SimCSE: Simple Contrastive Learning of Sentence Embeddings

Mohamad Arvin Fadriansyah
Faculty of Computer Science
Universitas Indonesia
Depok, Indonesia
Email: mohamad.arvin@ui.ac.id

***Abstract*—This paper presents SimCSE, a simple contrastive learning framework that greatly advances the state-of-the-art sentence embeddings. We first describe an unsupervised approach, which takes an input sentence and predicts *itself* in a contrastive objective, with only standard dropout used as noise. This simple method works surprisingly well, performing on par with previous supervised counterparts. We find that dropout acts as minimal data augmentation and removing it leads to a representation collapse. Then, we propose a supervised approach, which incorporates annotated pairs from natural language inference datasets into our contrastive learning framework, by using "entailment" pairs as positives and "contradiction" pairs as hard negatives. We evaluate SimCSE on standard semantic textual similarity (STS) tasks, and our unsupervised and supervised models using $\text{BERT}_{\text{base}}$ achieve an average of 76.3% and 81.6% Spearman's correlation respectively, a 4.2% and 2.2% improvement compared to previous best results. We also show—both theoretically and empirically—that contrastive learning objective regularizes pre-trained embeddings' anisotropic space to be more uniform, and it better aligns positive pairs when supervised signals are available.**

I. Cara Kerja Program

A. Penjelasan Parameter Pelatihan

```
(myenv) C:\Users\moham\OneDrive\Desktop\NLP\
    SimCSE>python train.py --
    model_name_or_path bert-base-uncased --
    train_file data/nli_for_simcse.csv --
    output_dir result/my-sup-simcse-bert-base-
    uncased --num_train_epochs 3 --
    per_device_train_batch_size 128 --
    learning_rate 5e-5 --max_seq_length 32 --
    evaluation_strategy steps --
    metric_for_best_model stsb_spearman --
    load_best_model_at_end --eval_steps 125 --
    pooler_type cls --overwrite_output_dir --
    temp 0.05 --do_train --do_eval --fp16
```

- `--model_name_or_path bert-base-uncased`
  Model yang digunakan sebagai backbone, dalam hal ini adalah `bert-base-uncased`, model BERT yang tidak case-sensitive.
- `--train_file data/nli_for_simcse.csv`
  Path ke file dataset yang akan digunakan untuk pelatihan. Dalam contoh ini, dataset berada di `data/nli_for_simcse.csv`.
- `--output_dir result/my-sup-simcse-bert-base-uncased`
  Direktori tempat model hasil pelatihan akan disimpan. Di sini, output akan disimpan di `result/my-sup-simcse-bert-base-uncased`.
- `--num_train_epochs 3`
  Jumlah epoch pelatihan, yaitu sebanyak 3 kali iterasi penuh terhadap dataset.
- `--per_device_train_batch_size 128`
  Ukuran batch untuk pelatihan per perangkat (misalnya per GPU/CPU). Ukuran batch ditetapkan menjadi 128.
- `--learning_rate 5e-5`
  Laju pembelajaran (*learning rate*) yang digunakan oleh optimizer selama pelatihan, dalam hal ini adalah 0.00005.
- `--max_seq_length 32`
  Panjang maksimum tokenisasi untuk setiap urutan teks. Teks yang lebih panjang dari 32 token akan dipotong.
- `--evaluation_strategy steps`
  Strategi evaluasi, dalam hal ini evaluasi dilakukan berdasarkan langkah (*steps*) tertentu.
- `--metric_for_best_model stsb_spearman`
  Metode evaluasi yang digunakan untuk menentukan model terbaik, yaitu *Spearman correlation* pada STS-B dataset.
- `--load_best_model_at_end`
  Instruksi untuk memuat model terbaik di akhir pelatihan berdasarkan metrik evaluasi.
- `--eval_steps 125`
  Interval evaluasi, yaitu setiap 125 langkah pelatihan akan dilakukan evaluasi.
- `--pooler_type cls`
  Tipe pooling yang digunakan untuk representasi akhir embedding, yaitu `cls` (menggunakan token [CLS]).
- `--overwrite_output_dir`
  Mengizinkan penimpaan (*overwrite*) direktori output jika sudah ada data sebelumnya.
- `--temp 0.05`
  Temperatur yang digunakan dalam loss fungsi kontrasif untuk SimCSE.
- `--do_train`
  Menjalankan proses pelatihan (*training*).
- `--do_eval`
  Menjalankan proses evaluasi (*evaluation*).
- `--fp16`
  Menggunakan *mixed precision training* untuk memper-

cepat pelatihan dan mengurangi penggunaan memori dengan representasi 16-bit (*float16*).

B. Arsitektur Model

```
[INFO|configuration_utils.py:481] 2024-11-18
    20:04:18,040 >> Model config BertConfig {
  "architectures": [
    "BertForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0.1,
  "gradient_checkpointing": false,
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.1,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "bert",
  "num_attention_heads": 12,
  "num_hidden_layers": 12,
  "pad_token_id": 0,
  "position_embedding_type": "absolute",
  "transformers_version": "4.2.1",
  "type_vocab_size": 2,
  "use_cache": true,
  "vocab_size": 30522
}
```

Berikut adalah penjelasan tiap parameter pada konfigurasi:

- **architectures**: Jenis model yang digunakan adalah `BertForMaskedLM`, yang dirancang untuk tugas Masked Language Modeling.
- **attention_probs_dropout_prob**: Probabilitas dropout pada mekanisme perhatian (*attention*) adalah 0.1.
- **hidden_act**: Fungsi aktivasi yang digunakan pada lapisan tersembunyi adalah `gelu`.
- **hidden_size**: Dimensi representasi tersembunyi (*hidden states*) adalah 768.
- **num_attention_heads**: Model memiliki 12 kepala perhatian (*attention heads*) pada setiap lapisan.
- **num_hidden_layers**: Terdapat 12 lapisan tersembunyi pada arsitektur BERT.
- **intermediate_size**: Ukuran lapisan *intermediate* adalah 3072.
- **max_position_embeddings**: Jumlah maksimum token yang dapat diproses adalah 512.
- **vocab_size**: Ukuran kosakata yang digunakan adalah 30,522 token.

Model ini merupakan implementasi dari arsitektur BERT dengan parameter-parameter standar untuk ukuran *base model*. Konfigurasi ini diatur menggunakan pustaka `transformers` versi 4.2.1.

C. Evaluasi Model Training

```
***** Eval results *****
   epoch = 3.0
   eval_CR = 88.03
   eval_MPQA = 88.53
   eval_MR = 81.64
   eval_MRPC = 73.5
   eval_SST2 = 86.93
   eval_SUBJ = 94.4
   eval_TREC = 81.69
   eval_avg_sts = 0.8134442269792448
   eval_avg_transfer = 84.96000000000001
   eval_sickr_spearman = 0.8028492266028334
   eval_stsb_spearman = 0.824039227355656
}
```

Berikut adalah interpretasi dari hasil evaluasi berdasarkan metrik-metrik yang digunakan:

- **Epoch (3.0)**: Model dievaluasi setelah 3 epoch, yang menunjukkan bahwa model telah melalui tiga kali proses pelatihan penuh terhadap seluruh dataset.
- **Eval_CR (Customer Reviews) = 88.03%**: Angka ini menunjukkan akurasi model dalam melakukan klasifikasi sentimen pada dataset ulasan pelanggan. Performa ini cukup tinggi, yang berarti model mampu menangkap konteks sentimen dalam teks pelanggan dengan baik.
- **Eval_MPQA (Multi-Perspective Question Answering) = 88.53%**: Angka ini merepresentasikan akurasi model dalam memahami polaritas opini (positif, negatif, netral) dalam teks dari berbagai perspektif. Performa ini menunjukkan model memiliki kemampuan yang baik untuk memahami teks opini dengan tingkat subjektivitas yang tinggi.
- **Eval_MR (Movie Reviews) = 81.64%**: Akurasi ini mengukur kemampuan model dalam menentukan sentimen (positif/negatif) pada ulasan film. Meskipun performanya lebih rendah dibanding dataset lainnya, angka ini tetap menunjukkan kinerja yang solid untuk tugas ini.
- **Eval_MRPC (Microsoft Research Paraphrase Corpus) = 73.5%**: Metrik ini mengevaluasi akurasi model dalam menentukan apakah dua kalimat merupakan parafrase (memiliki makna yang sama). Skor ini relatif rendah dibandingkan metrik lain, yang mengindikasikan model sedikit kesulitan menangkap kesamaan semantik antar kalimat.
- **Eval_SST2 (Stanford Sentiment Treebank 2) = 86.93%**: Akurasi ini mengukur performa model dalam klasifikasi sentimen biner (positif/negatif) pada dataset SST2. Skor ini mengindikasikan bahwa model cukup efektif dalam memahami sentimen sederhana dalam teks.
- **Eval_SUBJ (Subjectivity Dataset) = 94.4%**: Skor ini menunjukkan akurasi model dalam membedakan kalimat subjektif (opini, emosi) dan objektif (fakta). Performa yang sangat tinggi ini menunjukkan bahwa model sangat andal dalam membedakan kedua jenis teks.
- **Eval_TREC (Text REtrieval Conference) = 81.69%**: Angka ini merepresentasikan akurasi model dalam klasifikasi pertanyaan ke dalam kategori tertentu. Performa ini menunjukkan model cukup baik dalam memahami konteks dan tipe pertanyaan.
- **Eval_Avg_STS (Semantic Textual Similarity) = 0.8134**: Skor rata-rata ini mencerminkan kemampuan model dalam menentukan kesamaan semantik antara dua teks

menggunakan korelasi Spearman. Angka ini menunjukkan bahwa model mampu menangkap makna semantik dengan akurasi yang tinggi.

- **Eval_Avg_Transfer = 84.96%**: Skor rata-rata ini menunjukkan akurasi model dalam tugas transfer learning pada berbagai dataset. Nilai ini mengindikasikan bahwa representasi yang dihasilkan model bersifat general dan dapat diterapkan pada berbagai tugas.
- **Eval_SICKR_Spearman = 0.8028**: Korelasi Spearman ini mengukur kesamaan semantik pada dataset SICK-R (Semantic Relatedness). Skor ini menunjukkan bahwa model mampu memahami hubungan semantik antar kalimat dengan cukup baik.
- **Eval_STSB_Spearman = 0.8240**: Korelasi Spearman ini menunjukkan performa model dalam dataset STS-B (Semantic Textual Similarity Benchmark). Skor ini lebih tinggi dibanding SICK-R, yang berarti model lebih akurat dalam memahami kesamaan semantik pada dataset ini.

Secara keseluruhan, model menunjukkan performa yang baik pada sebagian besar metrik, terutama dalam tugas klasifikasi sentimen, subjektivitas, dan kesamaan semantik. Namun, performa pada tugas parafrase (MRPC) masih relatif lebih rendah dibandingkan tugas lainnya, yang bisa menjadi fokus untuk peningkatan di masa mendatang.

II. Hasil Uji Coba dan Analisis

A. Metriks Evaluasi

Spearman's Rank Correlation Coefficient

Spearman's rank correlation coefficient, yang disimbolkan dengan ρ atau r_s, digunakan untuk mengukur kekuatan dan arah hubungan monoton antara dua variabel. Berbeda dengan Pearson, Spearman tidak memerlukan asumsi hubungan linier antara variabel. Sebaliknya, ia mengukur apakah dua variabel bergerak dalam arah yang sama atau berlawanan (monotonik), terlepas dari apakah hubungan mereka linier atau tidak.

Spearman's ρ dihitung dengan cara mengonversi data mentah menjadi peringkat (ranking) dan kemudian menghitung korelasi antara peringkat tersebut. Rumus untuk Spearman's rank correlation coefficient adalah:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Dimana:

- d_i adalah selisih antara peringkat masing-masing pasangan nilai.
- n adalah jumlah data.

Korelasi Spearman memberikan nilai antara -1 dan 1, di mana:

- $\rho = 1$ menunjukkan korelasi positif sempurna (dua variabel bergerak dalam arah yang sama secara monotonik).
- $\rho = -1$ menunjukkan korelasi negatif sempurna (dua variabel bergerak dalam arah yang berlawanan secara monotonik).
- $\rho = 0$ menunjukkan tidak ada korelasi monotonik.

Pearson's Correlation Coefficient

Pearson's correlation coefficient, yang disimbolkan dengan r, digunakan untuk mengukur kekuatan dan arah hubungan linier antara dua variabel. Ini adalah ukuran yang paling umum digunakan untuk korelasi, dan mengasumsikan bahwa hubungan antara dua variabel adalah linier serta data mengikuti distribusi normal.

Rumus untuk Pearson's correlation coefficient adalah:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Dimana:

- x_i dan y_i adalah nilai individual dari dua variabel yang dihitung.
- $\bar{x}$ dan $\bar{y}$ adalah rata-rata dari variabel x dan y.

Nilai r berkisar antara -1 hingga 1, dengan interpretasi berikut:

- $r = 1$ menunjukkan hubungan linier positif sempurna.
- $r = -1$ menunjukkan hubungan linier negatif sempurna.
- $r = 0$ menunjukkan tidak ada hubungan linier.

Pearson sangat berguna ketika hubungan antara variabel adalah linier, tetapi tidak cocok jika hubungan yang ada bersifat non-linier atau ada outlier yang memengaruhi hasil.

B. Dataset

Berikut adalah dataset yang digunakan untuk Melatih model,

TABLE I
DATASET PELATIHAN

Nama Dataset	Deskripsi		
	Ukuran	***Bahasa***	***Referensi***
nli_for_simcse	10,000 pasang kalimat	Inggris	[Ref1]

[Ref1] https://huggingface.co/datasets/princeton-nlp/datasets-for-simcse/resolve/main/nli_for_simcse.csv

TABLE II
PENJELASAN KOLOM PADA DATASET PELATIHAN

Nama	Deskripsi
sent0	Kalimat pertama dalam pasangan yang digunakan untuk membangun hubungan semantik.
sent1	Merujuk ke kalimat kedua dalam pasangan yang digunakan, kalimat ini dibandingkan dengan sent0 untuk mengevaluasi kemiripan semantik.
hard_neg	Merujuk ke hard negative example, yaitu sebuah kalimat yang sengaja dipilih karena terlihat mirip secara sekilas dengan sent0 atau sent1, tetapi tidak memiliki kemiripan semantik sebenarnya

Berikut adalah dataset yang digunakan untuk Melakukan evaluasi pada model,

TABLE III
SEMANTIC TEXTUAL SIMILARITY-2012

Nama Dataset	Deskripsi		
	Ukuran	*Bahasa*	*Referensi*
MSR-Paraphrase	750 pasang kalimat	Inggris	[Ref1]
MSR-Video	750 pasang kalimat	Inggris	[Ref2]
SMTeuroparl	459 pasang kalimat	Inggris	[Ref3]
SMTnews	399 pasang kalimat	Inggris	[Ref4]
OnWN	750 pasang kalimat	Inggris	[Ref5]
Gabungan (ALL)	3,108 pasang kalimat	Inggris	-

[Ref1] http://research.microsoft.com/en-us/downloads/607d14d9-20cd-47e3-85bc-a2f65cd28042/
[Ref2] http://research.microsoft.com/en-us/downloads/38cf15fd-b8df-477e-a4e4-a4680caa75af/
[Ref3] http://www.statmt.org/wmt08/shared-evaluation-task.html
[Ref4] News conversation sentence pairs from WMT 399 pairs of sentences.
[Ref5] Pairs of sentences where the first comes from Ontonotes and the second from a WordNet definition.

TABLE IV
SEMANTIC TEXTUAL SIMILARITY-2013

Nama Dataset	*Ukuran*	*Bahasa*	*Referensi*
headlines	750 pasang kalimat	Inggris	[Ref1]
OnWN	561 pasang kalimat	Inggris	[Ref2]
FNWN	189 pasang kalimat	Inggris	[Ref3]
SMT	750 pasang kalimat	Inggris	[Ref4]

[Ref1] http://emm.newsexplorer.eu/NewsExplorer/home/en/latest.html
[Ref2] The sentences are sense definitions from WordNet and OntoNotes.
[Ref3] The sentences are sense definitions from WordNet and FrameNet. Note that some FrameNet definitions involve more than one sentence.
[Ref4] This SMT dataset comes from DARPA GALE HTER and HyTER, where one sentence is a MT output and the other is a reference translation where a reference is generated based on human post editing (provided by LDC) or an original human reference (provided by LDC) or a human generated reference based on FSM as described in (Dreyer and Marcu, NAACL 2012). The reference comes from post edited translations.

TABLE V
SEMANTIC TEXTUAL SIMILARITY-2014

Nama Dataset	*Ukuran*	*Bahasa*	*Referensi*
image	750 pasang kalimat	Inggris	[Ref1]
OnWN	750 pasang kalimat	Inggris	[Ref2]
tweet-news	750 pasang kalimat	Inggris	[Ref3]
deft-news	300 pasang kalimat	Inggris	[Ref4]
deft-forum	450 pasang kalimat	Inggris	[Ref5]
headlines	750 pasang kalimat	Inggris	[Ref6]

[Ref1] The Image Descriptions data set is a subset of the PASCAL VOC-2008 data set (Rashtchian et al., 2010) . PASCAL VOC-2008 data set consists of 1,000 images and has been used by a number of image description systems. The image captions of the data set are released under a CreativeCommons Attribution-ShareAlike license, the descriptions itself are free.
[Ref2] The sentences are sense definitions from WordNet and OntoNotes. 5 pairs of sentences.
[Ref3] The tweet-news data set is a subset of the Linking-Tweets-to-News data set (Guo et al., 2013), which consists of 34,888 tweets and 12,704 news articles. The tweets are the comments on the news articles. The news sentences are the titles of news articles. one sentence.
[Ref4] A subset of news article data in the DEFT project.
[Ref5] A subset of discussion forum data in the DEFT project.
[Ref6] http://emm.newsexplorer.eu/NewsExplorer/home/en/latest.html

TABLE VI
SEMANTIC TEXTUAL SIMILARITY-2015

Nama Dataset	*Ukuran*	*Bahasa*	*Referensi*
answers-forums	2000 pasang kalimat	Inggris	
answers-students	1500 pasang kalimat	Inggris	
belief	2000 pasang kalimat	Inggris	
headlines	1500 pasang kalimat	Inggris	
images	1500 pasang kalimat	Inggris	

TABLE VII
SEMANTIC TEXTUAL SIMILARITY-2016

Nama Dataset	*Ukuran*	*Bahasa*	*Referensi*
answer-answer	1572 pasang kalimat	Inggris	
headlines	1498 pasang kalimat	Inggris	
plagiarism	1271 pasang kalimat	Inggris	
postediting	3287 pasang kalimat	Inggris	
question-question	1555 pasang kalimat	Inggris	

TABLE VIII
SEMANTIC TEXTUAL SIMILARITY-BENCHMARK

Dataset	*file*	*years*	*Train*	*Dev*	*Test*
news	MSRpar	2012	1000	250	250
news	headlines	2013-16	1999	250	250
news	deft-news	2014	300	0	0
captions	MSRvid	2012	1000	250	250
captions	images	2014-15	1000	250	250
captions	track5.en-en	2017	0	125	125
forum	deft-forum	2014	450	0	0
forum	answers-forums	2015	0	375	0
forum	750 answer-answer	2016	0	0	254

TABLE IX
SENTENCES INVOLVING COMPOSITIONAL KNOWLEDGE

Nama Dataset	*Ukuran*	*Bahasa*	*Referensi*
SICK_test_annotated	4927 pasang kalimat	Inggris	
SICK_train	4500 pasang kalimat	Inggris	
SICK_trial	500 pasang kalimat	Inggris	

C. Skenario Eksperimen

Pada percobaan ini, kami menggunakan spesifikasi hardware dan software sebagai berikut untuk memastikan kelancaran pelatihan dan evaluasi model:

1) Spesifikasi Hardware:

- **CPU:** Intel Core i5-10300H
- **GPU:** NVIDIA GeForce RTX 2060 with Max-Q Design dengan 6GB VRAM
- **RAM:** 16 GB
- **Penyimpanan:** SSD dengan kapasitas 1TB

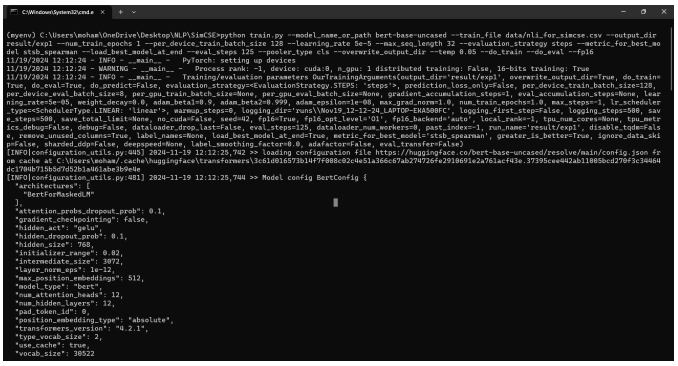
2) Spesifikasi Software dan library yang terinstall:

- **Sistem Operasi:** Windows 11
- **Bahasa Pemrograman:** Python 3.9.2
- **aiofiles**: 23.2.1
- **aiohappyeyeballs**: 2.4.3
- **aiohttp**: 3.10.10
- **aiosignal**: 1.3.1
- **annotated-types**: 0.7.0
- **anyio**: 4.6.2.post1
- **async-timeout**: 4.0.3
- **attrs**: 24.2.0
- **certifi**: 2024.8.30
- **charset-normalizer**: 3.4.0
- **click**: 8.1.7
- **colorama**: 0.4.6
- **contourpy**: 1.3.0
- **cycler**: 0.12.1
- **datasets**: 3.1.0
- **dill**: 0.3.8
- **exceptiongroup**: 1.2.2
- **fastapi**: 0.115.4
- **ffmpy**: 0.4.0
- **filelock**: 3.16.1
- **fonttools**: 4.54.1
- **frozenlist**: 1.5.0
- **fsspec**: 2024.9.0
- **gradio**: 4.44.1
- **gradio_client**: 1.3.0
- **h11**: 0.14.0
- **httpcore**: 1.0.6
- **httpx**: 0.27.2
- **huggingface-hub**: 0.26.2
- **idna**: 3.10
- **importlib_resources**: 6.4.5
- **Jinja2**: 3.1.4
- **joblib**: 1.4.2
- **kiwisolver**: 1.4.7
- **markdown-it-py**: 3.0.0
- **MarkupSafe**: 2.1.5
- **matplotlib**: 3.9.2
- **mdurl**: 0.1.2
- **mpmath**: 1.3.0
- **multidict**: 6.1.0
- **multiprocess**: 0.70.16
- **networkx**: 3.2.1
- **numpy**: 1.26.4
- **orjson**: 3.10.11
- **packaging**: 24.1
- **pandas**: 2.2.3
- **pillow**: 10.4.0
- **pip**: 24.2
- **prettytable**: 3.12.0
- **propcache**: 0.2.0
- **pyarrow**: 18.0.0
- **pydantic**: 2.9.2
- **pydantic_core**: 2.23.4
- **pydub**: 0.25.1
- **Pygments**: 2.18.0
- **pyparsing**: 3.2.0
- **python-dateutil**: 2.9.0.post0
- **python-multipart**: 0.0.17
- **pytz**: 2024.2
- **PyYAML**: 6.0.2
- **regex**: 2024.11.6
- **requests**: 2.32.3
- **rich**: 13.9.4
- **ruff**: 0.7.2
- **sacremoses**: 0.1.1
- **safetensors**: 0.4.5
- **scikit-learn**: 1.3.2
- **scipy**: 1.5.4
- **semantic-version**: 2.10.0
- **setuptools**: 75.1.0
- **shellingham**: 1.5.4
- **six**: 1.16.0
- **sniffio**: 1.3.1
- **starlette**: 0.41.2
- **sympy**: 1.13.1
- **threadpoolctl**: 3.5.0
- **tokenizers**: 0.9.4
- **tomlkit**: 0.12.0
- **torch**: 1.7.1+cu110
- **tqdm**: 4.67.0
- **transformers**: 4.2.1
- **typer**: 0.12.5
- **typing_extensions**: 4.12.2
- **tzdata**: 2024.2
- **urllib3**: 2.2.3
- **uvicorn**: 0.32.0
- **wcwidth**: 0.2.13
- **websockets**: 12.0
- **wheel**: 0.44.0
- **xxhash**: 3.5.0
- **yarl**: 1.17.1
- **zipp**: 3.20.2

3) Default Parameter:

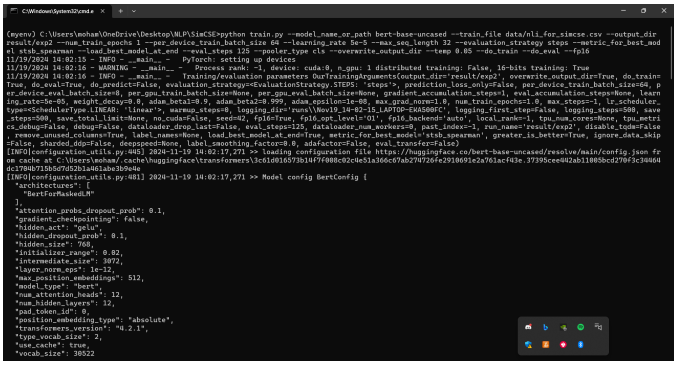
- `--model_name_or_path bert-base-uncased`
- `--train_file data/nli_for_simcse.csv`
- `--output_dir result/my-sup-simcse-bert-base-uncased`
- `--num_train_epochs 3`
- `--per_device_train_batch_size 128`
- `--learning_rate 5e-5`
- `--max_seq_length 32`
- `--evaluation_strategy steps`
- `--metric_for_best_model stsb_spearman`
- `--load_best_model_at_end`
- `--eval_steps 125`
- `--pooler_type cls`
- `--overwrite_output_dir`
- `--temp 0.05`
- `--do_train`
- `--do_eval`
- `--fp16`

4) Eksperimen 1: Mengubah Default hyperparameter dengan perubahan jumlah epoch training menjadi 1 dengan tujuan agar waktu eksekusi menjadi lebih cepat dan mengurangi resiko model mengalami overfitting



- `--model_name_or_path bert-base-uncased`
- `--train_file data/nli_for_simcse.csv`
- `--output_dir result/my-sup-simcse-bert-base-uncased`
- `--num_train_epochs 1`
- `--per_device_train_batch_size 128`
- `--learning_rate 5e-5`
- `--max_seq_length 32`
- `--evaluation_strategy steps`
- `--metric_for_best_model stsb_spearman`
- `--load_best_model_at_end`
- `--eval_steps 125`
- `--pooler_type cls`
- `--overwrite_output_dir`
- `--temp 0.05`
- `--do_train`
- `--do_eval`
- `--fp16`

5) Eksperimen 2: Mengubah Default hyperparameter dengan perubahan jumlah epoch menjadi 1 dan train batch size menjadi 64 (dari yang awalnya 128) dengan tujuan agar waktu eksekusi menjadi lebih cepat dan parameter akan diperbarui lebih sering dengan harapan bisa meningkatkan kecepatan konvergensi karena model mungkin akan lebih cepat mencapai global minima.



- `--model_name_or_path bert-base-uncased`
- `--train_file data/nli_for_simcse.csv`
- `--output_dir result/my-sup-simcse-bert-base-uncased`
- `--num_train_epochs 1`
- `--per_device_train_batch_size 64`
- `--learning_rate 5e-5`
- `--max_seq_length 32`
- `--evaluation_strategy steps`
- `--metric_for_best_model stsb_spearman`
- `--load_best_model_at_end`
- `--eval_steps 125`
- `--pooler_type cls`
- `--overwrite_output_dir`
- `--temp 0.05`
- `--do_train`
- `--do_eval`
- `--fp16`

6) Eksperimen 3: Mengubah Default hyperparameter dengan perubahan jumlah epoch menjadi 1 dan temperature menjadi 0.1 (dari yang awalnya 0.05) agar waktu eksekusi menjadi lebih cepat dan meningkatkan peforma model agar tidak mengalami overfitting, karena model tidak akan terlalu terpaku pada kelas tertentu dan akan memberikan lebih banyak bobot pada kelas-kelas yang memiliki probabilitas yang rendah sehingga meberikan hasil klasifikasi yang lebih beragam serta meningkatkan kinerja model pada data yang memiliki noise atau data yang lebih sulit untuk diprediksi dengan tegas.

- `--model_name_or_path bert-base-uncased`
- `--train_file data/nli_for_simcse.csv`
- `--output_dir result/my-sup-simcse-bert-base-uncased`
- `--num_train_epochs 1`
- `--per_device_train_batch_size 128`
- `--learning_rate 5e-5`
- `--max_seq_length 32`
- `--evaluation_strategy steps`
- `--metric_for_best_model stsb_spearman`
- `--load_best_model_at_end`
- `--eval_steps 125`
- `--pooler_type cls`
- `--overwrite_output_dir`
- `--temp 0.1`
- `--do_train`
- `--do_eval`
- `--fp16`

D. Hasil

TABLE X
HASIL EVALUASI MENGGUNAKAN METRIKS SPEARMAN'S RANK CORRELATION COEFFICIENT

Eksperimen	STS12	STS13	STS14	STS15	STS16	STSBenchmark	SICKRelatedness	Average
Default	73.28	80.91	73.69	80.40	78.47	79.53	78.15	**77.78**
Eksperimen 1	72.16	80.20	72.24	80.40	76.96	79.04	74.97	76.57
Eksperimen 2	72.98	80.82	72.85	79.76	78.48	79.05	77.70	77.38
Eksperimen 3	66.29	76.60	67.60	76.15	72.15	75.31	71.29	72.20

TABLE XI
HASIL EVALUASI MENGGUNAKAN METRIKS PEARSON'S CORRELATION COEFFICIENT

Eksperimen	STS12	STS13	STS14	STS15	STS16	STSBenchmark	SICKRelatedness	Average
Default	82.72	79.36	77.62	78.92	76.03	78.47	83.91	**79.58**
Eksperimen 1	81.78	79.12	77.44	79.16	74.93	78.61	82.58	79.09
Eksperimen 2	82.17	79.35	77.29	78.56	75.76	78.06	84.03	79.32
Eksperimen 3	77.67	76.34	74.70	75.16	69.59	76.70	80.49	75.81

- Eksperimen dengan parameter default memberikan performa terbaik secara keseluruhan baik dengan metriks spearman atau pearson, dengan rata-rata skor tertinggi baik dengan menggunakan spearman (rata-rata) sebesar 77.78 dan menggunakan pearson (rata-rata) sebesar 79.58 pada semua benchmark STS tasks.
- Eksperimen 1 memiliki skor spearman (rata-rata) 76.57 dan skor pearson (rata-rata) 79.09. Dengan pengurangan jumlah epoch menjadi 1, performa model sedikit menurun dibandingkan dengan default. Hal ini menunjukkan bahwa 3 epoch memberikan model lebih banyak kesempatan untuk mempelajari data, menghasilkan generalisasi yang lebih baik.
- Eksperimen 2 memiliki skor spearman (rata-rata) 77.38 dan skor pearson (rata-rata) 79.32. Mengurangi batch size menjadi 64 dengan 1 epoch hampir menyamai performa default. Hal ini menunjukkan bahwa batch size yang lebih kecil dapat mempercepat konvergensi ke global minima, meskipun membutuhkan waktu pelatihan yang lebih lama per epoch.
- Eksperimen 3 memiliki skor spearman (rata-rata) 72.20 dan skor pearson (rata-rata) 75.81. Meningkatkan temperature menjadi 0.1 dengan 1 epoch secara signifikan menurunkan performa. Temperature yang lebih tinggi tampaknya membuat distribusi prediksi model menjadi terlalu seragam, sehingga kurang fokus pada pasangan kalimat yang relevan.

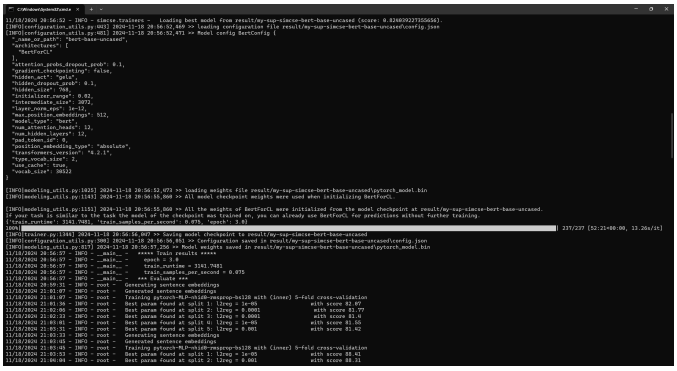
E. Diskusi

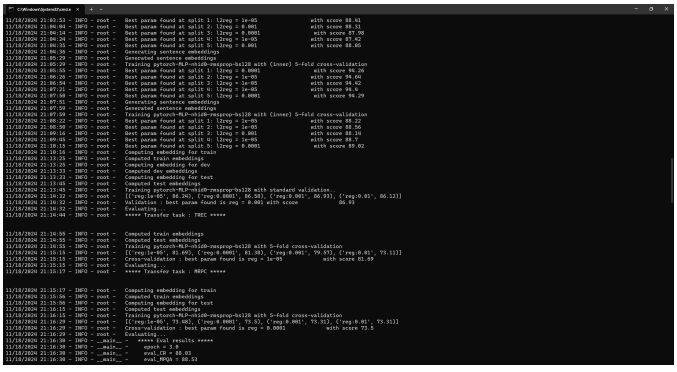
- Mengapa Default Parameter Memberikan Performa Terbaik?
  Jumlah Epoch: Dengan 3 epoch, model memiliki waktu lebih lama untuk mempelajari data. Ini memungkinkan pembelajaran yang lebih mendalam terhadap representasi semantik. Ukuran Batch size yang besar (128) memberikan stabilitas dalam pembaruan parameter, terutama pada model besar seperti bert-base-uncased. Temperature default (0.05) menghasilkan distribusi probabilitas yang cukup tajam, membantu model untuk menekankan pasangan kalimat yang relevan secara lebih efektif.
- Pengurangan Epoch (Eksperimen 1) membuat model memiliki iterasi pelatihan yang lebih sedikit, yang membatasi pembelajaran fitur secara mendalam. Hal ini mengindikasikan bahwa satu epoch tidak cukup untuk mengoptimalkan model dengan dataset ini.
- Pengurangan Batch Size (Eksperimen 2) menghasilkan pembaruan parameter yang lebih sering, tetapi juga meningkatkan volatilitas dalam pembelajaran. Hasil eksperimen menunjukkan bahwa pengurangan batch size tidak cukup untuk mengimbangi dampak pengurangan jumlah epoch
- Pengaruh Temperature yang Lebih Tinggi (Eksperimen 3) membuat model menjadi terlalu bebas dalam melakukan klasifikasi antar kalimat dengan tingkat kesamaan rendah. Ini menghasilkan distribusi prediksi yang kurang tajam, sehingga performa pada sebagian besar task STS menurun drastis.
- Performansi pada STS12 lebih rendah dibanding dataset lain di semua eksperimen ketikang meggunakan metriks spearman (Default: 73.28, Eksperimen 1: 72.16, Eksperimen 2: 72.98, Eksperimen 3: 66.29). Kemungkinan disebabkan dataset STS 12 mengandung lebih banyak pasangan teks yang sulit dibandingkan dataset lain, seperti pasangan yang membutuhkan reasoning atau generalisasi

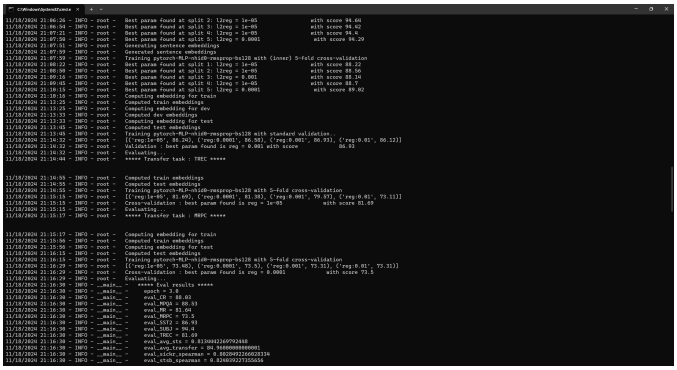
lebih dalam. salah satu solusi yang bisa diterapkan adalah melakukan augmentasi pada data untuk menambah variasi pasangan teks yang lebih kompleks.

- Ada risiko overfitting pada parameter default jika diterapkan pada dataset berbeda. Oleh karena itu, parameter ini mungkin tidak selalu optimal untuk dataset atau task lain. Hal ini terlihat dari jomplangnya skor spearman di dataset yang berbeda seperti STS 13 dengan STS 12.
- Kalimat yang panjang atau kalimat yang memiliki struktur yang kompleks cenderung lebih sulit dalam melakukan proses encoding. Contohnya terdapat pada dataset SICKrelatedness yang biasanya mengandung struktur multi klausa, Hal ini menyebabkan skor spearman yang lebih rendah pada model yang memiliki konfigurasi hyperparameter yang lebih murah.
- Dataset STS 12 dan STS 14 memiliki variasi yang lebih beragam di kalimat-kalimat yang memiliki kemiripan, Hal ini membuat proses pengklasifikasian lebih sulit pada model yang memiliki konfigurasi hyperparameter yang kurang optimal.

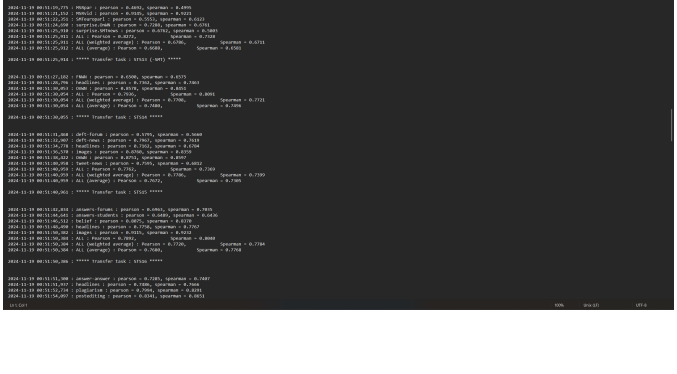
III. Screenshot Pelatihan Model

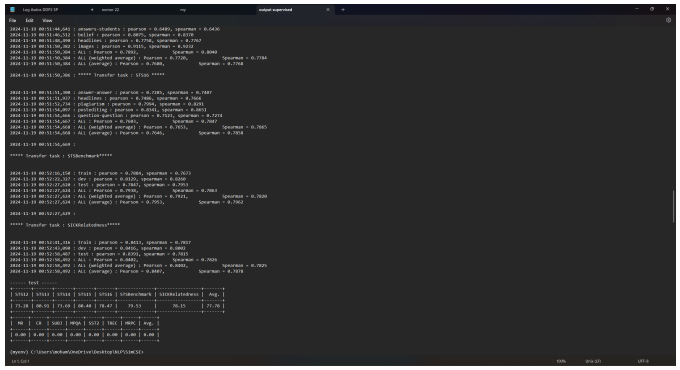




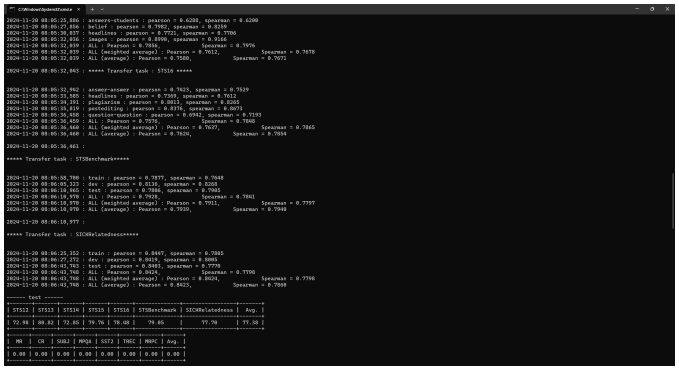


IV. Screenshot Evaluasi Dengan Default Parameter

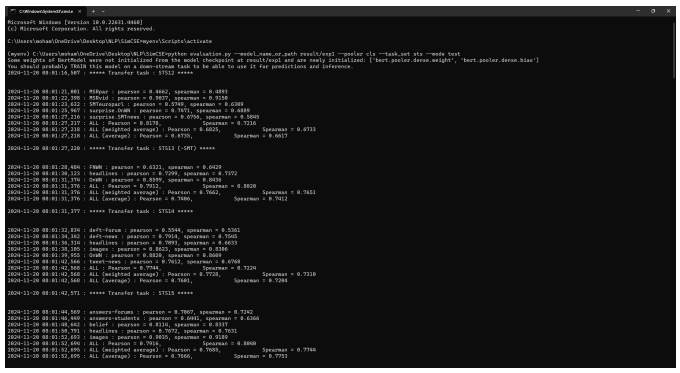


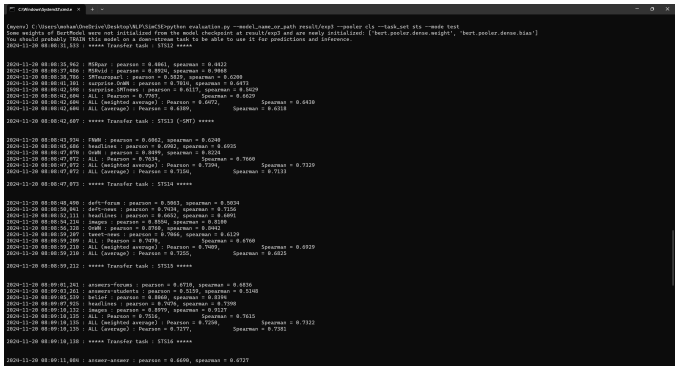


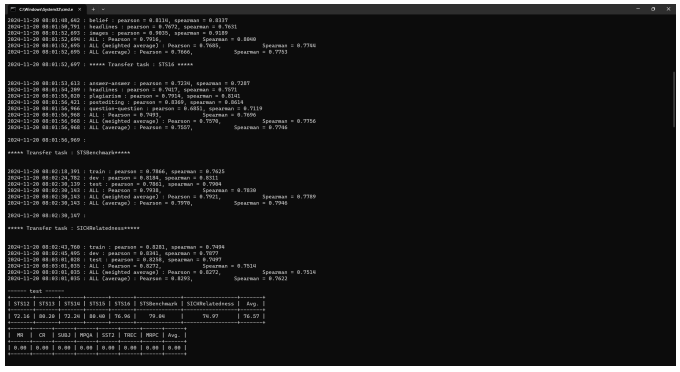
V. SCREENSHOT EVALUASI EKSPERIMEN 1

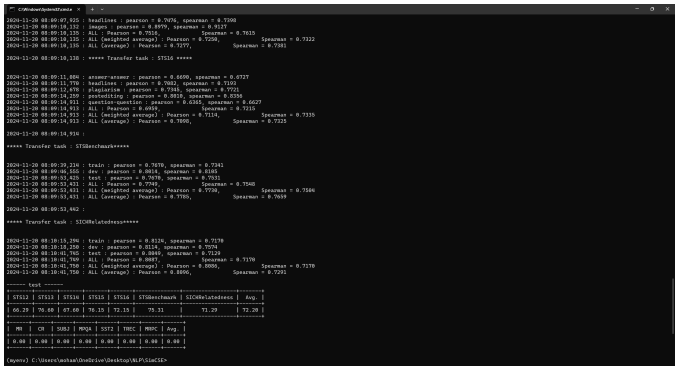


VII. SCREENSHOT EVALUASI EKSPERIMEN 3









VI. SCREENSHOT EVALUASI EKSPERIMEN 2