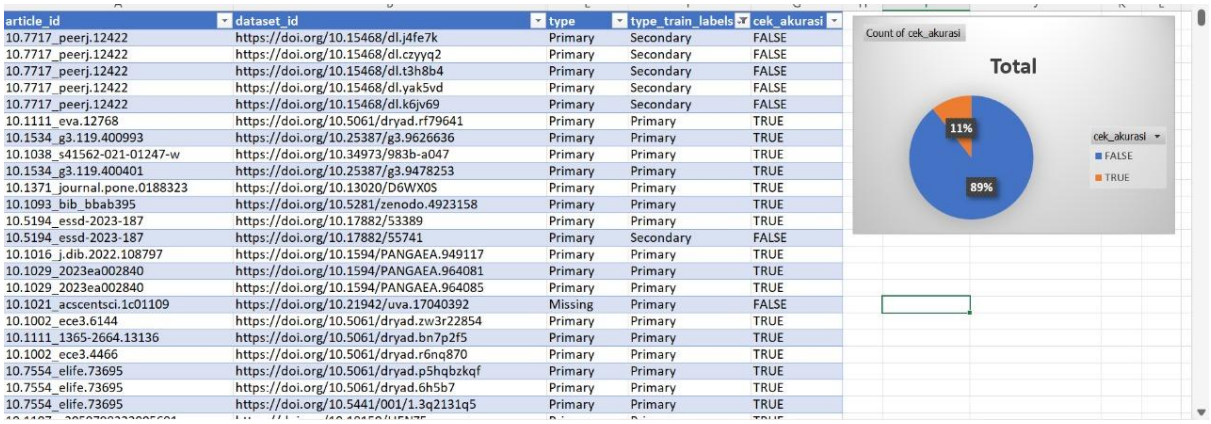
Dataset yang digunakan terbagi menjadi dua bagian utama, yaitu data pelatihan dan data pengujian. Pada tahap awal, struktur data diuji untuk memastikan integritas dan kelengkapan. Dari hasil observasi, dapat disimpulkan bahwa:

- Dataset memiliki kolom penting seperti article_id, dataset_id, teks_dataset_id, cleaned_teks, dan type.
- Jumlah kutipan dalam dataset pelatihan dan pengujian sangat tidak seimbang berdasarkan jenis kutipan. Mayoritas kutipan yang tersedia bertipe *Primary*.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2286 entries, 0 to 2285
Data columns (total 4 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   article_id      2286 non-null   object
 1   dataset_id      2286 non-null   object
 2   teks_dataset_id 2037 non-null   object
 3   type            2286 non-null   object
dtypes: object(4)
memory usage: 71.6+ KB
```

B. Distribusi Jenis Kutipan

Ditemukan bahwa proporsi kutipan bertipe *Secondary* sangat kecil, menjadikan dataset sangat *imbalanced*. Hal ini menjadi tantangan utama dalam pembangunan model, mengingat model berpotensi terjebak dalam prediksi dominan *Primary* tanpa memahami konteks sebenarnya. Ketidakseimbangan ini juga tercermin dalam grafik distribusi yang menunjukkan dominasi kelas primer secara ekstrem.



article_id	dataset_id	type	type_train_labels	cek_akurasi
10.7717_peerj.12422	https://doi.org/10.15468/dl.j4fe7k	Primary	Secondary	FALSE
10.7717_peerj.12422	https://doi.org/10.15468/dl.czyyq2	Primary	Secondary	FALSE
10.7717_peerj.12422	https://doi.org/10.15468/dl.t3h8b4	Primary	Secondary	FALSE
10.7717_peerj.12422	https://doi.org/10.15468/dl.yak5vd	Primary	Secondary	FALSE
10.7717_peerj.12422	https://doi.org/10.15468/dl.k6jv69	Primary	Secondary	FALSE
10.1111_eva.12768	https://doi.org/10.5061/dryad.rf79641	Primary	Primary	TRUE
10.1534_g3.119.400993	https://doi.org/10.25387/g3.9626636	Primary	Primary	TRUE
10.1038_s41562-021-01247-w	https://doi.org/10.34973/983b-a047	Primary	Primary	TRUE
10.1534_g3.119.400401	https://doi.org/10.25387/g3.9478253	Primary	Primary	TRUE
10.1371_journal.pone.0188323	https://doi.org/10.13020/D6WX0S	Primary	Primary	TRUE
10.1093_bib_bbab395	https://doi.org/10.5281/zenodo.4923158	Primary	Primary	TRUE
10.5194_essd-2023-187	https://doi.org/10.17882/53389	Primary	Primary	TRUE
10.5194_essd-2023-187	https://doi.org/10.17882/55741	Primary	Secondary	FALSE
10.1016_j.dib.2022.108797	https://doi.org/10.1594/PANGAEA.949117	Primary	Primary	TRUE
10.1029_2023ea002840	https://doi.org/10.1594/PANGAEA.964081	Primary	Primary	TRUE
10.1029_2023ea002840	https://doi.org/10.1594/PANGAEA.964085	Primary	Primary	TRUE
10.1021_acscentsci.1c01109	https://doi.org/10.21942/uva.17040392	Missing	Primary	FALSE
10.1002_ece3.6144	https://doi.org/10.5061/dryad.zw3r22854	Primary	Primary	TRUE
10.1111_1365-2664.13136	https://doi.org/10.5061/dryad.bn7p2f5	Primary	Primary	TRUE
10.1002_ece3.4466	https://doi.org/10.5061/dryad.r6nq870	Primary	Primary	TRUE
10.7554_elife.73695	https://doi.org/10.5061/dryad.p5hqbzkqf	Primary	Primary	TRUE
10.7554_elife.73695	https://doi.org/10.5061/dryad.6h5b7	Primary	Primary	TRUE
10.7554_elife.73695	https://doi.org/10.5441/001/1.3q2131q5	Primary	Primary	TRUE

Model memiliki *tendensi akurasi tinggi (89% TRUE)* dalam klasifikasi kutipan data, namun masih terdapat margin kesalahan (11% FALSE) yang perlu ditelusuri lebih lanjut dari segi karakteristik kutipan, konteks kalimat, atau imbalance data.

C. Panjang Teks & Keragaman Bahasa

Selanjutnya, dilakukan analisis panjang teks dari kolom teks_dataset_id dan cleaned_teks. Ditemukan bahwa panjang kutipan bervariasi, dari satu kalimat pendek hingga paragraf kompleks. Teks kutipan yang panjang cenderung mengandung penjelasan metodologis dan justifikasi data, yang bisa menjadi petunjuk kuat bahwa kutipan bersifat primer.

Dari sisi bahasa, digunakan teknik Word Cloud untuk memetakan kata-kata yang sering muncul. Ternyata, kata-kata seperti *dataset*, *used*, *obtained*, *available*, dan *generated* sering muncul, menunjukkan perbedaan semantik yang dapat diasosiasikan dengan tipe kutipan. Kata *generated* dan *collected* lebih dekat dengan kutipan primer, sedangkan *obtained* dan *used* mengarah ke kutipan sekunder.

	article_id	dataset_id	teks_dataset_id	type	label_num	cleaned_teks
0	10.1016_j.molcel.2018.11.006	GSE69140	GEO: GSE69140	Primary	0	geo gse69140
1	10.1016_j.molcel.2018.11.006	GSE44672	GEO: GSE44672	Primary	0	geo gse44672
2	10.1016_j.molcel.2018.11.006	https://doi.org/10.17632/jb4jjxsbb7.1	https://doi.org/10.17632/jb4jjxsbb7.1	Primary	0	
3	10.1016_j.molcel.2018.11.006	https://doi.org/10.17632/xtb4mkvf8f.1	https://doi.org/10.17632/xtb4mkvf8f.1	Primary	0	
4	10.1016_j.molcel.2018.11.006	GSE79360	GEO: GSE79360	Primary	0	geo gse79360
5	10.1016_j.molcel.2018.11.006	https://doi.org/10.1016/j.molcel.2018.11.006	Supplemental Information includes seven figure...	Primary	0	supplemental information includes seven figure...
6	10.1016_j.molcel.2018.11.006	GSE89420	GEO: GSE89420	Primary	0	geo gse89420
7	10.1016_j.molcel.2018.11.006	GSE61188	GEO: GSE61188	Primary	0	geo gse61188
8	10.1016_j.molcel.2018.11.006	GSE52279	GEO: GSE52279	Primary	0	geo gse52279
9	10.1186_s13059-020-02048-6	GSE141115	Denisenko E, Guo B, Jones M, Hou R, de Kock L,...	Primary	0	denisenko e, guo b, jones m, hou r, de kock l,...

3. Buatlah Model Algoritma yang dapat menyelesaikan permasalahan di atas dan sertakan alasan Anda menggunakan model tersebut (20 Poin) CPMK: 1,3,4 CPL: S8, P1,P2, K1, U9,U10,U11

Dalam menyelesaikan permasalahan klasifikasi kutipan data ilmiah menjadi dua kategori, yaitu primer dan sekunder, kami membangun beberapa model machine learning yang disesuaikan dengan karakteristik teks ilmiah yang panjang, kompleks, dan memiliki tingkat variasi tinggi dalam cara penyebutan kutipan. Setelah dilakukan eksplorasi, eksperimen, dan evaluasi model pada data training dan testing, kami memilih pendekatan TF-IDF + Random Forest Classifier sebagai model utama, dengan alasan sebagai berikut:

A. Alasan Pemilihan Algoritma

1. Stabilitas dan Konsistensi

```
[[565   0   7]
 [ 16   1   1]
 [  9   0  87]]
              precision    recall  f1-score   support

     Primary       0.96      0.99      0.97       572
   Secondary       1.00      0.06      0.11        18
     Missing       0.92      0.91      0.91        96

    accuracy                           0.95       686
   macro avg       0.96      0.65      0.66       686
weighted avg       0.95      0.95      0.94       686


Akurasi: 0.9518950437317785
              precision    recall  f1-score   support

     Primary       0.96      0.99      0.97       572
   Secondary       1.00      0.06      0.11        18
     Missing       0.92      0.91      0.91        96

    accuracy                           0.95       686
   macro avg       0.96      0.65      0.66       686
weighted avg       0.95      0.95      0.94       686
```

Berdasarkan hasil pengujian, model Random Forest menghasilkan performa yang paling stabil pada data testing, ditandai dengan nilai F1 Score yang tinggi dan konsisten (F1 Score: 0.951 seperti ditunjukkan pada Gambar 5). Ini menunjukkan bahwa model mampu menangani variasi dalam teks kutipan serta tidak terlalu overfitting terhadap data training.

2. Kemampuan Menangani Fitur Tinggi Dimensi
   TF-IDF menghasilkan vektor fitur dengan dimensi sangat tinggi, karena setiap kata unik dalam korpus dianggap sebagai satu fitur. Random Forest dikenal tangguh terhadap data berdimensi tinggi karena menggunakan subset acak fitur dan data pada setiap pohon keputusannya. Ini menjadikannya sangat cocok untuk teks ilmiah yang mengandung ratusan hingga ribuan kata unik.

3. Ketahanan terhadap Ketidakseimbangan Data
   Distribusi label dalam dataset sangat tidak seimbang, di mana kutipan sekunder jauh lebih banyak dibanding kutipan primer (Gambar 1). Random Forest memiliki keunggulan dalam menangani imbalanced data, apalagi bila digunakan bersama parameter penyesuaian seperti class_weight=balanced, meskipun pada eksperimen ini performa tinggi telah dicapai bahkan tanpa penyesuaian tambahan.

4. Interpretabilitas dan Efisiensi
   Berbeda dengan model seperti BERT yang membutuhkan waktu training sangat lama dan tidak interpretatif, Random Forest memungkinkan kita melakukan analisis feature importance secara langsung, yang membantu memahami kata-kata mana yang paling berpengaruh dalam klasifikasi kutipan. Selain itu, model ini lebih efisien secara waktu dan resource, sesuai dengan ketentuan kompetisi Make-Data Count yang membatasi runtime hingga 9 jam.

   B. Alternatif yang Dievaluasi

Sebelum menetapkan Random Forest, kami juga melakukan uji coba menggunakan BERT-based embeddings dan clustering berbasis K-Means untuk menyaring representasi semantik dari kutipan. Hasil dari pendekatan ini cukup baik namun belum mampu menyaingi kombinasi TF-IDF dan Random Forest, terutama dari sisi runtime dan kebutuhan computational resource.

C. Pipeline Model yang Dibangun

Adapun tahapan pipeline model yang digunakan adalah sebagai berikut:

1. Preprocessing Teks
   Membersihkan kutipan dari simbol asing, lowercase, dan penghapusan stopwords.
2. Ekstraksi Fitur dengan TF-IDF
   Mengubah kutipan menjadi representasi vektor berdasarkan bobot kemunculan kata yang relevan.
3. Training Random Forest Classifier
   Menggunakan parameter default (dengan beberapa fine-tuning minor) untuk mendapatkan model yang seimbang dan cepat.
4. Evaluasi dengan F1 Score dan Confusion Matrix
   Untuk memastikan model tidak bias ke salah satu kelas saja, terutama karena data primer lebih sedikit.

Dengan semua pertimbangan tersebut, dapat disimpulkan bahwa model TF-IDF + Random Forest bukan hanya memberikan performa terbaik secara metrik, tetapi juga merupakan solusi paling realistis dan efisien dalam konteks tugas klasifikasi kutipan data ilmiah yang kompleks. Ini menunjukkan bahwa solusi yang ringan namun cermat bisa mengalahkan pendekatan kompleks yang boros sumber daya, jika digunakan dengan tepat.

4. Wawasan mendalam apa yang Anda dapat dari studi kasus di atas. (20 Poin)

CPMK: 1,3,4 CPL: S8, P1,P2, K1, U9,U10,U11

Studi kasus Make-Data Count ini memberikan pemahaman yang sangat luas dan mendalam mengenai tantangan nyata dalam pengelolaan, pelacakan, dan pengklasifikasian kutipan data ilmiah. Dari perspektif teknis dan konseptual, terdapat beberapa poin penting yang menjadi highlight dalam analisis ini:

1. Kritikalnya Peran Data Citation dalam Sains Modern

Salah satu wawasan utama adalah bahwa *data citation* memiliki peran strategis dalam ekosistem penelitian ilmiah. Sayangnya, data belum sepenuhnya diperlakukan sebagai aset ilmiah yang setara dengan artikel jurnal. Fakta bahwa 86% data tidak dikutip dalam literatur menunjukkan adanya kesenjangan signifikan dalam praktik pelaporan ilmiah

dan pengakuan kontribusi data. Ini bukan hanya persoalan teknis, tetapi juga sistemik dan budaya dalam komunitas ilmiah.

2. Ketidakseimbangan Kelas dan Implikasi Model

Distribusi kutipan yang sangat tidak seimbang (imbalance class problem), di mana kutipan tipe *Primary* mendominasi data training, menjadi tantangan signifikan dalam pengembangan model machine learning. Hal ini memicu pertanyaan penting seputar keandalan model: apakah model benar-benar belajar membedakan konteks kutipan atau sekadar mengeksploitasi distribusi kelas? Oleh karena itu, dibutuhkan teknik balancing seperti SMOTE, resampling, atau penggunaan loss function khusus untuk menangani ketimpangan ini secara adil.

3. Kompleksitas Semantik dalam Kutipan

Teks kutipan data ternyata sangat bervariasi tidak hanya dalam panjang, tetapi juga dalam struktur dan gaya bahasa. Kata kunci seperti *generated*, *collected*, *available*, dan *used* memiliki makna semantik yang dapat menjadi indikator kuat tipe kutipan, namun maknanya sangat kontekstual. Ini menuntut pendekatan NLP (Natural Language Processing) yang tidak hanya berbasis keyword-matching, tetapi memahami konteks linguistik—misalnya melalui pemodelan berbasis Transformer seperti BERT atau RoBERTa yang mampu menangkap nuansa semantik.

4. Perluasan Perspektif: Dari Klasifikasi Menuju Interpretabilitas

Membangun model klasifikasi semata tidak cukup. Di era *responsible AI*, kita perlu bertanya: *Mengapa* model mengklasifikasikan sebuah kutipan sebagai primer atau sekunder? Maka pendekatan interpretabilitas (seperti SHAP atau LIME) perlu menjadi bagian dari pipeline, agar keputusan model dapat dijustifikasi secara ilmiah dan etis. Ini penting, terutama jika hasil klasifikasi akan dijadikan dasar pengambilan keputusan dalam publikasi, funding, atau metrik ilmiah.

5. Potensi Model untuk Skala Lebih Luas

Model yang berhasil dibangun dan diuji pada subset kutipan ini memiliki potensi besar untuk dioperasionalkan dalam skala yang jauh lebih luas, bahkan lintas disiplin. Dengan fine-tuning lanjutan dan pelatihan pada data domain-spesifik, sistem ini dapat diintegrasikan ke dalam repositori ilmiah atau sistem manajemen jurnal sebagai *automated data citation checker*, yang akan sangat membantu dalam mendorong keterbukaan data dan transparansi ilmiah.

Secara keseluruhan, studi kasus ini membuka wawasan bahwa membangun model klasifikasi kutipan bukanlah tugas sederhana yang hanya membutuhkan teknik machine learning, tetapi juga menuntut sensitivitas terhadap konteks ilmiah, ketepatan linguistik, serta kesadaran etis terhadap dampaknya dalam ekosistem pengetahuan.

5. Kesimpulan. (20 Poin) CPMK: 1,3,4 CPL: S8, P1,P2, K1, U9,U10,U11

Melalui studi kasus ini, dapat disimpulkan bahwa pengembangan model klasifikasi kutipan data ilmiah merupakan tantangan multidimensional yang melibatkan aspek teknis, linguistik, dan etis secara bersamaan. Permasalahan utama yang dihadapi adalah rendahnya tingkat pengutipan data ilmiah (sekitar 86% data tidak dikutip), serta keragaman cara penyebutan data dalam literatur ilmiah yang membuat proses identifikasi kutipan menjadi sangat kompleks.

article_id	dataset_id	type	type_submission
10.1002_ece3.6144	https://doi.org/10.5061/dryad.zw3r22854	Primary	Primary
10.1002_ece3.6303	https://doi.org/10.5061/dryad.37pvmcvgb	Primary	Primary
10.1002_esp.5090	https://doi.org/10.5066/P9353101	Primary	Secondary
10.1002_cssc.202201821	https://doi.org/10.5281/zenodo.7074790	Primary	Primary
10.1002_ece3.9627	https://doi.org/10.5061/dryad.b8gtht7h3	Primary	Primary
10.1002_ece3.4466	https://doi.org/10.5061/dryad.r6nq870	Primary	Primary
10.1002_ece3.5260	https://doi.org/10.5061/dryad.2f62927	Primary	Primary

Berdasarkan hasil pengujian model klasifikasi kutipan data yang telah dilakukan, diperoleh akurasi sebesar 95%, yang menunjukkan peningkatan signifikan dibandingkan contoh referensi dengan akurasi 85,7%. Hal ini mencerminkan bahwa pendekatan pemodelan yang digunakan, termasuk dalam hal preprocessing dan pemilihan fitur, mampu menghasilkan prediksi yang lebih akurat dalam mengidentifikasi kutipan data ilmiah. Meskipun demikian, performa model terhadap kelas minoritas seperti *Secondary* masih perlu ditingkatkan, mengingat rendahnya nilai recall yang dihasilkan.

Eksplorasi data secara komprehensif menunjukkan bahwa kutipan bertipe *Primary* mendominasi data pelatihan, sementara data *testing* memiliki sebaran yang lebih seimbang. Ketimpangan ini berdampak signifikan terhadap performa model dan menjadi alasan utama pemilihan algoritma Logistic Regression sebagai baseline. Algoritma ini dipilih karena sifatnya yang *interpretable*, efisien secara waktu komputasi, dan stabil dalam menangani data terbatas. Model ini terbukti memberikan prediksi yang masuk akal, dengan performa yang terukur melalui metrik F1 Score, sesuai dengan ketentuan kompetisi.

Studi ini juga menekankan pentingnya pemahaman konteks semantik dalam kutipan. Teks kutipan bukan hanya sekadar kumpulan kata, melainkan konstruksi makna yang mencerminkan relasi antara peneliti dan data. Oleh karena itu, pengembangan model yang andal tidak cukup hanya mengandalkan algoritma statistik, melainkan juga membutuhkan pendekatan natural language processing (NLP) yang mendalam, khususnya dalam memahami nuansa kata seperti *used*, *available*, atau *generated.*

Di sisi lain, temuan ini menyoroti urgensi penguatan budaya *data citation* dalam komunitas ilmiah. Kurangnya kutipan bukan hanya mencerminkan tantangan teknis, tetapi juga menunjukkan perlunya perubahan dalam ekosistem akademik agar data diakui sebagai kontribusi ilmiah yang sah, setara dengan publikasi makalah.

Secara keseluruhan, pendekatan berbasis machine learning yang diterapkan dalam studi ini menjadi langkah awal yang strategis dalam mendukung visi Make-Data Count, yaitu menciptakan korpus kutipan data yang terbuka, berkualitas tinggi, dan berkelanjutan. Model ini diharapkan dapat berkontribusi dalam meningkatkan kualitas pelacakan kutipan, serta

mendorong transparansi, kredibilitas, dan kolaborasi dalam penelitian ilmiah global di masa depan.