I. Dataset you used for fine-tuning

Dataset: "gender identity" subset from the LREC 2022 paper "Large-Scale Hate Speech Detection with Cross-Domain Transfer"

Name of Dataset: “gender-hate-speech”

Path for Dataset: ”ctoraman/ gender-hate-speech”

Genre of Text: The dataset consists of tweets in English that are related to "gender identity." The focus is on understanding and detecting hate speech in the context of gender identity.

Number of Classes: The dataset includes three classes for hate speech labels:

- 0: Normal
- 1: Offensive
- 2: Hate

Size of the Dataset:

- **Total:** 20,000 tweets
- **Train Split:** 18,000 tweets
- **Test Split:** 2,000 tweets

Data Split: The data is split into training and testing sets, following the first fold of a 10-fold cross-validation setup.

GitHub Repository: The dataset is available on GitHub at https://github.com/avaapm/hatespeech.

Citation:

- Toraman, C., Şahinuç, F., & Yilmaz, E. (2022, June). Large-Scale Hate Speech Detection with Cross-Domain Transfer. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 2215-2225).
- Şahinuç, F., Yilmaz, E. H., Toraman, C., & Koç, A. (2023). The effect of gender bias on hate speech detection. Signal, Image and Video Processing, 17(4), 1591-1597.

This dataset and its associated research provide valuable insights into the detection of hate speech with a focus on gender identity, and the provided links can be used to access the dataset and related papers. For fine tuning the model only 2,000 data were used for “train” and “test” each.

II. BERT model used

Model Name: distilbert-base-uncased

Framework: Transformers library by Hugging Face

Type: DistilBERT (a distilled version of BERT)

Variant: Base

Case: Uncased

Key Characteristics:

- ***DistilBERT:*** DistilBERT is a compressed version of the original BERT (Bidirectional Encoder Representations from Transformers) model. It is trained to be more lightweight and faster while preserving much of the performance of the larger BERT model. The model is trained over 3 epochs with batch size as “25”.
- ***Base Variant:*** The "base" variant typically refers to a smaller version of the model compared to larger variants like "large" or "xlarge." The base variant is often chosen when computational resources are a consideration.
- ***Uncased:*** The "uncased" version indicates that the model was trained on text where all letters have been converted to lowercase. This makes the model suitable for tasks where capitalization might not carry significant semantic meaning.

Purpose: The choice of distilbert-base-uncased suggests a preference for a more computationally efficient model while still maintaining good performance. This is especially useful in scenarios where resources are limited, and there is a need for faster inference.

Source: The model is part of the Hugging Face Transformers library, a popular repository of pre-trained transformer models for natural language processing tasks. You can find more information about this specific model on the Hugging Face Model Hub: distilbert-base-uncased. This model is employed in the code for fine-tuning on the hate speech detection task, leveraging its capabilities to understand and represent contextual information in textual data.

III. Result of Task-1

```
Epoch 1/3

64/64 [==============================] - 136s 2s/step - loss: 1.2087 - accuracy: 0.5800 - val_loss: 0.8664 - val_accuracy: 0.5725

Epoch 2/3

64/64 [==============================] - 136s 2s/step - loss: 0.8484 - accuracy: 0.6200 - val_loss: 0.8885 - val_accuracy: 0.5725
```

```
Epoch 3/3
64/64 [==============================] - 136s 2s/step - loss: 0.9065 - accuracy: 0.6169 - val_loss: 0.8841 - val_accuracy: 0.5725
```

Accuracy on test data: 0.6334999799728394

Epoch 1/3:

Loss: 1.2087

Accuracy: 58.00%

Validation Loss: 0.8664

Validation Accuracy: 57.25%

Epoch 2/3:

Loss: 0.8484

Accuracy: 62.00%

Validation Loss: 0.8885

Validation Accuracy: 57.25%

Epoch 3/3:

Loss: 0.9065

Accuracy: 61.69%

Validation Loss: 0.8841

Validation Accuracy: 57.25%

Accuracy on Test Data: 63.35%

Loss: The loss represents the error during training, and it appears to decrease from Epoch 1 to Epoch 2 but increases slightly in Epoch 3. A decrease in loss is generally positive, indicating that the model is improving in its ability to make predictions. However, the subsequent increase in Epoch 3 might suggest overfitting or a need for further optimization.

Accuracy: The training accuracy increases from 58% to 61.69% over the three epochs. Training accuracy measures the percentage of correct predictions on the training set. It's important to consider this alongside validation accuracy and test accuracy to assess the model's generalization performance.

Validation Loss and Accuracy: The validation loss and accuracy represent the model's performance on a separate dataset not used during training. In this case, the validation accuracy remains constant at 57.25%, suggesting that the model may not be improving significantly on unseen data. This could be an indication of model limitations or insufficient complexity.

Accuracy on Test Data: The accuracy on the test data is 63.35%, which is higher than the validation accuracy. It's essential to analyze test accuracy to gauge the model's performance on completely unseen data. The increase from validation accuracy could indicate that the model generalizes well to the test set.

IV. Network and training setting.

The provided code snippet contains a neural network architecture for fine-tuning a BERT-based model on a hate speech detection task. Here's a brief description of the network and the training settings:

A. Neural Network Architecture:

1. **Input Layer:**
    - Two input layers for token IDs (token_ids) and attention masks (attention masks), indicating the input text and attention mechanism for the BERT model.
2. **Pre-trained BERT Model:**
    - Utilizes the "distilbert-base-uncased" pre-trained BERT model from Hugging Face's Transformers library.
    - The BERT model is followed by a Masking layer to handle padding in the input sequences.
3. **Pooling Layer:**
    - A GlobalAveragePooling1D layer is applied to the output of the BERT model. This layer reduces the dimensionality of the representation by taking the average across the sequence dimension.
4. **Output Layer:**
    - A Dense layer with SoftMax activation is used as the output layer, producing probabilities for each class.
    - The number of output nodes is set to 3, corresponding to the three hate speech classes: Normal, Offensive, and Hate.

B. Training And Settings:

Optimizer: Adam optimizer is used with default parameters.

Loss Function: Categorical Cross entropy loss is employed, suitable for multi-class classification tasks.

Metrics: Accuracy is used as the evaluation metric during training.

Training Data: The training dataset is a subset of the "gender identity" dataset, consisting of 18,000 tweets in English.

Validation Split: 20% of the training data is used for validation during training.

Batch Size: Training is performed with a batch size of 25.

Epochs: The model is trained for 3 epochs.

Tokenization: The training and testing data are tokenized using the DistilBERT tokenizer with dynamic maximum length.

Fine-Tuning: The BERT model is fine-tuned on the hate speech detection task, adapting its weights to the specific characteristics of the provided dataset.

Data Splitting: The dataset is split into training and testing sets, with 80% of the data used for training and 20% for testing.

Evaluation: The model is evaluated on the test data using the accuracy metric.

Cosine Similarity Demonstration: The code includes a demonstration of cosine similarity between pairs of example texts using the fine-tuned BERT embeddings.

Analysis: The code further analyses correct and incorrect predictions on the test data, providing observations for both.

This architecture and training setting aim to leverage the power of pre-trained BERT representations for hate speech detection, with a focus on the specific "gender identity" dataset.

V. Comments On TASK-1

- Training accuracy starts around 58%, and validation accuracy is approximately 57.25% after the first epoch. Subsequent epochs show fluctuations in accuracy. After training, the model is evaluated on the test dataset.
- The reported accuracy on the test data is approximately 63.35%.Initial validation accuracy is around 57.25%, with slight increases and fluctuations in subsequent epochs. The final test accuracy of 63.35% suggests the model has learned patterns from the training data but may benefit from further improvement.
- Model performance can be influenced by hyperparameter choices (learning rate, batch size, epochs). Fine-tuning these parameters might lead to better results. Fluctuations in accuracy may indicate the need for regularization techniques or adjustments to the learning rate schedule.
- Experiment with different hyperparameter settings for potential improvements. Analyse misclassified examples to understand model weaknesses and enhance performance. Monitor loss and accuracy trends to determine if additional training epochs are beneficial.

VI. The 3 observations from Task-2

A. Observations for correct model prediction

1. Example 0:
   Predicted: 0

True Label: 0
Text: The least of things when we fight against the terrorism of non-reformed religions is not to imitate its non-inclusive discursive strategies ...
Gender equality is not a “mortal peril”, as these “white collar” terrorists chant!

2. Example 2:
   Predicted: 0
   True Label: 0
   Text: Transgender People Become Pawns in Culture War
   https://t.co/5ZpfgQJnFS

3. Example 3:
   Predicted: 0
   True Label: 0
   Text: “For years I’ve been hiding a secret. A secret that many people will not like. In fact they will probably hate me and unfollow me... but you know what... I don’t care at this point. I’m sick and tired of living a lie and to be perfectly honest I’m bisexual. Yes I said

4. Example 4:
   Predicted: 0
   True Label: 0
   Text: Me @ problematic tropes involving LGBT+: wtf why can't we be happy??? Why so much suffering/abuse/etc?
   Me @ actual healthy, happy LGBT+ representation: omg is that even real? Will I ever be able to experience that or is it a lie?

5. Example 5:
   Predicted: 0
   True Label: 0
   Text: I may be gay but my God @AlexaBliss_WWE is bloody beautiful and talented. I don't understand why @WWE wants to push stars that have had their time and not give it to their current stars who could be major players for the company given the chance!

6. Example 6:
   Predicted: 0
   True Label: 0
   Text: Detroit fam: LGBTQ+ discrimination sucks. Luckily, @FairandEqualMI is collecting petitions to expand MI's #ELCRA to include sexual orientation, gender identity and gender expression. Want to learn more and sign the petition? Come to their kickoff March 7!
   https://t.co/gYenk2bku1

7. Example 7:
   Predicted: 0
   True Label: 0
   Text: BREAKING: Local Gay Starts War On Christmas 🎄 Merry December, Fam! https://t.co/zCL8ly6OhS

8. Example 10:
   Predicted: 0
   True Label: 0
   Text: "LGBTQI people are still being persecuted, tortured and even killed because of sodomy laws… that were put in place by British imperialists."
   - Lady Phyll 💜 🙏🏿

   https://t.co/IufAx41GWI

9. Example 11:
   Predicted: 0
   True Label: 0
   Text: "Closeting is not a thing/Closeting is illegal" Okay Tiffany go and say that to the guy from my neighbourhood who is a well known celebrity here who is gay but he's being closeted, his record label making him a womaniser and so on. Stop fucking saying that, it makes my blood boil

10. Example 12:
    Predicted: 0
    True Label: 0
    Text: Telling a gay person that they're gonna go to hell from a young age and watch him internalise and embrace it. Don't get mad when it happens

B. 10 Examples for incorrect model prediction

1. Example 1:
   Predicted: 0
   True Label: 1
   Text: moot is straight and she’s saying she’s gay for lisa now ??? dumb bitch no youre not stfu

2. Example 8:
   Predicted: 0
   True Label: 2
   Text: On this #TransDayOfVisibility, we ask members to contact their state senators and tell them to oppose #SB29. This cruel bill is another attempt to discriminate against transgender students.
   More info here: https://t.co/e9VolN7q8k https://t.co/97iFvXihEv

3. Example 9:
   Predicted: 0
   True Label: 1
   Text: My 20s really were disgusting. The amount of negative disgusting behaviour I got from London gays makes me now not want to live and enjoy life. They were so disgusting and cruel. But I’m meant to just pick myself up and move un-phased

4. Example 13:
   Predicted: 0
   True Label: 1
   Text: CALLING ME GAY IS NOT AN INSULT 💯 That’s Like Saying Milk Is White Dumb Bitch😂 😂 😂

5. Example 14:
   Predicted: 0
   True Label: 1
   Text: made the mistake of reading * ****'s essay and twitter thread and screamed out loud at "men who feel they are women" bitch what are you SAYING??! read the thousands of gender theory resources you undoubtedly have access to!!

6. Example 20:
   Predicted: 0
   True Label: 1
   Text: cant eat until after i take this stupid blood test for this gay ass abilify smh

7. Example 21:
   Predicted: 0
   True Label: 1
   Text: jesus was hella gay and kinky, he literally told his 12 disciples to eat his body and drink his blood

8. Example 23:
   Predicted: 0
   True Label: 1
   Text: You ever look at some gay people and you’re like damn i don’t belong with these WEIRDOS

9. Example 24:
   Predicted: 0
   True Label: 1
   Text: Historia Reiss from Attack on Titan is a lesbian! https://t.co/z25RVjvPTO

10. Example 26:
    Predicted: 0
    True Label: 2
    Text: I’m having the hardest time finding out who I am, like sexuality wise. I am bisexual that’s a fact but I am romantically more attracted to women. Like I

dream about dating women more than I dream about dating men. But damn I like s*x with a man. Everything is so confusing

C. Observations of correct and incorrect examples

The provided text appears to be an assessment or evaluation of a language model's performance, specifically focusing on its sensitivity to context, challenges with negative tones, and its handling of individual expressions within the LGBTQ+ (Lesbian, Gay, Bisexual, Transgender, Queer/Questioning, and others) context. Let's break down each point for a more detailed description:

1. **Sensitivity to Context:**
   - Positive Recognition: The model is effective in recognizing and understanding positive and supportive content. This suggests that it can appropriately interpret and respond to messages that convey affirmative sentiments within the LGBTQ+ context.
   - Struggle with Nuanced or Sarcastic Language: However, the model faces challenges when confronted with nuanced or sarcastic language. Nuances and sarcasm often require a deeper understanding of context, tone, and subtle cues, which the model may find difficult to accurately grasp.

2. **Challenges with Negative Tone:**
   - Distinguishing Constructive Criticism: The model encounters difficulties in distinguishing between constructive criticism and derogatory language. This implies that it might struggle to discern the intention behind negative statements, making it challenging to differentiate between feedback aimed at improvement and outright harmful or offensive remarks.
   - Need for Additional Training: The suggestion of additional training implies that there's room for improvement in the model's ability to understand and appropriately respond to negative sentiments within the LGBTQ+ context.

3. **Individual Expressions:**
   - Understanding Individual Expressions: The model is reported to demonstrate an understanding of individual expressions and experiences within the LGBTQ+ community. This implies that it can recognize and respond to diverse narratives and personal stories.
   - Improvement Needed for a Wider Range: Despite its capability with understanding individual expressions, there is room for improvement in handling a wider range of expressions, including those with negative undertones. This suggests that the model may struggle when faced with more diverse or complex individual experiences within the LGBTQ+ community.
   - In summary, the model performs well in recognizing positive and supportive content but faces challenges with nuanced or sarcastic

language and distinguishing between different tones, particularly negative ones. It also shows an understanding of individual expressions within the LGBTQ+ community but requires improvement to handle a broader spectrum of experiences, including those with negative undertones. Additional training is recommended to enhance the model's performance in these areas.

VII. 5 Examples of Task-3 Results:

1. Text 1: he kangaroo is hopping roo-jumps.
   Text 2: The turtle is sunbathing.
   Cosine Similarity Score: 0.9999977946281433

2. Text 1: The penguin is sliding ice-slips.
   Text 2: The bear is hibernating.
   Cosine Similarity Score: 0.9999999403953552

3. Text 1: The giraffe is stretching neckyreach.
   Text 2: The monkey is swinging.
   Cosine Similarity Score: 0.999999463558197

4. Text 1: The dolphin is leaping sea-flips.
   Text 2: The whale is singing.
   Cosine Similarity Score: 1.0

5. Text 1: The owl is hooting night-tunes.
   Text 2: The rabbit is nibbling on grass.
   Cosine Similarity Score: 0.9999996423721313

Comments:

The provided examples seem to be pairs of texts describing different activities of animals, and a cosine similarity score is calculated for each pair. Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space, often used in natural language processing to determine the similarity between two texts. Let's examine each example along with comments:

1. Text 1: he kangaroo is hopping roo-jumps.
Text 2: The turtle is sunbathing.
Cosine Similarity Score: 0.9999977946281433

Comment:

- The cosine similarity score is very high, suggesting that the model perceives a significant similarity between the two sentences.

- However, the actual content of the texts describes completely different activities – one involving the kangaroo hopping and the other involving the turtle sunbathing.
- This discrepancy indicates that the model may struggle to accurately assess the semantic meaning or context of sentences, leading to misleading similarity scores.

2. Text 1: The penguin is sliding ice-slips.
Text 2: The bear is hibernating.
Cosine Similarity Score: 0.9999999403953552

Comment:

- The cosine similarity score is exceptionally high, indicating a very close similarity between the two sentences.
- However, the described activities are quite different – one involving a penguin sliding and the other involving a bear hibernating.
- This suggests a limitation in the model's ability to discern the actual content and context of the sentences, leading to inaccurate similarity scores.

3. Text 1: The giraffe is stretching neckyreach.
Text 2: The monkey is swinging.
Cosine Similarity Score: 0.999999463558197

Comment:

- The high cosine similarity score implies a strong perceived similarity between the two sentences.
- Nevertheless, the content of the texts describes distinct activities – one involving a giraffe stretching and the other involving a monkey swinging.
- This further indicates that the model may struggle to capture the nuances of different actions or behaviors performed by animals.

4. Text 1: The dolphin is leaping sea-flips.
Text 2: The whale is singing.
Cosine Similarity Score: 1.0

Comment:

- The perfect cosine similarity score (1.0) suggests identical or nearly identical meaning between the two sentences.
- However, the actual content of the texts describes different activities – one involving a dolphin leaping sea-flips and the other involving a whale singing.
- This raises concerns about the model's effectiveness in accurately capturing the diversity of animal behaviors.

5. Text 1: The owl is hooting night-tunes.
Text 2: The rabbit is nibbling on grass.
Cosine Similarity Score: 0.9999996423721313

Comment:

- The high cosine similarity score indicates a perceived strong similarity between the two sentences.
- Yet, the described activities are distinct – one involving an owl hooting night-tunes and the other involving a rabbit nibbling on grass.
- This underscores the model's difficulty in distinguishing between different actions performed by animals.

Overall, while the cosine similarity scores suggest high similarity between the text pairs, the actual content of the sentences reveals significant differences in the described activities. This may indicate a limitation in the model's understanding of context and semantics, particularly in distinguishing between diverse actions or behaviors.