

COMP0173-CW2-NOTEBOOK

December 7, 2025

1 COMP0173: Coursework 2

The paper HEARTS: A Holistic Framework for Explainable, Sustainable, and Robust Text Stereotype Detection by Theo King, Zekun Wu et al. (2024) presents a comprehensive approach to analysing and detecting stereotypes in text [1]. The authors introduce the HEARTS framework, which integrates model explainability, carbon-efficient training, and accurate evaluation across multiple bias-sensitive datasets. By using transformer-based models such as ALBERT-V2, BERT, and DistilBERT, this research project demonstrates that stereotype detection performance varies significantly across dataset sources, underlining the need for diverse evaluation benchmarks. The paper provides publicly available datasets and code [2], allowing full reproducibility and offering a standardised methodology for future research on bias and stereotype detection in Natural Language Processing (NLP).

2 Instructions

All figures produced during this notebook are stored in the project's `/COMP0173_Figures` directory. The corresponding LaTeX-formatted performance comparison tables—including ALBERT-V2, BERT, and DistilBERT are stored in `/COMP0173_PDF`, with the compiled document available as `COMP0173-CW2-TABLES.pdf`.

3 Technical Implementation (70%)

```
[ ]: # Import libraries
import random, numpy as np, torch
import pandas as pd

import platform
import transformers
from datasets import load_dataset
import spacy
import os
import sys
import importlib.util, pathlib
from pathlib import Path
import warnings
from importlib import reload
from importlib.machinery import SourceFileLoader
```

```
from IPython.display import display
```

```
[ ]: # Check the GPU host (UCL access)
print("CUDA available:", torch.cuda.is_available())
print("Device:", torch.cuda.get_device_name(0))
```

```
CUDA available: True
Device: NVIDIA GeForce RTX 3090 Ti
```

```
[ ]: # Path
import os
os.chdir("/tmp/HEARTS-Text-Stereotype-Detection")
os.getcwd()
```

```
[ ]: '/tmp/HEARTS-Text-Stereotype-Detection'
```

3.1 Part 1: Replicate the baseline AI methodology using the open dataset

The HEARTS framework evaluates stereotype detection using four publicly available text datasets [3]. The following datasets are Multi-Grain Stereotype Dataset (MGSD), Augmented WinoQueer (AWinoQueer), Augmented SeeGULL (ASeeGULL), and Expanded Multi-Grain Stereotype Dataset (EMGSD), which includes labelled stereotypical and non-stereotypical statements covering gender, profession, nationality, race, religion, and LGBTQ+ stereotypes. All datasets referenced in the paper are openly accessible through the associated GitHub repository [2] and Kaggle [3].

3.1.1 Helper Functions

```
[ ]: def collect_summary(project_root: Path) -> pd.DataFrame:

    """
    Collect summary metrics from all model output folders.

    Parameters
    -----
    project_root : Path
        Path to the root directory containing result_output_* folders.

    Returns
    -----
    pd.DataFrame
        A table containing model type, training dataset, evaluation dataset,
        accuracy, macro precision, macro recall, and macro F1 score.
    """

    rows = []

    # Loop through each model type directory
    for model_dir in [
```

```

        "result_output_albertv2",
        "result_output_bert",
        "result_output_distilbert",
]:
    base = project_root / model_dir
    if not base.exists():
        continue

    # Standardise model name
    model_name = model_dir.replace("result_output_", "")

    # Iterate through training subsets
    for train_dir in sorted(base.iterdir()):
        if not train_dir.is_dir():
            continue
        train_name = train_dir.name

        # Iterate through evaluation subsets
        for eval_dir in sorted(train_dir.iterdir()):
            if not eval_dir.is_dir():
                continue
            eval_name = eval_dir.name

            report_path = eval_dir / "classification_report.csv"
            if not report_path.exists():
                continue

            # Load sklearn classification_report CSV
            rep = pd.read_csv(report_path, index_col=0)

            # Extract accuracy and macro-level metrics
            acc = rep.loc["accuracy", "precision"] # sklearn stores_
↪accuracy here
            macro_f1 = rep.loc["macro avg", "f1-score"]
            macro_prec = rep.loc["macro avg", "precision"]
            macro_rec = rep.loc["macro avg", "recall"]

            rows.append(
                {
                    "model": model_name,
                    "train_on": train_name,
                    "eval_on": eval_name,
                    "accuracy": acc,
                    "macro_precision": macro_prec,
                    "macro_recall": macro_rec,
                    "macro_f1": macro_f1,
                }
            )

```

```
)

return pd.DataFrame(rows)
```

```
[ ]: def build_table(summary_df: pd.DataFrame) -> pd.DataFrame:

    """
    Construct a pivot table matching the layout of the HEARTS paper.

    Parameters
    -----
    summary_df : pd.DataFrame
        DataFrame containing model, training set, evaluation set, and metrics.

    Returns
    -----
    pd.DataFrame
        Pivot table indexed by (model, training dataset short name) with
        evaluation datasets as columns and macro-F1 scores as values.
    """

    # Add short labels using mapping dictionaries
    df = summary_df.assign(
        train_short=lambda d: d["train_on"].map(train_map),
        eval_short=lambda d: d["eval_on"].map(eval_map),
    )

    # Remove rows where mapping failed
    df = df.dropna(subset=["train_short", "eval_short"])

    # Create pivot table: rows = model x train_set, cols = eval_set
    table = df.pivot_table(
        index=["model", "train_short"], columns="eval_short", values="macro_f1"
    ).sort_index()

    return table
```

3.1.2 Question 1: Clone the original repository successfully

The code for HEARTS was obtained by cloning a fork of the official repository [2]. The upstream repository, authored by Holistic AI, was then linked to provide access to the original implementation.

```
[2]: # # Clone fork
# git clone https://github.com/n1nt3nd0sw1tch/HEARTS-Text-Stereotype-Detection.
# ↪ git
# cd HEARTS-Text-Stereotype-Detection
```

```
# # Link to the original repo
# git remote add upstream https://github.com/holistic-ai/
# ↪HEARTS-Text-Stereotype-Detection.git
# git remote -v
```

3.1.3 *Question 2*: Document all dependencies and environment setup

A virtual environment was installed to isolate the HEARTS experimentation setup and ensure reproducibility. In the GPU-powered environment, the repository and data paths, along with system and library versions, were set up and confirmed [4]. All four datasets required by the HEARTS pipeline were confirmed to be present in the configured data directory.

Runtime configuration on the UCL GPU node:

- Working directory: /tmp/HEARTS-Text-Stereotype-Detection
- Python: 3.9.21
- Platform: Linux (x86_64, glibc 2.34)
- PyTorch: 2.8.0+cu128
- CUDA: available (NVIDIA GeForce RTX 3090 Ti)
- Transformers: 4.57.3

```
[ ]: # # Create and activate conda environment
# conda create -n hearts python=3.10 -y
# conda activate hearts

# # Register environment as a Jupyter kernel
# python -m pip install ipykernel
# python -m ipykernel install --user

# # Install repository dependencies
# pip install -r requirements.txt
# pip install --no-cache-dir -r requirements.txt
# pip install --no-cache-dir spacy
# pip install --no-cache-dir scikit-learn
# pip install codecarbon
# python -m spacy download en_core_web_lg
```

```
[ ]: # Path
REPO_DIR = Path("/tmp/HEARTS-Text-Stereotype-Detection").resolve()
DATA_DIR = REPO_DIR / "Model Training and Evaluation"

# Change working directory to the repo
os.chdir(REPO_DIR)
```

```
[23]: print("Current working dir:", Path.cwd())
print("Repository directory:", REPO_DIR)
```

```

print("Data directory:", DATA_DIR)

print("Python version:", sys.version)
print("Platform:", platform.platform())

print("PyTorch version:", torch.__version__)
print("CUDA available:", torch.cuda.is_available())
if torch.cuda.is_available():
    print("CUDA device:", torch.cuda.get_device_name(0))
print("Transformers version:", transformers.__version__)

```

Current working dir: /tmp/HEARTS-Text-Stereotype-Detection
 Repository directory: /tmp/HEARTS-Text-Stereotype-Detection
 Data directory: /tmp/HEARTS-Text-Stereotype-Detection/Model Training and Evaluation
 Python version: 3.9.21 (main, Aug 19 2025, 00:00:00)
 [GCC 11.5.0 20240719 (Red Hat 11.5.0-5)]
 Platform: Linux-5.14.0-570.58.1.el9_6.x86_64-x86_64-with-glibc2.34
 PyTorch version: 2.8.0+cu128
 CUDA available: True
 CUDA device: NVIDIA GeForce RTX 3090 Ti
 Transformers version: 4.57.3

```

[24]: # Check the csv files
csv_files = [
    "MGSD.csv",
    "Winoqueer - GPT Augmentation.csv",
    "SeeGULL - GPT Augmentation.csv",

    # MGSD Expanded is located elsewhere
    "../Exploratory Data Analysis/MGSD - Expanded.csv",
]

print("\nChecking dataset files:\n")
for f in csv_files:
    path = (DATA_DIR / f).resolve()
    print(f"{f:45} -> {'FOUND' if path.exists() else 'MISSING'}")

```

Checking dataset files:

MGSD.csv	-> FOUND
Winoqueer - GPT Augmentation.csv	-> FOUND
SeeGULL - GPT Augmentation.csv	-> FOUND
../Exploratory Data Analysis/MGSD - Expanded.csv	-> FOUND

3.1.4 Question 3 : Reproduce baseline results within $\pm 5\%$ of original paper metrics

To maintain consistency with the HEARTS baseline, all experiments were conducted using a fixed random seed across Python, NumPy, and PyTorch (both CPU and CUDA). A dedicated module `BERT_Models_Fine_Tuning2.py` was duplicated and used from the original `BERT_Models_Fine_Tuning.py`, with minimal fixes to file paths, training arguments, and compatibility issues, and a single entry-point function.

The `run_full_hearts_pipeline()` function:

1. Trains ALBERT, BERT, and DistilBERT on each of the four training settings (`mgdsd_trained`, `winoqueer_gpt_augmentation_trained`, `seegull_gpt_augmentation_trained`, `merged_winoqueer_seegull_gpt_augmentation_trained`), and
2. Evaluates each model on the four evaluation sets (`mgdsd`, `winoqueer_gpt_augmentation`, `seegull_gpt_augmentation`, `merged_winoqueer_seegull_gpt_augmentation`).

For each configuration, the accuracy, macro-precision, macro-recall, and macro-F1 scores are recorded in a CSV file and summarised in a compact results table. The accuracy and macro-F1 scores achieved in this study for all three model families and most dataset combinations are within approximately ± 5 percentage points of the metrics reported in the HEARTS paper, thereby meeting the baseline replication requirement.

```
[25]: # Set a fixed seed
SEED = 42
random.seed(SEED)
np.random.seed(SEED)
torch.manual_seed(SEED)
torch.cuda.manual_seed_all(SEED)
torch.backends.cudnn.deterministic = True
torch.backends.cudnn.benchmark = False

[ ]: # Provide the path to the python file
module_path = pathlib.Path("Model Training and Evaluation/
↳BERT_Models_Fine_Tuning2.py")

# Upload models
spec = importlib.util.spec_from_file_location("hearts_models", module_path)
hearts_models = importlib.util.module_from_spec(spec)
spec.loader.exec_module(hearts_models)

# Run
hearts_models.run_full_hearts_pipeline()
```

Loading: /tmp/HEARTS-Text-Stereotype-Detection/Model Training and Evaluation/MGSD.csv

First few examples from the training data:

	group	text	label	\
38518	race	Syria is a well known area nowadays. There is ...	1	
12979	race	The people of germany are angry.	1	
35494	race	Lebanon is a small country in the Middle East ...	1	

42213	race	We have a french tutor. She is very humble and...	0
16293	gender	Every male acts dumb.	1

	data_name
38518	MGSD
12979	MGSD
35494	MGSD
42213	MGSD
16293	MGSD

First few examples from the testing data:

	group	text	label	\
17851	race	When i see an arab, i think they are a happy.	0	
37449	race	People in India dress differently. The types o...	0	
23634	gender	Everyone has a grandmother. The bridge is long.	0	
36047	race	Cameroon is a country. It is a popular tourist...	0	
28528	profession	The delivery man drove a truck. He was highly ...	0	

	data_name
17851	MGSD
37449	MGSD
23634	MGSD
36047	MGSD
28528	MGSD

Train data size: 41493

Test data size: 10374

Loading: /tmp/HEARTS-Text-Stereotype-Detection/Model Training and
Evaluation/Winoqueer - GPT Augmentation.csv

First few examples from the training data:

	text	label	group	\
68	Gay people are survivors.	1	Gay	
2269	Transgender people are doors.	0	Transgender	
3208	most people who are Pansexual are undermined i...	0	Pansexual	
826	most lesbians are sinful.	1	Lesbian	
2452	All LGBTQ people are unfit for spaceship travel.	0	LGBTQ	

	data_name
68	Winoqueer - GPT Augmentation
2269	Winoqueer - GPT Augmentation
3208	Winoqueer - GPT Augmentation
826	Winoqueer - GPT Augmentation
2452	Winoqueer - GPT Augmentation

First few examples from the testing data:

	text	label	group	\
3236	most people who are Gay are kites	0	Gay	
2749	we expect that LGBTQ people are paints.	0	LGBTQ	
2223	LGBTQ people are televisions.	0	LGBTQ	
2724	we expect that LGBTQ people are unfit for moon...	0	LGBTQ	
1506	all Transgender people face mislabeling but ar...	0	Transgender	

data_name
 3236 Winoqueer - GPT Augmentation
 2749 Winoqueer - GPT Augmentation
 2223 Winoqueer - GPT Augmentation
 2724 Winoqueer - GPT Augmentation
 1506 Winoqueer - GPT Augmentation

Train data size: 2611

Test data size: 653

Loading: /tmp/HEARTS-Text-Stereotype-Detection/Model Training and
 Evaluation/SeeGULL - GPT Augmentation.csv

First few examples from the training data:

	text	label	group \
1206	The Russian cautious approach to learning affe...	0	Russian
213	The United States tourist was rude to the locals	1	United States
1165	Ethiopian challenges were a symptom of economi...	0	Ethiopian
76	The Vietnamese factory workers were subjected ...	1	Vietnamese
707	The Mexican leader managed a large organization	0	Mexican

data_name
 1206 SeeGULL - GPT Augmentation
 213 SeeGULL - GPT Augmentation
 1165 SeeGULL - GPT Augmentation
 76 SeeGULL - GPT Augmentation
 707 SeeGULL - GPT Augmentation

First few examples from the testing data:

	text	label	group \
840	The Somali refugees were labeled as unacquaint...	0	Somalis
1956	The Swiss man was considered spaghetti by his ...	0	Swiss
348	Barundians are often criticized for their unre...	1	Barundi
1735	The Afghans stars had to beg for music	0	Afghans
1913	The Mexican popcorn popped energetically throu...	0	Mexican

data_name
 840 SeeGULL - GPT Augmentation
 1956 SeeGULL - GPT Augmentation
 348 SeeGULL - GPT Augmentation
 1735 SeeGULL - GPT Augmentation
 1913 SeeGULL - GPT Augmentation

Train data size: 1656

Test data size: 414

First few examples from merged training data:

	text	label	group \
0	Gay people are survivors.	1	Gay
1	Transgender people are doors.	0	Transgender
2	most people who are Pansexual are undermined i...	0	Pansexual
3	most lesbians are sinful.	1	Lesbian
4	All LGBTQ people are unfit for spaceship travel.	0	LGBTQ

```

data_name
0 Winoqueer - GPT Augmentation
1 Winoqueer - GPT Augmentation
2 Winoqueer - GPT Augmentation
3 Winoqueer - GPT Augmentation
4 Winoqueer - GPT Augmentation

```

First few examples from merged testing data:

	text	label	group \
0	most people who are Gay are kites	0	Gay
1	we expect that LGBTQ people are paints.	0	LGBTQ
2	LGBTQ people are televisions.	0	LGBTQ
3	we expect that LGBTQ people are unfit for moon...	0	LGBTQ
4	all Transgender people face mislabeling but ar...	0	Transgender

```

data_name
0 Winoqueer - GPT Augmentation
1 Winoqueer - GPT Augmentation
2 Winoqueer - GPT Augmentation
3 Winoqueer - GPT Augmentation
4 Winoqueer - GPT Augmentation

```

Train data merged size: 44104

Test data merged size: 11027

First few examples from merged training data:

	text	label	group \
0	The Russian cautious approach to learning affe...	0	Russian
1	The United States tourist was rude to the locals	1	United States
2	Ethiopian challenges were a symptom of economi...	0	Ethiopian
3	The Vietnamese factory workers were subjected ...	1	Vietnamese
4	The Mexican leader managed a large organization	0	Mexican

```

data_name
0 SeeGULL - GPT Augmentation
1 SeeGULL - GPT Augmentation
2 SeeGULL - GPT Augmentation
3 SeeGULL - GPT Augmentation
4 SeeGULL - GPT Augmentation

```

First few examples from merged testing data:

	text	label	group \
0	The Somali refugees were labeled as unacquaint...	0	Somalis
1	The Swiss man was considered spaghetti by his ...	0	Swiss
2	Barundians are often criticized for their unre...	1	Barundi
3	The Afghans stars had to beg for music	0	Afghans
4	The Mexican popcorn popped energetically throu...	0	Mexican

```

data_name
0 SeeGULL - GPT Augmentation
1 SeeGULL - GPT Augmentation

```

2 SeeGULL - GPT Augmentation
 3 SeeGULL - GPT Augmentation
 4 SeeGULL - GPT Augmentation

Train data merged size: 43149

Test data merged size: 10788

First few examples from merged training data:

	text	label	group \
0	The Russian cautious approach to learning affe...	0	Russian
1	The United States tourist was rude to the locals	1	United States
2	Ethiopian challenges were a symptom of economi...	0	Ethiopian
3	The Vietnamese factory workers were subjected ...	1	Vietnamese
4	The Mexican leader managed a large organization	0	Mexican

data_name
 0 SeeGULL - GPT Augmentation
 1 SeeGULL - GPT Augmentation
 2 SeeGULL - GPT Augmentation
 3 SeeGULL - GPT Augmentation
 4 SeeGULL - GPT Augmentation

First few examples from merged testing data:

	text	label	group \
0	The Somali refugees were labeled as unacquaint...	0	Somalis
1	The Swiss man was considered spaghetti by his ...	0	Swiss
2	Barundians are often criticized for their unre...	1	Barundi
3	The Afghans stars had to beg for music	0	Afghans
4	The Mexican popcorn popped energetically throu...	0	Mexican

data_name
 0 SeeGULL - GPT Augmentation
 1 SeeGULL - GPT Augmentation
 2 SeeGULL - GPT Augmentation
 3 SeeGULL - GPT Augmentation
 4 SeeGULL - GPT Augmentation

Train data merged size: 45760

Test data merged size: 11441

Number of unique labels: 2

loading configuration file config.json from cache at

/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--
 albert-base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/config.json

Model config AlbertConfig {

```

  "architectures": [
    "AlbertForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0,
  "bos_token_id": 2,
  "classifier_dropout_prob": 0.1,
  "down_scale_factor": 1,

```

```

"embedding_size": 128,
"eos_token_id": 3,
"gap_size": 0,
"hidden_act": "gelu_new",
"hidden_dropout_prob": 0,
"hidden_size": 768,
"initializer_range": 0.02,
"inner_group_num": 1,
"intermediate_size": 3072,
"layer_norm_eps": 1e-12,
"max_position_embeddings": 512,
"model_type": "albert",
"net_structure_type": 0,
"num_attention_heads": 12,
"num_hidden_groups": 1,
"num_hidden_layers": 12,
"num_memory_blocks": 0,
"pad_token_id": 0,
"position_embedding_type": "absolute",
"transformers_version": "4.57.3",
"type_vocab_size": 2,
"vocab_size": 30000
}

```

loading weights file model.safetensors from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--albert-base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/model.safetensors

Some weights of the model checkpoint at albert/albert-base-v2 were not used when initializing AlbertForSequenceClassification: ['predictions.LayerNorm.bias', 'predictions.LayerNorm.weight', 'predictions.bias', 'predictions.decoder.bias', 'predictions.dense.bias', 'predictions.dense.weight']

- This IS expected if you are initializing AlbertForSequenceClassification from the checkpoint of a model trained on another task or with another architecture (e.g. initializing a BertForSequenceClassification model from a BertForPreTraining model).
- This IS NOT expected if you are initializing AlbertForSequenceClassification from the checkpoint of a model that you expect to be exactly identical (initializing a BertForSequenceClassification model from a BertForSequenceClassification model).

Some weights of AlbertForSequenceClassification were not initialized from the model checkpoint at albert/albert-base-v2 and are newly initialized: ['classifier.bias', 'classifier.weight']

You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.

loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--albert-base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/config.json

```

Model config AlbertConfig {
  "architectures": [
    "AlbertForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0,
  "bos_token_id": 2,
  "classifier_dropout_prob": 0.1,
  "down_scale_factor": 1,
  "embedding_size": 128,
  "eos_token_id": 3,
  "gap_size": 0,
  "hidden_act": "gelu_new",
  "hidden_dropout_prob": 0,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "inner_group_num": 1,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "albert",
  "net_structure_type": 0,
  "num_attention_heads": 12,
  "num_hidden_groups": 1,
  "num_hidden_layers": 12,
  "num_memory_blocks": 0,
  "pad_token_id": 0,
  "position_embedding_type": "absolute",
  "transformers_version": "4.57.3",
  "type_vocab_size": 2,
  "vocab_size": 30000
}

```

```

loading file spiece.model from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--
albert-base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/spiece.model
loading file tokenizer.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--
albert-base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/tokenizer.json
loading file added_tokens.json from cache at None
loading file special_tokens_map.json from cache at None
loading file tokenizer_config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--
albert-
base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/tokenizer_config.json
loading file chat_template.jinja from cache at None
loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--
albert-base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/config.json

```

```

Model config AlbertConfig {
  "architectures": [
    "AlbertForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0,
  "bos_token_id": 2,
  "classifier_dropout_prob": 0.1,
  "down_scale_factor": 1,
  "embedding_size": 128,
  "eos_token_id": 3,
  "gap_size": 0,
  "hidden_act": "gelu_new",
  "hidden_dropout_prob": 0,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "inner_group_num": 1,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "albert",
  "net_structure_type": 0,
  "num_attention_heads": 12,
  "num_hidden_groups": 1,
  "num_hidden_layers": 12,
  "num_memory_blocks": 0,
  "pad_token_id": 0,
  "position_embedding_type": "absolute",
  "transformers_version": "4.57.3",
  "type_vocab_size": 2,
  "vocab_size": 30000
}

```

```

Map: 100%|          | 33194/33194 [00:01<00:00, 24264.01 examples/s]

```

```

Map: 100%|          | 33194/33194 [00:01<00:00, 30903.07 examples/s]

```

```

Sample tokenized input from train: {'group': 'religion', 'text': 'Have you ever
met someone named Mohammed? I need a new pair of pajamas.', 'label': 0,
'data_name': 'MGSD', '__index_level_0__': 36953, 'input_ids': [2, 57, 42, 462,
798, 737, 377, 13479, 60, 31, 376, 21, 78, 2146, 16, 23358, 18, 9, 3, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0], 'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0],
'labels': 0}

```

```

Map: 100%|          | 8299/8299 [00:00<00:00, 25066.94 examples/s]

```

```

Map: 100%|          | 8299/8299 [00:00<00:00, 25612.81 examples/s]

```


The following columns in the Evaluation set don't have a corresponding argument in `AlbertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `AlbertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 8299

Batch size = 64

Saving model checkpoint to model_output_albertv2/mgsd_trained/checkpoint-1038

Configuration saved in

model_output_albertv2/mgsd_trained/checkpoint-1038/config.json

Model weights saved in

model_output_albertv2/mgsd_trained/checkpoint-1038/model.safetensors

tokenizer config file saved in

model_output_albertv2/mgsd_trained/checkpoint-1038/tokenizer_config.json

Special tokens file saved in

model_output_albertv2/mgsd_trained/checkpoint-1038/special_tokens_map.json

Deleting older checkpoint [model_output_albertv2/mgsd_trained/checkpoint-519]

due to args.save_total_limit

The following columns in the Evaluation set don't have a corresponding argument in `AlbertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `AlbertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 8299

Batch size = 64

Saving model checkpoint to model_output_albertv2/mgsd_trained/checkpoint-1557

Configuration saved in

model_output_albertv2/mgsd_trained/checkpoint-1557/config.json

Model weights saved in

model_output_albertv2/mgsd_trained/checkpoint-1557/model.safetensors

tokenizer config file saved in

model_output_albertv2/mgsd_trained/checkpoint-1557/tokenizer_config.json

Special tokens file saved in

model_output_albertv2/mgsd_trained/checkpoint-1557/special_tokens_map.json

The following columns in the Evaluation set don't have a corresponding argument in `AlbertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `AlbertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 8299

Batch size = 64

Saving model checkpoint to model_output_albertv2/mgsd_trained/checkpoint-2076

Configuration saved in
model_output_albertv2/mgsd_trained/checkpoint-2076/config.json
Model weights saved in
model_output_albertv2/mgsd_trained/checkpoint-2076/model.safetensors
tokenizer config file saved in
model_output_albertv2/mgsd_trained/checkpoint-2076/tokenizer_config.json
Special tokens file saved in
model_output_albertv2/mgsd_trained/checkpoint-2076/special_tokens_map.json
Deleting older checkpoint [model_output_albertv2/mgsd_trained/checkpoint-1557]
due to args.save_total_limit
The following columns in the Evaluation set don't have a corresponding argument
in `AlbertForSequenceClassification.forward` and have been ignored: group,
__index_level_0__, data_name, text. If group, __index_level_0__, data_name, text
are not expected by `AlbertForSequenceClassification.forward`, you can safely
ignore this message.

***** Running Evaluation *****

Num examples = 8299

Batch size = 64

Saving model checkpoint to model_output_albertv2/mgsd_trained/checkpoint-2595

Configuration saved in

model_output_albertv2/mgsd_trained/checkpoint-2595/config.json

Model weights saved in

model_output_albertv2/mgsd_trained/checkpoint-2595/model.safetensors

tokenizer config file saved in

model_output_albertv2/mgsd_trained/checkpoint-2595/tokenizer_config.json

Special tokens file saved in

model_output_albertv2/mgsd_trained/checkpoint-2595/special_tokens_map.json

Deleting older checkpoint [model_output_albertv2/mgsd_trained/checkpoint-2076]

due to args.save_total_limit

The following columns in the Evaluation set don't have a corresponding argument

in `AlbertForSequenceClassification.forward` and have been ignored: group,

__index_level_0__, data_name, text. If group, __index_level_0__, data_name, text

are not expected by `AlbertForSequenceClassification.forward`, you can safely

ignore this message.

***** Running Evaluation *****

Num examples = 8299

Batch size = 64

Saving model checkpoint to model_output_albertv2/mgsd_trained/checkpoint-3114

Configuration saved in

model_output_albertv2/mgsd_trained/checkpoint-3114/config.json

Model weights saved in

model_output_albertv2/mgsd_trained/checkpoint-3114/model.safetensors

tokenizer config file saved in

model_output_albertv2/mgsd_trained/checkpoint-3114/tokenizer_config.json

Special tokens file saved in

model_output_albertv2/mgsd_trained/checkpoint-3114/special_tokens_map.json

Deleting older checkpoint [model_output_albertv2/mgsd_trained/checkpoint-2595]
due to args.save_total_limit

Training completed. Do not forget to share your model on huggingface.co/models
=)

Loading best model from model_output_albertv2/mgsd_trained/checkpoint-1038
(score: 0.39199861884117126).
Deleting older checkpoint [model_output_albertv2/mgsd_trained/checkpoint-3114]
due to args.save_total_limit
Saving model checkpoint to model_output_albertv2/mgsd_trained
Configuration saved in model_output_albertv2/mgsd_trained/config.json
Model weights saved in model_output_albertv2/mgsd_trained/model.safetensors
tokenizer config file saved in
model_output_albertv2/mgsd_trained/tokenizer_config.json
Special tokens file saved in
model_output_albertv2/mgsd_trained/special_tokens_map.json
loading configuration file model_output_albertv2/mgsd_trained/config.json
Model config AlbertConfig {
 "architectures": [
 "AlbertForSequenceClassification"
],
 "attention_probs_dropout_prob": 0,
 "bos_token_id": 2,
 "classifier_dropout_prob": 0.1,
 "down_scale_factor": 1,
 "dtype": "float32",
 "embedding_size": 128,
 "eos_token_id": 3,
 "gap_size": 0,
 "hidden_act": "gelu_new",
 "hidden_dropout_prob": 0,
 "hidden_size": 768,
 "initializer_range": 0.02,
 "inner_group_num": 1,
 "intermediate_size": 3072,
 "layer_norm_eps": 1e-12,
 "max_position_embeddings": 512,
 "model_type": "albert",
 "net_structure_type": 0,
 "num_attention_heads": 12,
 "num_hidden_groups": 1,
 "num_hidden_layers": 12,
 "num_memory_blocks": 0,
 "pad_token_id": 0,
 "position_embedding_type": "absolute",


```

    "hidden_size": 768,
    "initializer_range": 0.02,
    "inner_group_num": 1,
    "intermediate_size": 3072,
    "layer_norm_eps": 1e-12,
    "max_position_embeddings": 512,
    "model_type": "albert",
    "net_structure_type": 0,
    "num_attention_heads": 12,
    "num_hidden_groups": 1,
    "num_hidden_layers": 12,
    "num_memory_blocks": 0,
    "pad_token_id": 0,
    "position_embedding_type": "absolute",
    "problem_type": "single_label_classification",
    "transformers_version": "4.57.3",
    "type_vocab_size": 2,
    "vocab_size": 30000
}

loading weights file model_output_albertv2/mgsd_trained/model.safetensors
loading file spiece.model
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

Saved full results to:
result_output_albertv2/mgsd_trained/winoqueer_gpt_augmentation/full_results.csv
Saved classification report to: result_output_albertv2/mgsd_trained/winoqueer_gpt
augmentation/classification_report.csv
Number of unique labels: 2

Map: 100%|          | 414/414 [00:00<00:00, 26623.15 examples/s]
Map: 100%|          | 414/414 [00:00<00:00, 30635.88 examples/s]
Device set to use cuda:0

Sample tokenized input from test: {'text': 'The Somali refugees were labeled as
unacquainted by some', 'label': 0, 'group': 'Somalis', 'data_name': 'SeeGULL -
GPT Augmentation', '__index_level_0__': 840, 'input_ids': [2, 14, 16330, 8790,
46, 14348, 28, 367, 26692, 108, 1427, 34, 109, 3, 0, 0, 0, 0, 0, 0],
'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0],
'labels': 0}

/tmp/hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of

```

```

`return_all_scores=False`.
warnings.warn(
loading configuration file model_output_albertv2/mgsd_trained/config.json
Model config AlbertConfig {
  "architectures": [
    "AlbertForSequenceClassification"
  ],
  "attention_probs_dropout_prob": 0,
  "bos_token_id": 2,
  "classifier_dropout_prob": 0.1,
  "down_scale_factor": 1,
  "dtype": "float32",
  "embedding_size": 128,
  "eos_token_id": 3,
  "gap_size": 0,
  "hidden_act": "gelu_new",
  "hidden_dropout_prob": 0,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "inner_group_num": 1,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "albert",
  "net_structure_type": 0,
  "num_attention_heads": 12,
  "num_hidden_groups": 1,
  "num_hidden_layers": 12,
  "num_memory_blocks": 0,
  "pad_token_id": 0,
  "position_embedding_type": "absolute",
  "problem_type": "single_label_classification",
  "transformers_version": "4.57.3",
  "type_vocab_size": 2,
  "vocab_size": 30000
}

loading weights file model_output_albertv2/mgsd_trained/model.safetensors
loading file spiece.model
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

Saved full results to:
result_output_albertv2/mgsd_trained/seegull_gpt_augmentation/full_results.csv
Saved classification report to: result_output_albertv2/mgsd_trained/seegull_gpt_

```

```

augmentation/classification_report.csv
Number of unique labels: 2

Map: 100%|          | 10374/10374 [00:00<00:00, 25262.18 examples/s]
Map: 100%|          | 10374/10374 [00:00<00:00, 27014.61 examples/s]
Device set to use cuda:0
/tmp/hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.
  warnings.warn(

Sample tokenized input from test: {'group': 'race', 'text': 'When i see an arab,
i think they are a happy.', 'label': 0, 'data_name': 'MGSD',
'__index_level_0__': 17851, 'input_ids': [2, 76, 31, 196, 40, 3666, 15, 31, 277,
59, 50, 21, 1700, 9, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0], 'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0], 'labels': 0}

loading configuration file model_output_albertv2/mgsd_trained/config.json
Model config AlbertConfig {
  "architectures": [
    "AlbertForSequenceClassification"
  ],
  "attention_probs_dropout_prob": 0,
  "bos_token_id": 2,
  "classifier_dropout_prob": 0.1,
  "down_scale_factor": 1,
  "dtype": "float32",
  "embedding_size": 128,
  "eos_token_id": 3,
  "gap_size": 0,
  "hidden_act": "gelu_new",
  "hidden_dropout_prob": 0,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "inner_group_num": 1,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "albert",
  "net_structure_type": 0,
  "num_attention_heads": 12,
  "num_hidden_groups": 1,
  "num_hidden_layers": 12,

```

```

    "num_memory_blocks": 0,
    "pad_token_id": 0,
    "position_embedding_type": "absolute",
    "problem_type": "single_label_classification",
    "transformers_version": "4.57.3",
    "type_vocab_size": 2,
    "vocab_size": 30000
}

```

```

loading weights file model_output_albertv2/mgsd_trained/model.safetensors
loading file spiece.model
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

```

```

Saved full results to: result_output_albertv2/mgsd_trained/mgsd/full_results.csv
Saved classification report to:
result_output_albertv2/mgsd_trained/mgsd/classification_report.csv
Number of unique labels: 2

```

```

Map: 100%|          | 11441/11441 [00:00<00:00, 25754.99 examples/s]
Map: 100%|          | 11441/11441 [00:00<00:00, 29745.51 examples/s]

```

```

Device set to use cuda:0

```

```

/tmp/hearts_gpu/lib64/python3.9/site-

```

```

packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.

```

```

    warnings.warn(

```

```

Sample tokenized input from test: {'text': 'The Somali refugees were labeled as
unacquainted by some', 'label': 0, 'group': 'Somalis', 'data_name': 'SeeGULL -
GPT Augmentation', 'input_ids': [2, 14, 16330, 8790, 46, 14348, 28, 367, 26692,
108, 1427, 34, 109, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'token_type_ids':
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0], 'labels': 0}

```

```

loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--
albert-base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/config.json

```

```

Model config AlbertConfig {
  "architectures": [
    "AlbertForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0,
  "bos_token_id": 2,
  "classifier_dropout_prob": 0.1,

```

```

"down_scale_factor": 1,
"embedding_size": 128,
"eos_token_id": 3,
"gap_size": 0,
"hidden_act": "gelu_new",
"hidden_dropout_prob": 0,
"hidden_size": 768,
"initializer_range": 0.02,
"inner_group_num": 1,
"intermediate_size": 3072,
"layer_norm_eps": 1e-12,
"max_position_embeddings": 512,
"model_type": "albert",
"net_structure_type": 0,
"num_attention_heads": 12,
"num_hidden_groups": 1,
"num_hidden_layers": 12,
"num_memory_blocks": 0,
"pad_token_id": 0,
"position_embedding_type": "absolute",
"transformers_version": "4.57.3",
"type_vocab_size": 2,
"vocab_size": 30000
}

```

loading weights file model.safetensors from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--
albert-
base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/model.safetensors

Saved full results to: result_output_albertv2/mgsd_trained/merged_winoqueer_seeg
ull_gpt_augmentation/full_results.csv

Saved classification report to: result_output_albertv2/mgsd_trained/merged_winoq
ueer_seegull_gpt_augmentation/classification_report.csv

Number of unique labels: 2

Some weights of the model checkpoint at albert/albert-base-v2 were not used when
initializing AlbertForSequenceClassification: ['predictions.LayerNorm.bias',
'predictions.LayerNorm.weight', 'predictions.bias', 'predictions.decoder.bias',
'predictions.dense.bias', 'predictions.dense.weight']

- This IS expected if you are initializing AlbertForSequenceClassification from
the checkpoint of a model trained on another task or with another architecture
(e.g. initializing a BertForSequenceClassification model from a
BertForPreTraining model).

- This IS NOT expected if you are initializing AlbertForSequenceClassification
from the checkpoint of a model that you expect to be exactly identical
(initializing a BertForSequenceClassification model from a
BertForSequenceClassification model).

Some weights of AlbertForSequenceClassification were not initialized from the

model checkpoint at albert/albert-base-v2 and are newly initialized:
['classifier.bias', 'classifier.weight']
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.

loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--albert-base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/config.json

```
Model config AlbertConfig {
  "architectures": [
    "AlbertForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0,
  "bos_token_id": 2,
  "classifier_dropout_prob": 0.1,
  "down_scale_factor": 1,
  "embedding_size": 128,
  "eos_token_id": 3,
  "gap_size": 0,
  "hidden_act": "gelu_new",
  "hidden_dropout_prob": 0,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "inner_group_num": 1,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "albert",
  "net_structure_type": 0,
  "num_attention_heads": 12,
  "num_hidden_groups": 1,
  "num_hidden_layers": 12,
  "num_memory_blocks": 0,
  "pad_token_id": 0,
  "position_embedding_type": "absolute",
  "transformers_version": "4.57.3",
  "type_vocab_size": 2,
  "vocab_size": 30000
}
```

loading file spiece.model from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--albert-base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/spiece.model
loading file tokenizer.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--albert-base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/tokenizer.json
loading file added_tokens.json from cache at None
loading file special_tokens_map.json from cache at None
loading file tokenizer_config.json from cache at

```

/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--
albert-
base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/tokenizer_config.json
loading file chat_template.jinja from cache at None
loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--
albert-base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/config.json
Model config AlbertConfig {
  "architectures": [
    "AlbertForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0,
  "bos_token_id": 2,
  "classifier_dropout_prob": 0.1,
  "down_scale_factor": 1,
  "embedding_size": 128,
  "eos_token_id": 3,
  "gap_size": 0,
  "hidden_act": "gelu_new",
  "hidden_dropout_prob": 0,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "inner_group_num": 1,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "albert",
  "net_structure_type": 0,
  "num_attention_heads": 12,
  "num_hidden_groups": 1,
  "num_hidden_layers": 12,
  "num_memory_blocks": 0,
  "pad_token_id": 0,
  "position_embedding_type": "absolute",
  "transformers_version": "4.57.3",
  "type_vocab_size": 2,
  "vocab_size": 30000
}

```

```

Map: 100%|          | 2088/2088 [00:00<00:00, 39269.76 examples/s]
Map: 100%|          | 2088/2088 [00:00<00:00, 35202.47 examples/s]

```

```

Sample tokenized input from train: {'text': 'we expect that nonbinary people are
out to share their views.', 'label': 0, 'group': 'NB', 'data_name': 'Winoqueer -
GPT Augmentation', '__index_level_0__': 1788, 'input_ids': [2, 95, 4186, 30,
538, 4772, 1857, 148, 50, 70, 20, 1891, 66, 4146, 9, 3, 0, 0, 0, 0, 0, 0, 0,
0, 0], 'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,

```

```

1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'labels': 0}
Map: 100%|          | 523/523 [00:00<00:00, 33515.47 examples/s]
Map: 100%|          | 523/523 [00:00<00:00, 31945.78 examples/s]
PyTorch: setting up devices
Model Training and Evaluation/BERT_Models_Fine_Tuning2.py:209: FutureWarning:
`tokenizer` is deprecated and will be removed in version 5.0.0 for
`Trainer.__init__`. Use `processing_class` instead.
    trainer = Trainer(

Sample tokenized input from validation: {'text': 'all lesbians are studios.',
'label': 0, 'group': 'Lesbian', 'data_name': 'Winoqueer - GPT Augmentation',
'__index_level_0__': 2716, 'input_ids': [2, 65, 10564, 18, 50, 1120, 267, 9, 3,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'labels': 0}

The following columns in the Training set don't have a corresponding argument in
`AlbertForSequenceClassification.forward` and have been ignored: group,
__index_level_0__, data_name, text. If group, __index_level_0__, data_name, text
are not expected by `AlbertForSequenceClassification.forward`, you can safely
ignore this message.
***** Running training *****
    Num examples = 2,088
    Num Epochs = 6
    Instantaneous batch size per device = 64
    Total train batch size (w. parallel, distributed & accumulation) = 64
    Gradient Accumulation steps = 1
    Total optimization steps = 198
    Number of trainable parameters = 11,685,122

<IPython.core.display.HTML object>

The following columns in the Evaluation set don't have a corresponding argument
in `AlbertForSequenceClassification.forward` and have been ignored: group,
__index_level_0__, data_name, text. If group, __index_level_0__, data_name, text
are not expected by `AlbertForSequenceClassification.forward`, you can safely
ignore this message.

***** Running Evaluation *****
    Num examples = 523
    Batch size = 64
Saving model checkpoint to
model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-33
Configuration saved in model_output_albertv2/winoqueer_gpt_augmentation_trained/
checkpoint-33/config.json
Model weights saved in model_output_albertv2/winoqueer_gpt_augmentation_trained/
checkpoint-33/model.safetensors
tokenizer config file saved in model_output_albertv2/winoqueer_gpt_augmentation_

```

trained/checkpoint-33/tokenizer_config.json
Special tokens file saved in model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-33/special_tokens_map.json
The following columns in the Evaluation set don't have a corresponding argument in `AlbertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `AlbertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 523

Batch size = 64

Saving model checkpoint to

model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-66

Configuration saved in model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-66/config.json

Model weights saved in model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-66/model.safetensors

tokenizer config file saved in model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-66/tokenizer_config.json

Special tokens file saved in model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-66/special_tokens_map.json

Deleting older checkpoint

[model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-33] due to args.save_total_limit

The following columns in the Evaluation set don't have a corresponding argument in `AlbertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `AlbertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 523

Batch size = 64

Saving model checkpoint to

model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-99

Configuration saved in model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-99/config.json

Model weights saved in model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-99/model.safetensors

tokenizer config file saved in model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-99/tokenizer_config.json

Special tokens file saved in model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-99/special_tokens_map.json

Deleting older checkpoint

[model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-66] due to args.save_total_limit

The following columns in the Evaluation set don't have a corresponding argument

in `AlbertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `AlbertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 523

Batch size = 64

Saving model checkpoint to

model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-132

Configuration saved in model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-132/config.json

Model weights saved in model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-132/model.safetensors

tokenizer config file saved in model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-132/tokenizer_config.json

Special tokens file saved in model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-132/special_tokens_map.json

Deleting older checkpoint

[model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-99] due to args.save_total_limit

The following columns in the Evaluation set don't have a corresponding argument in `AlbertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `AlbertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 523

Batch size = 64

Saving model checkpoint to

model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-165

Configuration saved in model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-165/config.json

Model weights saved in model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-165/model.safetensors

tokenizer config file saved in model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-165/tokenizer_config.json

Special tokens file saved in model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-165/special_tokens_map.json

The following columns in the Evaluation set don't have a corresponding argument in `AlbertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `AlbertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 523

```

Batch size = 64
Saving model checkpoint to
model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-198
Configuration saved in model_output_albertv2/winoqueer_gpt_augmentation_trained/
checkpoint-198/config.json
Model weights saved in model_output_albertv2/winoqueer_gpt_augmentation_trained/
checkpoint-198/model.safetensors
tokenizer config file saved in model_output_albertv2/winoqueer_gpt_augmentation_
trained/checkpoint-198/tokenizer_config.json
Special tokens file saved in model_output_albertv2/winoqueer_gpt_augmentation_tr
ained/checkpoint-198/special_tokens_map.json
Deleting older checkpoint
[model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-165] due to
args.save_total_limit

```

Training completed. Do not forget to share your model on huggingface.co/models
=)

```

Loading best model from
model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-132 (score:
0.05476154759526253).
Deleting older checkpoint
[model_output_albertv2/winoqueer_gpt_augmentation_trained/checkpoint-198] due to
args.save_total_limit
Saving model checkpoint to
model_output_albertv2/winoqueer_gpt_augmentation_trained
Configuration saved in
model_output_albertv2/winoqueer_gpt_augmentation_trained/config.json
Model weights saved in
model_output_albertv2/winoqueer_gpt_augmentation_trained/model.safetensors
tokenizer config file saved in
model_output_albertv2/winoqueer_gpt_augmentation_trained/tokenizer_config.json
Special tokens file saved in
model_output_albertv2/winoqueer_gpt_augmentation_trained/special_tokens_map.json
loading configuration file
model_output_albertv2/winoqueer_gpt_augmentation_trained/config.json
Model config AlbertConfig {
  "architectures": [
    "AlbertForSequenceClassification"
  ],
  "attention_probs_dropout_prob": 0,
  "bos_token_id": 2,
  "classifier_dropout_prob": 0.1,
  "down_scale_factor": 1,
  "dtype": "float32",
  "embedding_size": 128,

```

```

    "eos_token_id": 3,
    "gap_size": 0,
    "hidden_act": "gelu_new",
    "hidden_dropout_prob": 0,
    "hidden_size": 768,
    "initializer_range": 0.02,
    "inner_group_num": 1,
    "intermediate_size": 3072,
    "layer_norm_eps": 1e-12,
    "max_position_embeddings": 512,
    "model_type": "albert",
    "net_structure_type": 0,
    "num_attention_heads": 12,
    "num_hidden_groups": 1,
    "num_hidden_layers": 12,
    "num_memory_blocks": 0,
    "pad_token_id": 0,
    "position_embedding_type": "absolute",
    "problem_type": "single_label_classification",
    "transformers_version": "4.57.3",
    "type_vocab_size": 2,
    "vocab_size": 30000
}

loading weights file
model_output_albertv2/winoqueer_gpt_augmentation_trained/model.safetensors
loading file spiece.model
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

Saved fine-tuned model to:
model_output_albertv2/winoqueer_gpt_augmentation_trained
Number of unique labels: 2

Map: 100%|          | 653/653 [00:00<00:00, 32616.59 examples/s]
Map: 100%|          | 653/653 [00:00<00:00, 32063.69 examples/s]
Device set to use cuda:0
/tmp/hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.
  warnings.warn(

Sample tokenized input from test: {'text': 'most people who are Gay are kites',
'label': 0, 'group': 'Gay', 'data_name': 'Winoqueer - GPT Augmentation',

```

```
'__index_level_0__': 3236, 'input_ids': [2, 127, 148, 72, 50, 3398, 50, 19196,
18, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'token_type_ids': [0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0], 'labels': 0}
```

loading configuration file

model_output_albertv2/winoqueer_gpt_augmentation_trained/config.json

Model config AlbertConfig {

```
"architectures": [
  "AlbertForSequenceClassification"
],
"attention_probs_dropout_prob": 0,
"bos_token_id": 2,
"classifier_dropout_prob": 0.1,
"down_scale_factor": 1,
"dtype": "float32",
"embedding_size": 128,
"eos_token_id": 3,
"gap_size": 0,
"hidden_act": "gelu_new",
"hidden_dropout_prob": 0,
"hidden_size": 768,
"initializer_range": 0.02,
"inner_group_num": 1,
"intermediate_size": 3072,
"layer_norm_eps": 1e-12,
"max_position_embeddings": 512,
"model_type": "albert",
"net_structure_type": 0,
"num_attention_heads": 12,
"num_hidden_groups": 1,
"num_hidden_layers": 12,
"num_memory_blocks": 0,
"pad_token_id": 0,
"position_embedding_type": "absolute",
"problem_type": "single_label_classification",
"transformers_version": "4.57.3",
"type_vocab_size": 2,
"vocab_size": 30000
```

}

loading weights file

model_output_albertv2/winoqueer_gpt_augmentation_trained/model.safetensors

loading file spiece.model

loading file tokenizer.json

loading file added_tokens.json

loading file special_tokens_map.json


```

loading file tokenizer_config.json
loading file chat_template.jinja

Saved full results to: result_output_albertv2/winoqueer_gpt_augmentation_trained
/winoqueer_gpt_augmentation/full_results.csv
Saved classification report to: result_output_albertv2/winoqueer_gpt_augmentatio
n_trained/winoqueer_gpt_augmentation/classification_report.csv
Number of unique labels: 2

Map: 100%|          | 414/414 [00:00<00:00, 26737.94 examples/s]
Map: 100%|          | 414/414 [00:00<00:00, 30124.07 examples/s]
Device set to use cuda:0

Sample tokenized input from test: {'text': 'The Somali refugees were labeled as
unacquainted by some', 'label': 0, 'group': 'Somalis', 'data_name': 'SeeGULL -
GPT Augmentation', '__index_level_0__': 840, 'input_ids': [2, 14, 16330, 8790,
46, 14348, 28, 367, 26692, 108, 1427, 34, 109, 3, 0, 0, 0, 0, 0, 0],
'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0],
'labels': 0}

/tmp/hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.
  warnings.warn(
loading configuration file
model_output_albertv2/winoqueer_gpt_augmentation_trained/config.json
Model config AlbertConfig {
  "architectures": [
    "AlbertForSequenceClassification"
  ],
  "attention_probs_dropout_prob": 0,
  "bos_token_id": 2,
  "classifier_dropout_prob": 0.1,
  "down_scale_factor": 1,
  "dtype": "float32",
  "embedding_size": 128,
  "eos_token_id": 3,
  "gap_size": 0,
  "hidden_act": "gelu_new",
  "hidden_dropout_prob": 0,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "inner_group_num": 1,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "albert",

```

```

"net_structure_type": 0,
"num_attention_heads": 12,
"num_hidden_groups": 1,
"num_hidden_layers": 12,
"num_memory_blocks": 0,
"pad_token_id": 0,
"position_embedding_type": "absolute",
"problem_type": "single_label_classification",
"transformers_version": "4.57.3",
"type_vocab_size": 2,
"vocab_size": 30000
}

loading weights file
model_output_albertv2/winoqueer_gpt_augmentation_trained/model.safetensors
loading file spiece.model
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

Saved full results to: result_output_albertv2/winoqueer_gpt_augmentation_trained
/seegull_gpt_augmentation/full_results.csv
Saved classification report to: result_output_albertv2/winoqueer_gpt_augmentatio
n_trained/seegull_gpt_augmentation/classification_report.csv
Number of unique labels: 2

Map: 100%|          | 10374/10374 [00:00<00:00, 24882.03 examples/s]
Map: 100%|          | 10374/10374 [00:00<00:00, 26299.02 examples/s]
Device set to use cuda:0
/tmp/hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.
  warnings.warn(

Sample tokenized input from test: {'group': 'race', 'text': 'When i see an arab,
i think they are a happy.', 'label': 0, 'data_name': 'MGSD',
'__index_level_0__': 17851, 'input_ids': [2, 76, 31, 196, 40, 3666, 15, 31, 277,
59, 50, 21, 1700, 9, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0], 'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0], 'labels': 0}

loading configuration file
model_output_albertv2/winoqueer_gpt_augmentation_trained/config.json

```

```

Model config AlbertConfig {
  "architectures": [
    "AlbertForSequenceClassification"
  ],
  "attention_probs_dropout_prob": 0,
  "bos_token_id": 2,
  "classifier_dropout_prob": 0.1,
  "down_scale_factor": 1,
  "dtype": "float32",
  "embedding_size": 128,
  "eos_token_id": 3,
  "gap_size": 0,
  "hidden_act": "gelu_new",
  "hidden_dropout_prob": 0,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "inner_group_num": 1,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "albert",
  "net_structure_type": 0,
  "num_attention_heads": 12,
  "num_hidden_groups": 1,
  "num_hidden_layers": 12,
  "num_memory_blocks": 0,
  "pad_token_id": 0,
  "position_embedding_type": "absolute",
  "problem_type": "single_label_classification",
  "transformers_version": "4.57.3",
  "type_vocab_size": 2,
  "vocab_size": 30000
}

```

```

loading weights file
model_output_albertv2/winoqueer_gpt_augmentation_trained/model.safetensors
loading file spiece.model
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

```

```

Saved full results to:
result_output_albertv2/winoqueer_gpt_augmentation_trained/mgsd/full_results.csv
Saved classification report to: result_output_albertv2/winoqueer_gpt_augmentatio
n_trained/mgsd/classification_report.csv
Number of unique labels: 2

```

```

Map: 100%|          | 11441/11441 [00:00<00:00, 26019.70 examples/s]
Map: 100%|          | 11441/11441 [00:00<00:00, 29392.62 examples/s]
Device set to use cuda:0
/tmp/hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.
  warnings.warn(

Sample tokenized input from test: {'text': 'The Somali refugees were labeled as
unacquainted by some', 'label': 0, 'group': 'Somalis', 'data_name': 'SeeGULL -
GPT Augmentation', 'input_ids': [2, 14, 16330, 8790, 46, 14348, 28, 367, 26692,
108, 1427, 34, 109, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'token_type_ids':
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'labels': 0}

loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--
albert-base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/config.json
Model config AlbertConfig {
  "architectures": [
    "AlbertForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0,
  "bos_token_id": 2,
  "classifier_dropout_prob": 0.1,
  "down_scale_factor": 1,
  "embedding_size": 128,
  "eos_token_id": 3,
  "gap_size": 0,
  "hidden_act": "gelu_new",
  "hidden_dropout_prob": 0,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "inner_group_num": 1,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "albert",
  "net_structure_type": 0,
  "num_attention_heads": 12,
  "num_hidden_groups": 1,
  "num_hidden_layers": 12,
  "num_memory_blocks": 0,
  "pad_token_id": 0,
  "position_embedding_type": "absolute",
  "transformers_version": "4.57.3",

```

```

    "type_vocab_size": 2,
    "vocab_size": 30000
}

```

loading weights file model.safetensors from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--albert-base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/model.safetensors
Some weights of the model checkpoint at albert/albert-base-v2 were not used when initializing AlbertForSequenceClassification: ['predictions.LayerNorm.bias', 'predictions.LayerNorm.weight', 'predictions.bias', 'predictions.decoder.bias', 'predictions.dense.bias', 'predictions.dense.weight']
- This IS expected if you are initializing AlbertForSequenceClassification from the checkpoint of a model trained on another task or with another architecture (e.g. initializing a BertForSequenceClassification model from a BertForPreTraining model).
- This IS NOT expected if you are initializing AlbertForSequenceClassification from the checkpoint of a model that you expect to be exactly identical (initializing a BertForSequenceClassification model from a BertForSequenceClassification model).
Some weights of AlbertForSequenceClassification were not initialized from the model checkpoint at albert/albert-base-v2 and are newly initialized: ['classifier.bias', 'classifier.weight']
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.

Saved full results to: result_output_albertv2/winoqueer_gpt_augmentation_trained/merged_winoqueer_seegull_gpt_augmentation/full_results.csv
Saved classification report to: result_output_albertv2/winoqueer_gpt_augmentation_trained/merged_winoqueer_seegull_gpt_augmentation/classification_report.csv
Number of unique labels: 2

loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--albert-base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/config.json

```

Model config AlbertConfig {
  "architectures": [
    "AlbertForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0,
  "bos_token_id": 2,
  "classifier_dropout_prob": 0.1,
  "down_scale_factor": 1,
  "embedding_size": 128,
  "eos_token_id": 3,
  "gap_size": 0,
  "hidden_act": "gelu_new",
  "hidden_dropout_prob": 0,
  "hidden_size": 768,

```

```

"initializer_range": 0.02,
"inner_group_num": 1,
"intermediate_size": 3072,
"layer_norm_eps": 1e-12,
"max_position_embeddings": 512,
"model_type": "albert",
"net_structure_type": 0,
"num_attention_heads": 12,
"num_hidden_groups": 1,
"num_hidden_layers": 12,
"num_memory_blocks": 0,
"pad_token_id": 0,
"position_embedding_type": "absolute",
"transformers_version": "4.57.3",
"type_vocab_size": 2,
"vocab_size": 30000
}

```

```

loading file spiece.model from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--
albert-base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/spiece.model
loading file tokenizer.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--
albert-base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/tokenizer.json
loading file added_tokens.json from cache at None
loading file special_tokens_map.json from cache at None
loading file tokenizer_config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--
albert-
base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/tokenizer_config.json
loading file chat_template.jinja from cache at None
loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--
albert-base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/config.json
Model config AlbertConfig {
  "architectures": [
    "AlbertForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0,
  "bos_token_id": 2,
  "classifier_dropout_prob": 0.1,
  "down_scale_factor": 1,
  "embedding_size": 128,
  "eos_token_id": 3,
  "gap_size": 0,
  "hidden_act": "gelu_new",
  "hidden_dropout_prob": 0,
  "hidden_size": 768,

```

```

"initializer_range": 0.02,
"inner_group_num": 1,
"intermediate_size": 3072,
"layer_norm_eps": 1e-12,
"max_position_embeddings": 512,
"model_type": "albert",
"net_structure_type": 0,
"num_attention_heads": 12,
"num_hidden_groups": 1,
"num_hidden_layers": 12,
"num_memory_blocks": 0,
"pad_token_id": 0,
"position_embedding_type": "absolute",
"transformers_version": "4.57.3",
"type_vocab_size": 2,
"vocab_size": 30000
}

Map: 100%|          | 1324/1324 [00:00<00:00, 32692.57 examples/s]
Map: 100%|          | 1324/1324 [00:00<00:00, 34885.13 examples/s]

Sample tokenized input from train: {'text': 'The average Chadian income is
modest compared to developed countries', 'label': 0, 'group': 'Chadian',
'data_name': 'SeeGULL - GPT Augmentation', '__index_level_0__': 1249,
'input_ids': [2, 14, 862, 8708, 806, 1587, 25, 8904, 2428, 20, 885, 1166, 3, 0,
0, 0, 0, 0, 0, 0, 0], 'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 0, 0, 0, 0, 0, 0, 0, 0], 'labels': 0}

Map: 100%|          | 332/332 [00:00<00:00, 28725.74 examples/s]
Map: 100%|          | 332/332 [00:00<00:00, 29558.04 examples/s]
PyTorch: setting up devices
Model Training and Evaluation/BERT_Models_Fine_Tuning2.py:209: FutureWarning:
`tokenizer` is deprecated and will be removed in version 5.0.0 for
`Trainer.__init__`. Use `processing_class` instead.
    trainer = Trainer(

Sample tokenized input from validation: {'text': 'The Ethiopian woman was short
and petite', 'label': 1, 'group': 'Ethiopian', 'data_name': 'SeeGULL - GPT
Augmentation', '__index_level_0__': 609, 'input_ids': [2, 14, 16659, 524, 23,
502, 17, 22471, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'token_type_ids': [0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1,
1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'labels': 1}

The following columns in the Training set don't have a corresponding argument in
`AlbertForSequenceClassification.forward` and have been ignored: group,
__index_level_0__, data_name, text. If group, __index_level_0__, data_name, text
are not expected by `AlbertForSequenceClassification.forward`, you can safely
ignore this message.
***** Running training *****

```

```
Num examples = 1,324
Num Epochs = 6
Instantaneous batch size per device = 64
Total train batch size (w. parallel, distributed & accumulation) = 64
Gradient Accumulation steps = 1
Total optimization steps = 126
Number of trainable parameters = 11,685,122
```

<IPython.core.display.HTML object>

The following columns in the Evaluation set don't have a corresponding argument in `AlbertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `AlbertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

```
Num examples = 332
Batch size = 64
Saving model checkpoint to
model_output_albertv2/seegull_gpt_augmentation_trained/checkpoint-21
Configuration saved in
model_output_albertv2/seegull_gpt_augmentation_trained/checkpoint-21/config.json
Model weights saved in model_output_albertv2/seegull_gpt_augmentation_trained/ch
eckpoint-21/model.safetensors
tokenizer config file saved in model_output_albertv2/seegull_gpt_augmentation_tr
ained/checkpoint-21/tokenizer_config.json
Special tokens file saved in model_output_albertv2/seegull_gpt_augmentation_trai
ned/checkpoint-21/special_tokens_map.json
The following columns in the Evaluation set don't have a corresponding argument
in `AlbertForSequenceClassification.forward` and have been ignored: group,
__index_level_0__, data_name, text. If group, __index_level_0__, data_name, text
are not expected by `AlbertForSequenceClassification.forward`, you can safely
ignore this message.
```

***** Running Evaluation *****

```
Num examples = 332
Batch size = 64
Saving model checkpoint to
model_output_albertv2/seegull_gpt_augmentation_trained/checkpoint-42
Configuration saved in
model_output_albertv2/seegull_gpt_augmentation_trained/checkpoint-42/config.json
Model weights saved in model_output_albertv2/seegull_gpt_augmentation_trained/ch
eckpoint-42/model.safetensors
tokenizer config file saved in model_output_albertv2/seegull_gpt_augmentation_tr
ained/checkpoint-42/tokenizer_config.json
Special tokens file saved in model_output_albertv2/seegull_gpt_augmentation_trai
ned/checkpoint-42/special_tokens_map.json
Deleting older checkpoint
```


[model_output_albertv2/seegull_gpt_augmentation_trained/checkpoint-21] due to args.save_total_limit
The following columns in the Evaluation set don't have a corresponding argument in `AlbertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `AlbertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 332

Batch size = 64

Saving model checkpoint to

model_output_albertv2/seegull_gpt_augmentation_trained/checkpoint-63

Configuration saved in

model_output_albertv2/seegull_gpt_augmentation_trained/checkpoint-63/config.json

Model weights saved in model_output_albertv2/seegull_gpt_augmentation_trained/checkpoint-63/model.safetensors

tokenizer config file saved in model_output_albertv2/seegull_gpt_augmentation_trained/checkpoint-63/tokenizer_config.json

Special tokens file saved in model_output_albertv2/seegull_gpt_augmentation_trained/checkpoint-63/special_tokens_map.json

The following columns in the Evaluation set don't have a corresponding argument in `AlbertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `AlbertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 332

Batch size = 64

Saving model checkpoint to

model_output_albertv2/seegull_gpt_augmentation_trained/checkpoint-84

Configuration saved in

model_output_albertv2/seegull_gpt_augmentation_trained/checkpoint-84/config.json

Model weights saved in model_output_albertv2/seegull_gpt_augmentation_trained/checkpoint-84/model.safetensors

tokenizer config file saved in model_output_albertv2/seegull_gpt_augmentation_trained/checkpoint-84/tokenizer_config.json

Special tokens file saved in model_output_albertv2/seegull_gpt_augmentation_trained/checkpoint-84/special_tokens_map.json

Deleting older checkpoint

[model_output_albertv2/seegull_gpt_augmentation_trained/checkpoint-63] due to args.save_total_limit

The following columns in the Evaluation set don't have a corresponding argument in `AlbertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `AlbertForSequenceClassification.forward`, you can safely ignore this message.

```

***** Running Evaluation *****
  Num examples = 332
  Batch size = 64
Saving model checkpoint to
model_output_albertv2/seegull_gpt_augmentation_trained/checkpoint-105
Configuration saved in model_output_albertv2/seegull_gpt_augmentation_trained/ch
eckpoint-105/config.json
Model weights saved in model_output_albertv2/seegull_gpt_augmentation_trained/ch
eckpoint-105/model.safetensors
tokenizer config file saved in model_output_albertv2/seegull_gpt_augmentation_tr
ained/checkpoint-105/tokenizer_config.json
Special tokens file saved in model_output_albertv2/seegull_gpt_augmentation_trai
ned/checkpoint-105/special_tokens_map.json
Deleting older checkpoint
[model_output_albertv2/seegull_gpt_augmentation_trained/checkpoint-42] due to
args.save_total_limit
Deleting older checkpoint
[model_output_albertv2/seegull_gpt_augmentation_trained/checkpoint-84] due to
args.save_total_limit
The following columns in the Evaluation set don't have a corresponding argument
in `AlbertForSequenceClassification.forward` and have been ignored: group,
__index_level_0__, data_name, text. If group, __index_level_0__, data_name, text
are not expected by `AlbertForSequenceClassification.forward`, you can safely
ignore this message.

***** Running Evaluation *****
  Num examples = 332
  Batch size = 64
Saving model checkpoint to
model_output_albertv2/seegull_gpt_augmentation_trained/checkpoint-126
Configuration saved in model_output_albertv2/seegull_gpt_augmentation_trained/ch
eckpoint-126/config.json
Model weights saved in model_output_albertv2/seegull_gpt_augmentation_trained/ch
eckpoint-126/model.safetensors
tokenizer config file saved in model_output_albertv2/seegull_gpt_augmentation_tr
ained/checkpoint-126/tokenizer_config.json
Special tokens file saved in model_output_albertv2/seegull_gpt_augmentation_trai
ned/checkpoint-126/special_tokens_map.json

Training completed. Do not forget to share your model on huggingface.co/models
=)

Loading best model from
model_output_albertv2/seegull_gpt_augmentation_trained/checkpoint-105 (score:
0.2618383467197418).

```

```

Deleting older checkpoint
[model_output_albertv2/seegull_gpt_augmentation_trained/checkpoint-126] due to
args.save_total_limit
Saving model checkpoint to
model_output_albertv2/seegull_gpt_augmentation_trained
Configuration saved in
model_output_albertv2/seegull_gpt_augmentation_trained/config.json
Model weights saved in
model_output_albertv2/seegull_gpt_augmentation_trained/model.safetensors
tokenizer config file saved in
model_output_albertv2/seegull_gpt_augmentation_trained/tokenizer_config.json
Special tokens file saved in
model_output_albertv2/seegull_gpt_augmentation_trained/special_tokens_map.json
loading configuration file
model_output_albertv2/seegull_gpt_augmentation_trained/config.json
Model config AlbertConfig {
  "architectures": [
    "AlbertForSequenceClassification"
  ],
  "attention_probs_dropout_prob": 0,
  "bos_token_id": 2,
  "classifier_dropout_prob": 0.1,
  "down_scale_factor": 1,
  "dtype": "float32",
  "embedding_size": 128,
  "eos_token_id": 3,
  "gap_size": 0,
  "hidden_act": "gelu_new",
  "hidden_dropout_prob": 0,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "inner_group_num": 1,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "albert",
  "net_structure_type": 0,
  "num_attention_heads": 12,
  "num_hidden_groups": 1,
  "num_hidden_layers": 12,
  "num_memory_blocks": 0,
  "pad_token_id": 0,
  "position_embedding_type": "absolute",
  "problem_type": "single_label_classification",
  "transformers_version": "4.57.3",
  "type_vocab_size": 2,
  "vocab_size": 30000
}

```



```

"inner_group_num": 1,
"intermediate_size": 3072,
"layer_norm_eps": 1e-12,
"max_position_embeddings": 512,
"model_type": "albert",
"net_structure_type": 0,
"num_attention_heads": 12,
"num_hidden_groups": 1,
"num_hidden_layers": 12,
"num_memory_blocks": 0,
"pad_token_id": 0,
"position_embedding_type": "absolute",
"problem_type": "single_label_classification",
"transformers_version": "4.57.3",
"type_vocab_size": 2,
"vocab_size": 30000
}

loading weights file
model_output_albertv2/seegull_gpt_augmentation_trained/model.safetensors
loading file spiece.model
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

Saved full results to: result_output_albertv2/seegull_gpt_augmentation_trained/w
inoqueer_gpt_augmentation/full_results.csv
Saved classification report to: result_output_albertv2/seegull_gpt_augmentation_
trained/winoqueer_gpt_augmentation/classification_report.csv
Number of unique labels: 2

Map: 100%|          | 414/414 [00:00<00:00, 26535.68 examples/s]
Map: 100%|          | 414/414 [00:00<00:00, 31447.60 examples/s]
Device set to use cuda:0

Sample tokenized input from test: {'text': 'The Somali refugees were labeled as
unacquainted by some', 'label': 0, 'group': 'Somalis', 'data_name': 'SeeGULL -
GPT Augmentation', '__index_level_0__': 840, 'input_ids': [2, 14, 16330, 8790,
46, 14348, 28, 367, 26692, 108, 1427, 34, 109, 3, 0, 0, 0, 0, 0, 0],
'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0],
'labels': 0}

/tmp/hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.

```

```

warnings.warn(
loading configuration file
model_output_albertv2/seegull_gpt_augmentation_trained/config.json
Model config AlbertConfig {
  "architectures": [
    "AlbertForSequenceClassification"
  ],
  "attention_probs_dropout_prob": 0,
  "bos_token_id": 2,
  "classifier_dropout_prob": 0.1,
  "down_scale_factor": 1,
  "dtype": "float32",
  "embedding_size": 128,
  "eos_token_id": 3,
  "gap_size": 0,
  "hidden_act": "gelu_new",
  "hidden_dropout_prob": 0,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "inner_group_num": 1,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "albert",
  "net_structure_type": 0,
  "num_attention_heads": 12,
  "num_hidden_groups": 1,
  "num_hidden_layers": 12,
  "num_memory_blocks": 0,
  "pad_token_id": 0,
  "position_embedding_type": "absolute",
  "problem_type": "single_label_classification",
  "transformers_version": "4.57.3",
  "type_vocab_size": 2,
  "vocab_size": 30000
}

```

```

loading weights file
model_output_albertv2/seegull_gpt_augmentation_trained/model.safetensors
loading file spiece.model
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

```

```

Saved full results to: result_output_albertv2/seegull_gpt_augmentation_trained/s
eegull_gpt_augmentation/full_results.csv

```

Saved classification report to: result_output_albertv2/seegull_gpt_augmentation_
trained/seegull_gpt_augmentation/classification_report.csv
Number of unique labels: 2

Map: 100%| | 10374/10374 [00:00<00:00, 25123.87 examples/s]

Map: 100%| | 10374/10374 [00:00<00:00, 26863.03 examples/s]

Device set to use cuda:0

/tmp/hearts_gpu/lib64/python3.9/site-

packages/transformers/pipelines/text_classification.py:111: UserWarning:

`return_all_scores` is now deprecated, if want a similar functionality use

`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of

`return_all_scores=False`.

warnings.warn(
Sample tokenized input from test: {'group': 'race', 'text': 'When i see an arab,

i think they are a happy.', 'label': 0, 'data_name': 'MGSD',
`__index_level_0__`: 17851, 'input_ids': [2, 76, 31, 196, 40, 3666, 15, 31, 277,

59, 50, 21, 1700, 9, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,

0, 0, 0, 0, 0, 0, 0, 0, 0], 'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,

0, 0,

0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,

0, 0,

0], 'labels': 0}

loading configuration file

model_output_albertv2/seegull_gpt_augmentation_trained/config.json

Model config AlbertConfig {

"architectures": [
"AlbertForSequenceClassification"

],
"attention_probs_dropout_prob": 0,

"bos_token_id": 2,

"classifier_dropout_prob": 0.1,

"down_scale_factor": 1,

"dtype": "float32",

"embedding_size": 128,

"eos_token_id": 3,

"gap_size": 0,

"hidden_act": "gelu_new",

"hidden_dropout_prob": 0,

"hidden_size": 768,

"initializer_range": 0.02,

"inner_group_num": 1,

"intermediate_size": 3072,

"layer_norm_eps": 1e-12,

"max_position_embeddings": 512,

"model_type": "albert",

"net_structure_type": 0,

"num_attention_heads": 12,

```

    "num_hidden_groups": 1,
    "num_hidden_layers": 12,
    "num_memory_blocks": 0,
    "pad_token_id": 0,
    "position_embedding_type": "absolute",
    "problem_type": "single_label_classification",
    "transformers_version": "4.57.3",
    "type_vocab_size": 2,
    "vocab_size": 30000
}

loading weights file
model_output_albertv2/seegull_gpt_augmentation_trained/model.safetensors
loading file spiece.model
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

Saved full results to:
result_output_albertv2/seegull_gpt_augmentation_trained/mgsd/full_results.csv
Saved classification report to: result_output_albertv2/seegull_gpt_augmentation_
trained/mgsd/classification_report.csv
Number of unique labels: 2

Map: 100%|          | 11441/11441 [00:00<00:00, 25824.60 examples/s]
Map: 100%|          | 11441/11441 [00:00<00:00, 29088.25 examples/s]
Device set to use cuda:0
/tmp/hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.
  warnings.warn(

Sample tokenized input from test: {'text': 'The Somali refugees were labeled as
unacquainted by some', 'label': 0, 'group': 'Somalis', 'data_name': 'SeeGULL -
GPT Augmentation', 'input_ids': [2, 14, 16330, 8790, 46, 14348, 28, 367, 26692,
108, 1427, 34, 109, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'token_type_ids':
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0], 'labels': 0}

loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--
albert-base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/config.json
Model config AlbertConfig {
  "architectures": [
    "AlbertForMaskedLM"

```



```

],
"attention_probs_dropout_prob": 0,
"bos_token_id": 2,
"classifier_dropout_prob": 0.1,
"down_scale_factor": 1,
"embedding_size": 128,
"eos_token_id": 3,
"gap_size": 0,
"hidden_act": "gelu_new",
"hidden_dropout_prob": 0,
"hidden_size": 768,
"initializer_range": 0.02,
"inner_group_num": 1,
"intermediate_size": 3072,
"layer_norm_eps": 1e-12,
"max_position_embeddings": 512,
"model_type": "albert",
"net_structure_type": 0,
"num_attention_heads": 12,
"num_hidden_groups": 1,
"num_hidden_layers": 12,
"num_memory_blocks": 0,
"pad_token_id": 0,
"position_embedding_type": "absolute",
"transformers_version": "4.57.3",
"type_vocab_size": 2,
"vocab_size": 30000
}

```

loading weights file model.safetensors from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--albert-base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/model.safetensors

Some weights of the model checkpoint at albert/albert-base-v2 were not used when initializing AlbertForSequenceClassification: ['predictions.LayerNorm.bias', 'predictions.LayerNorm.weight', 'predictions.bias', 'predictions.decoder.bias', 'predictions.dense.bias', 'predictions.dense.weight']

- This IS expected if you are initializing AlbertForSequenceClassification from the checkpoint of a model trained on another task or with another architecture (e.g. initializing a BertForSequenceClassification model from a BertForPreTraining model).
- This IS NOT expected if you are initializing AlbertForSequenceClassification from the checkpoint of a model that you expect to be exactly identical (initializing a BertForSequenceClassification model from a BertForSequenceClassification model).

Some weights of AlbertForSequenceClassification were not initialized from the model checkpoint at albert/albert-base-v2 and are newly initialized: ['classifier.bias', 'classifier.weight']

You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.

Saved full results to: result_output_albertv2/seegull_gpt_augmentation_trained/merged_winoqueer_seegull_gpt_augmentation/full_results.csv

Saved classification report to: result_output_albertv2/seegull_gpt_augmentation_trained/merged_winoqueer_seegull_gpt_augmentation/classification_report.csv

Number of unique labels: 2

loading configuration file config.json from cache at

/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--albert-base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/config.json

```
Model config AlbertConfig {
  "architectures": [
    "AlbertForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0,
  "bos_token_id": 2,
  "classifier_dropout_prob": 0.1,
  "down_scale_factor": 1,
  "embedding_size": 128,
  "eos_token_id": 3,
  "gap_size": 0,
  "hidden_act": "gelu_new",
  "hidden_dropout_prob": 0,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "inner_group_num": 1,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "albert",
  "net_structure_type": 0,
  "num_attention_heads": 12,
  "num_hidden_groups": 1,
  "num_hidden_layers": 12,
  "num_memory_blocks": 0,
  "pad_token_id": 0,
  "position_embedding_type": "absolute",
  "transformers_version": "4.57.3",
  "type_vocab_size": 2,
  "vocab_size": 30000
}
```

loading file spiece.model from cache at

/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--albert-base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/spiece.model

loading file tokenizer.json from cache at

/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--

```

albert-base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/tokenizer.json
loading file added_tokens.json from cache at None
loading file special_tokens_map.json from cache at None
loading file tokenizer_config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--
albert-
base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/tokenizer_config.json
loading file chat_template.jinja from cache at None
loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--albert--
albert-base-v2/snapshots/8e2f239c5f8a2c0f253781ca60135db913e5c80c/config.json
Model config AlbertConfig {
  "architectures": [
    "AlbertForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0,
  "bos_token_id": 2,
  "classifier_dropout_prob": 0.1,
  "down_scale_factor": 1,
  "embedding_size": 128,
  "eos_token_id": 3,
  "gap_size": 0,
  "hidden_act": "gelu_new",
  "hidden_dropout_prob": 0,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "inner_group_num": 1,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "albert",
  "net_structure_type": 0,
  "num_attention_heads": 12,
  "num_hidden_groups": 1,
  "num_hidden_layers": 12,
  "num_memory_blocks": 0,
  "pad_token_id": 0,
  "position_embedding_type": "absolute",
  "transformers_version": "4.57.3",
  "type_vocab_size": 2,
  "vocab_size": 30000
}

```

```

Map: 100%|          | 36608/36608 [00:01<00:00, 23008.99 examples/s]
Map: 100%|          | 36608/36608 [00:01<00:00, 31414.40 examples/s]

```

```

Sample tokenized input from train: {'text': 'The commander could tell the
soldier was disciplined.', 'label': 1, 'group': 'profession', 'data_name':

```

[illegible]

PyTorch: setting up devices

```
trainer = Trainer(
```

```
***** Running training *****
```

[illegible]

The following columns in the Evaluation set don't have a corresponding argument in ``AlbertForSequenceClassification.forward`` and have been ignored: `group`, `__index_level_0__`, `data_name`, `text`. If `group`, `__index_level_0__`, `data_name`, `text` are not expected by ``AlbertForSequenceClassification.forward``, you can safely ignore this message.

Num examples = 9152

Batch size = 64
Saving model checkpoint to model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-572
Configuration saved in model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-572/config.json
Model weights saved in model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-572/model.safetensors
tokenizer config file saved in model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-572/tokenizer_config.json
Special tokens file saved in model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-572/special_tokens_map.json
The following columns in the Evaluation set don't have a corresponding argument in `AlbertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `AlbertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 9152

Batch size = 64

Saving model checkpoint to model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1144
Configuration saved in model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1144/config.json
Model weights saved in model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1144/model.safetensors
tokenizer config file saved in model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1144/tokenizer_config.json
Special tokens file saved in model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1144/special_tokens_map.json
Deleting older checkpoint [model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-572] due to args.save_total_limit
The following columns in the Evaluation set don't have a corresponding argument in `AlbertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `AlbertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 9152

Batch size = 64

Saving model checkpoint to model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1716
Configuration saved in model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1716/config.json
Model weights saved in model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1716/model.safetensors
tokenizer config file saved in model_output_albertv2/merged_winoqueer_seegull_gpt

t_augmentation_trained/checkpoint-1716/tokenizer_config.json
Special tokens file saved in model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1716/special_tokens_map.json
The following columns in the Evaluation set don't have a corresponding argument in `AlbertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `AlbertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 9152

Batch size = 64

Saving model checkpoint to model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2288

Configuration saved in model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2288/config.json

Model weights saved in model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2288/model.safetensors

tokenizer config file saved in model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2288/tokenizer_config.json

Special tokens file saved in model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2288/special_tokens_map.json

Deleting older checkpoint [model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1716] due to args.save_total_limit

The following columns in the Evaluation set don't have a corresponding argument in `AlbertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `AlbertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 9152

Batch size = 64

Saving model checkpoint to model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2860

Configuration saved in model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2860/config.json

Model weights saved in model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2860/model.safetensors

tokenizer config file saved in model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2860/tokenizer_config.json

Special tokens file saved in model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2860/special_tokens_map.json

Deleting older checkpoint [model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2288] due to args.save_total_limit

The following columns in the Evaluation set don't have a corresponding argument in `AlbertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text

are not expected by `AlbertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 9152

Batch size = 64

Saving model checkpoint to model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-3432

Configuration saved in model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-3432/config.json

Model weights saved in model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-3432/model.safetensors

tokenizer config file saved in model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-3432/tokenizer_config.json

Special tokens file saved in model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-3432/special_tokens_map.json

Deleting older checkpoint [model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2860] due to args.save_total_limit

Training completed. Do not forget to share your model on huggingface.co/models
=)

Loading best model from model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1144 (score: 0.37396711111068726).

Deleting older checkpoint [model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-3432] due to args.save_total_limit

Saving model checkpoint to

model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained

Configuration saved in model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/config.json

Model weights saved in model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/model.safetensors

tokenizer config file saved in model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/tokenizer_config.json

Special tokens file saved in model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/special_tokens_map.json

loading configuration file model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/config.json

Model config AlbertConfig {

 "architectures": [

 "AlbertForSequenceClassification"

],

 "attention_probs_dropout_prob": 0,

 "bos_token_id": 2,

 "classifier_dropout_prob": 0.1,

 "down_scale_factor": 1,

```

"dtype": "float32",
"embedding_size": 128,
"eos_token_id": 3,
"gap_size": 0,
"hidden_act": "gelu_new",
"hidden_dropout_prob": 0,
"hidden_size": 768,
"initializer_range": 0.02,
"inner_group_num": 1,
"intermediate_size": 3072,
"layer_norm_eps": 1e-12,
"max_position_embeddings": 512,
"model_type": "albert",
"net_structure_type": 0,
"num_attention_heads": 12,
"num_hidden_groups": 1,
"num_hidden_layers": 12,
"num_memory_blocks": 0,
"pad_token_id": 0,
"position_embedding_type": "absolute",
"problem_type": "single_label_classification",
"transformers_version": "4.57.3",
"type_vocab_size": 2,
"vocab_size": 30000
}

```

loading weights file model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/model.safetensors

loading file spiece.model

loading file tokenizer.json

loading file added_tokens.json

loading file special_tokens_map.json

loading file tokenizer_config.json

loading file chat_template.jinja

Saved fine-tuned model to:

model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained

Number of unique labels: 2

Map: 100%| | 653/653 [00:00<00:00, 32820.62 examples/s]

Map: 100%| | 653/653 [00:00<00:00, 33051.52 examples/s]

Device set to use cuda:0

/tmp/hearts_gpu/lib64/python3.9/site-

packages/transformers/pipelines/text_classification.py:111: UserWarning:

`return_all_scores` is now deprecated, if want a similar functionality use

`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of

`return_all_scores=False`.

warnings.warn(


```
Sample tokenized input from test: {'text': 'most people who are Gay are kites',
'label': 0, 'group': 'Gay', 'data_name': 'Winoqueer - GPT Augmentation',
'__index_level_0__': 3236, 'input_ids': [2, 127, 148, 72, 50, 3398, 50, 19196,
18, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'token_type_ids': [0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'labels': 0}
```

loading configuration file model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/config.json

```
Model config AlbertConfig {
  "architectures": [
    "AlbertForSequenceClassification"
  ],
  "attention_probs_dropout_prob": 0,
  "bos_token_id": 2,
  "classifier_dropout_prob": 0.1,
  "down_scale_factor": 1,
  "dtype": "float32",
  "embedding_size": 128,
  "eos_token_id": 3,
  "gap_size": 0,
  "hidden_act": "gelu_new",
  "hidden_dropout_prob": 0,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "inner_group_num": 1,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "albert",
  "net_structure_type": 0,
  "num_attention_heads": 12,
  "num_hidden_groups": 1,
  "num_hidden_layers": 12,
  "num_memory_blocks": 0,
  "pad_token_id": 0,
  "position_embedding_type": "absolute",
  "problem_type": "single_label_classification",
  "transformers_version": "4.57.3",
  "type_vocab_size": 2,
  "vocab_size": 30000
}
```

loading weights file model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/model.safetensors

loading file spiece.model

loading file tokenizer.json

```

loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

Saved full results to: result_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/winoqueer_gpt_augmentation/full_results.csv
Saved classification report to: result_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/winoqueer_gpt_augmentation/classification_report.csv
Number of unique labels: 2

Map: 100%|          | 414/414 [00:00<00:00, 26342.45 examples/s]
Map: 100%|          | 414/414 [00:00<00:00, 30666.18 examples/s]
Device set to use cuda:0

Sample tokenized input from test: {'text': 'The Somali refugees were labeled as unacquainted by some', 'label': 0, 'group': 'Somalis', 'data_name': 'SeeGULL - GPT Augmentation', '__index_level_0__': 840, 'input_ids': [2, 14, 16330, 8790, 46, 14348, 28, 367, 26692, 108, 1427, 34, 109, 3, 0, 0, 0, 0, 0, 0], 'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0], 'labels': 0}

/tmp/hearts_gpu/lib64/python3.9/site-packages/transformers/pipelines/text_classification.py:111: UserWarning: `return_all_scores` is now deprecated, if want a similar functionality use `top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of `return_all_scores=False`.
  warnings.warn(
loading configuration file model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/config.json
Model config AlbertConfig {
  "architectures": [
    "AlbertForSequenceClassification"
  ],
  "attention_probs_dropout_prob": 0,
  "bos_token_id": 2,
  "classifier_dropout_prob": 0.1,
  "down_scale_factor": 1,
  "dtype": "float32",
  "embedding_size": 128,
  "eos_token_id": 3,
  "gap_size": 0,
  "hidden_act": "gelu_new",
  "hidden_dropout_prob": 0,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "inner_group_num": 1,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,

```

```

    "max_position_embeddings": 512,
    "model_type": "albert",
    "net_structure_type": 0,
    "num_attention_heads": 12,
    "num_hidden_groups": 1,
    "num_hidden_layers": 12,
    "num_memory_blocks": 0,
    "pad_token_id": 0,
    "position_embedding_type": "absolute",
    "problem_type": "single_label_classification",
    "transformers_version": "4.57.3",
    "type_vocab_size": 2,
    "vocab_size": 30000
}

loading weights file model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/model.safetensors
loading file spiece.model
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

Saved full results to: result_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/seegull_gpt_augmentation/full_results.csv
Saved classification report to: result_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/seegull_gpt_augmentation/classification_report.csv
Number of unique labels: 2

Map: 100%|          | 10374/10374 [00:00<00:00, 25022.22 examples/s]
Map: 100%|          | 10374/10374 [00:00<00:00, 26777.21 examples/s]
Device set to use cuda:0
/tmp/hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.
  warnings.warn(

Sample tokenized input from test: {'group': 'race', 'text': 'When i see an arab,
i think they are a happy.', 'label': 0, 'data_name': 'MGSD',
'__index_level_0__': 17851, 'input_ids': [2, 76, 31, 196, 40, 3666, 15, 31, 277,
59, 50, 21, 1700, 9, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0], 'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0], 'labels': 0}

```

loading configuration file model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/config.json

```
Model config AlbertConfig {
  "architectures": [
    "AlbertForSequenceClassification"
  ],
  "attention_probs_dropout_prob": 0,
  "bos_token_id": 2,
  "classifier_dropout_prob": 0.1,
  "down_scale_factor": 1,
  "dtype": "float32",
  "embedding_size": 128,
  "eos_token_id": 3,
  "gap_size": 0,
  "hidden_act": "gelu_new",
  "hidden_dropout_prob": 0,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "inner_group_num": 1,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "albert",
  "net_structure_type": 0,
  "num_attention_heads": 12,
  "num_hidden_groups": 1,
  "num_hidden_layers": 12,
  "num_memory_blocks": 0,
  "pad_token_id": 0,
  "position_embedding_type": "absolute",
  "problem_type": "single_label_classification",
  "transformers_version": "4.57.3",
  "type_vocab_size": 2,
  "vocab_size": 30000
}
```

loading weights file model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/model.safetensors

loading file spiece.model

loading file tokenizer.json

loading file added_tokens.json

loading file special_tokens_map.json

loading file tokenizer_config.json

loading file chat_template.jinja

Saved full results to: result_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/mgsd/full_results.csv

Saved classification report to: result_output_albertv2/merged_winoqueer_seegull_

```

gpt_augmentation_trained/mgsd/classification_report.csv
Number of unique labels: 2

Map: 100%|          | 11441/11441 [00:00<00:00, 25863.35 examples/s]
Map: 100%|          | 11441/11441 [00:00<00:00, 29440.71 examples/s]
Device set to use cuda:0
/tmp/hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.
  warnings.warn(

Sample tokenized input from test: {'text': 'The Somali refugees were labeled as
unacquainted by some', 'label': 0, 'group': 'Somalis', 'data_name': 'SeeGULL -
GPT Augmentation', 'input_ids': [2, 14, 16330, 8790, 46, 14348, 28, 367, 26692,
108, 1427, 34, 109, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'token_type_ids':
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0], 'labels': 0}
Saved full results to: result_output_albertv2/merged_winoqueer_seegull_gpt_augme
ntation_trained/merged_winoqueer_seegull_gpt_augmentation/full_results.csv
Saved classification report to: result_output_albertv2/merged_winoqueer_seegull_
gpt_augmentation_trained/merged_winoqueer_seegull_gpt_augmentation/classificatio
n_report.csv
Number of unique labels: 2

loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--
distilbert-base-
uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/config.json
Model config DistilBertConfig {
  "activation": "gelu",
  "architectures": [
    "DistilBertForMaskedLM"
  ],
  "attention_dropout": 0.1,
  "dim": 768,
  "dropout": 0.1,
  "hidden_dim": 3072,
  "initializer_range": 0.02,
  "max_position_embeddings": 512,
  "model_type": "distilbert",
  "n_heads": 12,
  "n_layers": 6,
  "pad_token_id": 0,
  "qa_dropout": 0.1,
  "seq_classif_dropout": 0.2,
  "sinusoidal_pos_embs": false,

```

```

    "tie_weights_": true,
    "transformers_version": "4.57.3",
    "vocab_size": 30522
}

```

loading weights file model.safetensors from cache at
 /cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--distilbert-base-

uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/model.safetensors
 Some weights of the model checkpoint at distilbert/distilbert-base-uncased were not used when initializing DistilBertForSequenceClassification:

```

['vocab_layer_norm.bias', 'vocab_layer_norm.weight', 'vocab_projector.bias',
'vocab_transform.bias', 'vocab_transform.weight']

```

- This IS expected if you are initializing DistilBertForSequenceClassification from the checkpoint of a model trained on another task or with another architecture (e.g. initializing a BertForSequenceClassification model from a BertForPreTraining model).

- This IS NOT expected if you are initializing DistilBertForSequenceClassification from the checkpoint of a model that you expect to be exactly identical (initializing a BertForSequenceClassification model from a BertForSequenceClassification model).

Some weights of DistilBertForSequenceClassification were not initialized from the model checkpoint at distilbert/distilbert-base-uncased and are newly initialized: ['classifier.bias', 'classifier.weight', 'pre_classifier.bias', 'pre_classifier.weight']

You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.

loading configuration file config.json from cache at
 /cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--distilbert-base-

uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/config.json

```

Model config DistilBertConfig {
  "activation": "gelu",
  "architectures": [
    "DistilBertForMaskedLM"
  ],
  "attention_dropout": 0.1,
  "dim": 768,
  "dropout": 0.1,
  "hidden_dim": 3072,
  "initializer_range": 0.02,
  "max_position_embeddings": 512,
  "model_type": "distilbert",
  "n_heads": 12,
  "n_layers": 6,
  "pad_token_id": 0,
  "qa_dropout": 0.1,
  "seq_classif_dropout": 0.2,

```

```

    "sinusoidal_pos_embs": false,
    "tie_weights_": true,
    "transformers_version": "4.57.3",
    "vocab_size": 30522
}

loading file vocab.txt from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--
distilbert-base-
uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/vocab.txt
loading file tokenizer.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--
distilbert-base-
uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/tokenizer.json
loading file added_tokens.json from cache at None
loading file special_tokens_map.json from cache at None
loading file tokenizer_config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--
distilbert-base-
uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/tokenizer_config.json
loading file chat_template.jinja from cache at None
loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--
distilbert-base-
uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/config.json
Model config DistilBertConfig {
  "activation": "gelu",
  "architectures": [
    "DistilBertForMaskedLM"
  ],
  "attention_dropout": 0.1,
  "dim": 768,
  "dropout": 0.1,
  "hidden_dim": 3072,
  "initializer_range": 0.02,
  "max_position_embeddings": 512,
  "model_type": "distilbert",
  "n_heads": 12,
  "n_layers": 6,
  "pad_token_id": 0,
  "qa_dropout": 0.1,
  "seq_classif_dropout": 0.2,
  "sinusoidal_pos_embs": false,
  "tie_weights_": true,
  "transformers_version": "4.57.3",
  "vocab_size": 30522
}

```

```
Map: 100%|          | 33194/33194 [00:01<00:00, 28775.37 examples/s]
Map: 100%|          | 33194/33194 [00:00<00:00, 36683.12 examples/s]
```

[illegible]

```
Map: 100%|          | 8299/8299 [00:00<00:00, 31421.08 examples/s]
Map: 100%|          | 8299/8299 [00:00<00:00, 26846.98 examples/s]
```

PyTorch: setting up devices

```
Model Training and Evaluation/BERT_Models_Fine_Tuning2.py:209: FutureWarning:
`tokenizer` is deprecated and will be removed in version 5.0.0 for
`Trainer._init_`. Use `processing_class` instead.
```

```
trainer = Trainer(
```

[illegible]

The following columns in the Training set don't have a corresponding argument in ``DistilBertForSequenceClassification.forward`` and have been ignored: `group`, `__index_level_0__`, `data_name`, `text`. If `group`, `__index_level_0__`, `data_name`, `text` are not expected by ``DistilBertForSequenceClassification.forward``, you can safely ignore this message.

```
***** Running training *****
```

Num examples = 33,194

Num Epochs = 6

Instantaneous batch size per device = 64

Total train batch size (w. parallel, distributed & accumulation) = 64

Gradient Accumulation steps = 1

Total optimization steps = 3,114

Number of trainable parameters = 66,955,010

```
<IPython.core.display.HTML object>
```

The following columns in the Evaluation set don't have a corresponding argument in ``DistilBertForSequenceClassification.forward`` and have been ignored: `group`, `__index_level_0__`, `data_name`, `text`. If `group`, `__index_level_0__`, `data_name`, `text` are not expected by ``DistilBertForSequenceClassification.forward``, you can safely ignore this message.


```
***** Running Evaluation *****
  Num examples = 8299
  Batch size = 64
Saving model checkpoint to model_output_distilbert/mgsd_trained/checkpoint-519
Configuration saved in
model_output_distilbert/mgsd_trained/checkpoint-519/config.json
Model weights saved in
model_output_distilbert/mgsd_trained/checkpoint-519/model.safetensors
tokenizer config file saved in
model_output_distilbert/mgsd_trained/checkpoint-519/tokenizer_config.json
Special tokens file saved in
model_output_distilbert/mgsd_trained/checkpoint-519/special_tokens_map.json
The following columns in the Evaluation set don't have a corresponding argument
in `DistilBertForSequenceClassification.forward` and have been ignored: group,
__index_level_0__, data_name, text. If group, __index_level_0__, data_name, text
are not expected by `DistilBertForSequenceClassification.forward`, you can
safely ignore this message.
```

```
***** Running Evaluation *****
  Num examples = 8299
  Batch size = 64
Saving model checkpoint to model_output_distilbert/mgsd_trained/checkpoint-1038
Configuration saved in
model_output_distilbert/mgsd_trained/checkpoint-1038/config.json
Model weights saved in
model_output_distilbert/mgsd_trained/checkpoint-1038/model.safetensors
tokenizer config file saved in
model_output_distilbert/mgsd_trained/checkpoint-1038/tokenizer_config.json
Special tokens file saved in
model_output_distilbert/mgsd_trained/checkpoint-1038/special_tokens_map.json
Deleting older checkpoint [model_output_distilbert/mgsd_trained/checkpoint-519]
due to args.save_total_limit
The following columns in the Evaluation set don't have a corresponding argument
in `DistilBertForSequenceClassification.forward` and have been ignored: group,
__index_level_0__, data_name, text. If group, __index_level_0__, data_name, text
are not expected by `DistilBertForSequenceClassification.forward`, you can
safely ignore this message.
```

```
***** Running Evaluation *****
  Num examples = 8299
  Batch size = 64
Saving model checkpoint to model_output_distilbert/mgsd_trained/checkpoint-1557
Configuration saved in
model_output_distilbert/mgsd_trained/checkpoint-1557/config.json
Model weights saved in
model_output_distilbert/mgsd_trained/checkpoint-1557/model.safetensors
tokenizer config file saved in
```

model_output_distilbert/mgsd_trained/checkpoint-1557/tokenizer_config.json
Special tokens file saved in
model_output_distilbert/mgsd_trained/checkpoint-1557/special_tokens_map.json
The following columns in the Evaluation set don't have a corresponding argument
in `DistilBertForSequenceClassification.forward` and have been ignored: group,
__index_level_0__, data_name, text. If group, __index_level_0__, data_name, text
are not expected by `DistilBertForSequenceClassification.forward`, you can
safely ignore this message.

***** Running Evaluation *****

Num examples = 8299

Batch size = 64

Saving model checkpoint to model_output_distilbert/mgsd_trained/checkpoint-2076

Configuration saved in

model_output_distilbert/mgsd_trained/checkpoint-2076/config.json

Model weights saved in

model_output_distilbert/mgsd_trained/checkpoint-2076/model.safetensors

tokenizer config file saved in

model_output_distilbert/mgsd_trained/checkpoint-2076/tokenizer_config.json

Special tokens file saved in

model_output_distilbert/mgsd_trained/checkpoint-2076/special_tokens_map.json

Deleting older checkpoint [model_output_distilbert/mgsd_trained/checkpoint-1557]
due to args.save_total_limit

The following columns in the Evaluation set don't have a corresponding argument
in `DistilBertForSequenceClassification.forward` and have been ignored: group,
__index_level_0__, data_name, text. If group, __index_level_0__, data_name, text
are not expected by `DistilBertForSequenceClassification.forward`, you can
safely ignore this message.

***** Running Evaluation *****

Num examples = 8299

Batch size = 64

Saving model checkpoint to model_output_distilbert/mgsd_trained/checkpoint-2595

Configuration saved in

model_output_distilbert/mgsd_trained/checkpoint-2595/config.json

Model weights saved in

model_output_distilbert/mgsd_trained/checkpoint-2595/model.safetensors

tokenizer config file saved in

model_output_distilbert/mgsd_trained/checkpoint-2595/tokenizer_config.json

Special tokens file saved in

model_output_distilbert/mgsd_trained/checkpoint-2595/special_tokens_map.json

Deleting older checkpoint [model_output_distilbert/mgsd_trained/checkpoint-2076]
due to args.save_total_limit

The following columns in the Evaluation set don't have a corresponding argument
in `DistilBertForSequenceClassification.forward` and have been ignored: group,
__index_level_0__, data_name, text. If group, __index_level_0__, data_name, text
are not expected by `DistilBertForSequenceClassification.forward`, you can
safely ignore this message.

```

***** Running Evaluation *****
  Num examples = 8299
  Batch size = 64
Saving model checkpoint to model_output_distilbert/mgsd_trained/checkpoint-3114
Configuration saved in
model_output_distilbert/mgsd_trained/checkpoint-3114/config.json
Model weights saved in
model_output_distilbert/mgsd_trained/checkpoint-3114/model.safetensors
tokenizer config file saved in
model_output_distilbert/mgsd_trained/checkpoint-3114/tokenizer_config.json
Special tokens file saved in
model_output_distilbert/mgsd_trained/checkpoint-3114/special_tokens_map.json
Deleting older checkpoint [model_output_distilbert/mgsd_trained/checkpoint-2595]
due to args.save_total_limit

```

Training completed. Do not forget to share your model on huggingface.co/models
=)

```

Loading best model from model_output_distilbert/mgsd_trained/checkpoint-1038
(score: 0.39408475160598755).
Deleting older checkpoint [model_output_distilbert/mgsd_trained/checkpoint-3114]
due to args.save_total_limit
Saving model checkpoint to model_output_distilbert/mgsd_trained
Configuration saved in model_output_distilbert/mgsd_trained/config.json
Model weights saved in model_output_distilbert/mgsd_trained/model.safetensors
tokenizer config file saved in
model_output_distilbert/mgsd_trained/tokenizer_config.json
Special tokens file saved in
model_output_distilbert/mgsd_trained/special_tokens_map.json
loading configuration file model_output_distilbert/mgsd_trained/config.json
Model config DistilBertConfig {
  "activation": "gelu",
  "architectures": [
    "DistilBertForSequenceClassification"
  ],
  "attention_dropout": 0.1,
  "dim": 768,
  "dropout": 0.1,
  "dtype": "float32",
  "hidden_dim": 3072,
  "initializer_range": 0.02,
  "max_position_embeddings": 512,
  "model_type": "distilbert",
  "n_heads": 12,
  "n_layers": 6,

```

```

"pad_token_id": 0,
"problem_type": "single_label_classification",
"qa_dropout": 0.1,
"seq_classif_dropout": 0.2,
"sinusoidal_pos_embs": false,
"tie_weights_": true,
"transformers_version": "4.57.3",
"vocab_size": 30522
}

```

Saved fine-tuned model to: model_output_distilbert/mgsd_trained

```
Map: 100%|          | 653/653 [00:00<00:00, 36817.36 examples/s]
```

```
Device set to use cuda:0
```

```
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.
```

```
Sample tokenized input from test: {'text': 'most people who are Gay are kites',
'label': 0, 'group': 'Gay', 'data_name': 'Winoqueer - GPT Augmentation',
'__index_level_0__': 3236, 'input_ids': [101, 2087, 2111, 2040, 2024, 5637,
2024, 20497, 2015, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0], 'labels': 0}
```

```
Model config DistilBertConfig {
```

```

    "max_position_embeddings": 512,
    "model_type": "distilbert",
    "n_heads": 12,
    "n_layers": 6,
    "pad_token_id": 0,
    "problem_type": "single_label_classification",
    "qa_dropout": 0.1,
    "seq_classif_dropout": 0.2,
    "sinusoidal_pos_embs": false,
    "tie_weights_": true,
    "transformers_version": "4.57.3",
    "vocab_size": 30522
}

loading weights file model_output_distilbert/mgsd_trained/model.safetensors
loading file vocab.txt
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

Saved full results to: result_output_distilbert/mgsd_trained/winoqueer_gpt_augmentation/full_results.csv
Saved classification report to: result_output_distilbert/mgsd_trained/winoqueer_gpt_augmentation/classification_report.csv
Number of unique labels: 2

Map: 100%|          | 414/414 [00:00<00:00, 33828.33 examples/s]
Map: 100%|          | 414/414 [00:00<00:00, 34445.69 examples/s]
Device set to use cuda:0

Sample tokenized input from test: {'text': 'The Somali refugees were labeled as
unacquainted by some', 'label': 0, 'group': 'Somalis', 'data_name': 'SeeGULL -
GPT Augmentation', '__index_level_0__': 840, 'input_ids': [101, 1996, 16831,
8711, 2020, 12599, 2004, 14477, 2278, 16211, 18447, 2098, 2011, 2070, 102, 0, 0,
0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0,
0, 0, 0, 0, 0], 'labels': 0}

/tmp/ hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.
  warnings.warn(
loading configuration file model_output_distilbert/mgsd_trained/config.json
Model config DistilBertConfig {
  "activation": "gelu",
  "architectures": [
    "DistilBertForSequenceClassification"

```

```

],
"attention_dropout": 0.1,
"dim": 768,
"dropout": 0.1,
"dtype": "float32",
"hidden_dim": 3072,
"initializer_range": 0.02,
"max_position_embeddings": 512,
"model_type": "distilbert",
"n_heads": 12,
"n_layers": 6,
"pad_token_id": 0,
"problem_type": "single_label_classification",
"qa_dropout": 0.1,
"seq_classif_dropout": 0.2,
"sinusoidal_pos_embs": false,
"tie_weights_": true,
"transformers_version": "4.57.3",
"vocab_size": 30522
}

```

```

loading weights file model_output_distilbert/mgsd_trained/model.safetensors
loading file vocab.txt
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

```

Saved full results to:

result_output_distilbert/mgsd_trained/seegull_gpt_augmentation/full_results.csv

Saved classification report to: result_output_distilbert/mgsd_trained/seegull_gpt
t_augmentation/classification_report.csv

Number of unique labels: 2

Map: 100%| | 10374/10374 [00:00<00:00, 31346.42 examples/s]

Map: 100%| | 10374/10374 [00:00<00:00, 39870.79 examples/s]

Device set to use cuda:0

/tmp/hearts_gpu/lib64/python3.9/site-

packages/transformers/pipelines/text_classification.py:111: UserWarning:

`return_all_scores` is now deprecated, if want a similar functionality use

`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of

`return_all_scores=False`.

warnings.warn(

Sample tokenized input from test: {'group': 'race', 'text': 'When i see an arab,
i think they are a happy.', 'label': 0, 'data_name': 'MGSD',

'__index_level_0__': 17851, 'input_ids': [101, 2043, 1045, 2156, 2019, 5424,
1010, 1045, 2228, 2027, 2024, 1037, 3407, 1012, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0,

```
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'attention_mask': [1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'labels': 0}
```

loading configuration file model_output_distilbert/mgsd_trained/config.json

```
Model config DistilBertConfig {
  "activation": "gelu",
  "architectures": [
    "DistilBertForSequenceClassification"
  ],
  "attention_dropout": 0.1,
  "dim": 768,
  "dropout": 0.1,
  "dtype": "float32",
  "hidden_dim": 3072,
  "initializer_range": 0.02,
  "max_position_embeddings": 512,
  "model_type": "distilbert",
  "n_heads": 12,
  "n_layers": 6,
  "pad_token_id": 0,
  "problem_type": "single_label_classification",
  "qa_dropout": 0.1,
  "seq_classif_dropout": 0.2,
  "sinusoidal_pos_embs": false,
  "tie_weights_": true,
  "transformers_version": "4.57.3",
  "vocab_size": 30522
}
```

loading weights file model_output_distilbert/mgsd_trained/model.safetensors

loading file vocab.txt

loading file tokenizer.json

loading file added_tokens.json

loading file special_tokens_map.json

loading file tokenizer_config.json

loading file chat_template.jinja

Saved full results to:

result_output_distilbert/mgsd_trained/mgsd/full_results.csv

Saved classification report to:

result_output_distilbert/mgsd_trained/mgsd/classification_report.csv

Number of unique labels: 2

Map: 100%| | 11441/11441 [00:00<00:00, 32204.35 examples/s]

Map: 100%| | 11441/11441 [00:00<00:00, 45824.66 examples/s]

Device set to use cuda:0

/tmp/hearts_gpu/lib64/python3.9/site-

packages/transformers/pipelines/text_classification.py:111: UserWarning:

`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.

```
warnings.warn(
```

```
Sample tokenized input from test: {'text': 'The Somali refugees were labeled as  
unacquainted by some', 'label': 0, 'group': 'Somalis', 'data_name': 'SeeGULL -  
GPT Augmentation', 'input_ids': [101, 1996, 16831, 8711, 2020, 12599, 2004,  
14477, 2278, 16211, 18447, 2098, 2011, 2070, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0,  
0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0,  
0, 0, 0, 0, 0, 0, 0, 0], 'labels': 0}
```

```
loading configuration file config.json from cache at  
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--  
distilbert-base-  
uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/config.json
```

```
Model config DistilBertConfig {  
  "activation": "gelu",  
  "architectures": [  
    "DistilBertForMaskedLM"  
  ],  
  "attention_dropout": 0.1,  
  "dim": 768,  
  "dropout": 0.1,  
  "hidden_dim": 3072,  
  "initializer_range": 0.02,  
  "max_position_embeddings": 512,  
  "model_type": "distilbert",  
  "n_heads": 12,  
  "n_layers": 6,  
  "pad_token_id": 0,  
  "qa_dropout": 0.1,  
  "seq_classif_dropout": 0.2,  
  "sinusoidal_pos_embs": false,  
  "tie_weights_": true,  
  "transformers_version": "4.57.3",  
  "vocab_size": 30522  
}
```

```
loading weights file model.safetensors from cache at  
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--  
distilbert-base-  
uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/model.safetensors
```

```
Saved full results to: result_output_distilbert/mgsd_trained/merged_winoqueer_se  
egull_gpt_augmentation/full_results.csv
```

```
Saved classification report to: result_output_distilbert/mgsd_trained/merged_wi  
noqueer_seegull_gpt_augmentation/classification_report.csv
```

```
Number of unique labels: 2
```


Some weights of the model checkpoint at distilbert/distilbert-base-uncased were not used when initializing DistilBertForSequenceClassification:

```
['vocab_layer_norm.bias', 'vocab_layer_norm.weight', 'vocab_projector.bias',  
'vocab_transform.bias', 'vocab_transform.weight']
```

- This IS expected if you are initializing DistilBertForSequenceClassification from the checkpoint of a model trained on another task or with another architecture (e.g. initializing a BertForSequenceClassification model from a BertForPreTraining model).

- This IS NOT expected if you are initializing DistilBertForSequenceClassification from the checkpoint of a model that you expect to be exactly identical (initializing a BertForSequenceClassification model from a BertForSequenceClassification model).

Some weights of DistilBertForSequenceClassification were not initialized from the model checkpoint at distilbert/distilbert-base-uncased and are newly initialized: ['classifier.bias', 'classifier.weight', 'pre_classifier.bias', 'pre_classifier.weight']

You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.

loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--distilbert-base-uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/config.json

```
Model config DistilBertConfig {  
  "activation": "gelu",  
  "architectures": [  
    "DistilBertForMaskedLM"  
  ],  
  "attention_dropout": 0.1,  
  "dim": 768,  
  "dropout": 0.1,  
  "hidden_dim": 3072,  
  "initializer_range": 0.02,  
  "max_position_embeddings": 512,  
  "model_type": "distilbert",  
  "n_heads": 12,  
  "n_layers": 6,  
  "pad_token_id": 0,  
  "qa_dropout": 0.1,  
  "seq_classif_dropout": 0.2,  
  "sinusoidal_pos_embs": false,  
  "tie_weights_": true,  
  "transformers_version": "4.57.3",  
  "vocab_size": 30522  
}
```

loading file vocab.txt from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--distilbert-base-

```

uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/vocab.txt
loading file tokenizer.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--
distilbert-base-
uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/tokenizer.json
loading file added_tokens.json from cache at None
loading file special_tokens_map.json from cache at None
loading file tokenizer_config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--
distilbert-base-
uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/tokenizer_config.json
loading file chat_template.jinja from cache at None
loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--
distilbert-base-
uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/config.json
Model config DistilBertConfig {
  "activation": "gelu",
  "architectures": [
    "DistilBertForMaskedLM"
  ],
  "attention_dropout": 0.1,
  "dim": 768,
  "dropout": 0.1,
  "hidden_dim": 3072,
  "initializer_range": 0.02,
  "max_position_embeddings": 512,
  "model_type": "distilbert",
  "n_heads": 12,
  "n_layers": 6,
  "pad_token_id": 0,
  "qa_dropout": 0.1,
  "seq_classif_dropout": 0.2,
  "sinusoidal_pos_embs": false,
  "tie_weights_": true,
  "transformers_version": "4.57.3",
  "vocab_size": 30522
}

```

```

Map: 100%|          | 2088/2088 [00:00<00:00, 43712.69 examples/s]
Map: 100%|          | 2088/2088 [00:00<00:00, 13366.44 examples/s]

```

```

Sample tokenized input from train: {'text': 'we expect that nonbinary people are
out to share their views.', 'label': 0, 'group': 'NB', 'data_name': 'Winoqueer -
GPT Augmentation', '__index_level_0__': 1788, 'input_ids': [101, 2057, 5987,
2008, 2512, 21114, 2854, 2111, 2024, 2041, 2000, 3745, 2037, 5328, 1012, 102, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'labels': 0}

```

```
Map: 100%|          | 523/523 [00:00<00:00, 23435.62 examples/s]
Map: 100%|          | 523/523 [00:00<00:00, 33046.41 examples/s]
PyTorch: setting up devices
Model Training and Evaluation/BERT_Models_Fine_Tuning2.py:209: FutureWarning:
`tokenizer` is deprecated and will be removed in version 5.0.0 for
`Trainer.__init__`. Use `processing_class` instead.
    trainer = Trainer(
```

```
Sample tokenized input from validation: {'text': 'all lesbians are studious.',
'label': 0, 'group': 'Lesbian', 'data_name': 'Winoqueer - GPT Augmentation',
'__index_level_0__': 2716, 'input_ids': [101, 2035, 11690, 2015, 2024, 2996,
2271, 1012, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'labels': 0}
```

The following columns in the Training set don't have a corresponding argument in `DistilBertForSequenceClassification.forward` and have been ignored: `group`, `__index_level_0__`, `data_name`, `text`. If `group`, `__index_level_0__`, `data_name`, `text` are not expected by `DistilBertForSequenceClassification.forward`, you can safely ignore this message.

```
***** Running training *****
```

```
Num examples = 2,088
Num Epochs = 6
Instantaneous batch size per device = 64
Total train batch size (w. parallel, distributed & accumulation) = 64
Gradient Accumulation steps = 1
Total optimization steps = 198
Number of trainable parameters = 66,955,010
```

```
<IPython.core.display.HTML object>
```

The following columns in the Evaluation set don't have a corresponding argument in ``DistilBertForSequenceClassification.forward`` and have been ignored: `group`, `__index_level_0__`, `data_name`, `text`. If `group`, `__index_level_0__`, `data_name`, `text` are not expected by ``DistilBertForSequenceClassification.forward``, you can safely ignore this message.

```
***** Running Evaluation *****
```

```
Num examples = 523
Batch size = 64
```

```

Saving model checkpoint to
model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-33
Configuration saved in model_output_distilbert/winoqueer_gpt_augmentation_train
d/checkpoint-33/config.json
Model weights saved in model_output_distilbert/winoqueer_gpt_augmentation_train
d/checkpoint-33/model.safetensors
tokenizer config file saved in model_output_distilbert/winoqueer_gpt_augmentatio
n_trained/checkpoint-33/tokenizer_config.json
Special tokens file saved in model_output_distilbert/winoqueer_gpt_augmentation_
trained/checkpoint-33/special_tokens_map.json

```

The following columns in the Evaluation set don't have a corresponding argument in `DistilBertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `DistilBertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 523

Batch size = 64

Saving model checkpoint to

model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-66

Configuration saved in model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-66/config.json

Model weights saved in model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-66/model.safetensors

tokenizer config file saved in model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-66/tokenizer_config.json

Special tokens file saved in model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-66/special_tokens_map.json

Deleting older checkpoint

[model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-33] due to args.save_total_limit

The following columns in the Evaluation set don't have a corresponding argument in `DistilBertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `DistilBertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 523

Batch size = 64

Saving model checkpoint to

model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-99

Configuration saved in model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-99/config.json

Model weights saved in model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-99/model.safetensors

tokenizer config file saved in model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-99/tokenizer_config.json

Special tokens file saved in model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-99/special_tokens_map.json

Deleting older checkpoint

[model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-66] due to args.save_total_limit

The following columns in the Evaluation set don't have a corresponding argument in `DistilBertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `DistilBertForSequenceClassification.forward`, you can

safely ignore this message.

***** Running Evaluation *****

Num examples = 523

Batch size = 64

Saving model checkpoint to

model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-132

Configuration saved in model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-132/config.json

Model weights saved in model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-132/model.safetensors

tokenizer config file saved in model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-132/tokenizer_config.json

Special tokens file saved in model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-132/special_tokens_map.json

Deleting older checkpoint

[model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-99] due to args.save_total_limit

The following columns in the Evaluation set don't have a corresponding argument in `DistilBertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `DistilBertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 523

Batch size = 64

Saving model checkpoint to

model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-165

Configuration saved in model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-165/config.json

Model weights saved in model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-165/model.safetensors

tokenizer config file saved in model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-165/tokenizer_config.json

Special tokens file saved in model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-165/special_tokens_map.json

Deleting older checkpoint

[model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-132] due to args.save_total_limit

The following columns in the Evaluation set don't have a corresponding argument in `DistilBertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `DistilBertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 523

```

Batch size = 64
Saving model checkpoint to
model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-198
Configuration saved in model_output_distilbert/winoqueer_gpt_augmentation_train_
d/checkpoint-198/config.json
Model weights saved in model_output_distilbert/winoqueer_gpt_augmentation_train_
d/checkpoint-198/model.safetensors
tokenizer config file saved in model_output_distilbert/winoqueer_gpt_augmentatio
n_trained/checkpoint-198/tokenizer_config.json
Special tokens file saved in model_output_distilbert/winoqueer_gpt_augmentation_
trained/checkpoint-198/special_tokens_map.json
Deleting older checkpoint
[model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-165] due
to args.save_total_limit

```

Training completed. Do not forget to share your model on huggingface.co/models
=)

```

Loading best model from
model_output_distilbert/winoqueer_gpt_augmentation_trained/checkpoint-198
(score: 0.06338918209075928).
Saving model checkpoint to
model_output_distilbert/winoqueer_gpt_augmentation_trained
Configuration saved in
model_output_distilbert/winoqueer_gpt_augmentation_trained/config.json
Model weights saved in
model_output_distilbert/winoqueer_gpt_augmentation_trained/model.safetensors
tokenizer config file saved in
model_output_distilbert/winoqueer_gpt_augmentation_trained/tokenizer_config.json
Special tokens file saved in model_output_distilbert/winoqueer_gpt_augmentation_
trained/special_tokens_map.json
loading configuration file
model_output_distilbert/winoqueer_gpt_augmentation_trained/config.json
Model config DistilBertConfig {
  "activation": "gelu",
  "architectures": [
    "DistilBertForSequenceClassification"
  ],
  "attention_dropout": 0.1,
  "dim": 768,
  "dropout": 0.1,
  "dtype": "float32",
  "hidden_dim": 3072,
  "initializer_range": 0.02,
  "max_position_embeddings": 512,
  "model_type": "distilbert",

```

```

    "n_heads": 12,
    "n_layers": 6,
    "pad_token_id": 0,
    "problem_type": "single_label_classification",
    "qa_dropout": 0.1,
    "seq_classif_dropout": 0.2,
    "sinusoidal_pos_embs": false,
    "tie_weights_": true,
    "transformers_version": "4.57.3",
    "vocab_size": 30522
}

```

```
loading weights file
model_output_distilbert/winoqueer_gpt_augmentation_trained/model.safetensors
loading file vocab.txt
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja
```

```
Saved fine-tuned model to:
model_output_distilbert/winoqueer_gpt_augmentation_trained
Number of unique labels: 2
```

```
Map: 100%|          | 653/653 [00:00<00:00, 37391.37 examples/s]
Map: 100%|          | 653/653 [00:00<00:00, 37079.54 examples/s]
Device set to use cuda:0
/tmp/hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.
  warnings.warn(
```

```
Sample tokenized input from test: {'text': 'most people who are Gay are kites',
'label': 0, 'group': 'Gay', 'data_name': 'Winoqueer - GPT Augmentation',
'__index_level_0__': 3236, 'input_ids': [101, 2087, 2111, 2040, 2024, 5637,
2024, 20497, 2015, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0], 'labels': 0}
```

```
loading configuration file
model_output_distilbert/winoqueer_gpt_augmentation_trained/config.json
Model config DistilBertConfig {
  "activation": "gelu",
  "architectures": [
    "DistilBertForSequenceClassification"
  ],
  "attention dropout": 0.1,
```

```

"dim": 768,
"dropout": 0.1,
"dtype": "float32",
"hidden_dim": 3072,
"initializer_range": 0.02,
"max_position_embeddings": 512,
"model_type": "distilbert",
"n_heads": 12,
"n_layers": 6,
"pad_token_id": 0,
"problem_type": "single_label_classification",
"qa_dropout": 0.1,
"seq_classif_dropout": 0.2,
"sinusoidal_pos_embs": false,
"tie_weights_": true,
"transformers_version": "4.57.3",
"vocab_size": 30522
}

loading weights file
model_output_distilbert/winoqueer_gpt_augmentation_trained/model.safetensors
loading file vocab.txt
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

Saved full results to: result_output_distilbert/winoqueer_gpt_augmentation_train
ed/winoqueer_gpt_augmentation/full_results.csv
Saved classification report to: result_output_distilbert/winoqueer_gpt_augmentat
ion_trained/winoqueer_gpt_augmentation/classification_report.csv
Number of unique labels: 2

Map: 100%|          | 414/414 [00:00<00:00, 34003.21 examples/s]
Map: 100%|          | 414/414 [00:00<00:00, 33467.13 examples/s]
Device set to use cuda:0

Sample tokenized input from test: {'text': 'The Somali refugees were labeled as
unacquainted by some', 'label': 0, 'group': 'Somalis', 'data_name': 'SeeGULL -
GPT Augmentation', '__index_level_0__': 840, 'input_ids': [101, 1996, 16831,
8711, 2020, 12599, 2004, 14477, 2278, 16211, 18447, 2098, 2011, 2070, 102, 0, 0,
0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0,
0, 0, 0, 0, 0], 'labels': 0}

/tmp/hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.

```



```

warnings.warn(
loading configuration file
model_output_distilbert/winoqueer_gpt_augmentation_trained/config.json
Model config DistilBertConfig {
  "activation": "gelu",
  "architectures": [
    "DistilBertForSequenceClassification"
  ],
  "attention_dropout": 0.1,
  "dim": 768,
  "dropout": 0.1,
  "dtype": "float32",
  "hidden_dim": 3072,
  "initializer_range": 0.02,
  "max_position_embeddings": 512,
  "model_type": "distilbert",
  "n_heads": 12,
  "n_layers": 6,
  "pad_token_id": 0,
  "problem_type": "single_label_classification",
  "qa_dropout": 0.1,
  "seq_classif_dropout": 0.2,
  "sinusoidal_pos_embs": false,
  "tie_weights_": true,
  "transformers_version": "4.57.3",
  "vocab_size": 30522
}

loading weights file
model_output_distilbert/winoqueer_gpt_augmentation_trained/model.safetensors
loading file vocab.txt
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

Saved full results to: result_output_distilbert/winoqueer_gpt_augmentation_train
ed/seegull_gpt_augmentation/full_results.csv
Saved classification report to: result_output_distilbert/winoqueer_gpt_augmentat
ion_trained/seegull_gpt_augmentation/classification_report.csv
Number of unique labels: 2

Map: 100%|          | 10374/10374 [00:00<00:00, 31307.60 examples/s]
Map: 100%|          | 10374/10374 [00:00<00:00, 40814.33 examples/s]
Device set to use cuda:0
/tmp/hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use

```



```

ed/mgsd/full_results.csv
Saved classification report to: result_output_distilbert/winoqueer_gpt_augmentat
ion_trained/mgsd/classification_report.csv
Number of unique labels: 2

Map: 100%|          | 11441/11441 [00:00<00:00, 24977.79 examples/s]
Map: 100%|          | 11441/11441 [00:00<00:00, 45280.82 examples/s]
Device set to use cuda:0
/tmp/hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.
  warnings.warn(

Sample tokenized input from test: {'text': 'The Somali refugees were labeled as
unacquainted by some', 'label': 0, 'group': 'Somalis', 'data_name': 'SeeGULL -
GPT Augmentation', 'input_ids': [101, 1996, 16831, 8711, 2020, 12599, 2004,
14477, 2278, 16211, 18447, 2098, 2011, 2070, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0], 'labels': 0}

loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--
distilbert-base-
uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/config.json
Model config DistilBertConfig {
  "activation": "gelu",
  "architectures": [
    "DistilBertForMaskedLM"
  ],
  "attention_dropout": 0.1,
  "dim": 768,
  "dropout": 0.1,
  "hidden_dim": 3072,
  "initializer_range": 0.02,
  "max_position_embeddings": 512,
  "model_type": "distilbert",
  "n_heads": 12,
  "n_layers": 6,
  "pad_token_id": 0,
  "qa_dropout": 0.1,
  "seq_classif_dropout": 0.2,
  "sinusoidal_pos_embs": false,
  "tie_weights_": true,
  "transformers_version": "4.57.3",
  "vocab_size": 30522
}

```

loading weights file model.safetensors from cache at
 /cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--
 distilbert-base-
 uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/model.safetensors
 Some weights of the model checkpoint at distilbert/distilbert-base-uncased were
 not used when initializing DistilBertForSequenceClassification:
 ['vocab_layer_norm.bias', 'vocab_layer_norm.weight', 'vocab_projector.bias',
 'vocab_transform.bias', 'vocab_transform.weight']
 - This IS expected if you are initializing DistilBertForSequenceClassification
 from the checkpoint of a model trained on another task or with another
 architecture (e.g. initializing a BertForSequenceClassification model from a
 BertForPreTraining model).
 - This IS NOT expected if you are initializing
 DistilBertForSequenceClassification from the checkpoint of a model that you
 expect to be exactly identical (initializing a BertForSequenceClassification
 model from a BertForSequenceClassification model).
 Some weights of DistilBertForSequenceClassification were not initialized from
 the model checkpoint at distilbert/distilbert-base-uncased and are newly
 initialized: ['classifier.bias', 'classifier.weight', 'pre_classifier.bias',
 'pre_classifier.weight']
 You should probably TRAIN this model on a down-stream task to be able to use it
 for predictions and inference.

Saved full results to: result_output_distilbert/winoqueer_gpt_augmentation_train
 ed/merged_winoqueer_seegull_gpt_augmentation/full_results.csv
 Saved classification report to: result_output_distilbert/winoqueer_gpt_augmentat
 ion_trained/merged_winoqueer_seegull_gpt_augmentation/classification_report.csv
 Number of unique labels: 2

loading configuration file config.json from cache at
 /cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--
 distilbert-base-
 uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/config.json

Model config DistilBertConfig {
 "activation": "gelu",
 "architectures": [
 "DistilBertForMaskedLM"
],
 "attention_dropout": 0.1,
 "dim": 768,
 "dropout": 0.1,
 "hidden_dim": 3072,
 "initializer_range": 0.02,
 "max_position_embeddings": 512,
 "model_type": "distilbert",
 "n_heads": 12,
 "n_layers": 6,
 "pad_token_id": 0,
 "qa_dropout": 0.1,

```

    "seq_classif_dropout": 0.2,
    "sinusoidal_pos_embs": false,
    "tie_weights_": true,
    "transformers_version": "4.57.3",
    "vocab_size": 30522
}

```

```

loading file vocab.txt from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--
distilbert-base-
uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/vocab.txt
loading file tokenizer.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--
distilbert-base-
uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/tokenizer.json
loading file added_tokens.json from cache at None
loading file special_tokens_map.json from cache at None
loading file tokenizer_config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--
distilbert-base-
uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/tokenizer_config.json
loading file chat_template.jinja from cache at None
loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--
distilbert-base-
uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/config.json
Model config DistilBertConfig {
  "activation": "gelu",
  "architectures": [
    "DistilBertForMaskedLM"
  ],
  "attention_dropout": 0.1,
  "dim": 768,
  "dropout": 0.1,
  "hidden_dim": 3072,
  "initializer_range": 0.02,
  "max_position_embeddings": 512,
  "model_type": "distilbert",
  "n_heads": 12,
  "n_layers": 6,
  "pad_token_id": 0,
  "qa_dropout": 0.1,
  "seq_classif_dropout": 0.2,
  "sinusoidal_pos_embs": false,
  "tie_weights_": true,
  "transformers_version": "4.57.3",
  "vocab_size": 30522
}

```

```
Map: 100%|          | 1324/1324 [00:00<00:00, 40238.67 examples/s]
Map: 100%|          | 1324/1324 [00:00<00:00, 39438.24 examples/s]
```

```
Sample tokenized input from train: {'text': 'The average Chadian income is
modest compared to developed countries', 'label': 0, 'group': 'Chadian',
'data_name': 'SeeGULL - GPT Augmentation', '__index_level_0__': 1249,
'input_ids': [101, 1996, 2779, 9796, 2937, 3318, 2003, 10754, 4102, 2000, 2764,
3032, 102, 0, 0, 0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 0, 0, 0, 0, 0, 0], 'labels': 0}
```

```
Map: 100%|          | 332/332 [00:00<00:00, 35049.31 examples/s]
Map: 100%|          | 332/332 [00:00<00:00, 33797.12 examples/s]
```

PyTorch: setting up devices

```
Model Training and Evaluation/BERT_Models_Fine_Tuning2.py:209: FutureWarning:
`tokenizer` is deprecated and will be removed in version 5.0.0 for
`Trainer.__init__`. Use `processing_class` instead.
    trainer = Trainer(
```

```
Sample tokenized input from validation: {'text': 'The Ethiopian woman was short
and petite', 'label': 1, 'group': 'Ethiopian', 'data_name': 'SeeGULL - GPT
Augmentation', '__index_level_0__': 609, 'input_ids': [101, 1996, 15101, 2450,
2001, 2460, 1998, 20146, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'labels': 1}
```

The following columns in the Training set don't have a corresponding argument in `DistilBertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `DistilBertForSequenceClassification.forward`, you can safely ignore this message.

***** Running training *****

```
    Num examples = 1,324
    Num Epochs = 6
    Instantaneous batch size per device = 64
    Total train batch size (w. parallel, distributed & accumulation) = 64
    Gradient Accumulation steps = 1
    Total optimization steps = 126
    Number of trainable parameters = 66,955,010
```

<IPython.core.display.HTML object>

The following columns in the Evaluation set don't have a corresponding argument in `DistilBertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `DistilBertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

```
    Num examples = 332
    Batch size = 64
```

```

/tmp/hearts_gpu/lib64/python3.9/site-
packages/sklearn/metrics/_classification.py:1565: UndefinedMetricWarning:
Precision is ill-defined and being set to 0.0 in labels with no predicted
samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
Saving model checkpoint to
model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-21
Configuration saved in model_output_distilbert/seegull_gpt_augmentation_trained/
checkpoint-21/config.json
Model weights saved in model_output_distilbert/seegull_gpt_augmentation_trained/
checkpoint-21/model.safetensors
tokenizer config file saved in model_output_distilbert/seegull_gpt_augmentation_
trained/checkpoint-21/tokenizer_config.json
Special tokens file saved in model_output_distilbert/seegull_gpt_augmentation_tr
ained/checkpoint-21/special_tokens_map.json
The following columns in the Evaluation set don't have a corresponding argument
in `DistilBertForSequenceClassification.forward` and have been ignored: group,
__index_level_0__, data_name, text. If group, __index_level_0__, data_name, text
are not expected by `DistilBertForSequenceClassification.forward`, you can
safely ignore this message.

```

***** Running Evaluation *****

```

  Num examples = 332
  Batch size = 64
Saving model checkpoint to
model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-42
Configuration saved in model_output_distilbert/seegull_gpt_augmentation_trained/
checkpoint-42/config.json
Model weights saved in model_output_distilbert/seegull_gpt_augmentation_trained/
checkpoint-42/model.safetensors
tokenizer config file saved in model_output_distilbert/seegull_gpt_augmentation_
trained/checkpoint-42/tokenizer_config.json
Special tokens file saved in model_output_distilbert/seegull_gpt_augmentation_tr
ained/checkpoint-42/special_tokens_map.json
Deleting older checkpoint
[model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-21] due to
args.save_total_limit
The following columns in the Evaluation set don't have a corresponding argument
in `DistilBertForSequenceClassification.forward` and have been ignored: group,
__index_level_0__, data_name, text. If group, __index_level_0__, data_name, text
are not expected by `DistilBertForSequenceClassification.forward`, you can
safely ignore this message.

```

***** Running Evaluation *****

```

  Num examples = 332
  Batch size = 64
Saving model checkpoint to
model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-63

```

Configuration saved in model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-63/config.json
Model weights saved in model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-63/model.safetensors
tokenizer config file saved in model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-63/tokenizer_config.json
Special tokens file saved in model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-63/special_tokens_map.json
Deleting older checkpoint
[model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-42] due to args.save_total_limit
The following columns in the Evaluation set don't have a corresponding argument in `DistilBertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `DistilBertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 332

Batch size = 64

Saving model checkpoint to

model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-84

Configuration saved in model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-84/config.json

Model weights saved in model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-84/model.safetensors

tokenizer config file saved in model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-84/tokenizer_config.json

Special tokens file saved in model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-84/special_tokens_map.json

Deleting older checkpoint

[model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-63] due to args.save_total_limit

The following columns in the Evaluation set don't have a corresponding argument in `DistilBertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `DistilBertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 332

Batch size = 64

Saving model checkpoint to

model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-105

Configuration saved in model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-105/config.json

Model weights saved in model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-105/model.safetensors

tokenizer config file saved in model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-105/tokenizer_config.json
Special tokens file saved in model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-105/special_tokens_map.json
Deleting older checkpoint
[model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-84] due to args.save_total_limit
The following columns in the Evaluation set don't have a corresponding argument in `DistilBertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `DistilBertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 332

Batch size = 64

Saving model checkpoint to

model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-126

Configuration saved in model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-126/config.json

Model weights saved in model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-126/model.safetensors

tokenizer config file saved in model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-126/tokenizer_config.json

Special tokens file saved in model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-126/special_tokens_map.json

Training completed. Do not forget to share your model on huggingface.co/models
=)

Loading best model from

model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-105 (score: 0.22793379426002502).

Deleting older checkpoint

[model_output_distilbert/seegull_gpt_augmentation_trained/checkpoint-126] due to args.save_total_limit

Saving model checkpoint to

model_output_distilbert/seegull_gpt_augmentation_trained

Configuration saved in

model_output_distilbert/seegull_gpt_augmentation_trained/config.json

Model weights saved in

model_output_distilbert/seegull_gpt_augmentation_trained/model.safetensors

tokenizer config file saved in

model_output_distilbert/seegull_gpt_augmentation_trained/tokenizer_config.json

Special tokens file saved in

model_output_distilbert/seegull_gpt_augmentation_trained/special_tokens_map.json

```

loading configuration file
model_output_distilbert/seegull_gpt_augmentation_trained/config.json
Model config DistilBertConfig {
  "activation": "gelu",
  "architectures": [
    "DistilBertForSequenceClassification"
  ],
  "attention_dropout": 0.1,
  "dim": 768,
  "dropout": 0.1,
  "dtype": "float32",
  "hidden_dim": 3072,
  "initializer_range": 0.02,
  "max_position_embeddings": 512,
  "model_type": "distilbert",
  "n_heads": 12,
  "n_layers": 6,
  "pad_token_id": 0,
  "problem_type": "single_label_classification",
  "qa_dropout": 0.1,
  "seq_classif_dropout": 0.2,
  "sinusoidal_pos_embs": false,
  "tie_weights_": true,
  "transformers_version": "4.57.3",
  "vocab_size": 30522
}

loading weights file
model_output_distilbert/seegull_gpt_augmentation_trained/model.safetensors
loading file vocab.txt
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

Saved fine-tuned model to:
model_output_distilbert/seegull_gpt_augmentation_trained
Number of unique labels: 2

Map: 100%|          | 653/653 [00:00<00:00, 36814.89 examples/s]
Map: 100%|          | 653/653 [00:00<00:00, 36601.37 examples/s]
Device set to use cuda:0
/tmp/hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.
  warnings.warn(

```

```

Sample tokenized input from test: {'text': 'most people who are Gay are kites',
'label': 0, 'group': 'Gay', 'data_name': 'Winoqueer - GPT Augmentation',
'__index_level_0__': 3236, 'input_ids': [101, 2087, 2111, 2040, 2024, 5637,
2024, 20497, 2015, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0], 'labels': 0}

```

loading configuration file

model_output_distilbert/seegull_gpt_augmentation_trained/config.json

Model config DistilBertConfig {

```

    "activation": "gelu",
    "architectures": [
        "DistilBertForSequenceClassification"
    ],
    "attention_dropout": 0.1,
    "dim": 768,
    "dropout": 0.1,
    "dtype": "float32",
    "hidden_dim": 3072,
    "initializer_range": 0.02,
    "max_position_embeddings": 512,
    "model_type": "distilbert",
    "n_heads": 12,
    "n_layers": 6,
    "pad_token_id": 0,
    "problem_type": "single_label_classification",
    "qa_dropout": 0.1,
    "seq_classif_dropout": 0.2,
    "sinusoidal_pos_embs": false,
    "tie_weights_": true,
    "transformers_version": "4.57.3",
    "vocab_size": 30522

```

}

loading weights file

model_output_distilbert/seegull_gpt_augmentation_trained/model.safetensors

loading file vocab.txt

loading file tokenizer.json

loading file added_tokens.json

loading file special_tokens_map.json

loading file tokenizer_config.json

loading file chat_template.jinja

Saved full results to: result_output_distilbert/seegull_gpt_augmentation_trained/winoqueer_gpt_augmentation/full_results.csv

Saved classification report to: result_output_distilbert/seegull_gpt_augmentation_trained/winoqueer_gpt_augmentation/classification_report.csv

Number of unique labels: 2

```
Map: 100%|          | 414/414 [00:00<00:00, 33272.18 examples/s]
Map: 100%|          | 414/414 [00:00<00:00, 33910.91 examples/s]
Device set to use cuda:0
```

```
Sample tokenized input from test: {'text': 'The Somali refugees were labeled as
unacquainted by some', 'label': 0, 'group': 'Somalis', 'data_name': 'SeeGULL -
GPT Augmentation', '__index_level_0__': 840, 'input_ids': [101, 1996, 16831,
8711, 2020, 12599, 2004, 14477, 2278, 16211, 18447, 2098, 2011, 2070, 102, 0, 0,
0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0,
0, 0, 0, 0, 0], 'labels': 0}
```

```
/tmp/hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.
```

```
warnings.warn(
loading configuration file
model_output_distilbert/seegull_gpt_augmentation_trained/config.json
Model config DistilBertConfig {
```

```
  "activation": "gelu",
  "architectures": [
    "DistilBertForSequenceClassification"
  ],
  "attention_dropout": 0.1,
  "dim": 768,
  "dropout": 0.1,
  "dtype": "float32",
  "hidden_dim": 3072,
  "initializer_range": 0.02,
  "max_position_embeddings": 512,
  "model_type": "distilbert",
  "n_heads": 12,
  "n_layers": 6,
  "pad_token_id": 0,
  "problem_type": "single_label_classification",
  "qa_dropout": 0.1,
  "seq_classif_dropout": 0.2,
  "sinusoidal_pos_embs": false,
  "tie_weights_": true,
  "transformers_version": "4.57.3",
  "vocab_size": 30522
}
```

```
loading weights file
model_output_distilbert/seegull_gpt_augmentation_trained/model.safetensors
loading file vocab.txt
loading file tokenizer.json
loading file added_tokens.json
```



```

    "tie_weights_": true,
    "transformers_version": "4.57.3",
    "vocab_size": 30522
}

loading weights file
model_output_distilbert/seegull_gpt_augmentation_trained/model.safetensors
loading file vocab.txt
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

Saved full results to:
result_output_distilbert/seegull_gpt_augmentation_trained/mgsd/full_results.csv
Saved classification report to: result_output_distilbert/seegull_gpt_augmentatio
n_trained/mgsd/classification_report.csv
Number of unique labels: 2

Map: 100%|          | 11441/11441 [00:00<00:00, 24997.10 examples/s]
Map: 100%|          | 11441/11441 [00:00<00:00, 44625.57 examples/s]
Device set to use cuda:0
/tmp/hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.
  warnings.warn(

Sample tokenized input from test: {'text': 'The Somali refugees were labeled as
unacquainted by some', 'label': 0, 'group': 'Somalis', 'data_name': 'SeeGULL -
GPT Augmentation', 'input_ids': [101, 1996, 16831, 8711, 2020, 12599, 2004,
14477, 2278, 16211, 18447, 2098, 2011, 2070, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0], 'labels': 0}

loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--
distilbert-base-
uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/config.json
Model config DistilBertConfig {
  "activation": "gelu",
  "architectures": [
    "DistilBertForMaskedLM"
  ],
  "attention_dropout": 0.1,
  "dim": 768,
  "dropout": 0.1,
  "hidden_dim": 3072,

```

```

"initializer_range": 0.02,
"max_position_embeddings": 512,
"model_type": "distilbert",
"n_heads": 12,
"n_layers": 6,
"pad_token_id": 0,
"qa_dropout": 0.1,
"seq_classif_dropout": 0.2,
"sinusoidal_pos_embs": false,
"tie_weights_": true,
"transformers_version": "4.57.3",
"vocab_size": 30522
}

```

loading weights file model.safetensors from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--
distilbert-base-
uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/model.safetensors
Some weights of the model checkpoint at distilbert/distilbert-base-uncased were
not used when initializing DistilBertForSequenceClassification:

```

['vocab_layer_norm.bias', 'vocab_layer_norm.weight', 'vocab_projector.bias',
'vocab_transform.bias', 'vocab_transform.weight']
- This IS expected if you are initializing DistilBertForSequenceClassification
from the checkpoint of a model trained on another task or with another
architecture (e.g. initializing a BertForSequenceClassification model from a
BertForPreTraining model).

```

```

- This IS NOT expected if you are initializing
DistilBertForSequenceClassification from the checkpoint of a model that you
expect to be exactly identical (initializing a BertForSequenceClassification
model from a BertForSequenceClassification model).

```

Some weights of DistilBertForSequenceClassification were not initialized from
the model checkpoint at distilbert/distilbert-base-uncased and are newly
initialized: ['classifier.bias', 'classifier.weight', 'pre_classifier.bias',
'pre_classifier.weight']

You should probably TRAIN this model on a down-stream task to be able to use it
for predictions and inference.

Saved full results to: result_output_distilbert/seegull_gpt_augmentation_trained
/merged_winoqueer_seegull_gpt_augmentation/full_results.csv

Saved classification report to: result_output_distilbert/seegull_gpt_augmentatio
n_trained/merged_winoqueer_seegull_gpt_augmentation/classification_report.csv

Number of unique labels: 2

loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--
distilbert-base-
uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/config.json

```

Model config DistilBertConfig {
  "activation": "gelu",

```

```

    "architectures": [
        "DistilBertForMaskedLM"
    ],
    "attention_dropout": 0.1,
    "dim": 768,
    "dropout": 0.1,
    "hidden_dim": 3072,
    "initializer_range": 0.02,
    "max_position_embeddings": 512,
    "model_type": "distilbert",
    "n_heads": 12,
    "n_layers": 6,
    "pad_token_id": 0,
    "qa_dropout": 0.1,
    "seq_classif_dropout": 0.2,
    "sinusoidal_pos_embs": false,
    "tie_weights": true,
    "transformers_version": "4.57.3",
    "vocab_size": 30522
}

```

```

loading file vocab.txt from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--
distilbert-base-
uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/vocab.txt
loading file tokenizer.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--
distilbert-base-
uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/tokenizer.json
loading file added_tokens.json from cache at None
loading file special_tokens_map.json from cache at None
loading file tokenizer_config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--
distilbert-base-
uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/tokenizer_config.json
loading file chat_template.jinja from cache at None
loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--distilbert--
distilbert-base-
uncased/snapshots/12040accade4e8a0f71eabdb258fecc2e7e948be/config.json
Model config DistilBertConfig {
    "activation": "gelu",
    "architectures": [
        "DistilBertForMaskedLM"
    ],
    "attention_dropout": 0.1,
    "dim": 768,
    "dropout": 0.1,

```



```

    "hidden_dim": 3072,
    "initializer_range": 0.02,
    "max_position_embeddings": 512,
    "model_type": "distilbert",
    "n_heads": 12,
    "n_layers": 6,
    "pad_token_id": 0,
    "qa_dropout": 0.1,
    "seq_classif_dropout": 0.2,
    "sinusoidal_pos_embs": false,
    "tie_weights_": true,
    "transformers_version": "4.57.3",
    "vocab_size": 30522
}

```

```

Map: 100%|          | 36608/36608 [00:01<00:00, 29421.41 examples/s]
Map: 100%|          | 36608/36608 [00:01<00:00, 33418.00 examples/s]

```

```

Sample tokenized input from train: {'text': 'The commander could tell the
soldier was disciplined.', 'label': 1, 'group': 'profession', 'data_name':
'MGSD', '__index_level_0__': 7380, 'input_ids': [101, 1996, 3474, 2071, 2425,
1996, 5268, 2001, 28675, 1012, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0], 'labels': 1}

```

```

Map: 100%|          | 9152/9152 [00:00<00:00, 31388.93 examples/s]
Map: 100%|          | 9152/9152 [00:00<00:00, 28018.19 examples/s]

```

PyTorch: setting up devices

```

Model Training and Evaluation/BERT_Models_Fine_Tuning2.py:209: FutureWarning:
`tokenizer` is deprecated and will be removed in version 5.0.0 for
`Trainer.__init__`. Use `processing_class` instead.
    trainer = Trainer(

```

```

Sample tokenized input from validation: {'text': 'we expect that Queer people
are artistic.', 'label': 0, 'group': 'Queer', 'data_name': 'Winoqueer - GPT
Augmentation', '__index_level_0__': 3705, 'input_ids': [101, 2057, 5987, 2008,
19483, 2111, 2024, 6018, 1012, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0], 'labels': 0}

```

The following columns in the Training set don't have a corresponding argument in `DistilBertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `DistilBertForSequenceClassification.forward`, you can safely ignore this message.

***** Running training *****

```
Num examples = 36,608
Num Epochs = 6
Instantaneous batch size per device = 64
Total train batch size (w. parallel, distributed & accumulation) = 64
Gradient Accumulation steps = 1
Total optimization steps = 3,432
Number of trainable parameters = 66,955,010
```

<IPython.core.display.HTML object>

The following columns in the Evaluation set don't have a corresponding argument in `DistilBertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `DistilBertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

```
Num examples = 9152
Batch size = 64
Saving model checkpoint to model_output_distilbert/merged_winoqueer_seegull_gpt_
augmentation_trained/checkpoint-572
Configuration saved in model_output_distilbert/merged_winoqueer_seegull_gpt_augm
entation_trained/checkpoint-572/config.json
Model weights saved in model_output_distilbert/merged_winoqueer_seegull_gpt_augm
entation_trained/checkpoint-572/model.safetensors
tokenizer config file saved in model_output_distilbert/merged_winoqueer_seegull_
gpt_augmentation_trained/checkpoint-572/tokenizer_config.json
Special tokens file saved in model_output_distilbert/merged_winoqueer_seegull_gp
t_augmentation_trained/checkpoint-572/special_tokens_map.json
The following columns in the Evaluation set don't have a corresponding argument
in `DistilBertForSequenceClassification.forward` and have been ignored: group,
__index_level_0__, data_name, text. If group, __index_level_0__, data_name, text
are not expected by `DistilBertForSequenceClassification.forward`, you can
safely ignore this message.
```

***** Running Evaluation *****

```
Num examples = 9152
Batch size = 64
Saving model checkpoint to model_output_distilbert/merged_winoqueer_seegull_gpt_
augmentation_trained/checkpoint-1144
Configuration saved in model_output_distilbert/merged_winoqueer_seegull_gpt_augm
entation_trained/checkpoint-1144/config.json
Model weights saved in model_output_distilbert/merged_winoqueer_seegull_gpt_augm
entation_trained/checkpoint-1144/model.safetensors
tokenizer config file saved in model_output_distilbert/merged_winoqueer_seegull_
gpt_augmentation_trained/checkpoint-1144/tokenizer_config.json
Special tokens file saved in model_output_distilbert/merged_winoqueer_seegull_gp
t_augmentation_trained/checkpoint-1144/special_tokens_map.json
Deleting older checkpoint [model_output_distilbert/merged_winoqueer_seegull_gpt_
```

augmentation_trained/checkpoint-572] due to args.save_total_limit
The following columns in the Evaluation set don't have a corresponding argument in `DistilBertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `DistilBertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 9152

Batch size = 64

Saving model checkpoint to model_output_distilbert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1716

Configuration saved in model_output_distilbert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1716/config.json

Model weights saved in model_output_distilbert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1716/model.safetensors

tokenizer config file saved in model_output_distilbert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1716/tokenizer_config.json

Special tokens file saved in model_output_distilbert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1716/special_tokens_map.json

The following columns in the Evaluation set don't have a corresponding argument in `DistilBertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `DistilBertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 9152

Batch size = 64

Saving model checkpoint to model_output_distilbert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2288

Configuration saved in model_output_distilbert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2288/config.json

Model weights saved in model_output_distilbert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2288/model.safetensors

tokenizer config file saved in model_output_distilbert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2288/tokenizer_config.json

Special tokens file saved in model_output_distilbert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2288/special_tokens_map.json

Deleting older checkpoint [model_output_distilbert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1716] due to args.save_total_limit

The following columns in the Evaluation set don't have a corresponding argument in `DistilBertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `DistilBertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

```
Num examples = 9152
Batch size = 64
Saving model checkpoint to model_output_distilbert/merged_winoqueer_seegull_gpt_
augmentation_trained/checkpoint-2860
Configuration saved in model_output_distilbert/merged_winoqueer_seegull_gpt_augm
entation_trained/checkpoint-2860/config.json
Model weights saved in model_output_distilbert/merged_winoqueer_seegull_gpt_augm
entation_trained/checkpoint-2860/model.safetensors
tokenizer config file saved in model_output_distilbert/merged_winoqueer_seegull_
gpt_augmentation_trained/checkpoint-2860/tokenizer_config.json
Special tokens file saved in model_output_distilbert/merged_winoqueer_seegull_gp
t_augmentation_trained/checkpoint-2860/special_tokens_map.json
Deleting older checkpoint [model_output_distilbert/merged_winoqueer_seegull_gpt_
augmentation_trained/checkpoint-2288] due to args.save_total_limit
The following columns in the Evaluation set don't have a corresponding argument
in `DistilBertForSequenceClassification.forward` and have been ignored: group,
__index_level_0__, data_name, text. If group, __index_level_0__, data_name, text
are not expected by `DistilBertForSequenceClassification.forward`, you can
safely ignore this message.
```

***** Running Evaluation *****

```
Num examples = 9152
Batch size = 64
Saving model checkpoint to model_output_distilbert/merged_winoqueer_seegull_gpt_
augmentation_trained/checkpoint-3432
Configuration saved in model_output_distilbert/merged_winoqueer_seegull_gpt_augm
entation_trained/checkpoint-3432/config.json
Model weights saved in model_output_distilbert/merged_winoqueer_seegull_gpt_augm
entation_trained/checkpoint-3432/model.safetensors
tokenizer config file saved in model_output_distilbert/merged_winoqueer_seegull_
gpt_augmentation_trained/checkpoint-3432/tokenizer_config.json
Special tokens file saved in model_output_distilbert/merged_winoqueer_seegull_gp
t_augmentation_trained/checkpoint-3432/special_tokens_map.json
Deleting older checkpoint [model_output_distilbert/merged_winoqueer_seegull_gpt_
augmentation_trained/checkpoint-2860] due to args.save_total_limit
```

Training completed. Do not forget to share your model on huggingface.co/models
=)

```
Loading best model from model_output_distilbert/merged_winoqueer_seegull_gpt_aug
mentation_trained/checkpoint-1144 (score: 0.3716101348400116).
Deleting older checkpoint [model_output_distilbert/merged_winoqueer_seegull_gpt_
augmentation_trained/checkpoint-3432] due to args.save_total_limit
Saving model checkpoint to
model_output_distilbert/merged_winoqueer_seegull_gpt_augmentation_trained
Configuration saved in model_output_distilbert/merged_winoqueer_seegull_gpt_augm
```

```

mentation_trained/config.json
Model weights saved in model_output_distilbert/merged_winoqueer_seegull_gpt_augm
mentation_trained/model.safetensors
tokenizer config file saved in model_output_distilbert/merged_winoqueer_seegull_
gpt_augmentation_trained/tokenizer_config.json
Special tokens file saved in model_output_distilbert/merged_winoqueer_seegull_gp
t_augmentation_trained/special_tokens_map.json
loading configuration file model_output_distilbert/merged_winoqueer_seegull_gpt_
augmentation_trained/config.json
Model config DistilBertConfig {
  "activation": "gelu",
  "architectures": [
    "DistilBertForSequenceClassification"
  ],
  "attention_dropout": 0.1,
  "dim": 768,
  "dropout": 0.1,
  "dtype": "float32",
  "hidden_dim": 3072,
  "initializer_range": 0.02,
  "max_position_embeddings": 512,
  "model_type": "distilbert",
  "n_heads": 12,
  "n_layers": 6,
  "pad_token_id": 0,
  "problem_type": "single_label_classification",
  "qa_dropout": 0.1,
  "seq_classif_dropout": 0.2,
  "sinusoidal_pos_embs": false,
  "tie_weights_": true,
  "transformers_version": "4.57.3",
  "vocab_size": 30522
}

```

```

loading weights file model_output_distilbert/merged_winoqueer_seegull_gpt_augmen
tation_trained/model.safetensors
loading file vocab.txt
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

```

Saved fine-tuned model to:

model_output_distilbert/merged_winoqueer_seegull_gpt_augmentation_trained

Number of unique labels: 2

```

Map: 100%|          | 653/653 [00:00<00:00, 38347.86 examples/s]
Map: 100%|          | 653/653 [00:00<00:00, 35948.50 examples/s]

```

```

Device set to use cuda:0
/tmp/hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.
  warnings.warn(

Sample tokenized input from test: {'text': 'most people who are Gay are kites',
'label': 0, 'group': 'Gay', 'data_name': 'Winoqueer - GPT Augmentation',
'__index_level_0__': 3236, 'input_ids': [101, 2087, 2111, 2040, 2024, 5637,
2024, 20497, 2015, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0], 'labels': 0}

loading configuration file model_output_distilbert/merged_winoqueer_seegull_gpt_
augmentation_trained/config.json
Model config DistilBertConfig {
  "activation": "gelu",
  "architectures": [
    "DistilBertForSequenceClassification"
  ],
  "attention_dropout": 0.1,
  "dim": 768,
  "dropout": 0.1,
  "dtype": "float32",
  "hidden_dim": 3072,
  "initializer_range": 0.02,
  "max_position_embeddings": 512,
  "model_type": "distilbert",
  "n_heads": 12,
  "n_layers": 6,
  "pad_token_id": 0,
  "problem_type": "single_label_classification",
  "qa_dropout": 0.1,
  "seq_classif_dropout": 0.2,
  "sinusoidal_pos_embs": false,
  "tie_weights_": true,
  "transformers_version": "4.57.3",
  "vocab_size": 30522
}

loading weights file model_output_distilbert/merged_winoqueer_seegull_gpt_augmen
tation_trained/model.safetensors
loading file vocab.txt
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json

```

loading file chat_template.jinja

Saved full results to: result_output_distilbert/merged_winoqueer_seegull_gpt_augmentation_trained/winoqueer_gpt_augmentation/full_results.csv

Saved classification report to: result_output_distilbert/merged_winoqueer_seegull_gpt_augmentation_trained/winoqueer_gpt_augmentation/classification_report.csv

Number of unique labels: 2

Map: 100%| | 414/414 [00:00<00:00, 34410.20 examples/s]

Map: 100%| | 414/414 [00:00<00:00, 34705.93 examples/s]

Device set to use cuda:0

Sample tokenized input from test: {'text': 'The Somali refugees were labeled as unacquainted by some', 'label': 0, 'group': 'Somalis', 'data_name': 'SeeGULL - GPT Augmentation', '__index_level_0__': 840, 'input_ids': [101, 1996, 16831, 8711, 2020, 12599, 2004, 14477, 2278, 16211, 18447, 2098, 2011, 2070, 102, 0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0], 'labels': 0}

/tmp/hearts_gpu/lib64/python3.9/site-

packages/transformers/pipelines/text_classification.py:111: UserWarning:

`return_all_scores` is now deprecated, if want a similar functionality use `top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of `return_all_scores=False`.

warnings.warn(

loading configuration file model_output_distilbert/merged_winoqueer_seegull_gpt_augmentation_trained/config.json

Model config DistilBertConfig {

 "activation": "gelu",

 "architectures": [

 "DistilBertForSequenceClassification"

],

 "attention_dropout": 0.1,

 "dim": 768,

 "dropout": 0.1,

 "dtype": "float32",

 "hidden_dim": 3072,

 "initializer_range": 0.02,

 "max_position_embeddings": 512,

 "model_type": "distilbert",

 "n_heads": 12,

 "n_layers": 6,

 "pad_token_id": 0,

 "problem_type": "single_label_classification",

 "qa_dropout": 0.1,

 "seq_classif_dropout": 0.2,

 "sinusoidal_pos_embs": false,

 "tie_weights_": true,

 "transformers_version": "4.57.3",

 "vocab_size": 30522

```

}

loading weights file model_output_distilbert/merged_winoqueer_seegull_gpt_augmen
tation_trained/model.safetensors
loading file vocab.txt
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

Saved full results to: result_output_distilbert/merged_winoqueer_seegull_gpt_aug
mentation_trained/seegull_gpt_augmentation/full_results.csv
Saved classification report to: result_output_distilbert/merged_winoqueer_seegul
l_gpt_augmentation_trained/seegull_gpt_augmentation/classification_report.csv
Number of unique labels: 2

Map: 100%|          | 10374/10374 [00:00<00:00, 31747.00 examples/s]
Map: 100%|          | 10374/10374 [00:00<00:00, 39195.75 examples/s]
Device set to use cuda:0
/tmp/hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.
  warnings.warn(

Sample tokenized input from test: {'group': 'race', 'text': 'When i see an arab,
i think they are a happy.', 'label': 0, 'data_name': 'MGSD',
'__index_level_0__': 17851, 'input_ids': [101, 2043, 1045, 2156, 2019, 5424,
1010, 1045, 2228, 2027, 2024, 1037, 3407, 1012, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'attention_mask': [1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'labels': 0}

loading configuration file model_output_distilbert/merged_winoqueer_seegull_gpt_
augmentation_trained/config.json
Model config DistilBertConfig {
  "activation": "gelu",
  "architectures": [
    "DistilBertForSequenceClassification"
  ],
  "attention_dropout": 0.1,
  "dim": 768,
  "dropout": 0.1,
  "dtype": "float32",
  "hidden_dim": 3072,
  "initializer_range": 0.02,
  "max_position_embeddings": 512,
  "model_type": "distilbert",

```



```

    "n_heads": 12,
    "n_layers": 6,
    "pad_token_id": 0,
    "problem_type": "single_label_classification",
    "qa_dropout": 0.1,
    "seq_classif_dropout": 0.2,
    "sinusoidal_pos_embs": false,
    "tie_weights_": true,
    "transformers_version": "4.57.3",
    "vocab_size": 30522
}

```

loading weights file model_output_distilbert/merged_winoqueer_seegull_gpt_augmentation_trained/model.safetensors

loading file vocab.txt

loading file tokenizer.json

loading file added_tokens.json

loading file special_tokens_map.json

loading file tokenizer_config.json

loading file chat_template.jinja

Saved full results to: result_output_distilbert/merged_winoqueer_seegull_gpt_augmentation_trained/mgsd/full_results.csv

Saved classification report to: result_output_distilbert/merged_winoqueer_seegull_gpt_augmentation_trained/mgsd/classification_report.csv

Number of unique labels: 2

Map: 100%| | 11441/11441 [00:00<00:00, 32604.98 examples/s]

Map: 100%| | 11441/11441 [00:00<00:00, 44820.72 examples/s]

Device set to use cuda:0

/tmp/hearts_gpu/lib64/python3.9/site-

packages/transformers/pipelines/text_classification.py:111: UserWarning:

`return_all_scores` is now deprecated, if want a similar functionality use

`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of

`return_all_scores=False`.

warnings.warn(

Sample tokenized input from test: {'text': 'The Somali refugees were labeled as unacquainted by some', 'label': 0, 'group': 'Somalis', 'data_name': 'SeeGULL - GPT Augmentation', 'input_ids': [101, 1996, 16831, 8711, 2020, 12599, 2004, 14477, 2278, 16211, 18447, 2098, 2011, 2070, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'labels': 0}

Saved full results to: result_output_distilbert/merged_winoqueer_seegull_gpt_augmentation_trained/merged_winoqueer_seegull_gpt_augmentation/full_results.csv

Saved classification report to: result_output_distilbert/merged_winoqueer_seegull_gpt_augmentation_trained/merged_winoqueer_seegull_gpt_augmentation/classification_report.csv

Number of unique labels: 2

loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--bert-base-uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/config.json

```
Model config BertConfig {
  "architectures": [
    "BertForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0.1,
  "classifier_dropout": null,
  "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": "4.57.3",
  "type_vocab_size": 2,
  "use_cache": true,
  "vocab_size": 30522
}
```

loading weights file model.safetensors from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--bert-base-

uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/model.safetensors

A pretrained model of type `BertForSequenceClassification` contains parameters that have been renamed internally (a few are listed below but more are present in the model):

```
* `cls.predictions.transform.LayerNorm.beta` ->
`cls.predictions.transform.LayerNorm.bias`
* `cls.predictions.transform.LayerNorm.gamma` ->
`cls.predictions.transform.LayerNorm.weight`
```

If you are using a model from the Hub, consider submitting a PR to adjust these weights and help future users.

Some weights of the model checkpoint at google-bert/bert-base-uncased were not used when initializing BertForSequenceClassification: ['cls.predictions.bias', 'cls.predictions.transform.LayerNorm.bias', 'cls.predictions.transform.LayerNorm.weight', 'cls.predictions.transform.dense.bias', 'cls.predictions.transform.dense.weight', 'cls.seq_relationship.bias', 'cls.seq_relationship.weight']

- This IS expected if you are initializing BertForSequenceClassification from the checkpoint of a model trained on another task or with another architecture (e.g. initializing a BertForSequenceClassification model from a BertForPreTraining model).

- This IS NOT expected if you are initializing BertForSequenceClassification from the checkpoint of a model that you expect to be exactly identical (initializing a BertForSequenceClassification model from a BertForSequenceClassification model).

Some weights of BertForSequenceClassification were not initialized from the model checkpoint at google-bert/bert-base-uncased and are newly initialized:

['classifier.bias', 'classifier.weight']

You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.

loading configuration file config.json from cache at

/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--bert-base-uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/config.json

```
Model config BertConfig {
  "architectures": [
    "BertForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0.1,
  "classifier_dropout": null,
  "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": "4.57.3",
  "type_vocab_size": 2,
  "use_cache": true,
  "vocab_size": 30522
}
```

loading file vocab.txt from cache at

/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--bert-base-uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/vocab.txt

loading file tokenizer.json from cache at

/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--bert-base-uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/tokenizer.json

```
loading file added_tokens.json from cache at None
loading file special_tokens_map.json from cache at None
loading file tokenizer_config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--
bert-base-
uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/tokenizer_config.json
loading file chat_template.jinja from cache at None
loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--
bert-base-uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/config.json
Model config BertConfig {
  "architectures": [
    "BertForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0.1,
  "classifier_dropout": null,
  "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": "4.57.3",
  "type_vocab_size": 2,
  "use_cache": true,
  "vocab_size": 30522
}
```

[illegible]

```

0, 0, 0, 0, 0, 0, 0, 0, 0], 'labels': 0}

Map: 100%|          | 8299/8299 [00:00<00:00, 28170.47 examples/s]
Map: 100%|          | 8299/8299 [00:00<00:00, 25554.07 examples/s]
PyTorch: setting up devices
Model Training and Evaluation/BERT_Models_Fine_Tuning2.py:209: FutureWarning:
`tokenizer` is deprecated and will be removed in version 5.0.0 for
`Trainer.__init__`. Use `processing_class` instead.
    trainer = Trainer(

Sample tokenized input from validation: {'group': 'race', 'text': 'My friend is
German and sometimes we hand out. He gets very angry sometimes.', 'label': 1,
'data_name': 'MGSD', '__index_level_0__': 36651, 'input_ids': [101, 2026, 2767,
2003, 2446, 1998, 2823, 2057, 2192, 2041, 1012, 2002, 4152, 2200, 4854, 2823,
1012, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0], 'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'labels': 1}

The following columns in the Training set don't have a corresponding argument in
`BertForSequenceClassification.forward` and have been ignored: group,
__index_level_0__, data_name, text. If group, __index_level_0__, data_name, text
are not expected by `BertForSequenceClassification.forward`, you can safely
ignore this message.
***** Running training *****
    Num examples = 33,194
    Num Epochs = 6
    Instantaneous batch size per device = 64
    Total train batch size (w. parallel, distributed & accumulation) = 64
    Gradient Accumulation steps = 1
    Total optimization steps = 3,114
    Number of trainable parameters = 109,483,778

<IPython.core.display.HTML object>

The following columns in the Evaluation set don't have a corresponding argument
in `BertForSequenceClassification.forward` and have been ignored: group,
__index_level_0__, data_name, text. If group, __index_level_0__, data_name, text
are not expected by `BertForSequenceClassification.forward`, you can safely
ignore this message.

***** Running Evaluation *****
    Num examples = 8299
    Batch size = 64
Saving model checkpoint to model_output_bert/mgsd_trained/checkpoint-519
Configuration saved in model_output_bert/mgsd_trained/checkpoint-519/config.json
Model weights saved in
model_output_bert/mgsd_trained/checkpoint-519/model.safetensors

```

tokenizer config file saved in
model_output_bert/mgsd_trained/checkpoint-519/tokenizer_config.json
Special tokens file saved in
model_output_bert/mgsd_trained/checkpoint-519/special_tokens_map.json
The following columns in the Evaluation set don't have a corresponding argument
in `BertForSequenceClassification.forward` and have been ignored: group,
__index_level_0__, data_name, text. If group, __index_level_0__, data_name, text
are not expected by `BertForSequenceClassification.forward`, you can safely
ignore this message.

***** Running Evaluation *****

Num examples = 8299

Batch size = 64

Saving model checkpoint to model_output_bert/mgsd_trained/checkpoint-1038

Configuration saved in

model_output_bert/mgsd_trained/checkpoint-1038/config.json

Model weights saved in

model_output_bert/mgsd_trained/checkpoint-1038/model.safetensors

tokenizer config file saved in

model_output_bert/mgsd_trained/checkpoint-1038/tokenizer_config.json

Special tokens file saved in

model_output_bert/mgsd_trained/checkpoint-1038/special_tokens_map.json

Deleting older checkpoint [model_output_bert/mgsd_trained/checkpoint-519] due to
args.save_total_limit

The following columns in the Evaluation set don't have a corresponding argument
in `BertForSequenceClassification.forward` and have been ignored: group,
__index_level_0__, data_name, text. If group, __index_level_0__, data_name, text
are not expected by `BertForSequenceClassification.forward`, you can safely
ignore this message.

***** Running Evaluation *****

Num examples = 8299

Batch size = 64

Saving model checkpoint to model_output_bert/mgsd_trained/checkpoint-1557

Configuration saved in

model_output_bert/mgsd_trained/checkpoint-1557/config.json

Model weights saved in

model_output_bert/mgsd_trained/checkpoint-1557/model.safetensors

tokenizer config file saved in

model_output_bert/mgsd_trained/checkpoint-1557/tokenizer_config.json

Special tokens file saved in

model_output_bert/mgsd_trained/checkpoint-1557/special_tokens_map.json

The following columns in the Evaluation set don't have a corresponding argument
in `BertForSequenceClassification.forward` and have been ignored: group,
__index_level_0__, data_name, text. If group, __index_level_0__, data_name, text
are not expected by `BertForSequenceClassification.forward`, you can safely
ignore this message.

```

***** Running Evaluation *****
  Num examples = 8299
  Batch size = 64
Saving model checkpoint to model_output_bert/mgsd_trained/checkpoint-2076
Configuration saved in
model_output_bert/mgsd_trained/checkpoint-2076/config.json
Model weights saved in
model_output_bert/mgsd_trained/checkpoint-2076/model.safetensors
tokenizer config file saved in
model_output_bert/mgsd_trained/checkpoint-2076/tokenizer_config.json
Special tokens file saved in
model_output_bert/mgsd_trained/checkpoint-2076/special_tokens_map.json
Deleting older checkpoint [model_output_bert/mgsd_trained/checkpoint-1557] due
to args.save_total_limit
The following columns in the Evaluation set don't have a corresponding argument
in `BertForSequenceClassification.forward` and have been ignored: group,
__index_level_0__, data_name, text. If group, __index_level_0__, data_name, text
are not expected by `BertForSequenceClassification.forward`, you can safely
ignore this message.

```

```

***** Running Evaluation *****
  Num examples = 8299
  Batch size = 64
Saving model checkpoint to model_output_bert/mgsd_trained/checkpoint-2595
Configuration saved in
model_output_bert/mgsd_trained/checkpoint-2595/config.json
Model weights saved in
model_output_bert/mgsd_trained/checkpoint-2595/model.safetensors
tokenizer config file saved in
model_output_bert/mgsd_trained/checkpoint-2595/tokenizer_config.json
Special tokens file saved in
model_output_bert/mgsd_trained/checkpoint-2595/special_tokens_map.json
Deleting older checkpoint [model_output_bert/mgsd_trained/checkpoint-2076] due
to args.save_total_limit
The following columns in the Evaluation set don't have a corresponding argument
in `BertForSequenceClassification.forward` and have been ignored: group,
__index_level_0__, data_name, text. If group, __index_level_0__, data_name, text
are not expected by `BertForSequenceClassification.forward`, you can safely
ignore this message.

```

```

***** Running Evaluation *****
  Num examples = 8299
  Batch size = 64
Saving model checkpoint to model_output_bert/mgsd_trained/checkpoint-3114
Configuration saved in
model_output_bert/mgsd_trained/checkpoint-3114/config.json
Model weights saved in
model_output_bert/mgsd_trained/checkpoint-3114/model.safetensors

```

tokenizer config file saved in
model_output_bert/mgsd_trained/checkpoint-3114/tokenizer_config.json
Special tokens file saved in
model_output_bert/mgsd_trained/checkpoint-3114/special_tokens_map.json
Deleting older checkpoint [model_output_bert/mgsd_trained/checkpoint-2595] due
to args.save_total_limit

Training completed. Do not forget to share your model on huggingface.co/models
=)

Loading best model from model_output_bert/mgsd_trained/checkpoint-1038 (score:
0.3644278049468994).

Deleting older checkpoint [model_output_bert/mgsd_trained/checkpoint-3114] due
to args.save_total_limit

Saving model checkpoint to model_output_bert/mgsd_trained

Configuration saved in model_output_bert/mgsd_trained/config.json

Model weights saved in model_output_bert/mgsd_trained/model.safetensors

tokenizer config file saved in

model_output_bert/mgsd_trained/tokenizer_config.json

Special tokens file saved in

model_output_bert/mgsd_trained/special_tokens_map.json

loading configuration file model_output_bert/mgsd_trained/config.json

Model config BertConfig {

```
"architectures": [
  "BertForSequenceClassification"
],
"attention_probs_dropout_prob": 0.1,
"classifier_dropout": null,
"dtype": "float32",
"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",
"problem_type": "single_label_classification",
"transformers_version": "4.57.3",
"type_vocab_size": 2,
"use_cache": true,
```



```

    "vocab_size": 30522
}

loading weights file model_output_bert/mgsd_trained/model.safetensors
loading file vocab.txt
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

Saved fine-tuned model to: model_output_bert/mgsd_trained
Number of unique labels: 2

Map: 100%|          | 653/653 [00:00<00:00, 33534.30 examples/s]
Map: 100%|          | 653/653 [00:00<00:00, 32467.73 examples/s]
Device set to use cuda:0
/tmp/hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.
  warnings.warn(

Sample tokenized input from test: {'text': 'most people who are Gay are kites',
'label': 0, 'group': 'Gay', 'data_name': 'Winoqueer - GPT Augmentation',
'__index_level_0__': 3236, 'input_ids': [101, 2087, 2111, 2040, 2024, 5637,
2024, 20497, 2015, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'labels': 0}

loading configuration file model_output_bert/mgsd_trained/config.json
Model config BertConfig {
  "architectures": [
    "BertForSequenceClassification"
  ],
  "attention_probs_dropout_prob": 0.1,
  "classifier_dropout": null,
  "dtype": "float32",
  "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",
    "problem_type": "single_label_classification",
    "transformers_version": "4.57.3",
    "type_vocab_size": 2,
    "use_cache": true,
    "vocab_size": 30522
}

```

```

loading weights file model_output_bert/mgsd_trained/model.safetensors
loading file vocab.txt
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

```

Saved full results to:

result_output_bert/mgsd_trained/winoqueer_gpt_augmentation/full_results.csv

Saved classification report to: result_output_bert/mgsd_trained/winoqueer_gpt_augmentation/classification_report.csv

Number of unique labels: 2

Map: 100%| | 414/414 [00:00<00:00, 30932.22 examples/s]

Map: 100%| | 414/414 [00:00<00:00, 30669.97 examples/s]

Device set to use cuda:0

Sample tokenized input from test: {'text': 'The Somali refugees were labeled as unacquainted by some', 'label': 0, 'group': 'Somalis', 'data_name': 'SeeGULL - GPT Augmentation', '__index_level_0__': 840, 'input_ids': [101, 1996, 16831, 8711, 2020, 12599, 2004, 14477, 2278, 16211, 18447, 2098, 2011, 2070, 102, 0, 0, 0, 0, 0], 'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0], 'labels': 0}

/tmp/hearts_gpu/lib64/python3.9/site-

packages/transformers/pipelines/text_classification.py:111: UserWarning:

`return_all_scores` is now deprecated, if want a similar functionality use `top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of `return_all_scores=False`.

warnings.warn(

loading configuration file model_output_bert/mgsd_trained/config.json

Model config BertConfig {

```

    "architectures": [
        "BertForSequenceClassification"
    ],

```

```

    "attention_probs_dropout_prob": 0.1,
    "classifier_dropout": null,

```

```

    "dtype": "float32",

```



```
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,  
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1,  
1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,  
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'labels': 0}
```

```
Model config BertConfig {
  "architectures": [
    "BertForSequenceClassification"
  ],
  "attention_probs_dropout_prob": 0.1,
  "classifier_dropout": null,
  "dtype": "float32",
  "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",
  "problem_type": "single_label_classification",
  "transformers_version": "4.57.3",
  "type_vocab_size": 2,
  "use_cache": true,
  "vocab_size": 30522
}
```

```
loading file vocab.txt
```

```
loading file added_tokens.json
```

```
loading file tokenizer_config.json
```

Saved full results to: result_output_bert/mgsd_trained/mgsd/full_results.csv

```
result output bert/mgsd trained/mgsd/classification report.csv
```

```
Map: 100%|          | 11441/11441 [00:00<00:00, 29222.57 examples/s]
```

```
Device set to use cuda:0
```

```

/tmp/hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.
  warnings.warn(

Sample tokenized input from test: {'text': 'The Somali refugees were labeled as
unacquainted by some', 'label': 0, 'group': 'Somalis', 'data_name': 'SeeGULL -
GPT Augmentation', 'input_ids': [101, 1996, 16831, 8711, 2020, 12599, 2004,
14477, 2278, 16211, 18447, 2098, 2011, 2070, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0], 'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'labels': 0}
Saved full results to: result_output_bert/mgsd_trained/merged_winoqueer_seegull_
gpt_augmentation/full_results.csv
Saved classification report to: result_output_bert/mgsd_trained/merged_winoqueer
_seegull_gpt_augmentation/classification_report.csv
Number of unique labels: 2

loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--
bert-base-uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/config.json
Model config BertConfig {
  "architectures": [
    "BertForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0.1,
  "classifier_dropout": null,
  "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": "4.57.3",
  "type_vocab_size": 2,
  "use_cache": true,
  "vocab_size": 30522
}

loading weights file model.safetensors from cache at

```

```

/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--
bert-base-
uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/model.safetensors
Some weights of the model checkpoint at google-bert/bert-base-uncased were not
used when initializing BertForSequenceClassification: ['cls.predictions.bias',
'cls.predictions.transform.LayerNorm.bias',
'cls.predictions.transform.LayerNorm.weight',
'cls.predictions.transform.dense.bias',
'cls.predictions.transform.dense.weight', 'cls.seq_relationship.bias',
'cls.seq_relationship.weight']
- This IS expected if you are initializing BertForSequenceClassification from
the checkpoint of a model trained on another task or with another architecture
(e.g. initializing a BertForSequenceClassification model from a
BertForPreTraining model).
- This IS NOT expected if you are initializing BertForSequenceClassification
from the checkpoint of a model that you expect to be exactly identical
(initializing a BertForSequenceClassification model from a
BertForSequenceClassification model).
Some weights of BertForSequenceClassification were not initialized from the
model checkpoint at google-bert/bert-base-uncased and are newly initialized:
['classifier.bias', 'classifier.weight']
You should probably TRAIN this model on a down-stream task to be able to use it
for predictions and inference.
loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--
bert-base-uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/config.json
Model config BertConfig {
  "architectures": [
    "BertForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0.1,
  "classifier_dropout": null,
  "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": "4.57.3",
  "type_vocab_size": 2,
  "use_cache": true,

```

```

    "vocab_size": 30522
}

loading file vocab.txt from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--
bert-base-uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/vocab.txt
loading file tokenizer.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--
bert-base-
uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/tokenizer.json
loading file added_tokens.json from cache at None
loading file special_tokens_map.json from cache at None
loading file tokenizer_config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--
bert-base-
uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/tokenizer_config.json
loading file chat_template.jinja from cache at None
loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--
bert-base-uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/config.json
Model config BertConfig {
  "architectures": [
    "BertForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0.1,
  "classifier_dropout": null,
  "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": "4.57.3",
  "type_vocab_size": 2,
  "use_cache": true,
  "vocab_size": 30522
}

Map: 100%|          | 2088/2088 [00:00<00:00, 40277.73 examples/s]
Map: 100%|          | 2088/2088 [00:00<00:00, 35950.14 examples/s]

```


Batch size = 64
Saving model checkpoint to
model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-33
Configuration saved in
model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-33/config.json
Model weights saved in model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-33/model.safetensors
tokenizer config file saved in model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-33/tokenizer_config.json
Special tokens file saved in model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-33/special_tokens_map.json
The following columns in the Evaluation set don't have a corresponding argument in `BertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `BertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 523
Batch size = 64
Saving model checkpoint to
model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-66
Configuration saved in
model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-66/config.json
Model weights saved in model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-66/model.safetensors
tokenizer config file saved in model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-66/tokenizer_config.json
Special tokens file saved in model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-66/special_tokens_map.json
Deleting older checkpoint
[model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-33] due to args.save_total_limit
The following columns in the Evaluation set don't have a corresponding argument in `BertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `BertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 523
Batch size = 64
Saving model checkpoint to
model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-99
Configuration saved in
model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-99/config.json
Model weights saved in model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-99/model.safetensors

tokenizer config file saved in model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-99/tokenizer_config.json
Special tokens file saved in model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-99/special_tokens_map.json
Deleting older checkpoint
[model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-66] due to args.save_total_limit
The following columns in the Evaluation set don't have a corresponding argument in `BertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `BertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 523

Batch size = 64

Saving model checkpoint to

model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-132

Configuration saved in

model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-132/config.json

Model weights saved in model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-132/model.safetensors

tokenizer config file saved in model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-132/tokenizer_config.json

Special tokens file saved in model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-132/special_tokens_map.json

Deleting older checkpoint

[model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-99] due to args.save_total_limit

The following columns in the Evaluation set don't have a corresponding argument in `BertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `BertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 523

Batch size = 64

Saving model checkpoint to

model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-165

Configuration saved in

model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-165/config.json

Model weights saved in model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-165/model.safetensors

tokenizer config file saved in model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-165/tokenizer_config.json

Special tokens file saved in model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-165/special_tokens_map.json

The following columns in the Evaluation set don't have a corresponding argument in `BertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `BertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 523

Batch size = 64

Saving model checkpoint to

model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-198

Configuration saved in

model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-198/config.json

Model weights saved in model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-198/model.safetensors

tokenizer config file saved in model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-198/tokenizer_config.json

Special tokens file saved in model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-198/special_tokens_map.json

Deleting older checkpoint

[model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-165] due to args.save_total_limit

Training completed. Do not forget to share your model on huggingface.co/models =>

Loading best model from

model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-132 (score: 0.03348271921277046).

Deleting older checkpoint

[model_output_bert/winoqueer_gpt_augmentation_trained/checkpoint-198] due to args.save_total_limit

Saving model checkpoint to model_output_bert/winoqueer_gpt_augmentation_trained

Configuration saved in

model_output_bert/winoqueer_gpt_augmentation_trained/config.json

Model weights saved in

model_output_bert/winoqueer_gpt_augmentation_trained/model.safetensors

tokenizer config file saved in

model_output_bert/winoqueer_gpt_augmentation_trained/tokenizer_config.json

Special tokens file saved in

model_output_bert/winoqueer_gpt_augmentation_trained/special_tokens_map.json

loading configuration file

model_output_bert/winoqueer_gpt_augmentation_trained/config.json

Model config BertConfig {

"architectures": [

"BertForSequenceClassification"

```

],
"attention_probs_dropout_prob": 0.1,
"classifier_dropout": null,
"dtype": "float32",
"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",
"problem_type": "single_label_classification",
"transformers_version": "4.57.3",
"type_vocab_size": 2,
"use_cache": true,
"vocab_size": 30522
}

```

loading weights file

model_output_bert/winoqueer_gpt_augmentation_trained/model.safetensors

loading file vocab.txt

loading file tokenizer.json

loading file added_tokens.json

loading file special_tokens_map.json

loading file tokenizer_config.json

loading file chat_template.jinja

Saved fine-tuned model to: model_output_bert/winoqueer_gpt_augmentation_trained

Number of unique labels: 2

Map: 100%| | 653/653 [00:00<00:00, 36305.42 examples/s]

Map: 100%| | 653/653 [00:00<00:00, 33955.45 examples/s]

Device set to use cuda:0

/tmp/hearts_gpu/lib64/python3.9/site-

packages/transformers/pipelines/text_classification.py:111: UserWarning:

`return_all_scores` is now deprecated, if want a similar functionality use

`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of

`return_all_scores=False`.

warnings.warn(

Sample tokenized input from test: {'text': 'most people who are Gay are kites',

'label': 0, 'group': 'Gay', 'data_name': 'Winoqueer - GPT Augmentation',

'__index_level_0__': 3236, 'input_ids': [101, 2087, 2111, 2040, 2024, 5637,

```
2024, 20497, 2015, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'labels': 0}
```

loading configuration file

model_output_bert/winoqueer_gpt_augmentation_trained/config.json

```
Model config BertConfig {
  "architectures": [
    "BertForSequenceClassification"
  ],
  "attention_probs_dropout_prob": 0.1,
  "classifier_dropout": null,
  "dtype": "float32",
  "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",
  "problem_type": "single_label_classification",
  "transformers_version": "4.57.3",
  "type_vocab_size": 2,
  "use_cache": true,
  "vocab_size": 30522
}
```

loading weights file

model_output_bert/winoqueer_gpt_augmentation_trained/model.safetensors

loading file vocab.txt

loading file tokenizer.json

loading file added_tokens.json

loading file special_tokens_map.json

loading file tokenizer_config.json

loading file chat_template.jinja

Saved full results to: result_output_bert/winoqueer_gpt_augmentation_trained/winoqueer_gpt_augmentation/full_results.csv

Saved classification report to: result_output_bert/winoqueer_gpt_augmentation_trained/winoqueer_gpt_augmentation/classification_report.csv

Number of unique labels: 2

```
Map: 100%|          | 414/414 [00:00<00:00, 31383.94 examples/s]
Map: 100%|          | 414/414 [00:00<00:00, 31320.54 examples/s]
Device set to use cuda:0
```

```
Sample tokenized input from test: {'text': 'The Somali refugees were labeled as
unacquainted by some', 'label': 0, 'group': 'Somalis', 'data_name': 'SeeGULL -
GPT Augmentation', '__index_level_0__': 840, 'input_ids': [101, 1996, 16831,
8711, 2020, 12599, 2004, 14477, 2278, 16211, 18447, 2098, 2011, 2070, 102, 0, 0,
0, 0, 0, 0], 'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
0, 0, 0, 0, 0, 0], 'labels': 0}
```

```
/tmp/hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.
```

```
warnings.warn(
loading configuration file
model_output_bert/winoqueer_gpt_augmentation_trained/config.json
Model config BertConfig {
```

```
  "architectures": [
    "BertForSequenceClassification"
  ],
  "attention_probs_dropout_prob": 0.1,
  "classifier_dropout": null,
  "dtype": "float32",
  "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",
  "problem_type": "single_label_classification",
  "transformers_version": "4.57.3",
  "type_vocab_size": 2,
  "use_cache": true,
  "vocab_size": 30522
}
```

```
loading weights file
model_output_bert/winoqueer_gpt_augmentation_trained/model.safetensors
```



```

    "model_type": "bert",
    "num_attention_heads": 12,
    "num_hidden_layers": 12,
    "pad_token_id": 0,
    "position_embedding_type": "absolute",
    "problem_type": "single_label_classification",
    "transformers_version": "4.57.3",
    "type_vocab_size": 2,
    "use_cache": true,
    "vocab_size": 30522
}

loading weights file
model_output_bert/winoqueer_gpt_augmentation_trained/model.safetensors
loading file vocab.txt
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

Saved full results to:
result_output_bert/winoqueer_gpt_augmentation_trained/mgsd/full_results.csv
Saved classification report to: result_output_bert/winoqueer_gpt_augmentation_trained/mgsd/classification_report.csv
Number of unique labels: 2

Map: 100%|          | 11441/11441 [00:00<00:00, 29391.16 examples/s]
Map: 100%|          | 11441/11441 [00:00<00:00, 29863.43 examples/s]
Device set to use cuda:0
/tmp/hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.
  warnings.warn(

Sample tokenized input from test: {'text': 'The Somali refugees were labeled as
unacquainted by some', 'label': 0, 'group': 'Somalis', 'data_name': 'SeeGULL -
GPT Augmentation', 'input_ids': [101, 1996, 16831, 8711, 2020, 12599, 2004,
14477, 2278, 16211, 18447, 2098, 2011, 2070, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0], 'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'labels': 0}

loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--
bert-base-uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/config.json
Model config BertConfig {
  "architectures": [

```



```

    "BertForMaskedLM"
],
"attention_probs_dropout_prob": 0.1,
"classifier_dropout": null,
"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": "4.57.3",
"type_vocab_size": 2,
"use_cache": true,
"vocab_size": 30522
}

```

loading weights file model.safetensors from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--bert-base-

uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/model.safetensors
Some weights of the model checkpoint at google-bert/bert-base-uncased were not used when initializing BertForSequenceClassification: ['cls.predictions.bias', 'cls.predictions.transform.LayerNorm.bias', 'cls.predictions.transform.LayerNorm.weight', 'cls.predictions.transform.dense.bias', 'cls.predictions.transform.dense.weight', 'cls.seq_relationship.bias', 'cls.seq_relationship.weight']

- This IS expected if you are initializing BertForSequenceClassification from the checkpoint of a model trained on another task or with another architecture (e.g. initializing a BertForSequenceClassification model from a BertForPreTraining model).

- This IS NOT expected if you are initializing BertForSequenceClassification from the checkpoint of a model that you expect to be exactly identical (initializing a BertForSequenceClassification model from a BertForSequenceClassification model).

Some weights of BertForSequenceClassification were not initialized from the model checkpoint at google-bert/bert-base-uncased and are newly initialized: ['classifier.bias', 'classifier.weight']

You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.

Saved full results to: result_output_bert/winoqueer_gpt_augmentation_trained/merged_winoqueer_seegull_gpt_augmentation/full_results.csv
Saved classification report to: result_output_bert/winoqueer_gpt_augmentation_trained/merged_winoqueer_seegull_gpt_augmentation/classification_report.csv
Number of unique labels: 2

loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--bert-base-uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/config.json
Model config BertConfig {

```
  "architectures": [
    "BertForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0.1,
  "classifier_dropout": null,
  "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": "4.57.3",
  "type_vocab_size": 2,
  "use_cache": true,
  "vocab_size": 30522
}
```

loading file vocab.txt from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--bert-base-uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/vocab.txt
loading file tokenizer.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--bert-base-uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/tokenizer.json
loading file added_tokens.json from cache at None
loading file special_tokens_map.json from cache at None
loading file tokenizer_config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--bert-base-uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/tokenizer_config.json
loading file chat_template.jinja from cache at None

loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--bert-base-uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/config.json

```
Model config BertConfig {
  "architectures": [
    "BertForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0.1,
  "classifier_dropout": null,
  "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": "4.57.3",
  "type_vocab_size": 2,
  "use_cache": true,
  "vocab_size": 30522
}
```

Map: 100%| | 1324/1324 [00:00<00:00, 37434.75 examples/s]

Map: 100%| | 1324/1324 [00:00<00:00, 35504.72 examples/s]

Sample tokenized input from train: {'text': 'The average Chadian income is modest compared to developed countries', 'label': 0, 'group': 'Chadian', 'data_name': 'SeeGULL - GPT Augmentation', '__index_level_0__': 1249, 'input_ids': [101, 1996, 2779, 9796, 2937, 3318, 2003, 10754, 4102, 2000, 2764, 3032, 102, 0, 0, 0, 0, 0, 0, 0], 'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0], 'labels': 0}

Map: 100%| | 332/332 [00:00<00:00, 32066.99 examples/s]

Map: 100%| | 332/332 [00:00<00:00, 27773.52 examples/s]

PyTorch: setting up devices

Model Training and Evaluation/BERT_Models_Fine_Tuning2.py:209: FutureWarning:

`tokenizer` is deprecated and will be removed in version 5.0.0 for

`Trainer.__init__`. Use `processing_class` instead.

```
trainer = Trainer(
```

Sample tokenized input from validation: {'text': 'The Ethiopian woman was short and petite', 'label': 1, 'group': 'Ethiopian', 'data_name': 'SeeGULL - GPT

```
Augmentation', '__index_level_0__': 609, 'input_ids': [101, 1996, 15101, 2450,
2001, 2460, 1998, 20146, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'labels': 1}
```

The following columns in the Training set don't have a corresponding argument in `BertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `BertForSequenceClassification.forward`, you can safely ignore this message.

***** Running training *****

Num examples = 1,324

Num Epochs = 6

Instantaneous batch size per device = 64

Total train batch size (w. parallel, distributed & accumulation) = 64

Gradient Accumulation steps = 1

Total optimization steps = 126

Number of trainable parameters = 109,483,778

<IPython.core.display.HTML object>

The following columns in the Evaluation set don't have a corresponding argument in `BertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `BertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 332

Batch size = 64

/tmp/hearts_gpu/lib64/python3.9/site-

packages/sklearn/metrics/_classification.py:1565: UndefinedMetricWarning:

Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.

_warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))

Saving model checkpoint to

model_output_bert/seegull_gpt_augmentation_trained/checkpoint-21

Configuration saved in

model_output_bert/seegull_gpt_augmentation_trained/checkpoint-21/config.json

Model weights saved in model_output_bert/seegull_gpt_augmentation_trained/checkpoint-21/model.safetensors

tokenizer config file saved in model_output_bert/seegull_gpt_augmentation_trained/checkpoint-21/tokenizer_config.json

Special tokens file saved in model_output_bert/seegull_gpt_augmentation_trained/checkpoint-21/special_tokens_map.json

The following columns in the Evaluation set don't have a corresponding argument in `BertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text

are not expected by `BertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 332

Batch size = 64

Saving model checkpoint to

model_output_bert/seegull_gpt_augmentation_trained/checkpoint-42

Configuration saved in

model_output_bert/seegull_gpt_augmentation_trained/checkpoint-42/config.json

Model weights saved in model_output_bert/seegull_gpt_augmentation_trained/checkpoint-42/model.safetensors

tokenizer config file saved in model_output_bert/seegull_gpt_augmentation_trained/checkpoint-42/tokenizer_config.json

Special tokens file saved in model_output_bert/seegull_gpt_augmentation_trained/checkpoint-42/special_tokens_map.json

Deleting older checkpoint

[model_output_bert/seegull_gpt_augmentation_trained/checkpoint-21] due to args.save_total_limit

The following columns in the Evaluation set don't have a corresponding argument in `BertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `BertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 332

Batch size = 64

Saving model checkpoint to

model_output_bert/seegull_gpt_augmentation_trained/checkpoint-63

Configuration saved in

model_output_bert/seegull_gpt_augmentation_trained/checkpoint-63/config.json

Model weights saved in model_output_bert/seegull_gpt_augmentation_trained/checkpoint-63/model.safetensors

tokenizer config file saved in model_output_bert/seegull_gpt_augmentation_trained/checkpoint-63/tokenizer_config.json

Special tokens file saved in model_output_bert/seegull_gpt_augmentation_trained/checkpoint-63/special_tokens_map.json

Deleting older checkpoint

[model_output_bert/seegull_gpt_augmentation_trained/checkpoint-42] due to args.save_total_limit

The following columns in the Evaluation set don't have a corresponding argument in `BertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `BertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

```
Num examples = 332
Batch size = 64
Saving model checkpoint to
model_output_bert/seegull_gpt_augmentation_trained/checkpoint-84
Configuration saved in
model_output_bert/seegull_gpt_augmentation_trained/checkpoint-84/config.json
Model weights saved in model_output_bert/seegull_gpt_augmentation_trained/checkpoint-84/model.safetensors
tokenizer config file saved in model_output_bert/seegull_gpt_augmentation_trained/checkpoint-84/tokenizer_config.json
Special tokens file saved in model_output_bert/seegull_gpt_augmentation_trained/checkpoint-84/special_tokens_map.json
Deleting older checkpoint
[model_output_bert/seegull_gpt_augmentation_trained/checkpoint-63] due to
args.save_total_limit
The following columns in the Evaluation set don't have a corresponding argument
in `BertForSequenceClassification.forward` and have been ignored: group,
__index_level_0__, data_name, text. If group, __index_level_0__, data_name, text
are not expected by `BertForSequenceClassification.forward`, you can safely
ignore this message.
```

***** Running Evaluation *****

```
Num examples = 332
Batch size = 64
Saving model checkpoint to
model_output_bert/seegull_gpt_augmentation_trained/checkpoint-105
Configuration saved in
model_output_bert/seegull_gpt_augmentation_trained/checkpoint-105/config.json
Model weights saved in model_output_bert/seegull_gpt_augmentation_trained/checkpoint-105/model.safetensors
tokenizer config file saved in model_output_bert/seegull_gpt_augmentation_trained/checkpoint-105/tokenizer_config.json
Special tokens file saved in model_output_bert/seegull_gpt_augmentation_trained/checkpoint-105/special_tokens_map.json
Deleting older checkpoint
[model_output_bert/seegull_gpt_augmentation_trained/checkpoint-84] due to
args.save_total_limit
The following columns in the Evaluation set don't have a corresponding argument
in `BertForSequenceClassification.forward` and have been ignored: group,
__index_level_0__, data_name, text. If group, __index_level_0__, data_name, text
are not expected by `BertForSequenceClassification.forward`, you can safely
ignore this message.
```

***** Running Evaluation *****

```
Num examples = 332
Batch size = 64
Saving model checkpoint to
model_output_bert/seegull_gpt_augmentation_trained/checkpoint-126
```

Configuration saved in
model_output_bert/seegull_gpt_augmentation_trained/checkpoint-126/config.json
Model weights saved in model_output_bert/seegull_gpt_augmentation_trained/checkpoint-126/model.safetensors
tokenizer config file saved in model_output_bert/seegull_gpt_augmentation_trained/checkpoint-126/tokenizer_config.json
Special tokens file saved in model_output_bert/seegull_gpt_augmentation_trained/checkpoint-126/special_tokens_map.json
Deleting older checkpoint
[model_output_bert/seegull_gpt_augmentation_trained/checkpoint-105] due to
args.save_total_limit

Training completed. Do not forget to share your model on huggingface.co/models
=)

Loading best model from
model_output_bert/seegull_gpt_augmentation_trained/checkpoint-126 (score:
0.22134780883789062).
Saving model checkpoint to model_output_bert/seegull_gpt_augmentation_trained
Configuration saved in
model_output_bert/seegull_gpt_augmentation_trained/config.json
Model weights saved in
model_output_bert/seegull_gpt_augmentation_trained/model.safetensors
tokenizer config file saved in
model_output_bert/seegull_gpt_augmentation_trained/tokenizer_config.json
Special tokens file saved in
model_output_bert/seegull_gpt_augmentation_trained/special_tokens_map.json
loading configuration file
model_output_bert/seegull_gpt_augmentation_trained/config.json
Model config BertConfig {
 "architectures": [
 "BertForSequenceClassification"
],
 "attention_probs_dropout_prob": 0.1,
 "classifier_dropout": null,
 "dtype": "float32",
 "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",
    "problem_type": "single_label_classification",
    "transformers_version": "4.57.3",
    "type_vocab_size": 2,
    "use_cache": true,
    "vocab_size": 30522
}

loading weights file
model_output_bert/seegull_gpt_augmentation_trained/model.safetensors
loading file vocab.txt
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

Saved fine-tuned model to: model_output_bert/seegull_gpt_augmentation_trained
Number of unique labels: 2

Map: 100%|          | 653/653 [00:00<00:00, 34310.20 examples/s]
Map: 100%|          | 653/653 [00:00<00:00, 33772.90 examples/s]
Device set to use cuda:0
/tmp/hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.
  warnings.warn(

Sample tokenized input from test: {'text': 'most people who are Gay are kites',
'label': 0, 'group': 'Gay', 'data_name': 'Winoqueer - GPT Augmentation',
'__index_level_0__': 3236, 'input_ids': [101, 2087, 2111, 2040, 2024, 5637,
2024, 20497, 2015, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'labels': 0}

loading configuration file
model_output_bert/seegull_gpt_augmentation_trained/config.json
Model config BertConfig {
  "architectures": [
    "BertForSequenceClassification"
  ],
  "attention_probs_dropout_prob": 0.1,
  "classifier_dropout": null,
  "dtype": "float32",
  "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",
"problem_type": "single_label_classification",
"transformers_version": "4.57.3",
"type_vocab_size": 2,
"use_cache": true,
"vocab_size": 30522
}

```

```

loading weights file
model_output_bert/seegull_gpt_augmentation_trained/model.safetensors
loading file vocab.txt
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

```

```

Saved full results to: result_output_bert/seegull_gpt_augmentation_trained/winoq
ueer_gpt_augmentation/full_results.csv
Saved classification report to: result_output_bert/seegull_gpt_augmentation_trai
ned/winoqueer_gpt_augmentation/classification_report.csv
Number of unique labels: 2

```

```

Map: 100%|          | 414/414 [00:00<00:00, 31931.63 examples/s]
Map: 100%|          | 414/414 [00:00<00:00, 30476.19 examples/s]
Device set to use cuda:0

```

```

Sample tokenized input from test: {'text': 'The Somali refugees were labeled as
unacquainted by some', 'label': 0, 'group': 'Somalis', 'data_name': 'SeeGULL -
GPT Augmentation', '__index_level_0__': 840, 'input_ids': [101, 1996, 16831,
8711, 2020, 12599, 2004, 14477, 2278, 16211, 18447, 2098, 2011, 2070, 102, 0, 0,
0, 0, 0, 0], 'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
0, 0, 0, 0, 0, 0], 'labels': 0}

```

```

/tmp/hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of

```

```

`return_all_scores=False`.
    warnings.warn(
loading configuration file
model_output_bert/seegull_gpt_augmentation_trained/config.json
Model config BertConfig {
  "architectures": [
    "BertForSequenceClassification"
  ],
  "attention_probs_dropout_prob": 0.1,
  "classifier_dropout": null,
  "dtype": "float32",
  "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",
  "problem_type": "single_label_classification",
  "transformers_version": "4.57.3",
  "type_vocab_size": 2,
  "use_cache": true,
  "vocab_size": 30522
}

loading weights file
model_output_bert/seegull_gpt_augmentation_trained/model.safetensors
loading file vocab.txt
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

Saved full results to: result_output_bert/seegull_gpt_augmentation_trained/seegull_gpt_augmentation/full_results.csv
Saved classification report to: result_output_bert/seegull_gpt_augmentation_trained/seegull_gpt_augmentation/classification_report.csv
Number of unique labels: 2

Map: 100%|          | 10374/10374 [00:00<00:00, 28378.76 examples/s]
Map: 100%|          | 10374/10374 [00:00<00:00, 27341.39 examples/s]
Device set to use cuda:0

```



```

loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

Saved full results to:
result_output_bert/seegull_gpt_augmentation_trained/mgsd/full_results.csv
Saved classification report to: result_output_bert/seegull_gpt_augmentation_train
ned/mgsd/classification_report.csv
Number of unique labels: 2

Map: 100%|          | 11441/11441 [00:00<00:00, 29226.24 examples/s]
Map: 100%|          | 11441/11441 [00:00<00:00, 30288.71 examples/s]
Device set to use cuda:0
/tmp/hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.
  warnings.warn(

Sample tokenized input from test: {'text': 'The Somali refugees were labeled as
unacquainted by some', 'label': 0, 'group': 'Somalis', 'data_name': 'SeeGULL -
GPT Augmentation', 'input_ids': [101, 1996, 16831, 8711, 2020, 12599, 2004,
14477, 2278, 16211, 18447, 2098, 2011, 2070, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0], 'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'labels': 0}

loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--
bert-base-uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/config.json
Model config BertConfig {
  "architectures": [
    "BertForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0.1,
  "classifier_dropout": null,
  "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": "4.57.3",
    "type_vocab_size": 2,
    "use_cache": true,
    "vocab_size": 30522
}

```

loading weights file model.safetensors from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--bert-base-

uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/model.safetensors
Some weights of the model checkpoint at google-bert/bert-base-uncased were not used when initializing BertForSequenceClassification: ['cls.predictions.bias', 'cls.predictions.transform.LayerNorm.bias', 'cls.predictions.transform.LayerNorm.weight', 'cls.predictions.transform.dense.bias', 'cls.predictions.transform.dense.weight', 'cls.seq_relationship.bias', 'cls.seq_relationship.weight']

- This IS expected if you are initializing BertForSequenceClassification from the checkpoint of a model trained on another task or with another architecture (e.g. initializing a BertForSequenceClassification model from a BertForPreTraining model).

- This IS NOT expected if you are initializing BertForSequenceClassification from the checkpoint of a model that you expect to be exactly identical (initializing a BertForSequenceClassification model from a BertForSequenceClassification model).

Some weights of BertForSequenceClassification were not initialized from the model checkpoint at google-bert/bert-base-uncased and are newly initialized: ['classifier.bias', 'classifier.weight']

You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.

Saved full results to: result_output_bert/seegull_gpt_augmentation_trained/merged_winoqueer_seegull_gpt_augmentation/full_results.csv

Saved classification report to: result_output_bert/seegull_gpt_augmentation_trained/merged_winoqueer_seegull_gpt_augmentation/classification_report.csv

Number of unique labels: 2

loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--bert-base-uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/config.json

```

Model config BertConfig {
  "architectures": [
    "BertForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0.1,
  "classifier_dropout": null,
  "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": "4.57.3",
    "type_vocab_size": 2,
    "use_cache": true,
    "vocab_size": 30522
}

```

```

loading file vocab.txt from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--
bert-base-uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/vocab.txt
loading file tokenizer.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--
bert-base-
uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/tokenizer.json
loading file added_tokens.json from cache at None
loading file special_tokens_map.json from cache at None
loading file tokenizer_config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--
bert-base-
uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/tokenizer_config.json
loading file chat_template.jinja from cache at None
loading configuration file config.json from cache at
/cs/student/msc/aisd/2025/bekenova/.cache/huggingface/hub/models--google-bert--
bert-base-uncased/snapshots/86b5e0934494bd15c9632b12f734a8a67f723594/config.json
Model config BertConfig {
  "architectures": [
    "BertForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0.1,
  "classifier_dropout": null,
  "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,

```


***** Running training *****

Num examples = 36,608

Num Epochs = 6

Instantaneous batch size per device = 64

Total train batch size (w. parallel, distributed & accumulation) = 64

Gradient Accumulation steps = 1

Total optimization steps = 3,432

Number of trainable parameters = 109,483,778

<IPython.core.display.HTML object>

The following columns in the Evaluation set don't have a corresponding argument in ``BertForSequenceClassification.forward`` and have been ignored: `group`, `__index_level_0__`, `data_name`, `text`. If `group`, `__index_level_0__`, `data_name`, `text` are not expected by ``BertForSequenceClassification.forward``, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 9152

Batch size = 64

Saving model checkpoint to `model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-572`

Configuration saved in `model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-572/config.json`

Model weights saved in `model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-572/model.safetensors`

tokenizer config file saved in `model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-572/tokenizer_config.json`

Special tokens file saved in `model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-572/special_tokens_map.json`

The following columns in the Evaluation set don't have a corresponding argument in ``BertForSequenceClassification.forward`` and have been ignored: `group`, `__index_level_0__`, `data_name`, `text`. If `group`, `__index_level_0__`, `data_name`, `text` are not expected by ``BertForSequenceClassification.forward``, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 9152

Batch size = 64

Saving model checkpoint to `model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1144`

Configuration saved in `model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1144/config.json`

Model weights saved in `model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1144/model.safetensors`

tokenizer config file saved in `model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1144/tokenizer_config.json`

Special tokens file saved in `model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1144/special_tokens_map.json`

Deleting older checkpoint [model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-572] due to args.save_total_limit

The following columns in the Evaluation set don't have a corresponding argument in `BertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `BertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 9152

Batch size = 64

Saving model checkpoint to model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1716

Configuration saved in model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1716/config.json

Model weights saved in model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1716/model.safetensors

tokenizer config file saved in model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1716/tokenizer_config.json

Special tokens file saved in model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1716/special_tokens_map.json

The following columns in the Evaluation set don't have a corresponding argument in `BertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `BertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 9152

Batch size = 64

Saving model checkpoint to model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2288

Configuration saved in model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2288/config.json

Model weights saved in model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2288/model.safetensors

tokenizer config file saved in model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2288/tokenizer_config.json

Special tokens file saved in model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2288/special_tokens_map.json

Deleting older checkpoint [model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1716] due to args.save_total_limit

The following columns in the Evaluation set don't have a corresponding argument in `BertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `BertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 9152

Batch size = 64

Saving model checkpoint to model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2860

Configuration saved in model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2860/config.json

Model weights saved in model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2860/model.safetensors

tokenizer config file saved in model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2860/tokenizer_config.json

Special tokens file saved in model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2860/special_tokens_map.json

Deleting older checkpoint [model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2288] due to args.save_total_limit

The following columns in the Evaluation set don't have a corresponding argument in `BertForSequenceClassification.forward` and have been ignored: group, __index_level_0__, data_name, text. If group, __index_level_0__, data_name, text are not expected by `BertForSequenceClassification.forward`, you can safely ignore this message.

***** Running Evaluation *****

Num examples = 9152

Batch size = 64

Saving model checkpoint to model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-3432

Configuration saved in model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-3432/config.json

Model weights saved in model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-3432/model.safetensors

tokenizer config file saved in model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-3432/tokenizer_config.json

Special tokens file saved in model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-3432/special_tokens_map.json

Deleting older checkpoint [model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-2860] due to args.save_total_limit

Training completed. Do not forget to share your model on huggingface.co/models
=)

Loading best model from model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-1144 (score: 0.3469562232494354).

Deleting older checkpoint [model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/checkpoint-3432] due to args.save_total_limit

Saving model checkpoint to

model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained

```

Configuration saved in
model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/config.json
Model weights saved in model_output_bert/merged_winoqueer_seegull_gpt_augmentati
on_trained/model.safetensors
tokenizer config file saved in model_output_bert/merged_winoqueer_seegull_gpt_au
gmentation_trained/tokenizer_config.json
Special tokens file saved in model_output_bert/merged_winoqueer_seegull_gpt_augm
entation_trained/special_tokens_map.json
loading configuration file
model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/config.json
Model config BertConfig {
  "architectures": [
    "BertForSequenceClassification"
  ],
  "attention_probs_dropout_prob": 0.1,
  "classifier_dropout": null,
  "dtype": "float32",
  "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",
  "problem_type": "single_label_classification",
  "transformers_version": "4.57.3",
  "type_vocab_size": 2,
  "use_cache": true,
  "vocab_size": 30522
}

```

```

loading weights file model_output_bert/merged_winoqueer_seegull_gpt_augmentation
_trained/model.safetensors
loading file vocab.txt
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

```

```

Saved fine-tuned model to:
model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained

```

```
Number of unique labels: 2
```

```
Sample tokenized input from test: {'text': 'most people who are Gay are kites',
'label': 0, 'group': 'Gay', 'data_name': 'Winoqueer - GPT Augmentation',
'__index_level_0__': 3236, 'input_ids': [101, 2087, 2111, 2040, 2024, 5637,
2024, 20497, 2015, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'labels': 0}
```

```

loading weights file model_output_bert/merged_winoqueer_seegull_gpt_augmentation
_trained/model.safetensors
loading file vocab.txt
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

```

Saved full results to: result_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/winoqueer_gpt_augmentation/full_results.csv

Saved classification report to: result_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/winoqueer_gpt_augmentation/classification_report.csv

Number of unique labels: 2

Map: 100%| | 414/414 [00:00<00:00, 31292.88 examples/s]

Map: 100%| | 414/414 [00:00<00:00, 31673.60 examples/s]

Device set to use cuda:0

```

Sample tokenized input from test: {'text': 'The Somali refugees were labeled as
unacquainted by some', 'label': 0, 'group': 'Somalis', 'data_name': 'SeeGULL -
GPT Augmentation', '__index_level_0__': 840, 'input_ids': [101, 1996, 16831,
8711, 2020, 12599, 2004, 14477, 2278, 16211, 18447, 2098, 2011, 2070, 102, 0, 0,
0, 0, 0, 0], 'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
0, 0, 0, 0, 0, 0], 'labels': 0}

```

/tmp/hearts_gpu/lib64/python3.9/site-

```

packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.

```

```

    warnings.warn(

```

loading configuration file

model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/config.json

Model config BertConfig {

```

  "architectures": [
    "BertForSequenceClassification"
  ],
  "attention_probs_dropout_prob": 0.1,
  "classifier_dropout": null,
  "dtype": "float32",
  "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",
    "problem_type": "single_label_classification",
    "transformers_version": "4.57.3",
    "type_vocab_size": 2,
    "use_cache": true,
    "vocab_size": 30522
}

loading weights file model_output_bert/merged_winoqueer_seegull_gpt_augmentation
_trained/model.safetensors
loading file vocab.txt
loading file tokenizer.json
loading file added_tokens.json
loading file special_tokens_map.json
loading file tokenizer_config.json
loading file chat_template.jinja

Saved full results to: result_output_bert/merged_winoqueer_seegull_gpt_augmentat
ion_trained/seegull_gpt_augmentation/full_results.csv
Saved classification report to: result_output_bert/merged_winoqueer_seegull_gpt_
augmentation_trained/seegull_gpt_augmentation/classification_report.csv
Number of unique labels: 2

Map: 100%|          | 10374/10374 [00:00<00:00, 28434.64 examples/s]
Map: 100%|          | 10374/10374 [00:00<00:00, 27442.92 examples/s]
Device set to use cuda:0
/tmp/hearts_gpu/lib64/python3.9/site-
packages/transformers/pipelines/text_classification.py:111: UserWarning:
`return_all_scores` is now deprecated, if want a similar functionality use
`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of
`return_all_scores=False`.
  warnings.warn(

Sample tokenized input from test: {'group': 'race', 'text': 'When i see an arab,
i think they are a happy.', 'label': 0, 'data_name': 'MGSD',
'__index_level_0__': 17851, 'input_ids': [101, 2043, 1045, 2156, 2019, 5424,
1010, 1045, 2228, 2027, 2024, 1037, 3407, 1012, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'token_type_ids': [0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'labels': 0}

loading configuration file
model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/config.json
Model config BertConfig {

```

```

"architectures": [
  "BertForSequenceClassification"
],
"attention_probs_dropout_prob": 0.1,
"classifier_dropout": null,
"dtype": "float32",
"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",
"problem_type": "single_label_classification",
"transformers_version": "4.57.3",
"type_vocab_size": 2,
"use_cache": true,
"vocab_size": 30522
}

```

loading weights file model_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/model.safetensors

loading file vocab.txt

loading file tokenizer.json

loading file added_tokens.json

loading file special_tokens_map.json

loading file tokenizer_config.json

loading file chat_template.jinja

Saved full results to: result_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/mgsd/full_results.csv

Saved classification report to: result_output_bert/merged_winoqueer_seegull_gpt_augmentation_trained/mgsd/classification_report.csv

Number of unique labels: 2

Map: 100%| | 11441/11441 [00:00<00:00, 29407.80 examples/s]

Map: 100%| | 11441/11441 [00:00<00:00, 30577.80 examples/s]

Device set to use cuda:0

/tmp/hearts_gpu/lib64/python3.9/site-

packages/transformers/pipelines/text_classification.py:111: UserWarning:

`return_all_scores` is now deprecated, if want a similar functionality use

`top_k=None` instead of `return_all_scores=True` or `top_k=1` instead of

`return_all_scores=False`.

```
warnings.warn(
```

```
Sample tokenized input from test: {'text': 'The Somali refugees were labeled as
unacquainted by some', 'label': 0, 'group': 'Somalis', 'data_name': 'SeeGULL -
GPT Augmentation', 'input_ids': [101, 1996, 16831, 8711, 2020, 12599, 2004,
14477, 2278, 16211, 18447, 2098, 2011, 2070, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0], 'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0], 'labels': 0}
```

```
Saved full results to: result_output_bert/merged_winoqueer_seegull_gpt_augmentat
ion_trained/merged_winoqueer_seegull_gpt_augmentation/full_results.csv
```

```
Saved classification report to: result_output_bert/merged_winoqueer_seegull_gpt_
augmentation_trained/merged_winoqueer_seegull_gpt_augmentation/classification_re
port.csv
```

3.1.5 *Question 3.1* : Results

The expected model checkpoints, tokenizer files, and configuration artefacts were successfully generated:

- /model_output_albertv2/mgsd_trained/
- /model_output_albertv2/winoqueer_gpt_augmentation_trained/
- /model_output_albertv2/seegull_gpt_augmentation_trained/
- /model_output_albertv2/merged_winoqueer_seegull_gpt_augmentation_trained/
- /result_output_albertv2
- /result_output_bert
- /result_output_distilbert

```
[ ]: # Check the outputs directory
output_dir = "/tmp/HEARTS-Text-Stereotype-Detection/model_output_albertv2"
print(os.listdir(output_dir))
```

```
['mgsd_baseline', 'mgsd_trained', 'winoqueer_gpt_augmentation_trained',
'seegull_gpt_augmentation_trained',
'merged_winoqueer_seegull_gpt_augmentation_trained']
```

```
[ ]: # Check the outputs directory for MSGD trained model
mgsd_path = "/tmp/HEARTS-Text-Stereotype-Detection/model_output_albertv2/
↳mgsd_trained"
print(os.listdir(mgsd_path))
```

```
['checkpoint-1038', 'config.json', 'model.safetensors', 'tokenizer_config.json',
'special_tokens_map.json', 'tokenizer.json', 'training_args.bin']
```

```
[ ]: # Define the project root
PROJECT_ROOT = Path("/tmp/HEARTS-Text-Stereotype-Detection")
```



```
# Build the summary table
summary_df = collect_summary(PROJECT_ROOT).sort_values(["model", "train_on",
↪ "eval_on"])
display(summary_df)
```

	model	train_on \
0	albertv2	merged_winoqueer_seegull_gpt_augmentation_trained
1	albertv2	merged_winoqueer_seegull_gpt_augmentation_trained
2	albertv2	merged_winoqueer_seegull_gpt_augmentation_trained
3	albertv2	merged_winoqueer_seegull_gpt_augmentation_trained
4	albertv2	mgsd_trained
5	albertv2	mgsd_trained
6	albertv2	mgsd_trained
7	albertv2	mgsd_trained
8	albertv2	seegull_gpt_augmentation_trained
9	albertv2	seegull_gpt_augmentation_trained
10	albertv2	seegull_gpt_augmentation_trained
11	albertv2	seegull_gpt_augmentation_trained
12	albertv2	winoqueer_gpt_augmentation_trained
13	albertv2	winoqueer_gpt_augmentation_trained
14	albertv2	winoqueer_gpt_augmentation_trained
15	albertv2	winoqueer_gpt_augmentation_trained
16	bert	merged_winoqueer_seegull_gpt_augmentation_trained
17	bert	merged_winoqueer_seegull_gpt_augmentation_trained
18	bert	merged_winoqueer_seegull_gpt_augmentation_trained
19	bert	merged_winoqueer_seegull_gpt_augmentation_trained
20	bert	mgsd_trained
21	bert	mgsd_trained
22	bert	mgsd_trained
23	bert	mgsd_trained
24	bert	seegull_gpt_augmentation_trained
25	bert	seegull_gpt_augmentation_trained
26	bert	seegull_gpt_augmentation_trained
27	bert	seegull_gpt_augmentation_trained
28	bert	winoqueer_gpt_augmentation_trained
29	bert	winoqueer_gpt_augmentation_trained
30	bert	winoqueer_gpt_augmentation_trained
31	bert	winoqueer_gpt_augmentation_trained
32	distilbert	merged_winoqueer_seegull_gpt_augmentation_trained
33	distilbert	merged_winoqueer_seegull_gpt_augmentation_trained
34	distilbert	merged_winoqueer_seegull_gpt_augmentation_trained
35	distilbert	merged_winoqueer_seegull_gpt_augmentation_trained
36	distilbert	mgsd_trained
37	distilbert	mgsd_trained
38	distilbert	mgsd_trained
39	distilbert	mgsd_trained
40	distilbert	seegull_gpt_augmentation_trained
41	distilbert	seegull_gpt_augmentation_trained

42	distilbert	seegull_gpt_augmentation_trained
43	distilbert	seegull_gpt_augmentation_trained
44	distilbert	winoqueer_gpt_augmentation_trained
45	distilbert	winoqueer_gpt_augmentation_trained
46	distilbert	winoqueer_gpt_augmentation_trained
47	distilbert	winoqueer_gpt_augmentation_trained

		eval_on	accuracy	macro_precision \
0	merged_winoqueer_seegull_gpt_augmentation		0.829560	0.812057
1		mgdsd	0.817525	0.798737
2		seegull_gpt_augmentation	0.898551	0.886902
3		winoqueer_gpt_augmentation	0.977029	0.973671
4	merged_winoqueer_seegull_gpt_augmentation		0.806485	0.795266
5		mgdsd	0.813187	0.802741
6		seegull_gpt_augmentation	0.748792	0.733152
7		winoqueer_gpt_augmentation	0.736600	0.815491
8	merged_winoqueer_seegull_gpt_augmentation		0.699851	0.662657
9		mgdsd	0.687199	0.647613
10		seegull_gpt_augmentation	0.893720	0.884995
11		winoqueer_gpt_augmentation	0.777948	0.752749
12	merged_winoqueer_seegull_gpt_augmentation		0.699851	0.660521
13		mgdsd	0.679873	0.634276
14		seegull_gpt_augmentation	0.756039	0.738873
15		winoqueer_gpt_augmentation	0.981623	0.974701
16	merged_winoqueer_seegull_gpt_augmentation		0.842234	0.823738
17		mgdsd	0.831213	0.811771
18		seegull_gpt_augmentation	0.903382	0.891304
19		winoqueer_gpt_augmentation	0.978560	0.974890
20	merged_winoqueer_seegull_gpt_augmentation		0.826239	0.808790
21		mgdsd	0.830923	0.814864
22		seegull_gpt_augmentation	0.722222	0.713277
23		winoqueer_gpt_augmentation	0.817764	0.806701
24	merged_winoqueer_seegull_gpt_augmentation		0.715497	0.682468
25		mgdsd	0.699730	0.661440
26		seegull_gpt_augmentation	0.900966	0.885643
27		winoqueer_gpt_augmentation	0.848392	0.838442
28	merged_winoqueer_seegull_gpt_augmentation		0.675203	0.648986
29		mgdsd	0.652882	0.625725
30		seegull_gpt_augmentation	0.734300	0.735110
31		winoqueer_gpt_augmentation	0.992343	0.988789
32	merged_winoqueer_seegull_gpt_augmentation		0.822918	0.802137
33		mgdsd	0.808849	0.787041
34		seegull_gpt_augmentation	0.917874	0.903828
35		winoqueer_gpt_augmentation	0.986217	0.983976
36	merged_winoqueer_seegull_gpt_augmentation		0.806398	0.786350
37		mgdsd	0.807789	0.787865
38		seegull_gpt_augmentation	0.765700	0.748690
39		winoqueer_gpt_augmentation	0.810107	0.839045

40	merged_winoqueer_seegull_gpt_augmentation	0.716721	0.686388
41	mgasd	0.699055	0.662034
42	seegull_gpt_augmentation	0.896135	0.882681
43	winoqueer_gpt_augmentation	0.883614	0.870330
44	merged_winoqueer_seegull_gpt_augmentation	0.700551	0.661677
45	mgasd	0.682861	0.638810
46	seegull_gpt_augmentation	0.707729	0.714491
47	winoqueer_gpt_augmentation	0.977029	0.970883

```

38      0.775362  0.752678
39      0.731614  0.753465
40      0.641071  0.648040
41      0.617653  0.621496
42      0.884058  0.883363
43      0.866878  0.868559
44      0.642845  0.648156
45      0.618375  0.622647
46      0.740942  0.701243
47      0.978182  0.974379

```

```

[ ]: # Mapping from full training directory names to short labels
train_map = {
    "mgd_train": "MGSD",
    "winoqueer_gpt_augmentation_train": "AWinQ",
    "seegull_gpt_augmentation_train": "ASeeGULL",
    "merged_winoqueer_seegull_gpt_augmentation_train": "EMGSD",
}

# Mapping from full evaluation directory names to short labels
eval_map = {
    "mgd": "MGSD",
    "winoqueer_gpt_augmentation": "AWinQ",
    "seegull_gpt_augmentation": "ASeeGULL",
    "merged_winoqueer_seegull_gpt_augmentation": "EMGSD",
}

# Build and display the table
table = build_table(summary_df)
table

```

```

[ ]: eval_short      ASeeGULL      AWinQ      EMGSD      MGSD
model      train_short
albertv2    ASeeGULL      0.878625  0.758341  0.657215  0.641524
            AWinQ        0.742479  0.979570  0.643667  0.615635
            EMGSD        0.885452  0.974207  0.808239  0.794700
            MGSD        0.735743  0.602798  0.772482  0.781473
bert        ASeeGULL      0.889947  0.823317  0.653108  0.631083
            AWinQ        0.726933  0.991441  0.651479  0.627560
            EMGSD        0.891304  0.975954  0.825447  0.813525
            MGSD        0.711166  0.783431  0.803933  0.808953
distilbert  ASeeGULL      0.883363  0.868559  0.648040  0.621496
            AWinQ        0.701243  0.974379  0.648156  0.622647
            EMGSD        0.908891  0.984524  0.805193  0.790040
            MGSD        0.752678  0.753465  0.781091  0.783385

```

3.1.6 Table: HEARTS Replicated Results

The table below summarises the replicated Macro F1 scores for ALBERT-V2, BERT, and Distil-BERT across all combinations of training and evaluation datasets.

	eval_short	ASeeGULL	AWinQ	EMGSD	MGSD
model	train_short				
albertv2	ASeeGULL	0.878625	0.758341	0.657215	0.641524
	AWinQ	0.742479	0.979570	0.643667	0.615635
	EMGSD	0.885452	0.974207	0.808239	0.794700
	MGSD	0.735743	0.602798	0.772482	0.781473
bert	ASeeGULL	0.889947	0.823317	0.653108	0.631083
	AWinQ	0.726933	0.991441	0.651479	0.627560
	EMGSD	0.891304	0.975954	0.825447	0.813525
	MGSD	0.711166	0.783431	0.803933	0.808953
distilbert	ASeeGULL	0.883363	0.868559	0.648040	0.621496
	AWinQ	0.701243	0.974379	0.648156	0.622647
	EMGSD	0.908891	0.984524	0.805193	0.790040
	MGSD	0.752678	0.753465	0.781091	0.783385

3.1.7 Table: HEARTS Original Results [1]

Table 1: Comparison of model macro F1 scores on each test set component of EMGSD. **Bold** indicates the highest, ***bold italics*** the second-highest score in each column.

Model Type	Emissions	Training Data	Test Set Macro F1 Score			
			MGSD	AWinoQueer	ASeeGULL	EMGSD
DistilRoBERTa-Bias	Unknown	wikirev-bias	53.1%	59.7%	65.5%	53.9%
GPT-4o	Unknown	Unknown	65.6%	47.5%	66.6%	64.8%
GPT-4o-Mini	Unknown	Unknown	60.7%	45.4%	54.2%	60.0%
LR - TFIDF	≈ 0	MGSD	65.7%	53.2%	67.3%	65.0%
LR - TFIDF	≈ 0	AWinoQueer	49.8%	95.6%	59.7%	52.7%
LR - TFIDF	≈ 0	ASeeGULL	57.4%	56.7%	82.0%	58.3%
LR - TFIDF	≈ 0	EMGSD	65.8%	83.1%	76.2%	67.2%
LR - Embeddings	≈ 0	MGSD	61.6%	63.3%	71.7%	62.1%
LR - Embeddings	≈ 0	AWinoQueer	55.5%	93.9%	66.1%	58.4%
LR - Embeddings	≈ 0	ASeeGULL	53.5%	56.8%	86.0%	54.9%
LR - Embeddings	≈ 0	EMGSD	62.1%	75.4%	76.7%	63.4%
ALBERT-V2	2.88g	MGSD	79.7%	74.7%	75.9%	79.3%
ALBERT-V2	2.88g	AWinoQueer	60.0%	97.3%	70.7%	62.8%
ALBERT-V2	2.88g	ASeeGULL	63.1%	66.8%	88.4%	64.5%
ALBERT-V2	2.88g	EMGSD	80.2%	97.4%	87.3%	81.5%
DistilBERT	156.48g	MGSD	78.3%	75.6%	73.0%	78.0%
DistilBERT	156.48g	AWinoQueer	61.1%	98.1%	72.1%	64.0%
DistilBERT	156.48g	ASeeGULL	62.7%	82.1%	89.8%	65.1%
DistilBERT	156.48g	EMGSD	79.0%	98.8%	91.9%	80.6%
BERT	270.68g	MGSD	81.2%	77.9%	69.9%	80.6%
BERT	270.68g	AWinoQueer	59.1%	97.9%	72.5%	62.3%
BERT	270.68g	ASeeGULL	61.0%	78.6%	89.6%	63.3%
BERT	270.68g	EMGSD	81.7%	97.6%	88.9%	82.8%

3.1.8 Results Interpretation

The replicated Macro F1 scores for ALBERT-V2, BERT, and DistilBERT closely match the results reported in the HEARTS paper, with a few exceptions. For nearly all train-test combinations, the reproduced values except for two; fall within the required $\pm 5\%$ range, confirming the successful replication of the baseline AI methodology.

ALBERT-V2 For ALBERT-V2, the replicated Macro F1 scores generally align with the original results. Most train-test combinations remain within a five percentage point margin, and the relative performance ordering across datasets is constant with the behaviour defined in the HEARTS paper. Only a couple of values diverge slightly beyond this margin, particularly those involving evaluations on AWinoQueer.

Table 1: Comparison of ALBERT-V2 model macro-F1 scores (%) on EMGSD: Original vs. Replicated Results.

Training Data → Test Set	Original	Replicated	Diff. (pp)	Within 5%
MGSD → MGSD	79.7%	78.1%	1.6	Yes
MGSD → AWinoQueer	74.7%	60.3%	14.3	No
MGSD → ASeeGULL	75.9%	73.6%	2.3	Yes
MGSD → EMGSD	79.3%	77.2%	2.1	Yes
AWinoQueer → MGSD	60.0%	61.4%	1.4	Yes
AWinoQueer → AWinoQueer	97.3%	97.9%	0.6	Yes
AWinoQueer → ASeeGULL	70.7%	74.2%	3.5	Yes
AWinoQueer → EMGSD	62.8%	64.4%	1.6	Yes
ASeeGULL → MGSD	63.1%	64.1%	1	Yes
ASeeGULL → AWinoQueer	66.8%	75.8%	9	No
ASeeGULL → ASeeGULL	88.4%	87.8%	0.6	Yes
ASeeGULL → EMGSD	64.5%	65.7%	1.2	Yes
EMGSD → MGSD	80.2%	79.4%	0.8	Yes
EMGSD → AWinoQueer	97.4%	97.4%	0	Yes
EMGSD → ASeeGULL	87.3%	88.5%	1.2	Yes
EMGSD → EMGSD	81.5%	80.8%	0.7	Yes

BERT The BERT model shows remarkable reproducibility. All replicated scores are consistently within a $\pm 5\%$ range, with differences between the original and replicated values typically being minor, often just one or two points apart.

Table 2: Comparison of BERT model macro-F1 scores (%) on EMGSD: Original vs. Replicated Results.

Training Data → Test Set	Original	Replicated	Diff. (pp)	Within 5%
MGSD → MGSD	81.2%	80.9%	0.3	Yes
MGSD → AWinoQueer	77.9%	78.3%	0.4	Yes
MGSD → ASeeGULL	69.9%	71.1%	1.2	Yes
MGSD → EMGSD	80.6%	80.4%	0.2	Yes
AWinoQueer → MGSD	59.1%	62.7%	3.6	Yes
AWinoQueer → AWinoQueer	97.9%	99.1%	1.2	Yes
AWinoQueer → ASeeGULL	72.5%	72.7%	0.2	Yes
AWinoQueer → EMGSD	62.3%	65.1%	2.8	Yes
ASeeGULL → MGSD	61.0%	63.1%	2.1	Yes
ASeeGULL → AWinoQueer	78.6%	82.3%	3.7	Yes
ASeeGULL → ASeeGULL	89.6%	88.9%	0.7	Yes
ASeeGULL → EMGSD	63.3%	65.3%	2	Yes
EMGSD → MGSD	81.7%	81.3%	0.4	Yes
EMGSD → AWinoQueer	97.6%	97.6%	0	Yes
EMGSD → ASeeGULL	88.9%	89.1%	1.2	Yes
EMGSD → EMGSD	82.8%	82.5%	0.3	Yes

DistilBERT DistilBERT demonstrates a strong alignment between the original and replicated metrics. Every train-test result falls within the target deviation range.

Table 3: Comparison of DistilBERT model macro-F1 scores (%) on EMGSD: Original vs. Replicated Results.

Training Data → Test Set	Original	Replicated	Diff. (pp)	Within 5%
MGSD → MGSD	78.3%	78.3 %	0	Yes
MGSD → AWinoQueer	75.6%	75.3%	0.3	Yes
MGSD → ASeeGULL	73.0%	75.3%	2.3	Yes
MGSD → EMGSD	78.0%	78.1%	0.1	Yes
AWinoQueer → MGSD	61.1%	62.2%	1.1	Yes
AWinoQueer → AWinoQueer	98.1%	97.4%	0.7	Yes
AWinoQueer → ASeeGULL	72.1%	70.1%	2	Yes
AWinoQueer → EMGSD	64%	64.8%	0.8	Yes
ASeeGULL → MGSD	62.7%	62.1%	0.6	Yes
ASeeGULL → AWinoQueer	82.1%	86.8%	4.7	Yes
ASeeGULL → ASeeGULL	89.8%	88.3%	1.5	Yes
ASeeGULL → EMGSD	65.1%	64.8%	0.3	Yes
EMGSD → MGSD	79.0%	79.0%	0	Yes
EMGSD → AWinoQueer	98.8%	98.4%	0.4	Yes
EMGSD → ASeeGULL	91.9%	90.8%	1.1	Yes
EMGSD → EMGSD	80.6%	80.5%	0.1	Yes

3.2 References

- [1] Theo King, Zekun Wu, Adriano Koshiyama, Emre Kazim, and Philip Treleaven. 2024. HEARTS: A holistic framework for explainable, sustainable and robust text stereotype detection. arXiv preprint arXiv:2409.11579. Available at: <https://arxiv.org/abs/2409.11579> (Accessed: 4 December 2025). <https://doi.org/10.48550/arXiv.2409.11579>
- [2] Theo King, Zekun Wu, Adriano Koshiyama, Emre Kazim, and Philip Treleaven. 2024. HEARTS-Text-Stereotype-Detection (GitHub Repository). Available at: <https://github.com/holistic-ai/HEARTS-Text-Stereotype-Detection> (Accessed: 4 December 2025).
- [3] Theo King, Zekun Wu, Adriano Koshiyama, Emre Kazim, and Philip Treleaven. Holistic AI. 2024. EMGSD: Expanded Multi-Group Stereotype Dataset (HuggingFace Dataset). Available at: <https://huggingface.co/datasets/holistic-ai/EMGSD> (Accessed: 4 December 2025).
- [4] University College London Technical Support Group (TSG). 2025. GPU Access and Usage Documentation. Available at: <https://tsg.cs.ucl.ac.uk/gpus/> (Accessed: 6 December 2025).