TUGAS PEMROSESAN BAHASA ALAMI

Nama	NIM
Kanaya Dea Thalita Akhmad	121450001
Anasthashya Rachman	121450013
Kirana Ratu Malhanny	121450082
Ghozi Alvin Karim	121450123
Claudhea Angeliani	121450124

Jurusan Sains Data
Fakultas Sains
Institut Teknologi Sumatera
2024

SOAL

Kerjakan eksperimen dengan data yahoo answer berikut dengan Framework NLP Pytorch. buat dalam file .py dan modul. Gunakan word embedding Word2Vec atau GloVe. Model yang diuji adalah sebagai berikut:

- LSTM atau GRU (pilih salah satu)
- Fast Text
- Transformer
- BERT

Buat Analisis Perbandingan model di atas dengan parameter:

- Dataset (Apakah membutuhkan yang lebih besar?)
- Waktu dan Sumber Daya Komputasi
- Jelaskan Generalisasi

ANALISIS

1. **GRU**
    - Dataset

      Dataset yang digunakan menggunakan data yahoo answer dengan didapatkan akurasi pada data train sebesar 0.8142 namun pada akurasi untuk data latih (validasi) didapatkan 0.7621 dengan selisihnya adalah sekitar 0.0521 atau 5.21% selisih ini menunjukkan bahwa model tidak mengalami overfitting yang signifikan menunjukkan bahwa model masih dapat dikatakan dapat belajar dengan baik dengan jumlah data yang telah ditentukan pada dataset tersebut, untuk pada saat proses pelatihan sendiri menghasilkan nilai yang tidak terlalu jauh pada nilai training mengindikasikan hasil proses pembelajaran masih cukup signifikan terhadap proses pembelajaran model, sehingga dapat dikatakan dataset sudah cukup, walaupun berkemungkinan akan lebih baik jika dataset ditambah dalam proses training data.
    - Waktu dan sumber daya komputasi

      Waktu komputasi yang digunakan dengan model GRU adalah 332.312 detik dengan daya komputasi yang tergolong sedang karena tergantung pada ukuran jaringan yang digunakan dan cocok untuk penggunaan pada GPU atau CPU dengan spesifikasi umum.
    - Generalisasi

      GRU dapat menangkap pola sekuensial dengan baik, sehingga memiliki kemampuan generalisasi yang baik untuk teks bersifat time-dependent. Dengan dataset Yahoo Answers, GRU dapat menghindari overfitting pada data yang moderat berdasarkan parameter yang diatur dan ketepatan parameter.

2. **Fast Text**
    - Dataset

      Dataset yang digunakan menggunakan data yahoo answer dengan didapatkan akurasi pada data train sebesar 0.7995 namun pada akurasi untuk data latih (validasi) didapatkan 0.7433 selisih ini masih cukup kecil untuk menunjukkan bahwa model tidak mengalami overfitting, meskipun ada ruang untuk meningkatkan performa jika pada training data ditambah. Model dengan jumah saat ini sudah cukup pada training data dan data testing (validasi) dapat mengikuti dengan signifikan terhadap data train, sehingga masih dikatakan baik.
    - Waktu dan sumber daya komputasi

      Waktu komputasi yang digunakan dengan model Fasttext adalah 319.242 detik dengan daya komputasi yang tergolong rendah karena dapat digunakan pada mesin dengan spesifikasi rendah, seperti CPU standar.
    - Generalisasi

      FastText sangat baik dalam menangkap representasi kata sederhana dan bekerja efektif untuk teks pendek. Generalisasinya baik meski datasetnya tidak besar, sehingga cocok untuk kasus Yahoo Answers yang memiliki beragam jenis pertanyaan.

3. **Transformer**
    - Dataset

      Dataset yang digunakan menggunakan data yahoo answer dengan didapatkan akurasi pada data train sebesar 0.7939 namun pada akurasi untuk data latih (validasi) didapatkan 0.7410 dengan selisih ini menunjukkan bahwa model tidak mengalami overfitting dan menunjukkan bahwa model masih dapat dikatakan dapat belajar dengan baik dengan jumlah data yang telah ditentukan pada dataset tersebut, untuk pada saat proses pelatihan sendiri menghasilkan nilai yang tidak terlalu jauh pada nilai training mengindikasikan hasil proses pembelajaran masih cukup signifikan terhadap proses pembelajaran model, sehingga dapat dikatakan dataset sudah cukup, namun dari hasil tersebut walau tidak dapat dikatakan overfit akan lebih baik jika dataset ditambah.
    - Waktu dan sumber daya komputasi

      Waktu komputasi yang digunakan dengan model Transformer adalah 374.333 detik dengan daya komputasi yang tergolong tinggi karena membutuhkan GPU atau TPU untuk efisiensi serta akan terjadi peningkatan eksponensial pada kebutuhan memori dengan bertambahnya panjang sekuens input yang digunakan.
    - Generalisasi

      Transformer unggul dalam menangkap konteks global dari teks panjang. Generalisasinya sangat baik pada dataset besar seperti Yahoo Answers, namun pada dataset kecil berisiko underfitting atau overfitting tanpa regulasi yang memadai.

4. **BERT**
    - Dataset
      Dataset yang digunakan menggunakan data yahoo answer dengan didapatkan akurasi pada data train sebesar 0.7697 namun pada akurasi untuk data latih (validasi) didapatkan 0.7297 dengan selisih ini menunjukkan bahwa model tidak mengalami overfitting dan menunjukkan bahwa model masih dapat dikatakan dapat belajar dengan baik dengan jumlah data yang telah ditentukan pada dataset tersebut, untuk pada saat proses pelatihan sendiri menghasilkan nilai yang tidak terlalu jauh pada nilai training mengindikasikan hasil proses pembelajaran masih cukup signifikan terhadap proses pembelajaran model, sehingga dapat dikatakan dataset sudah cukup, namun dari hasil tersebut walau tidak dapat dikatakan overfit akan lebih baik jika dataset ditambah agar model BERT dapat melatih model dengan baik jika terdapat banyak variabel data.
    - Waktu dan sumber daya komputasi
      Waktu komputasi yang digunakan dengan model BERT adalah 491.344 detik dengan daya komputasi yang tergolong tinggi karena membutuhkan GPU/TPU dengan RAM besar untuk menangani dimensi embedding dan parameter serta membutuhkan distributed training untuk efisiensi dalam pretraining.
    - Generalisasi
      BERT mampu menangkap konteks dua arah dari teks dengan presisi tinggi. Generalisasinya kuat, bahkan pada dataset moderat seperti Yahoo Answers, dengan syarat dilakukan fine-tuning yang baik. Namun, performa dapat lebih optimal dengan dataset yang lebih besar.

KESIMPULAN

Secara keseluruhan, berdasarkan dari segi dataset, waktu dan sumber daya komputasi. Model yang unggul dalam menerapkan komputasi ini adalah model GRU berdasarkan dengan performa evaluasi pemodelan yang dilakukan karena memiliki akurasi validasi dan latih yang paling tinggi serta dari segi selisih evaluasi memiliki generalisasi pemodelan yang cukup baik,yaitu 5,21%. Selain itu,model dengan BERT memiliki generalisasi yang paling baik karena selisih evaluasinya terkecil.Berdasarkan dengan hal tersebut, model GRU masih tetap lebih baik karena memiliki evaluasi terbaik.

Metode	Dataset	Waktu dan Sumber Daya Komputasi	Generalisasi
GRU	Penggunaan dataset Yahoo Answers dinilai cukup baik dengan akurasi data	Waktu komputasi 332.312 detik, dengan daya	Baik untuk data sekuensial khususnya teks yang

	latih sebesar 0.8142 dan validasi 0.7621 (selisih 5.21%), menunjukkan model tidak overfitting dan belajar dengan baik. Penambahan data dapat meningkatkan performa.	komputasi sedang	bersifat time-dependent
Fast Text	Penggunaan dataset Yahoo Answers dinilai cukup baik dengan akurasi data latih sebesar 0.7995 dan validasi sebesar 0.7433 (selisih 5.62%), dimana selisih ini masih cukup kecil yang menunjukkan bahwa model tidak mengalami overfitting.	Waktu komputasi 319.242 detik, dengan daya komputasi rendah	Baik untuk teks singkat
Transformer	Penggunaan dataset Yahoo Answers dinilai cukup baik dengan akurasi data latih sebesar 0.7939 dan validasi sebesar 0.7410 (selisih 5.29%), dimana selisih ini menunjukkan bahwa model tidak mengalami overfitting.	Waktu komputasi 374.333 detik, dengan daya komputasi tinggi	Sangat baik pada dataset besar
BERT	Penggunaan dataset Yahoo Answers dinilai cukup baik dengan akurasi data latih sebesar 0.7697 dan validasi sebesar 0.7297 (selisih 4%), dimana selisih ini menunjukkan bahwa model tidak mengalami overfitting.	Waktu komputasi 491.344 detik, dengan daya komputasi tinggi	Sangat baik untuk teks kontekstual