**📝 LAPORAN PROYEK MACHINE LEARNING: KLASIFIKASI BERITA MENGGUNAKAN RNN (LSTM)**

**📌 1. Judul Proyek**

**Klasifikasi Teks Berita Indonesia ke dalam Kategori Politik dan Olahraga Menggunakan Recurrent Neural Network (LSTM)**

**📚 2. Latar Belakang**

Dengan meningkatnya jumlah berita online, pengelompokan otomatis berita berdasarkan topik menjadi hal penting. Pekerjaan ini bertujuan untuk membangun model machine learning berbasis RNN (LSTM) untuk mengklasifikasikan berita Indonesia menjadi dua kategori utama: **politik** dan **olahraga**.

**🎯 3. Tujuan**

* Mengklasifikasikan teks berita berdasarkan kontennya menjadi dua kategori.
* Menerapkan arsitektur RNN (LSTM) pada data teks berbahasa Indonesia.
* Mengatasi ketidakseimbangan kelas dengan class\_weight.

**🧾 4. Deskripsi Dataset**

* **Sumber:** Kaggle - Indonesian News Dataset (/kaggle/input/indonesian-news-dataset/data.csv)
* **Kolom digunakan:** title, content
* **Labeling Otomatis:** Berdasarkan kata kunci di konten:
  + Kata kunci “politik”, “presiden”, dll. → Label 0 (politik)
  + Kata kunci “sepak bola”, “timnas”, dll. → Label 1 (olahraga)

**⚙️ 5. Pra-pemrosesan Data**

* Penggabungan kolom title dan content.
* Tokenisasi menggunakan Tokenizer dari Keras.
* Padding hingga max\_len = 200.
* Split data: 80% latih, 20% uji.
* Penanganan ketidakseimbangan dengan compute\_class\_weight.

**🧠 6. Arsitektur Model**

python

SalinEdit

Sequential([

Embedding(input\_dim=10000, output\_dim=64),

LSTM(64),

Dropout(0.5),

Dense(32, activation='relu'),

Dense(1, activation='sigmoid')

])

* **Optimizer:** Adam
* **Loss Function:** Binary Crossentropy
* **Epochs:** 5
* **Batch Size:** 32

**📈 7. Hasil Training**

| **Epoch** | **Train Accuracy** | **Val Accuracy** | **Train Loss** | **Val Loss** |
| --- | --- | --- | --- | --- |
| 1 | 0.8496 | 0.9113 | 0.4359 | 0.2984 |
| 2 | 0.9225 | 0.9124 | 0.2588 | 0.2912 |
| 3 | 0.9272 | 0.9379 | 0.2163 | 0.2393 |
| 4 | 0.9662 | 0.9349 | 0.1201 | 0.2262 |
| 5 | 0.9747 | 0.9397 | 0.0904 | 0.2231 |

**📊 8. Evaluasi Model (Test Set)**

**📋 Classification Report:**

markdown

SalinEdit

precision recall f1-score support

0 0.96 0.97 0.97 2879

1 0.81 0.77 0.79 460

accuracy 0.94 3339

macro avg 0.89 0.87 0.88

weighted avg 0.94 0.94 0.94

**🧮 Confusion Matrix:**

lua

SalinEdit

[[2794 85]

[ 104 356]]

**🔍 9. Analisis**

* **Akurasi total** sangat tinggi: **94%**
* **Kelas mayoritas (politik)** sangat akurat.
* **Kelas minoritas (olahraga)** sedikit lebih rendah recall-nya (77%) → masih ada kesalahan klasifikasi ke politik.
* Model berhasil menangani ketidakseimbangan data secara signifikan dengan **class\_weight**.

**✅ 10. Kesimpulan**

Model RNN (LSTM) berhasil membedakan dua kategori berita secara efektif, terutama pada data teks berbahasa Indonesia. Performa sangat tinggi untuk kategori politik, dan cukup baik untuk kategori olahraga. Dengan penambahan data olahraga dan/atau fine-tuning lebih lanjut, performa dapat ditingkatkan lebih jauh.