TUGAS KELOMPOK

Pemrosesan Bahasa Alami – SD4126

Disusun Oleh

1. **M Gilang Martiansyah M (121450056)**
2. **M. Fahrul Aditya (121450156)**
3. **Rangga Adi Putra (121450106)**
4. **Shula Thalita Ardhya Putri (121450087)**
5. **Nadia Silvani (121450054)**

Dosen Pengampu : Luluk Muthoharoh S.Si,. M.Si., Ardika Satria, S.Si., M.Si., Ayu Mawadda Warohmah

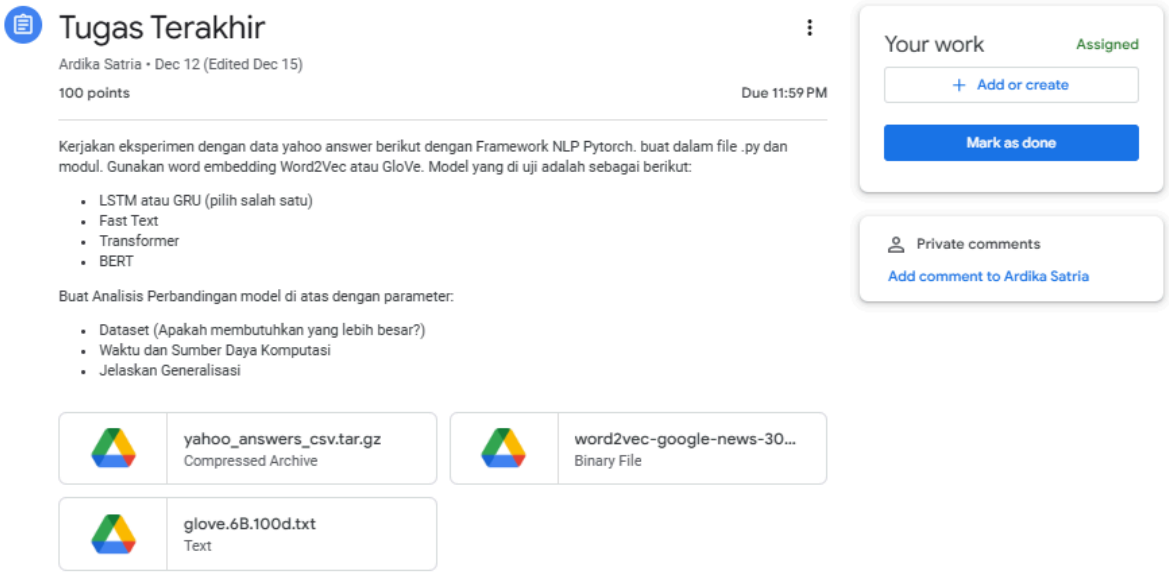
PROGRAM STUDI SAINS DATA

JURUSAN SAINS

INSTITUT TEKNOLOGI SUMATERA

2024/2025

Tugas Analisis



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 di uji 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

Hasil Analisis

Dari kelima eksperimen yang disarankan, kami memilih 4 diantaranya, yakni model :

- GRU (Gated Recurrent Unit)
- Fast Text
- Transformer
- BERT (Bidirectional Encoder Representations from Transformers)

Setelah eksperimen yang dilakukan, hasil analisis yang didapatkan adalah sebagai berikut:

1. **Dataset (Apakah membutuhkan yang lebih besar?)**

    a. GRU (Gated Recurrent Unit)

    Model GRU efektif untuk dataset yang berukuran sedang hingga besar, seperti yang digunakan dalam eksperimen ini dengan 120,000 entri untuk pelatihan dan 7,600 untuk pengujian. GRU sangat cocok untuk data sekuensial dan teks karena kemampuannya dalam memproses informasi sekuensial dan mempertahankan informasi penting melalui gerbangnya yang memungkinkan model untuk mengatasi masalah vanishing gradient yang sering terjadi pada RNN tradisional. Dataset yang cukup besar memungkinkan GRU untuk menangkap dependensi dan konteks yang cukup dalam teks, yang penting untuk tugas klasifikasi berdasarkan konten teks yang kompleks.

    b. Fast Text

    FastText adalah model yang dirancang oleh Facebook untuk klasifikasi teks dan representasi kata yang efisien. FastText unik karena tidak hanya mempelajari fitur pada level kata, tetapi juga pada level n-gram karakter, memungkinkannya memahami morfologi kata lebih baik. Ini membuat FastText sangat adaptif terhadap dataset yang memiliki banyak kata yang jarang muncul atau bahasa dengan aglutinasi tinggi. FastText membutuhkan dataset yang lebih kecil dibandingkan model seperti BERT atau Transformer untuk mencapai kinerja yang kompetitif.

    c. Transformer

Transformer memanfaatkan mekanisme attention untuk mengatasi keterbatasan RNN/LSTM/GRU dalam pemrosesan sekuensial dan memungkinkan paralelisasi yang lebih besar dalam pemrosesan data. Model Transformer diketahui memerlukan dataset yang cukup besar dan beragam untuk mengoptimalkan kemampuannya dalam memahami konteks dan dependensi jarak jauh dalam teks. Dalam eksperimen ini, dataset yang digunakan diolah dengan embeddings GloVe yang membantu mengurangi dimensi spesifikasi model namun tetap membutuhkan data yang cukup untuk menangkap nuansa bahasa secara efektif. Penggunaan embeddings dari GloVe secara tidak langsung mengurangi kebutuhan untuk dataset yang lebih besar daripada jika model harus mempelajari embeddings dari nol, namun dataset yang besar tetap diutamakan untuk meningkatkan efektivitas model. Transformer adalah dasar dari model-model lanjutan seperti BERT dan telah merevolusi pemahaman konteks dalam teks.

d. Bidirectional Encoder Representations from Transformers (BERT)

BERT (Bidirectional Encoder Representations from Transformers) mengambil ide dasar Transformer dan menerapkannya dalam model yang mengkondisikan kedua sisi konteks secara bersamaan. BERT melakukan pre-training dengan dua tugas: masked language modeling dan next sentence prediction, sehingga model BERT sangat bergantung pada dataset yang besar dan beragam untuk melatih kecanggihannya dalam memahami bahasa secara mendalam. BERT dioptimalkan melalui pre-training pada dataset seperti Wikipedia dan BookCorpus, yang mencakup konten luas dari berbagai topik dan domain pengetahuan. Untuk aplikasi klasifikasi teks, BERT membutuhkan dataset yang besar tidak hanya untuk mempertahankan tingkat pemahaman ini tetapi juga untuk mengadaptasikannya ke konteks spesifik yang diperlukan dalam klasifikasi. Dataset yang besar dan berkualitas tinggi memungkinkan BERT menangkap nuansa dan variabilitas dalam penggunaan bahasa yang berbeda, yang kritis untuk menghasilkan prediksi yang akurat dan relevan.

Walaupun umumnya BERT membutuhkan dataset yang besar untuk melakukan fine-tuning secara efektif, sebagaimana dirancang untuk memanfaatkan pre-training ekstensif dari data yang luas untuk memahami konteks dan nuansa bahasa secara

mendalam. Dalam eksperimen ini dataset dengan 120,000 data pelatihan dan 7,600 data uji merupakan skala yang cukup untuk BERT bekerja dengan baik. Ukuran dataset ini memungkinkan BERT untuk mengadaptasi pengetahuan linguistik umum yang diperoleh selama pre-training ke nuansa spesifik dari tugas klasifikasi yang dihadapi. BERT cenderung berperforma lebih baik dengan dataset yang lebih besar karena kemampuannya untuk menggeneralisasi dari konteks yang lebih luas ke kasus spesifik.

2. Waktu dan Sumber Daya Komputasi

Waktu dan sumber daya komputasi sangat dipengaruhi oleh arsitektur model serta ukuran dan kompleksitas dataset:

a. **GRU**

Eksperimen GRU menunjukkan efisiensi yang cukup tinggi dalam hal waktu dan sumber daya komputasi karena dijalankan di infrastruktur yang berbeda, yakni melalui kaggle. Dengan pelatihan yang membutuhkan sekitar 3 menit per epoch dan pengujian yang berlangsung dalam hitungan detik, GRU menawarkan solusi yang efisien untuk dataset berukuran besar. Penggunaan batch yang besar dan parameter model yang disederhanakan dibandingkan dengan model seperti BERT atau Transformer memungkinkan GRU untuk menjalankan pelatihan dan validasi dengan cepat, menjadikannya pilihan yang praktis untuk aplikasi yang membutuhkan respon cepat atau di lingkungan dengan sumber daya komputasi yang terbatas. GRU lebih efisien daripada LSTM dalam hal waktu dan penggunaan memori karena memiliki struktur yang lebih sederhana, tetapi masih membutuhkan sumber daya yang signifikan untuk dataset besar karena sifat pemrosesan sekuensialnya.

b. **Fast Text**:

Salah satu model tercepat dan paling efisien dari segi sumber daya. Keefektifan Fast Text dalam pemrosesan teks membuatnya ideal untuk aplikasi yang memerlukan respons cepat dan penggunaan sumber daya yang rendah.

Keunggulan utama FastText dalam hal sumber daya komputasi adalah efisiensinya. FastText dapat dilatih dengan cepat pada perangkat keras standar dan tidak memerlukan GPU yang mahal, yang membuatnya ideal untuk pengembangan model di lingkungan dengan sumber daya terbatas.

c. **Transformer**:

Dalam penerapan model Transformer yang dianalisis, observasi menunjukkan bahwa pelatihan model membutuhkan waktu yang signifikan meskipun hanya dijalankan selama tiga epoch. Dengan waktu pelatihan yang mencatat sekitar beberapa menit per epoch dan kecepatan proses batch yang efisien, ini menandakan bahwa sumber daya komputasi yang digunakan cukup memadai untuk menjalankan model. Namun, seperti kebanyakan model berbasis Transformer, terdapat kebutuhan substansial akan sumber daya komputasi, terutama ketika menggunakan GPU, untuk memproses operasi yang intensif dalam model yang memiliki banyak parameter dan kompleksitas tinggi.

d. **BERT**:

Sama seperti Transformer dalam hal kebutuhan hardware tetapi lebih intensif karena skala pre-training. Pelatihan BERT sangat bergantung pada GPU atau TPU untuk mengelola kompleksitas dan ukuran modelnya.

Proses training BERT untuk eksperimen ini membutuhkan waktu yang cukup signifikan, sekitar 50 menit per epoch dengan total 5 epoch, yang menghasilkan total waktu pelatihan sekitar 4-5 jam. Validation, di sisi lain, terjadi jauh lebih cepat, selesai dalam waktu sekitar 1 menit. Kecepatan batch per detik selama training berkisar antara 77-79, dan selama validation antara 110-160, menunjukkan efisiensi yang cukup tinggi dalam proses komputasi. Kinerja ini menunjukkan bahwa, meskipun BERT memerlukan sumber daya komputasi yang signifikan karena arsitekturnya yang kompleks dan jumlah parameter yang besar, infrastruktur yang digunakan cukup mampu mengatasi tantangan ini, menjadikan proses pelatihan dan validasi cukup efisien.

3. Jelaskan Generalisasi

a. **GRU**

GRU menunjukkan kemampuan generalisasi yang sangat baik dalam eksperimen ini, dengan performa konsisten antara data pelatihan dan pengujian. Model mencapai akurasi validasi yang tinggi, menunjukkan kemampuannya untuk tidak hanya mempelajari tetapi juga menggeneralisasikan dari data pelatihan ke data yang tidak terlihat. Kinerja yang stabil dan tingkat akurasi yang tinggi menunjukkan bahwa GRU mampu mengatasi overfitting, yang sering menjadi masalah pada model pembelajaran mendalam lainnya.

b. **Fast Text**

Performa model pada data validasi dan pengujian menunjukkan bahwa FastText mampu menggeneralisasi, namun tingkat akurasinya yang sekitar 37.5% menunjukkan bahwa masih ada banyak ruang untuk peningkatan. Meskipun model mampu mengenali dan mengelompokkan teks ke dalam kategori yang benar, akurasi yang relatif rendah bisa jadi menandakan bahwa model belum sepenuhnya mengerti variasi atau nuansa dalam data.

Untuk meningkatkan kemampuan generalisasi bisa dilakukan dengan cara menambah jumlah data pelatihan atau menggunakan teknik augmentasi data untuk memberikan model eksposur yang lebih luas terhadap variasi dalam cara pertanyaan ditanyakan atau dijawab. Selain itu, eksperimen dengan parameter seperti ukuran batch, jumlah epoch, dan tingkat pembelajaran juga dapat membantu meningkatkan kinerja model.

c. Transformer

Dari hasil eksperimen, model Transformer menunjukkan kemampuan generalisasi yang baik dengan mencapai akurasi validasi 83.80%. Ini mengindikasikan bahwa model cukup baik dalam menangani overfitting, mengingat bahwa loss validasi dan akurasi menunjukkan kinerja yang stabil sepanjang proses pelatihan. Kemampuan generalisasi yang tinggi ini penting dalam aplikasi dunia nyata di mana model perlu menangani data yang beragam dan tidak terlihat sebelumnya.

d. **BERT**:

Hasil dari eksperimen menunjukkan bahwa BERT memiliki kemampuan generalisasi yang sangat baik. Konsistensi antara akurasi pelatihan dan validasi menunjukkan bahwa model ini tidak hanya mempelajari karakteristik dari data pelatihan tetapi juga mampu menerapkannya pada data baru yang tidak terlihat sebelumnya dengan efektivitas yang tinggi. Skor presisi, recall, dan F1 yang tinggi memperkuat ini, menunjukkan bahwa model berhasil mempertahankan keseimbangan antara sensitivitas dan spesifisitas di berbagai kelas, yang penting untuk kinerja yang adil dan objektif dalam aplikasi dunia nyata.