UNIVERSITAS INDONESIA
Veritas, Probitas, Iustitia

FAKULTAS ILMU KOMPUTER

CSIE604284 • Analitika Media Sosial
Semester Gasal 2025/2026
Fakultas Ilmu Komputer, Universitas Indonesia

Tutorial 2: *Text Classification* dan *Sentiment Analysis*

Deadline: 9 Oktober 2025 pukul 23.59 WIB

Keterkaitan dengan Sub-CPMK

- **Sub-CPMK4**: Mampu mendemonstrasikan teknik pengumpulan data dari media sosial
- **Sub-CPMK5**: Mampu menerapkan teknik pra-pemrosesan dan representasi teks media sosial
- **Sub-CPMK7**: Mampu menerapkan teknik classification dan clustering pada data terstruktur, teks, dan jejaring sosial.
- **Sub-CPMK8**: Mampu menganalisis sentimen dan topik pembicaraan dari teks media sosial.

1 Latar Belakang

Analisis teks merupakan komponen fundamental dalam analitika media sosial. Dengan volume data teks yang masif dari berbagai platform digital, kemampuan untuk mengotomatisasi pemahaman dan kategorisasi teks menjadi sangat penting. Tutorial ini akan memperkenalkan dua teknik fundamental dalam *text analytics*: ***Text Classification*** dan ***Sentiment Analysis***.

1.1 *Text Classification*

Text classification adalah proses mengkategorikan dokumen teks ke dalam kelas atau kategori yang telah ditentukan sebelumnya. Aplikasinya meliputi:

- Kategorisasi berita (politik, ekonomi, olahraga, teknologi)
- Deteksi *spam email*
- Klasifikasi topik diskusi di media sosial
- Deteksi *fake news*

1.2 *Sentiment Analysis*

Sentiment analysis adalah subset dari *text classification* yang fokus pada identifikasi dan ekstraksi opini atau sentimen dari teks. Aplikasinya meliputi:

- Analisis *review* produk
- *Monitoring* reputasi *brand* di media sosial
- Analisis sentimen publik terhadap kebijakan pemerintah
- Prediksi tren pasar berdasarkan sentimen investor

2 Deskripsi Tugas

Tutorial ini terdiri dari tiga bagian. Kode contoh akan didemonstrasikan oleh asisten dosen pada sesi tutorial dan akan tersedia di SCELE. Kehadiran pada sesi tutorial **adalah kewajiban**.

2.1 Bagian 1: *Web Scraping* + *Pretrained Model* (35 poin)

Mahasiswa diminta untuk melakukan *web scraping* dari salah satu situs berita Indonesia (selain Kompas), mengumpulkan minimal 10 artikel, lalu menganalisisnya menggunakan *pretrained model* untuk **text classification**. Hasil klasifikasi perlu divisualisasikan dalam bentuk grafik distribusi dan tabel ringkas.

2.1.1 Langkah-langkah

1. ***Web Scraping* Berita Indonesia**
   - Pilih satu situs berita (contoh: CNN Indonesia, Tempo.co, Liputan6, Tribunnews, Republika, Antara News).
   - Data yang dikumpulkan: `title`, `content`, dan `url` (minimal 10 artikel).
   - Simpan hasil scraping ke file: `scraped_news.csv`.
   - Disarankan menggunakan: BeautifulSoup + *requests*, Scrapy, atau Selenium.
2. **Analisis dengan *Pretrained Model***
   - Gunakan salah satu model dari HuggingFace (contoh: IndoBERT, mBERT, XLM-RoBERTa).
   - Simpan hasil klasifikasi ke file: `classification_results.csv` dengan kolom `title`, `content`, dan `predicted_label`.
   - Buat visualisasi:
     - Distribusi artikel per label (bar chart).
     - Contoh artikel beserta label hasil klasifikasi (tabel ringkas).

2.1.2 Contoh Implementasi

```
from transformers import pipeline

classifier = pipeline(
    "text-classification",
    model="indobenchmark/indobert-base-p1"
)

```

```
texts = [
    "Pemerintah umumkan kebijakan baru terkait subsidi BBM",
    "Persija Jakarta menang dramatis melawan Persib Bandung"
]

results = classifier(texts)
print(results)
```

Listing 1: Contoh penggunaan pretrained model untuk text classification

2.1.3 File yang Dikumpulkan

- `part1_webscraping.ipynb` — notebook dengan kode lengkap.
- `scraped_news.csv` — data hasil scraping.
- `classification_results.csv` — hasil klasifikasi.

2.2 Bagian 2: *Text Classification* dengan Dataset Publik (35 poin)

Pada bagian ini, mahasiswa diminta untuk membangun *pipeline text classification* lengkap, mulai dari eksplorasi data, preprocessing, ekstraksi fitur, pelatihan model, hingga evaluasi. Berbeda dengan Bagian 1 yang menggunakan *pretrained model*, di sini mahasiswa harus melatih model sendiri dengan dataset publik.

2.2.1 Langkah-langkah

1. **Pemilihan Dataset**
   - Cari dan pilih dataset *text classification* **bukan sentiment analysis**.
   - Sumber yang disarankan: Kaggle, UCI ML Repository, HuggingFace Datasets, atau Papers with Code.
   - Simpan informasi dataset ke file: `dataset_info.txt`.
   - Isi minimal dalam `dataset_info.txt`:
     - Sumber dataset (URL/link).
     - Jumlah total data.
     - Jumlah kelas.
     - Deskripsi singkat isi dataset.
2. **Exploratory Data Analysis (EDA)**
   - Analisis distribusi kelas (bar chart).
   - Statistik panjang teks (histogram).
   - Frekuensi kata (top 20 kata).
3. **Text Preprocessing**
   - Lowercasing, hapus karakter khusus, tokenisasi.
   - Hapus stopwords.
   - Stemming atau lemmatization.
4. **Feature Representation**
   - Representasi fitur bebas, contoh yang dapat digunakan:

- TF–IDF
- Count Vectorizer
- Word2Vec
- FastText

5. **Model Training**
   - Latih minimal 3 algoritma berbeda (contoh: Naive Bayes, Logistic Regression, SVM, Random Forest, XGBoost).
   - Tidak perlu melakukan hyperparameter tuning.
6. **Model Evaluation**
   - Hitung metrik: Accuracy, Precision, Recall, F1-Score (per class dan rata-rata).
   - Visualisasi: confusion matrix heatmap, classification report, ROC curve (jika biner), dan perbandingan performa model (bar chart).
   - Simpan hasil evaluasi ke file: `model_results.csv` dengan kolom:
     - `model_name`
     - `accuracy`
     - `precision`
     - `recall`
     - `f1_score`

2.2.2 File yang Dikumpulkan

- `part2_text_classification.ipynb` — notebook dengan implementasi lengkap.
- `dataset_info.txt` — deskripsi dataset (minimal: sumber, jumlah data, jumlah kelas, deskripsi isi).
- `model_results.csv` — hasil perbandingan performa model (dengan kolom: `model_name`, `accuracy`, `precision`, `recall`, `f1_score`).

2.3 Bagian 3: *Sentiment Analysis* dengan Dataset Publik (30 poin)

Pada bagian ini, mahasiswa diminta untuk membandingkan dua pendekatan dalam *sentiment analysis*: (1) *traditional machine learning* dengan representasi fitur sederhana, dan (2) *pretrained models*. Fokus utama adalah membandingkan hasil, bukan membuat sistem sempurna.

2.3.1 Langkah-langkah

1. **Pemilihan Dataset**
   - WAJIB menggunakan dataset publik *sentiment analysis* (binary atau multi-class).
   - Contoh: IMDb Reviews, Amazon Product Reviews, Twitter Sentiment, Restaurant/Hotel Reviews, atau dataset Indonesia (misalnya IndoNLU).
   - Kriteria dataset: minimal 1000 sampel.
   - Buat file `dataset_info.txt` berisi minimal:
     - Sumber dataset (URL/link).
     - Jumlah total data.
     - Jumlah kelas.

- Deskripsi singkat isi dataset.

2. **Pendekatan Traditional ML**
   - Lakukan preprocessing dasar.
   - Representasi fitur bebas, contoh yang dapat digunakan:
     - TF–IDF
     - Count Vectorizer
     - Word2Vec
     - FastText
   - Latih minimal 2 algoritma (contoh: Naive Bayes, Logistic Regression, SVM).
   - Simpan hasil evaluasi ke `traditional_ml_results.csv` dengan kolom:
     - `model_name`
     - `accuracy`
     - `precision`
     - `recall`
     - `f1_score`
3. **Pendekatan Pretrained Model**
   - Gunakan minimal 1 model pretrained (contoh: BERT, RoBERTa, DistilBERT, IndoBERT).
   - Laporkan prediksi beserta *confidence score*.
   - Simpan hasil evaluasi ke `pretrained_results.csv` dengan kolom:
     - `model_name`
     - `accuracy`
     - `precision`
     - `recall`
     - `f1_score`
4. **Comparative Analysis**
   - Buat tabel perbandingan hasil (accuracy, precision, recall, f1-score).
   - Sertakan visualisasi: confusion matrix untuk masing-masing pendekatan, serta grafik perbandingan performa.
   - Tuliskan analisis singkat: kapan traditional ML lebih baik, kapan pretrained model lebih baik pada *Notebook* Anda.

2.3.2 File yang Dikumpulkan

- `part3_sentiment_analysis.ipynb` — notebook dengan implementasi lengkap.
- `dataset_info.txt` — deskripsi dataset (sumber, jumlah data, jumlah kelas, deskripsi isi).
- `traditional_ml_results.csv` — hasil evaluasi traditional ML.
- `pretrained_results.csv` — hasil evaluasi pretrained model.

3 *Deliverables*

3.1 Struktur Folder dan Penamaan File

File yang dikumpulkan harus berupa satu berkas `.zip` dengan nama:
`NPM_NamaLengkap_Tutorial2.zip`
Contoh: `2106123456_BudiSantoso_Tutorial2.zip`

Isi dari file `.zip` tersebut adalah **tiga folder utama** berikut:

- **part1_webscraping**
  - `part1_webscraping.ipynb`
  - `scraped_news.csv`
  - `classification_results.csv`
- **part2_classification**
  - `part2_text_classification.ipynb`
  - `dataset_info.txt`
  - `model_results.csv`
- **part3_sentiment**
  - `part3_sentiment_analysis.ipynb`
  - `dataset_info.txt`
  - `traditional_ml_results.csv`
  - `pretrained_results.csv`

4 Ketentuan Pengumpulan

INFORMASI PENGUMPULAN

- **Deadline**: 9 Oktober 2025 pukul 23.59 WIB
- **Platform**: SCELE
- **Format File**: ZIP dengan nama `NPM_NamaLengkap_Tutorial2.zip`
- ***Late Penalty***: 10% per hari (maksimal 3 hari)
- **Ukuran File**: Maksimal 50 MB (jangan *include large datasets*)

5 *Resources* dan Referensi

5.1 Dokumentasi

- Scikit-learn: `https://scikit-learn.org/stable/`
- HuggingFace Transformers: `https://huggingface.co/docs/transformers`
- NLTK: `https://www.nltk.org/`
- BeautifulSoup: `https://www.crummy.com/software/BeautifulSoup/bs4/doc/`

5.2 Datasets

- Kaggle Datasets: `https://www.kaggle.com/datasets`
- HuggingFace Datasets: `https://huggingface.co/datasets`
- UCI ML Repository: `https://archive.ics.uci.edu/ml/index.php`
- Papers with Code: `https://paperswithcode.com/datasets`

6 Kontak

Untuk pertanyaan terkait tugas:

- **Asisten Dosen:**
  - Syahrul Apriansyah (apriansyah.syahrul@gmail.com)
  - Sheryl Ivana W. (sherylivana99@gmail.com)
  - Kelvin Saputra (kelvinsaputra599@gmail.com)
- **Dosen:**
  - Prof. Dr. Indra Budi (indra@cs.ui.ac.id)
  - Satrio Yudhoatmojo, Ph.D. (satrio.baskoro@cs.ui.ac.id)

Good luck!!