UNIVERSITAS
GADJAH MADA

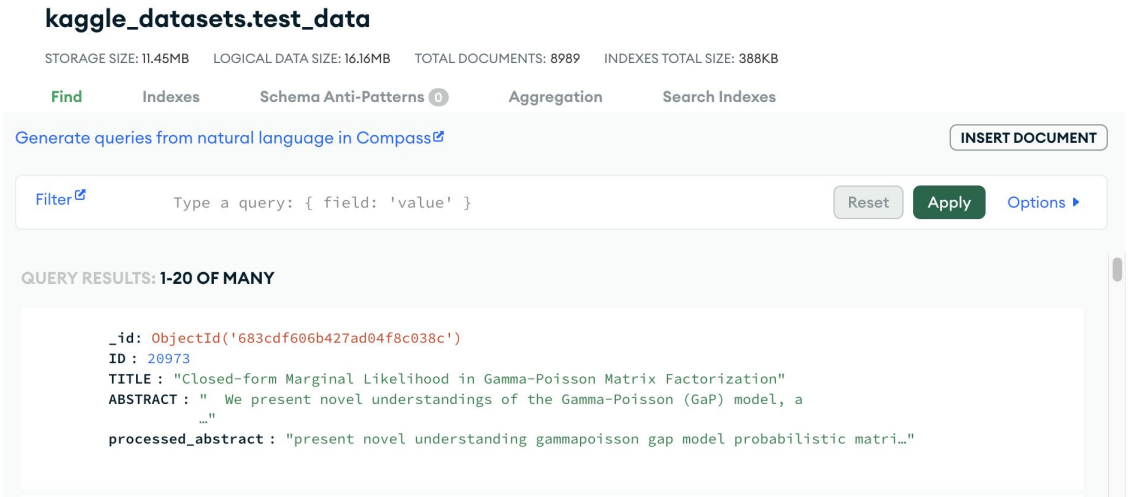
Proyek Berbasis Text Mining

KLASIFIKASI TOPIK ARTIKEL ILMIAH BERBASIS ABSTRAK DAN JUDUL

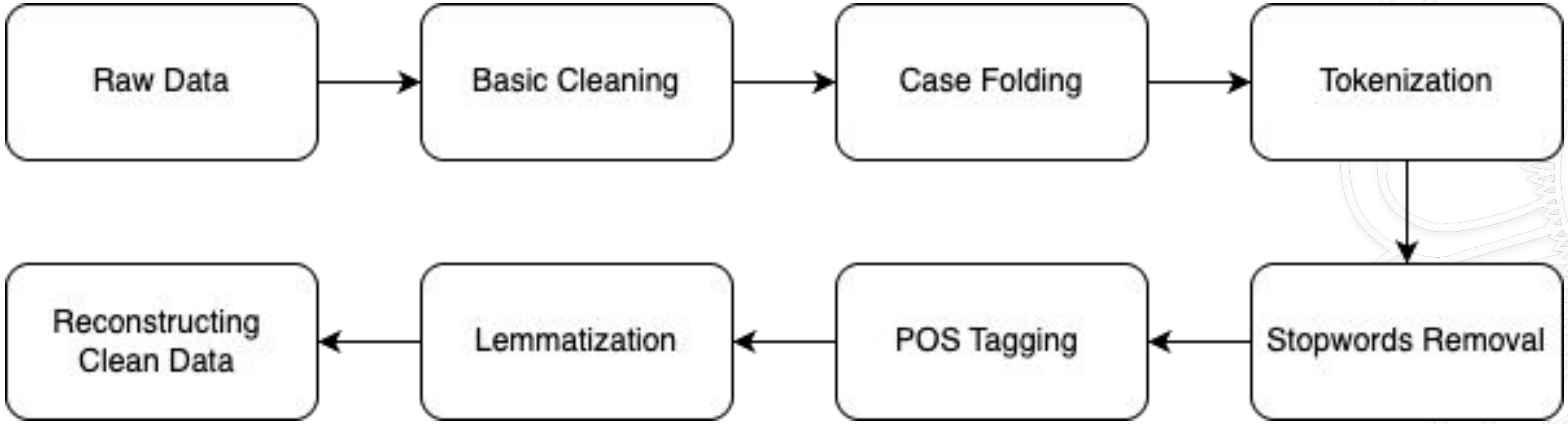
Shevrilla Vilnafa Bilbina S.	22/492511/PA/21118
Priskilla N. P. Br Silalahi	22/493324/PA/21176
Intan Dwi Febryanti	22/494760/PA/21285
Dhanada Santika Putri	22/497239/PA/21407
Yessica Thipandona	22/497660/PA/21441
Febriana Nur Syifa Rizqi	22/499532/PA/21541

Dataset

- Bersumber dari Kaggle
- Terdiri atas 30.000 dokumen
- Merupakan kumpulan artikel penelitian dari topik Computer Science, Physics, Mathematics, Statistics, Quantitative Biology, dan Quantitative Finance
- Data train.csv dan test.csv diimpor ke MongoDB Atlas
- Disimpan di database kaggle_datasets dan pada koleksi train_data dan test_data



Preprocessing



Ekstraksi FItur

Tujuan: Mengubah teks menjadi representasi numerik yang bisa diproses oleh model machine learning.

Metode:

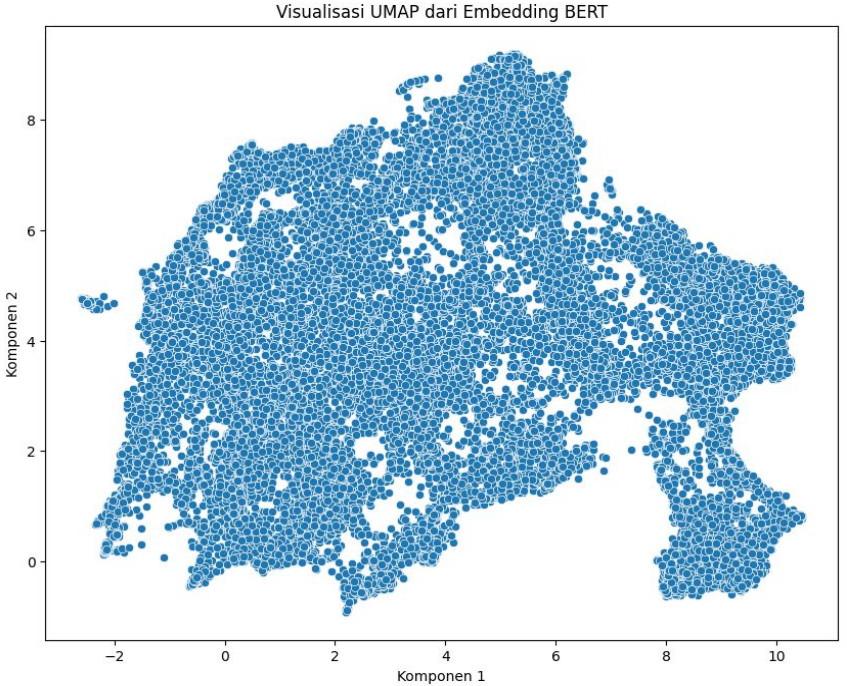
- Bag of Words (BoW): Hitung frekuensi kata, sederhana tapi abaikan konteks.
- TF-IDF: Bobot kata berdasarkan frekuensi dalam dokumen & korpus; lebih informatif dari BoW.
- Word2Vec (CBOW): Representasi kata berbasis konteks, rata-rata vektor kata untuk dokumen.
- BERT (MiniLM): Embedding kalimat berdimensi 384; tangkap konteks lokal & global.

Reduksi Dimensi & Visualisasi

Tujuan: Kurangi dimensi vektor tinggi (TF-IDF/BERT) untuk efisiensi dan visualisasi.

Metode:

- Truncated SVD: Reduksi linier untuk TF-IDF; cocok untuk data sparse.
- PCA: Proyeksi linier berdasarkan variansi terbesar; baseline cepat dan interpretatif.
- t-SNE: Proyeksi non-linier; jaga struktur lokal, tapi lambat & sensitif parameter.
- UMAP: Cepat & stabil; jaga struktur lokal-global, hasil visual lebih informatif.



Bigram & Trigram

```
Top 15 Bigram:
neural network: 2452
machine learn: 1006
result show: 926
paper propose: 877
paper present: 798
propose method: 767
et al: 699
magnetic field: 627
data set: 597
deep learning: 590
deep neural: 558
allow u: 539
experimental result: 535
optimization problem: 528
propose novel: 512
```

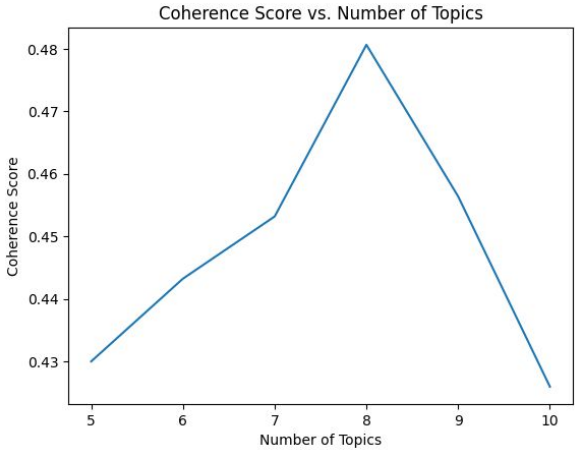
Bigram yang muncul dalam korpus cenderung berupa frasa yang lebih **umum** dan **general.**

```
Top 15 Trigram:
deep neural network: 538
convolutional neural network: 461
recurrent neural network: 239
paper propose novel: 166
experimental result show: 165
generative adversarial network: 163
play important role: 133
stochastic gradient descent: 122
partial differential equation: 120
support vector machine: 118
density functional theory: 113
paper propose new: 110
neural network cnn: 108
markov chain monte: 103
deep reinforcement learn: 103
```

Trigram memberikan **istilah teknis yang lebih spesifik** seperti nama arsitektur, algoritma, atau metode ilmiah.

Topic Modeling LDA

- Dokumen dikonversi menjadi BoW.
- Model LDA dilatih dengan berbagai jumlah topik (5–10).
- Koherensi diukur untuk menentukan jumlah topik optimal.
- Model akhir dilatih dengan **num_topics=8.**




```
Topik-Topik:
Topik 1: 0.019*"graph" + 0.014*"group" + 0.012*"n" + 0.011*"give" + 0.011*"show" + 0.010
Topik 2: 0.006*"magnetic" + 0.006*"temperature" + 0.006*"star" + 0.006*"surface" + 0.005
Topik 3: 0.014*"system" + 0.014*"state" + 0.011*"model" + 0.010*"quantum" + 0.009*"phase
Topik 4: 0.020*"model" + 0.016*"data" + 0.015*"method" + 0.015*"network" + 0.014*"learn"
Topik 5: 0.020*"model" + 0.012*"cluster" + 0.011*"galaxy" + 0.010*"data" + 0.009*"mass"
Topik 6: 0.013*"system" + 0.012*"network" + 0.008*"model" + 0.007*"use" + 0.007*"paper"
Topik 7: 0.015*"use" + 0.008*"propose" + 0.008*"design" + 0.007*"performance" + 0.007*"m
Topik 8: 0.013*"problem" + 0.012*"function" + 0.010*"result" + 0.009*"method" + 0.008*"s
```

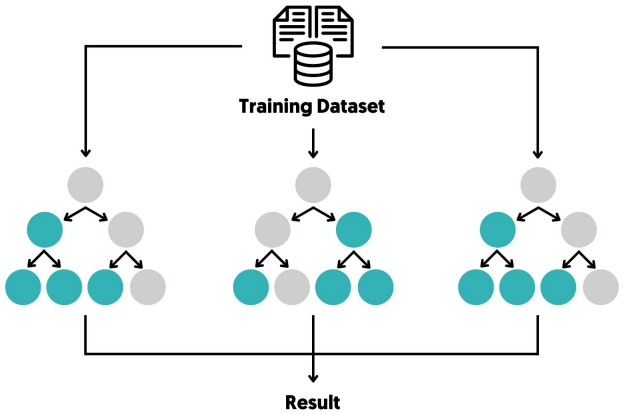
Evaluasi Topic Modeling LDA

```
Perplexity: -8.292649049775749
Coherence: 0.4362839143982079 6
```

- **Perplexity** mengukur seberapa baik model LDA memprediksi data baru. Semakin rendah nilainya, semakin baik modelnya. Nilai **log perplexity** sebesar **–8.29** menunjukkan bahwa model **cukup baik** dalam memodelkan distribusi kata, meskipun metrik ini tidak mencerminkan kualitas semantik topik.
- **Coherence score** sebesar **0.436** menunjukkan bahwa topik-topik yang dihasilkan **cukup koheren secara semantik.** Meskipun belum optimal, nilainya masih dapat diterima, terutama untuk korpus ilmiah atau kompleks.

Modelling - Klasifikasi Topik

- Fitur teks direpresentasikan menggunakan **TF-IDF** (maksimal 5000 fitur).
- Data dibagi menjadi data latih dan data uji (80:20) dengan stratifikasi label.
- Teknik **SMOTE** diterapkan pada data latih untuk mengatasi ketidakseimbangan kelas.
- Model yang digunakan adalah **Random Forest Classifier** dengan 100 pohon keputusan.

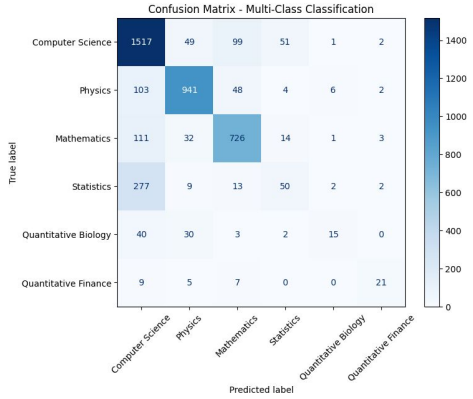


Evaluasi

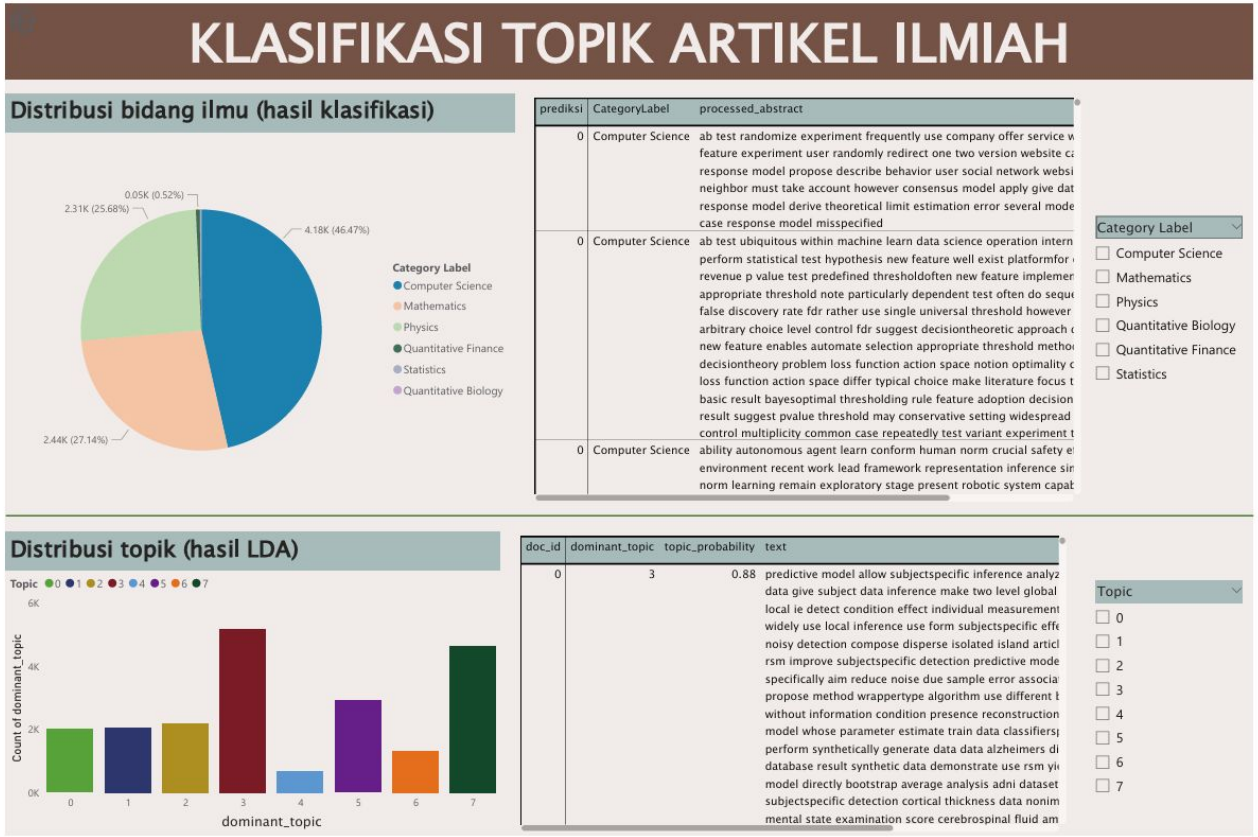
Model klasifikasi menunjukkan **akurasi keseluruhan sebesar 78%** dengan **performa tinggi pada kelas mayoritas** seperti Physics, Mathematics, dan Computer Science (F1-score di atas 0.80). Namun, **performa menurun signifikan pada kelas minor** seperti Statistics dan Quantitative Biology dengan F1-score masing-masing hanya 0.21 dan 0.26. Confusion matrix mengungkapkan banyak kesalahan klasifikasi ke kelas mayoritas, terutama untuk Statistics yang sering dikira Computer Science. Hal ini menunjukkan bahwa meskipun SMOTE telah diterapkan, model masih kurang sensitif terhadap kelas dengan jumlah data kecil.

Classification Report:

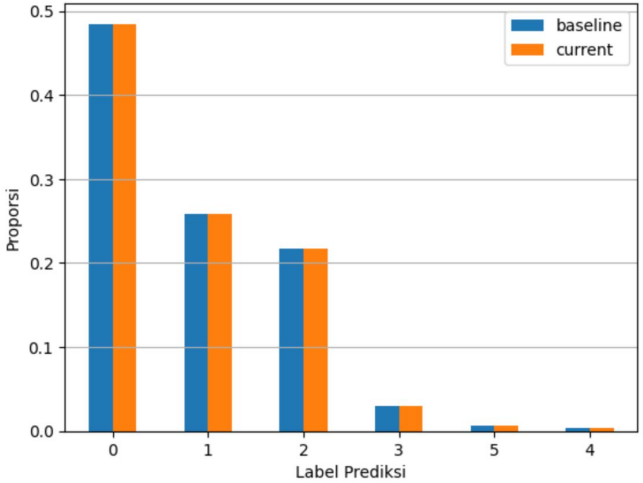
	precision	recall	f1-score	support
Computer Science	0.74	0.88	0.80	1719
Physics	0.88	0.85	0.87	1104
Mathematics	0.81	0.82	0.81	887
Statistics	0.41	0.14	0.21	353
Quantitative Biology	0.60	0.17	0.26	90
Quantitative Finance	0.70	0.50	0.58	42
accuracy			0.78	4195
macro avg	0.69	0.56	0.59	4195
weighted avg	0.76	0.78	0.76	4195



Dashboard Visualisasi Hasil



Monitoring Model Drift



Hasil monitoring distribusi prediksi antar batch menunjukkan bahwa proporsi prediksi antar label relatif stabil. Monitoring model drift dilakukan dengan membandingkan distribusi prediksi antar batch, yaitu antara batch baseline (prediksi awal saat model pertama kali di-deploy) dan batch current (prediksi terbaru). Tidak terdapat perbedaan signifikan (>10%) antar distribusi baseline dan batch saat ini. Hal ini menunjukkan bahwa model masih berjalan stabil tanpa adanya indikasi model drift.

Batch Scoring

```
Distribusi hasil prediksi:
prediksi
0    4357
1    2319
2    1948
3     273
5      59
4      33
Name: count, dtype: int64
```

	processed_abstract	prediksi
0	present novel understanding gammapoisson gap m...	0
1	meteorite contain mineral solar system asteroi...	1
2	frame aggregation mechanism multiple frame com...	0
3	milky way open cluster diverse term age chemic...	1
4	prove cryptographic protocol correct secrecy h...	0
5	paper propose regularized pairwise difference ...	0
6	central issue theory extreme value focus suita...	0
7	astrophysics cosmology rich data advent widear...	0
8	number recent work propose technique endtoend ...	0
9	use hydrodynamical galaxy formation simulation...	1

Mayoritas prediksi jatuh pada kelas 0 (4.357 data), kemudian 1 (2.319 data) dan 2 (1.948 data). Kelas 3, 4, dan 5 masing-masing memiliki jumlah prediksi yang jauh lebih kecil, menunjukkan ketidakseimbangan dalam distribusi hasil prediksi.

UNIVERSITAS
GADJAH MADA

TERIMA KASIH