TUGAS NATURAL LANGUAGE PROCESSING

Eksperimen dengan Model LSTM, Fast text, Transformer, dan BERT

Dibuat oleh :

Syakira Tsania Muthmainnah
121450110
Pemrosesan Bahasa Alami

PROGRAM STUDI SAINS DATA
JURUSAN SAINS

INSTITUT TEKNOLOGI SUMATERA
2024

PENDAHULUAN

Pemrosesan Bahasa Alami atau Natural Language Processing (NLP) merupakan salah satu cabang kecerdasan buatan yang fokus pada interaksi antara komputer dan bahasa manusia. NLP memungkinkan komputer untuk memahami, menganalisis, dan menghasilkan teks atau ucapan manusia. Dalam era digital saat ini, penerapan NLP semakin luas, mencakup berbagai bidang seperti analisis sentimen, chatbot, terjemahan bahasa, hingga klasifikasi teks. Salah satu tantangan utama dalam NLP adalah bagaimana memahami konteks dalam teks dan mengekstraksi informasi relevan dari data berukuran besar secara efisien.

Dalam tugas ini, dilakukan eksplorasi berbagai metode dan model untuk pemrosesan data teks pada dataset AG News, yang terdiri dari artikel berita dalam berbagai kategori. Tugas ini bertujuan untuk membangun model klasifikasi teks yang mampu mengelompokkan berita ke dalam empat kategori berbeda. Dataset yang digunakan mencakup berita dengan label, judul, dan konten, sehingga memberikan data yang kaya untuk diterapkan dalam berbagai teknik NLP.

Laporan ini bertujuan untuk memaparkan tahapan implementasi, termasuk analisis data, pra-pemrosesan, pemuatan model embedding, hingga pelatihan dan evaluasi model klasifikasi teks. Dengan mengimplementasikan berbagai pendekatan ini, diharapkan dapat ditemukan metode terbaik untuk memecahkan masalah klasifikasi berita berbasis teks.

METODE

Tahapan yang dilakukan dalam penelitian ini melibatkan beberapa proses, dimulai dari pra-pemrosesan teks seperti pembersihan data, tokenisasi, dan penghapusan stopwords untuk meningkatkan kualitas data masukan. Selanjutnya, digunakan representasi vektor seperti Word2Vec dan GloVe untuk mengubah kata-kata menjadi vektor numerik, yang diperlukan untuk model pembelajaran mesin. Untuk klasifikasi, beberapa arsitektur model digunakan, termasuk Long Short-Term Memory (LSTM), FastText, dan Transformer. Masing-masing model ini memiliki keunggulan tersendiri dalam menangkap informasi konteks dan pola dari teks input.

1. Dataset

Dataset yang digunakan adalah AG News, yang berisi data berita dalam 4 kategori: *World, Sports, Business, dan Science/Technology.* Dataset diekstrak dari file kompresi dan di-load dalam format CSV.

2. Tahapan Implementasi

a. Ekstraksi dan Loading Dataset

- Dataset diekstrak menggunakan modul tarfile.

- File CSV *train.csv* dan *test.csv* dimuat ke dalam DataFrame menggunakan *library pandas*.
- Kolom dataset diberi nama: *label, title,* dan *content.*

b. Preprocessing Teks
Langkah preprocessing mencakup:

- Menghapus tanda baca dan spasi berlebih menggunakan regex.
- Konversi teks menjadi huruf kecil.
- Menghapus stopwords menggunakan modul nltk.
- Tokenisasi dilakukan secara manual untuk menghasilkan kata-kata yang bersih.

c. Word Embedding

- **Word2Vec**: Model pretrained dari *Google News* digunakan untuk memperoleh representasi kata dalam bentuk vektor berdimensi 300.
- **GloVe**: Model *pretrained GloVe 6B* digunakan untuk representasi kata dengan dimensi 100.

d. Dataset untuk PyTorch DataLoader
Dataset diproses untuk kompatibilitas dengan *PyTorch*. Setiap teks diubah menjadi vektor dengan rata-rata embedding kata. Label dikodekan menjadi angka menggunakan *LabelEncoder.*

e. Model Neural Networks

1. **LSTM**:
    - Model LSTM terdiri atas satu layer LSTM dengan dimensi hidden sebesar 256 dan *layer fully connected* untuk klasifikasi.
    - Loss function yang digunakan adalah *CrossEntropyLoss*, dan *optimizer* adalah *AdamW* dengan *learning rate* 0.0001.
2. **FastText**:
    - Model *FastText* terdiri dari *layer fully connected* sederhana yang mengoperasikan rata-rata embedding.
    - Pelatihan menggunakan parameter yang sama seperti LSTM.
3. **Transformer**:
    - Model Transformer memanfaatkan arsitektur transformer encoder dengan 6 layer, 8 heads, dan embedding dimensi 300.
    - Input berupa token yang di-embed menjadi representasi numerik.

f. Pelatihan Model
Pelatihan dilakukan selama maksimal 10 epoch dengan *early stopping* untuk menghentikan pelatihan jika loss tidak membaik selama 3 epoch berturut-turut.

g. Evaluasi Model
Dilihat akurasi dari setiap model untuk mengetahui performance model terbaik.

HASIL DAN ANALISIS

Hasil yang didapatkan pada percobaan keempat model menunjukkan perbedaan performa yang signifikan dalam mempelajari data. Keempat model yang digunakan, yaitu LSTM, FastText, Transformer, dan BERT, dievaluasi berdasarkan tiga parameter utama, yakni nilai loss, epoch terbaik, dan waktu pelatihan. Hasil dari keempat model tersebut disajikan pada Tabel 1.

Tabel 1. Hasil Lost Model

Model	Lost	Epoch Terbaik	Waktu (menit)
LSTM	0.3303	10	16,8
Fast Text	0.5508	3	3,1
Transformer	1,9853	9	25,6
BERT	0,00401	6	110,98

Hasil performance model juga diukur dari metrik evaluasi pada setiap model pada Gambar 1,2,3, dan 4.

Gambar 1. Metrik Evaluasi LSTM

```
Evaluasi LSTM:
Accuracy: 0.9517
              precision    recall  f1-score   support

           1       0.97      0.83      0.90      1900
           2       0.95      0.99      0.97      5700

    accuracy                           0.95      7600
   macro avg       0.96      0.91      0.93      7600
weighted avg       0.95      0.95      0.95      7600
```

Gambar 2. Metrik Evaluasi Fast Text

	precision	recall	f1-score	support
0	0.00	0.00	0.00	0
1	0.25	0.80	0.38	30000
2	0.75	0.12	0.20	90000
3	0.00	0.00	0.00	0
accuracy			0.29	120000
macro avg	0.25	0.23	0.15	120000
weighted avg	0.63	0.29	0.25	120000

Gambar 3. Metrik Evaluasi Transformer

Gambar 4. Metrik Evaluasi BERT

	precision	recall	f1-score	support
0	0.00	0.00	0.00	225
1	0.00	0.00	0.00	254
2	0.00	0.00	0.00	275
3	0.25	1.00	0.39	246
accuracy			0.25	1000
macro avg	0.06	0.25	0.10	1000
weighted avg	0.06	0.25	0.10	1000

Evaluasi BERT:
Accuracy: 0.9645

	precision	recall	f1-score	support
1	0.94	0.91	0.93	1900
2	0.97	0.98	0.98	5700
accuracy			0.96	7600
macro avg	0.96	0.95	0.95	7600
weighted avg	0.96	0.96	0.96	7600

Adapun analisis ini dibagi dalam beberapa bagian, yaitu :

1. **Dataset**

   Hasil analisis menunjukkan bahwa performa empat model yang digunakan, yaitu LSTM, FastText, Transformer, dan BERT, dipengaruhi oleh kualitas dataset yang digunakan. LSTM dan BERT mampu menunjukkan performa yang baik dengan nilai loss masing-masing 0.3303 dan 0.00401, yang menunjukkan bahwa dataset ini sudah cukup representatif untuk kedua model tersebut. Distribusi data yang memadai memungkinkan kedua model ini belajar pola dengan baik dan melakukan generalisasi secara optimal.

   Sebaliknya, FastText dan Transformer mencatatkan performa yang jauh lebih rendah dengan nilai loss 0.5508 dan 1.9853. Ini mengindikasikan bahwa model tersebut kesulitan dalam menangkap pola dari dataset yang digunakan. Rendahnya performa pada FastText dan Transformer bisa saja disebabkan oleh distribusi data yang tidak merata atau tidak seimbang, di mana terdapat kelas yang memiliki jumlah sampel sangat sedikit. Hal ini terlihat dari adanya Recall 0.0 dan peringatan seperti *UndefinedMetricWarning*, yang menunjukkan beberapa label tidak terprediksi sama sekali. Oleh karena itu, kualitas distribusi dan preprocessing dataset, training serta pemodelan datanya perlu diperbaiki perlu menjadi perhatian lebih lanjut untuk memaksimalkan performa kedua model ini.

2. **Waktu dan Sumber Daya Komputasi**

   Perbandingan waktu pelatihan antar model menunjukkan efisiensi dan kebutuhan sumber daya yang berbeda-beda. FastText memiliki waktu pelatihan tercepat, yakni hanya 3,1 menit pada epoch ke-3, tetapi performanya sangat rendah. Hal ini menandakan bahwa meskipun FastText unggul dalam kecepatan pelatihan, model ini tidak mampu menangkap kompleksitas dataset yang lebih tinggi.

   LSTM mencatat waktu pelatihan 16,8 menit pada epoch ke-10 dengan nilai loss yang cukup rendah, menjadikannya model yang efisien baik dari segi waktu maupun

performa. Model ini mampu mencapai keseimbangan antara akurasi dan efisiensi komputasi, sehingga cocok digunakan jika sumber daya terbatas.

Berbeda dengan LSTM dan FastText, Transformer membutuhkan waktu pelatihan lebih lama, yakni 25,6 menit, namun menghasilkan nilai loss yang jauh lebih tinggi (1.9853). Hal ini menunjukkan bahwa Transformer belum dapat belajar optimal dari dataset dan memerlukan parameter tuning lebih lanjut dan training serta pemodelan datanya perlu diperbaiki.

BERT memiliki waktu pelatihan paling lama, yaitu 110,98 menit pada epoch ke-6, tetapi performa yang dicapai sangat baik dengan nilai loss 0.00401. Ini menegaskan bahwa BERT membutuhkan sumber daya komputasi yang jauh lebih besar dibandingkan model lain, namun memberikan hasil terbaik untuk menangani dataset yang kompleks.

Secara keseluruhan, jika efisiensi waktu dan sumber daya menjadi prioritas, LSTM adalah pilihan yang paling baik. Namun, jika performa optimal yang diutamakan dan sumber daya komputasi memadai, BERT menjadi model yang lebih unggul.

3. Generalisasi

Analisis generalisasi menunjukkan bahwa LSTM dan BERT memiliki kemampuan generalisasi yang baik, ditunjukkan melalui nilai loss yang rendah dan konsistensi performa pada epoch tertentu. Kedua model ini mampu memahami pola dalam dataset dan memprediksi label secara akurat meskipun memiliki kompleksitas model yang berbeda.

Di sisi lain, FastText mengalami kesulitan dalam melakukan generalisasi. Nilai loss yang masih tinggi serta adanya indikasi early stopping pada epoch ke-3 menunjukkan bahwa model ini tidak mampu belajar lebih jauh dari dataset. Hal ini diperparah dengan performa yang buruk pada kelas tertentu akibat distribusi data yang tidak merata atau kelas dengan jumlah sampel sangat sedikit.

Transformer juga memiliki kemampuan generalisasi yang buruk, terlihat dari nilai loss tertinggi dan prediksi yang hanya benar pada kelas tertentu. Model ini mengalami kesulitan dalam menangani distribusi data yang tidak seimbang atau jumlah data yang mungkin kurang mencukupi untuk kompleksitas arsitekturnya. Akibatnya, Transformer tidak mampu melakukan prediksi dengan akurat dan generalisasi dengan baik.

Secara keseluruhan, generalisasi model sangat dipengaruhi oleh kualitas dataset serta kompleksitas arsitektur model. LSTM dan BERT terbukti mampu melakukan generalisasi dengan baik, sedangkan FastText dan Transformer memerlukan perbaikan

lebih lanjut, terutama dalam menangani distribusi data yang tidak merata dan training datanya.

KESIMPULAN

Penelitian ini mengevaluasi empat model klasifikasi teks, yakni LSTM, FastText, Transformer, dan BERT, menggunakan dataset AG News dengan tahapan preprocessing, embedding, dan evaluasi performa. Hasil menunjukkan bahwa BERT menghasilkan performa terbaik dengan nilai loss terendah (0.00401) meskipun memerlukan waktu dan sumber daya komputasi tertinggi (110,98 menit), sementara LSTM menjadi alternatif efisien dengan keseimbangan performa dan waktu (loss 0.3303 dalam 16,8 menit). FastText, meskipun cepat (3,1 menit), gagal menangkap kompleksitas data dengan baik, sedangkan Transformer membutuhkan parameter tuning lebih lanjut karena nilai loss tertinggi (1.9853). Analisis generalisasi menunjukkan bahwa LSTM dan BERT mampu belajar pola data dengan baik, sedangkan FastText dan Transformer mengalami kesulitan akibat distribusi, dan pengolahan data yang tidak merata dan kurang tepat. Dengan demikian, BERT direkomendasikan untuk performa optimal, sementara LSTM cocok digunakan dalam sumber daya komputasi yang terbatas.

SARAN

Pada penelitian selanjutnya, untuk percobaan pada model Fast Text dan Transformer dapat diperbaiki lagi setiap langkahnya dari awal (pre-processing), training modelnya hingga evaluasi modelnya agar hasilnya lebih baik.