Bert-TweetEval: Natural Language Classification

Bence Danko
dept. of applied data science
San Jose State University
San Jose, USA
bence.danko@sjsu.edu
0009-0000-5824-9296

***Abstract*—Extracting sentiment and intent from human natural language holds immense value in strategic decision-making in many domains. A variety of transformer architectures and base models have emerged as notable language processors, but they vary widely in training scale, vocabulary, and parametric counts. In real-world deployment, models can be hindered due to their cost and latency constraints. Models deployed to production are also subject to real-world stress cases, such as lexical diversity, unknown symbols and vocabulary, and dataset imbalance from the training data. In this work, we analyze lightweight variants of base and fine-tuned Bidirectional Encoder Representations from Transformers (BERT) models performance on the TweetEval emotion classification task. We compare and train DistilBERT and DistilRoBERTa variants and the suitability of their tokenizer architectures (WordPiece, BPE) for the emotion classification domain and their impact on performance. We construct a framework to stress-test distribution shifts and corrupted inputs, and conduct structured error analysis and interpret model confidence and calibration. We also benchmark additional competitive LLM models, Qwen3-4B-Instruct-2507 and GPT 4o-mini, under consistent prompting strategies on the same classification task. All code is released to the public at https://github.com/bencejdanko/bert-tweeteval. Models are released to the public at https://huggingface.co/bdanko.**

***Index Terms*—Emotion analysis, natural language understanding, transformer, DistilBERT, RoBERTa, tokenizer**

I. Introduction and Related Work

TweetEval [1] consists of seven Twitter-specific classification tasks, including emoji prediction, emotion recognition, hate speech detection, irony detection, offensive language identification, sentiment analysis, and stance detection. TweetEval and BERT-variant combinations have already been extensively explored. BERTweet [2], a prior RoBERTa-based model trained on a corpus of 850 million English tweets, established the state-of-the-art (SOTA) baseline across most of TweetEval's subtasks and proved the value in domain-specific pre-training, outperforming the original BERT and RoBERTa. TimeLMs [3] later introduced models continuously trained on fresh Twitter data, outperforming BERTweet in all TweetEval domains except irony detection. SuperTweetEval [4] has also been since released, adding several more NLP task domains that TweetEval lacked.

II. Task Description

In this work, we will be targeting emotion classification task from the original TweetEval. Each sample is labeled with one of 4 classes: "anger", "sadness", "joy", or "optimism". The dataset is heavily imbalanced, with an overrepresentation of "anger" classes at 42.98% of the dataset, and underrepresentation of "optimism" at 9.03%, while joy and sadness sit at 21.74% and 26.25% respectively. In total, we have a training and validation set of sizes 3257 and 374. We test our results on 1421 samples.

We compare models across these key metrics:

- **Accuracy**: Number of overall correct classifications over the validation sample.
- **Macro F1**: Macro F1 is the arithmetic mean of the F1 scores calculated for each individual class. It treats all classes equally, regardless of how many samples each class contains, meaning unbalanced classes are given the same weight:

$$\text{Macro F1} = \frac{1}{n}\sum_{i=1}^{n} F1_i$$

- **Precision**: Number of true positives calculated over all classifications marked as positives. Measures the quality of a positive prediction. We are taking the macro average.
- **Recall**: The true positive rate. Out of all positive cases in the data, how many can the classifier identify. We are taking the macro average.
- **ms/100 samples**: Test of throughput. We test this by running evaluation inference in batches of 100 on all our models, except for GPT 4o-mini, in which we instead wall-clock for 100 async API requests (up to 20 concurrent).
- **Expected Calibration Error (ECE)**: A measure of how well a model's confidence aligns with its actual accuracy. If a model assigns a probability of 0.90 to a prediction, it should be correct 90% of the time. This tests for overconfident or underconfident models, and typically an ECE score of 0.01 or 0.02 is considered reliable.

To ensure reproducability, we set all configurable seeds to 15179996. No further data processing was done. For example, in future work, it would be possible to experiment with NLP augmentation techniques like Easy Data Augmentation (EDA) [5], back-translation [6], random masking or MixUp [7] and variants. However, these are currently out of scope for our tasks, so we will omit them. All local tests were done evaluated using the NVIDIA L4 GPU as hardware.

TABLE I
EVALUATION REFERENCE

Question	Location
Part A – Zero-Shot Baseline	Final Results at Table II
Part B – Fine-Tuning Transformers	Final Results at Table II. Training strategy located at Section V. Loss curves and explanation at Appendix G Deployment Stress tests at Section V-A
Part C – Error Analysis	Error Analysis at Section VI
Part D – LLMs	Minimal Prompt at Appendix A Strucured Prompt at Appendix B Final Results at Table II Prompt Explanation at Section VII
Other Requirements	Summary of hyperparameters at Section V. Final output table at Table II Training screenshots located at Appendix D Confusion matrices located at Appendix C

A reference guide for evaluators

III. Summarized Results

In our studies, we reaffirmed that domain specific pre-training provides BERT models state of the art results on NLP classification tasks. We also demonstrate the arising competitive performance of open-source decoder models against closed-source.

IV. Baseline Analysis

On our baseline analysis, we found that the original DistilBert and DistilRoBERTa models severely underperformed on the TweetEval dataset. From Figure 8 and Figure 9, we see an extreme bias towards particular classifications and distributions that do not match our dataset and domain.

V. Training Strategy

On both models, we initialize for 20 training epochs, a batch size of 16, AdamW optimization using a learning rate of 2e-5 and weight decay of 0.01. We employ EarlyStoppingCallback, early stopping based on the Macro F1 score on the validation set, and then select for the model with the best Macro F1 score after 3 failed improvements (patience of 3). We choose early stopping as our models experimentally overfit very easily, and we select Macro F1 as our stopping metric as it weighs each class equally, which is particularly important due to our imbalanced dataset.

A. Corruption Stress Testing

To simulate data-corruption stress tests, we randomly introduce typos, hashtag splitting, and emoji removal.

- **Typos**: We randomly swap, delete, or insert characters into words with chance $p = 0.1$. This tests the model robustness against misspelled but recognizable words.
- **Split Hashtags**: We identify hashtags and split CamelCase words or remove the hashtag. This tests if the model relies on the hashtag or underlying semantic content better.
- **Remove Emoji**: Emojis strongly indicate emotion, and removing them tests if the model is robust enough to understand the other semantic cues.

We'll also conduct domain shift simulation. We create a shift by filtering tweets without mentions, links, or hashtags and compare performance.

- **Mentions**: We compare performance once we strip all @user mentions from the evaluation.
- **Links**: We compare performance once we strip all http links.
- **Hashtags**: We compare performance once we strip all hashtags.

VI. Error Analysis

8 Missclassified samples from both models can be seen in Appendix F. We can see how misspellings, such as "Deppression", cause unrecognizable token fragments '#', 'de', '##pp', '##ress', '##ion'. These can't be clearly mapped to any particular emotional state, and may be unrecognizable from the pre-trained vocabulary.

For misspellings, we would need to implement character-level data augmentation to simulate typos. We would inject random character insertions, deletions, and keyboard-distance typos, especially targetting emotional keywords. By forcing the model to see misspelled variants, we train the attention heads to recognize the pattern of the fragments. This may improve the score.

Other missclassified sentences have words that seem semantically biased for one class, where subtle, but important tokens change the meaning significantly. For example, both tokenizers correctly break down "revolting", but neither model can weigh "i am" enough to overcome the "angry" connotations and classifications. Thus "i am revolting" is missclassified as "angry".

In order to counteract this, we need to increase our training samples or introduce more robust data augmentation to increase the semantic representation and understanding for our models. The best technique in this case would be Counterfactual data augmentation (CDA), where we generate more samples consisting of the word "revolting" in all scenarios, and thus we can diversify the model's interpretation of "revolting" across a greater number of samples and classes.

VII. Open Source and Closed Models

We evaluated two LLM models, Qwen3-4B-Instruct-2507 and GPT 4o-mini, on prompts from Appendix A and Appendix B. They achieved SOTA results without manual tuning, and responses aligned with the distribution of the training data.

A structured result demonstrated greater performance in all metrics over minimal prompts, except on throughput for Qwen. Longer prompts tap further into the parametric memory of models, they can elicit the pre-trained memory to produce more aligned responses. Our longer prompt thus increased the statistical chances of the model producing accurate emotion-classification assessment. Without such context priming, the model is not parametrically activated in the same specialized manner, and thus was more unlikely to produce an aligned response.

TABLE II
SUMMARIZED RESULTS FROM TWEETEVAL EMOTION CLASSIFICATION

Model	Accuracy	Macro F	Macro Precision	Macro Recall	ms/100 samples	ECE
DistilBERT (WordPiece)	0.083744	0.064219	0.384379	0.192035	312.318607	0.167510
DistilRoBERTa (BPE)	0.217452	0.155317	0.177579	0.200315	**299.071175**	0.056845
bdanko/bert-tweeteval-distilbert	0.79803	0.761196	0.767879	0.756296		0.0364
bdanko/bert-tweeteval-distilroberta	0.788881	0.750644	0.799548	0.728545		0.0364
GPT-4o-mini (Minimal Prompt)	0.800141	0.601466	0.653766	0.579754	5060.112734	
GPT-4o-mini (Structured Prompt)	**0.821956**	**0.781499**	0.791218	**0.773544**	3823.308211	
Qwen3-4B-Instruct-2507 (Minimal Prompt)	0.751583	0.584864	0.594974	0.581751	2571.471757	
Qwen3-4B-Instruct-2507 (Structured Prompt)	0.812104	0.758003	**0.793805**	0.741873	4931.231370	

Final comparison chart across all model evaluations on TweetEval Emotion Classification. While we see that LLMs dominate on performance benchmarks, BERT-architecture models can reach nearly the same performance.

TABLE III
CORRUPTION ABLATIONS

Dataset Shift / Corruption	Accuracy (DistilBERT)	ECE (DistilBERT)	Macro F1 (DistilBERT)	Macro Precision (DistilBERT)	Macro Recall (DistilBERT)	Accuracy (DistilRoBERTa)	ECE (DistilRoBERTa)	Macro F1 (DistilRoBERTa)	Macro Precision (DistilRoBERTa)	Macro Recall (DistilRoBERTa)
All corruptions	0.777621	0.186762	0.731208	0.748215	0.719606	0.769880	0.027936	0.729321	0.785784	0.704836
Emoji Removal	0.796622	0.169403	0.759471	0.765310	0.754525	0.788177	0.044574	0.754264	0.799723	0.730673
Hashtag Splitting	0.797326	0.169666	0.758076	0.767174	0.751265	0.790289	0.049607	0.750437	0.804039	0.727826
Typos	0.774806	0.192172	0.735679	0.750336	0.726397	0.762139	0.032455	0.719843	0.767177	0.700259
Baseline	0.798030	0.169843	0.761196	0.767879	0.756296	0.788881	0.042115	0.750644	0.799548	0.728545
All Domain Shifts	0.798030	0.169843	0.761196	0.767879	0.756296	0.788881	0.042115	0.750644	0.799548	0.728545
No Hashtags	0.779412	0.183594	0.729214	0.740608	0.721849	0.783422	0.055099	0.730822	0.795115	0.706223
No HTTP Links	0.798030	0.169843	0.761196	0.767879	0.756296	0.788881	0.042115	0.750644	0.799548	0.728545
No @ Mentions	0.786865	0.178595	0.764952	0.763810	0.766616	0.788104	0.041246	0.763955	0.796607	0.749939

All corruption ablations and domain shifting ablations. While typos and corruption_all impact base model's accuracy (0.77), the distilroberta model maintains higher calibration.

The final results for each tested model has been collected and summarized in Table II.

REFERENCES

[1] F. Barbieri, J. Camacho-Collados, L. Neves, and L. Espinosa-Anke, "Tweeteval: Unified benchmark and comparative evaluation for tweet classification," 2020. [Online]. Available: https://arxiv.org/abs/2010.12421

[2] D. Q. Nguyen, T. Vu, and A. Tuan Nguyen, "BERTweet: A pre-trained language model for English tweets," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Q. Liu and D. Schlangen, Eds. Online: Association for Computational Linguistics, Oct. 2020, pp. 9–14. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.2/

[3] D. Loureiro, F. Barbieri, L. Neves, L. E. Anke, and J. Camacho-Collados, "Timelms: Diachronic language models from twitter," 2022. [Online]. Available: https://arxiv.org/abs/2202.03829

[4] D. Antypas, A. Ushio, F. Barbieri, L. Neves, K. Rezaee, L. Espinosa-Anke, J. Pei, and J. Camacho-Collados, "Supertweeteval: A challenging, unified and heterogeneous benchmark for social media nlp research," 2023. [Online]. Available: https://arxiv.org/abs/2310.14757

[5] J. Wei and K. Zou, "Eda: Easy data augmentation techniques for boosting performance on text classification tasks," 2019. [Online]. Available: https://arxiv.org/abs/1901.11196

[6] S. Edunov, M. Ott, M. Auli, and D. Grangier, "Understanding back-translation at scale," 2018. [Online]. Available: https://arxiv.org/abs/1808.09381

[7] L. Sun, C. Xia, W. Yin, T. Liang, P. S. Yu, and L. He, "Mixup-transformer: Dynamic data augmentation for nlp tasks," 2020. [Online]. Available: https://arxiv.org/abs/2010.02394

Appendix A
Minimal Prompt for TweetEval Emotion Classification

```
Classify the following tweet into one of these emotions: anger, joy, optimism,
 ↪  sadness.
Tweet: {text}
Emotion:
```

Appendix B
Structured Prompt for TweetEval Emotion Classification

```
Task: Sentiment classification for tweets.
Labels:
- anger: The tweet expresses frustration, resentment, or rage.
- joy: The tweet expresses happiness, pleasure, or satisfaction.
- optimism: The tweet expresses hopefulness, confidence about the future, or positive
 ↪  anticipation.
- sadness: The tweet expresses sorrow, disappointment, or unhappiness.

Instructions:
1. Read the tweet provided below.
2. Select the most appropriate label from the list above.
3. Output ONLY the label name. Do not include any other text or explanation.

Tweet: {text}

Label:
```

Appendix C
Confusion matrices comparison across all evaluated models.

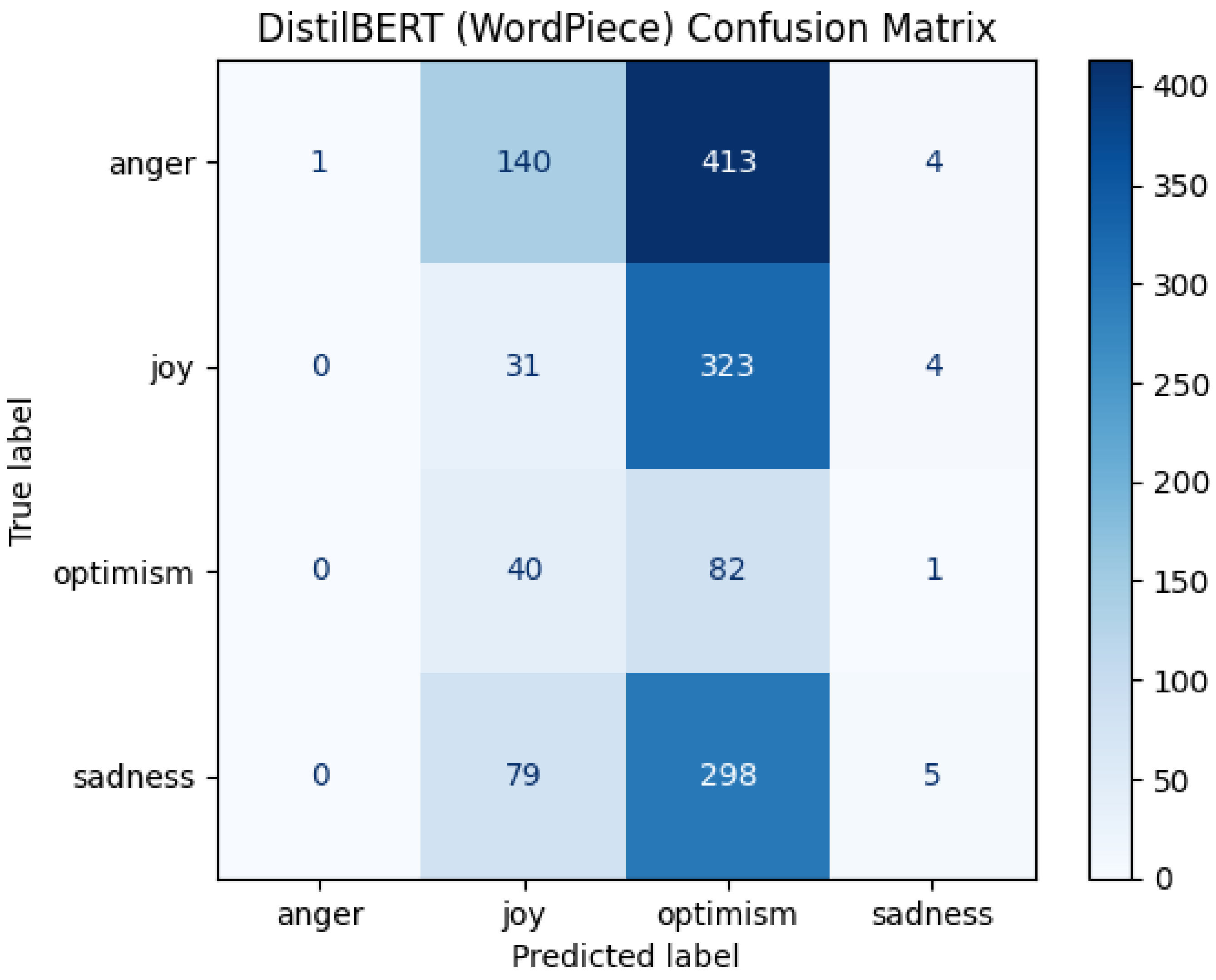


Fig. 1. Baseline DistilBERT confusion matrices on the TweetEval emotion prediction task. Notice how the model appears to be heavily biased for classifying for optimism by default.

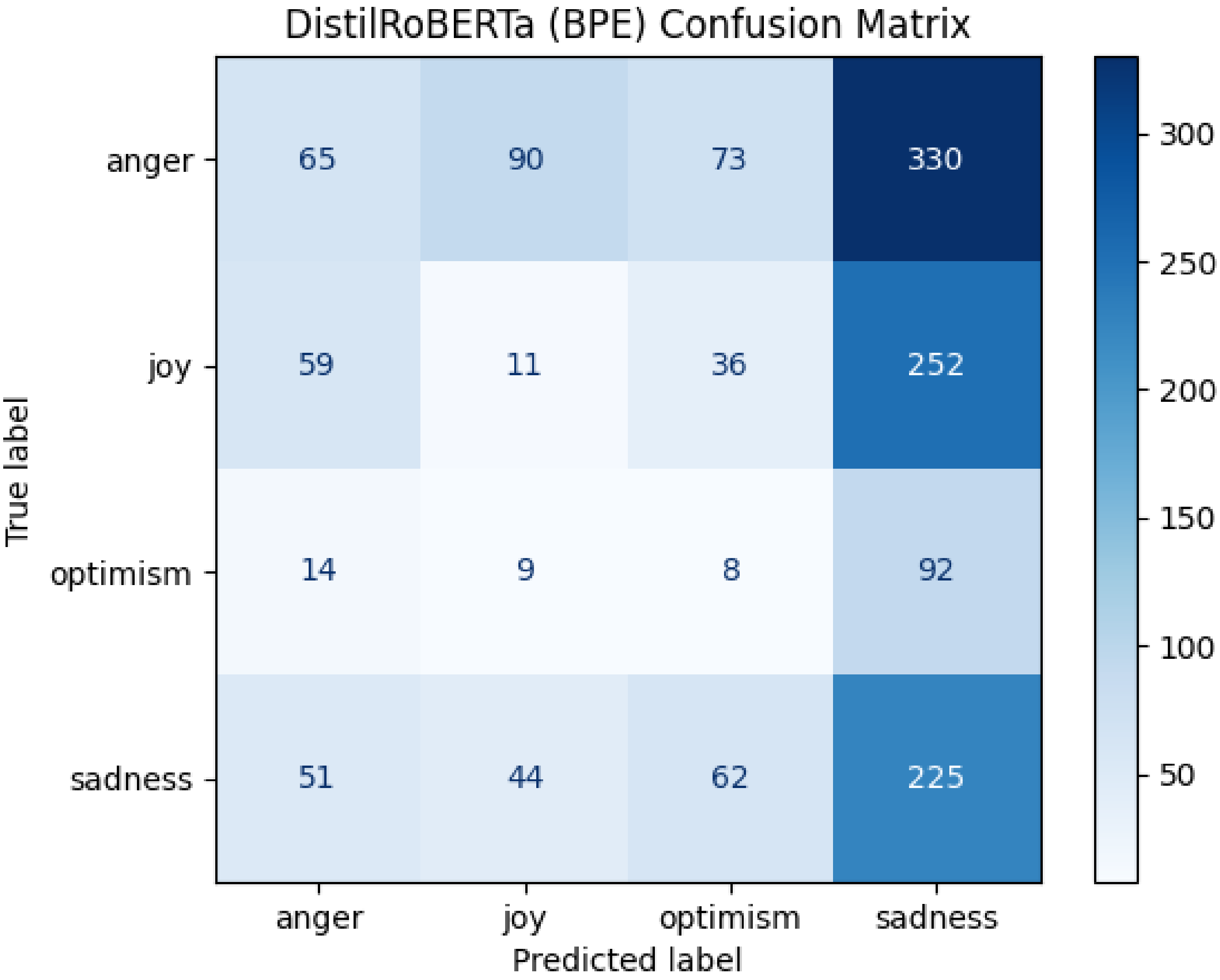


Fig. 2. Baseline DistilRoBERTa confusion matrices on the TweetEval emotion prediction task. Notice how the model is slightly more distributed in classifying then DistilBERT base, but still tends to classify for sadness by default.

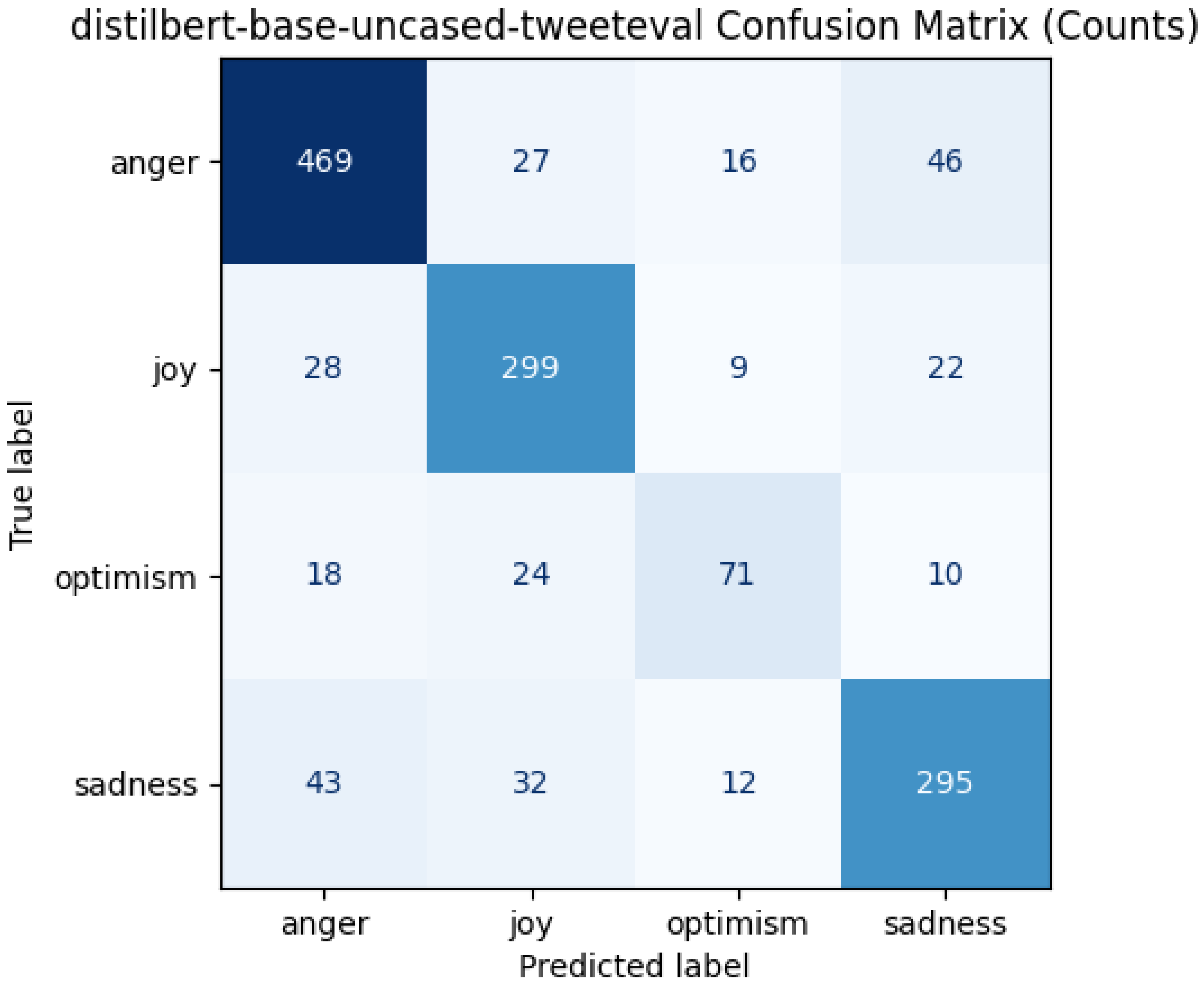


Fig. 3. Confusion Matrices for bert-tweeteval-distilbert. Notice how the classifications now tend towards the original training distributions, in comparison to the original base classifications.

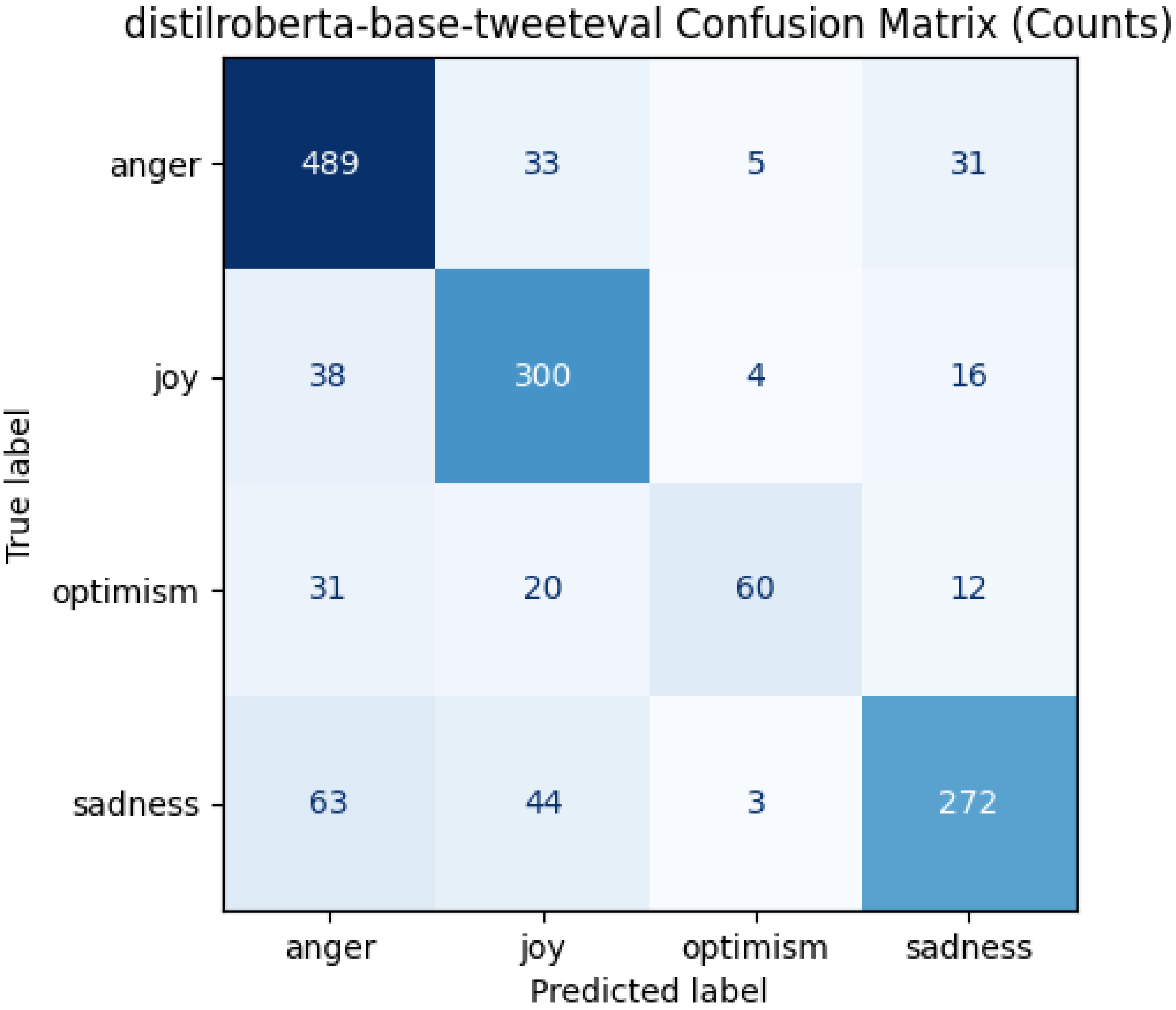


Fig. 4. Confusion Matrices for bert-tweeteval-distilroberta. Notice how the classifications now tend towards the original training distributions, in comparison to the original base classifications.

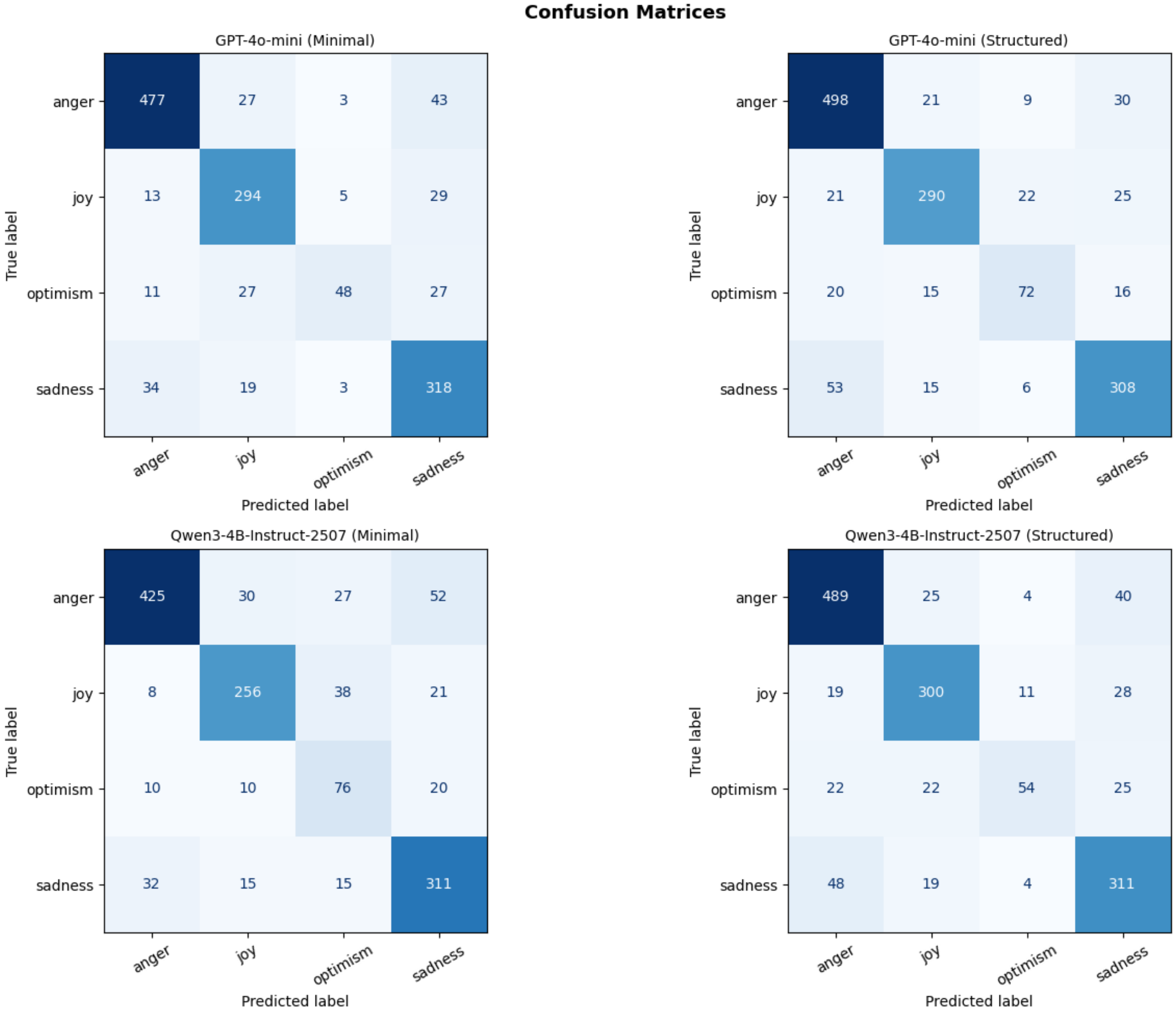
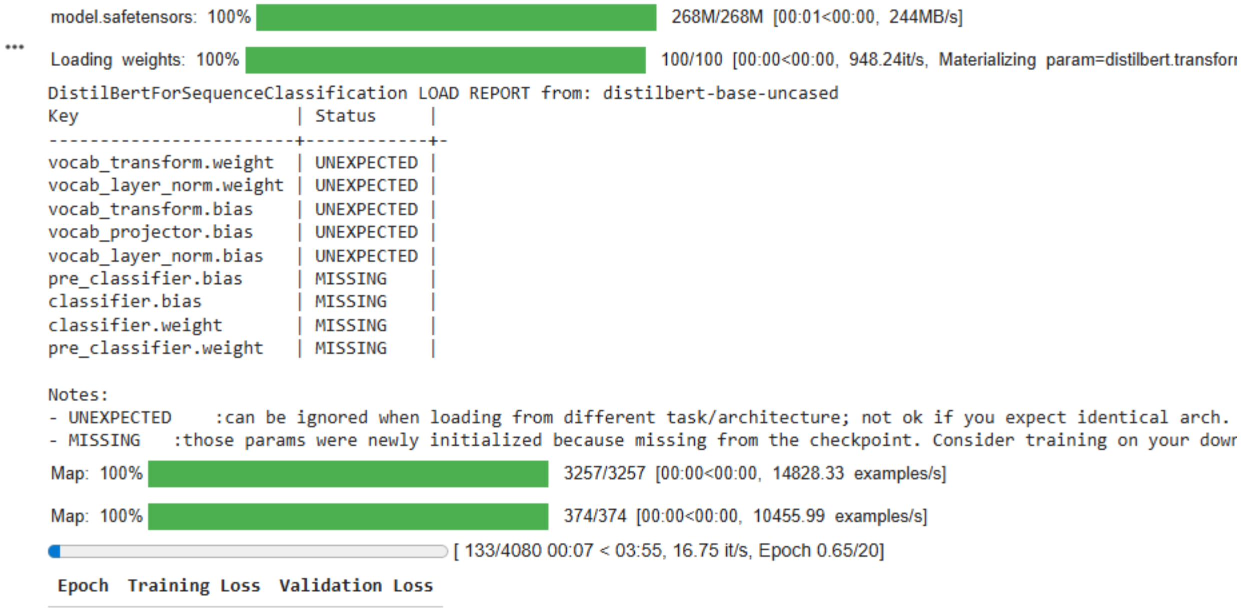


Fig. 5. Confusion Matrices for all LLM models and prompting strategies. Structured prompts clearly yielded better results.

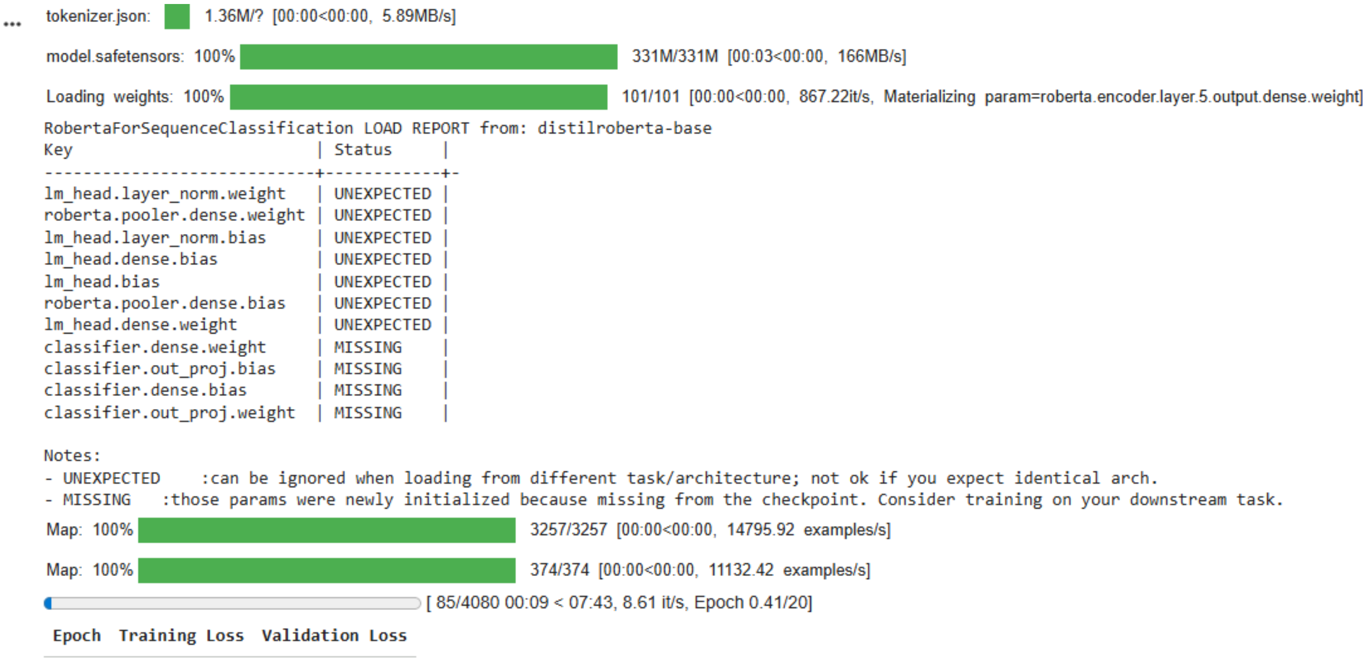
Appendix D
Screenshots for Run Verification



```
model.safetensors: 100% 268M/268M [00:01<00:00, 244MB/s]
Loading weights: 100% 100/100 [00:00<00:00, 948.24it/s, Materializing param=distilbert.transforı
DistilBertForSequenceClassification LOAD REPORT from: distilbert-base-uncased
Key                     | Status     |
------------------------+------------+-
vocab_transform.weight  | UNEXPECTED |
vocab_layer_norm.weight | UNEXPECTED |
vocab_transform.bias    | UNEXPECTED |
vocab_projector.bias    | UNEXPECTED |
vocab_layer_norm.bias   | UNEXPECTED |
pre_classifier.bias     | MISSING    |
classifier.bias         | MISSING    |
classifier.weight       | MISSING    |
pre_classifier.weight   | MISSING    |

Notes:
- UNEXPECTED    :can be ignored when loading from different task/architecture; not ok if you expect identical arch.
- MISSING   :those params were newly initialized because missing from the checkpoint. Consider training on your dowr
Map: 100% 3257/3257 [00:00<00:00, 14828.33 examples/s]
Map: 100% 374/374 [00:00<00:00, 10455.99 examples/s]
[ 133/4080 00:07 < 03:55, 16.75 it/s, Epoch 0.65/20]
Epoch  Training Loss  Validation Loss
```

Fig. 6. Screenshot of the beginning of DistilBERT training



```
tokenizer.json: 1.36M/? [00:00<00:00, 5.89MB/s]
model.safetensors: 100% 331M/331M [00:03<00:00, 166MB/s]
Loading weights: 100% 101/101 [00:00<00:00, 867.22it/s, Materializing param=roberta.encoder.layer.5.output.dense.weight]
RobertaForSequenceClassification LOAD REPORT from: distilroberta-base
Key                         | Status     |
----------------------------+------------+-
lm_head.layer_norm.weight   | UNEXPECTED |
roberta.pooler.dense.weight | UNEXPECTED |
lm_head.layer_norm.bias     | UNEXPECTED |
lm_head.dense.bias          | UNEXPECTED |
lm_head.bias                | UNEXPECTED |
roberta.pooler.dense.bias   | UNEXPECTED |
lm_head.dense.weight        | UNEXPECTED |
classifier.dense.weight     | MISSING    |
classifier.out_proj.bias    | MISSING    |
classifier.dense.bias       | MISSING    |
classifier.out_proj.weight  | MISSING    |

Notes:
- UNEXPECTED    :can be ignored when loading from different task/architecture; not ok if you expect identical arch.
- MISSING   :those params were newly initialized because missing from the checkpoint. Consider training on your downstream task.
Map: 100% 3257/3257 [00:00<00:00, 14795.92 examples/s]
Map: 100% 374/374 [00:00<00:00, 11132.42 examples/s]
[ 85/4080 00:09 < 07:43, 8.61 it/s, Epoch 0.41/20]
Epoch  Training Loss  Validation Loss
```

Fig. 7. Screenshot for the beginning of DistilRoBERTa

Appendix E
Misclassified Base Models TweetEval Samples

```
Text: @user Interesting choice of words... Are you confirming that governments fund
 ↪  #terrorism? Bit of an open door, but still...
True: anger | Fine-Tuned DistilBERT Pred: label_0 | Fine-Tuned DistilRoBERTa Pred:
 ↪  label_0
WordPiece (DistilBERT): ['@', 'user', 'interesting', 'choice', 'of', 'words', '.',
 ↪  '.', '.', 'are', 'you', 'confirming', 'that', 'governments', 'fund', '#',
 ↪  'terrorism', '?', 'bit', 'of', 'an', 'open', 'door', ',', 'but', 'still', '.',
 ↪  '.', '.']
BPE (DistilRoBERTa): ['@', 'user', 'ĠInteresting', 'Ġchoice', 'Ġof', 'Ġwords', '...',
 ↪  'ĠAre', 'Ġyou', 'Ġconfirming', 'Ġthat', 'Ġgovernments', 'Ġfund', 'Ġ#',
 ↪  'terrorism', '?', 'ĠBit', 'Ġof', 'Ġan', 'Ġopen', 'Ġdoor', ',', 'Ġbut', 'Ġstill',
 ↪  '...']
--------------------------------------------------
Text: @user Welcome to #MPSVT! We are delighted to have you! #grateful #MPSVT
 ↪  #relationships
True: joy | Fine-Tuned DistilBERT Pred: label_1 | Fine-Tuned DistilRoBERTa Pred:
 ↪  label_1
WordPiece (DistilBERT): ['@', 'user', 'welcome', 'to', '#', 'mps', '##v', '##t', '!',
 ↪  'we', 'are', 'delighted', 'to', 'have', 'you', '!', '#', 'grateful', '#', 'mps',
 ↪  '##v', '##t', '#', 'relationships']
BPE (DistilRoBERTa): ['@', 'user', 'ĠWelcome', 'Ġto', 'Ġ#', 'M', 'PS', 'VT', '!',
 ↪  'ĠWe', 'Ġare', 'Ġdelighted', 'Ġto', 'Ġhave', 'Ġyou', '!', 'Ġ#', 'gr', 'ateful',
 ↪  'Ġ#', 'M', 'PS', 'VT', 'Ġ#', 'relations', 'hips']
--------------------------------------------------
Text: i am revolting.
True: anger | Fine-Tuned DistilBERT Pred: label_0 | Fine-Tuned DistilRoBERTa Pred:
 ↪  label_0
WordPiece (DistilBERT): ['i', 'am', 'revolt', '##ing', '.']
BPE (DistilRoBERTa): ['i', 'Ġam', 'Ġrevol', 'ting', '.']
--------------------------------------------------
Text: Rin might ever appeared gloomy but to be a melodramatic person was not her
 ↪  thing.\n\nBut honestly, she missed her old friend. The special one.
True: sadness | Fine-Tuned DistilBERT Pred: label_3 | Fine-Tuned DistilRoBERTa Pred:
 ↪  label_3
WordPiece (DistilBERT): ['ri', '##n', 'might', 'ever', 'appeared', 'gloom', '##y',
 ↪  'but', 'to', 'be', 'a', 'mel', '##od', '##rama', '##tic', 'person', 'was', 'not',
 ↪  'her', 'thing', '.', '\\', 'n', '\\', 'n', '##bu', '##t', 'honestly', ',', 'she',
 ↪  'missed', 'her', 'old', 'friend', '.', 'the', 'special', 'one', '.']
BPE (DistilRoBERTa): ['R', 'in', 'Ġmight', 'Ġever', 'Ġappeared', 'Ġgloomy', 'Ġbut',
 ↪  'Ġto', 'Ġbe', 'Ġa', 'Ġmel', 'od', 'ram', 'atic', 'Ġperson', 'Ġwas', 'Ġnot',
 ↪  'Ġher', 'Ġthing', '.', '\\', 'n', '\\', 'n', 'But', 'Ġhonestly', ',', 'Ġshe',
 ↪  'Ġmissed', 'Ġher', 'Ġold', 'Ġfriend', '.', 'ĠThe', 'Ġspecial', 'Ġone', '.']
--------------------------------------------------
Text: @user Get Donovan out of your soccer booth. He's awful. He's bitter. He makes me
 ↪  want to mute the tv. #horrid
True: anger | Fine-Tuned DistilBERT Pred: label_0 | Fine-Tuned DistilRoBERTa Pred:
 ↪  label_0
WordPiece (DistilBERT): ['@', 'user', 'get', 'donovan', 'out', 'of', 'your', 'soccer',
 ↪  'booth', '.', 'he', "'", 's', 'awful', '.', 'he', "'", 's', 'bitter', '.', 'he',
 ↪  'makes', 'me', 'want', 'to', 'mute', 'the', 'tv', '.', '#', 'ho', '##rri', '##d']
BPE (DistilRoBERTa): ['@', 'user', 'ĠGet', 'ĠDonovan', 'Ġout', 'Ġof', 'Ġyour',
 ↪  'Ġsoccer', 'Ġbooth', '.', 'ĠHe', "'s", 'Ġawful', '.', 'ĠHe', "'s", 'Ġbitter', '.',
 ↪  'ĠHe', 'Ġmakes', 'Ġme', 'Ġwant', 'Ġto', 'Ġmute', 'Ġthe', 'Ġtv', '.', 'Ġ#', 'hor',
 ↪  'rid']
```

```
--------------------------------------------------
Text: @user how can u have sold so many copies but ur game has so many fucking bugs
 ↪  and mad lag issues. Optimize ur shit soon.
True: anger | Fine-Tuned DistilBERT Pred: label_0 | Fine-Tuned DistilRoBERTa Pred:
 ↪  label_0
WordPiece (DistilBERT): ['@', 'user', 'how', 'can', 'u', 'have', 'sold', 'so', 'many',
 ↪  'copies', 'but', 'ur', 'game', 'has', 'so', 'many', 'fucking', 'bugs', 'and',
 ↪  'mad', 'la', '##g', 'issues', '.', 'opt', '##imi', '##ze', 'ur', 'shit', 'soon',
 ↪  '.']
BPE (DistilRoBERTa): ['@', 'user', 'Ġhow', 'Ġcan', 'Ġu', 'Ġhave', 'Ġsold', 'Ġso',
 ↪  'Ġmany', 'Ġcopies', 'Ġbut', 'Ġur', 'Ġgame', 'Ġhas', 'Ġso', 'Ġmany', 'Ġfucking',
 ↪  'Ġbugs', 'Ġand', 'Ġmad', 'Ġlag', 'Ġissues', '.', 'ĠOptim', 'ize', 'Ġur', 'Ġshit',
 ↪  'Ġsoon', '.']
--------------------------------------------------
Text: People who say nmu are the worst, something has to be going on, tell me I wanna
 ↪  know bout your life that's why I fucking asked, I care
True: anger | Fine-Tuned DistilBERT Pred: label_0 | Fine-Tuned DistilRoBERTa Pred:
 ↪  label_0
WordPiece (DistilBERT): ['people', 'who', 'say', 'nm', '##u', 'are', 'the', 'worst',
 ↪  ',', 'something', 'has', 'to', 'be', 'going', 'on', ',', 'tell', 'me', 'i',
 ↪  'wanna', 'know', 'bout', 'your', 'life', 'that', "'", 's', 'why', 'i', 'fucking',
 ↪  'asked', ',', 'i', 'care', '[UNK]']
BPE (DistilRoBERTa): ['People', 'Ġwho', 'Ġsay', 'Ġnm', 'u', 'Ġare', 'Ġthe', 'Ġworst',
 ↪  ',', 'Ġsomething', 'Ġhas', 'Ġto', 'Ġbe', 'Ġgoing', 'Ġon', ',', 'Ġtell', 'Ġme',
 ↪  'ĠI', 'Ġwanna', 'Ġknow', 'Ġbout', 'Ġyour', 'Ġlife', 'Ġthat', "'s", 'Ġwhy', 'ĠI',
 ↪  'Ġfucking', 'Ġasked', ',', 'ĠI', 'Ġcare', 'ĠðŁĺ', '¤']
--------------------------------------------------
Text: @user The hatred from the Left ought to concern everyone----who wants a police
 ↪  state-the left, so than can spy on all of us.
True: anger | Fine-Tuned DistilBERT Pred: label_0 | Fine-Tuned DistilRoBERTa Pred:
 ↪  label_0
WordPiece (DistilBERT): ['@', 'user', 'the', 'hatred', 'from', 'the', 'left', 'ought',
 ↪  'to', 'concern', 'everyone', '-', '-', '-', '-', 'who', 'wants', 'a', 'police',
 ↪  'state', '-', 'the', 'left', ',', 'so', 'than', 'can', 'spy', 'on', 'all', 'of',
 ↪  'us', '.']
BPE (DistilRoBERTa): ['@', 'user', 'ĠThe', 'Ġhatred', 'Ġfrom', 'Ġthe', 'ĠLeft',
 ↪  'Ġought', 'Ġto', 'Ġconcern', 'Ġeveryone', '----', 'who', 'Ġwants', 'Ġa',
 ↪  'Ġpolice', 'Ġstate', '-', 'the', 'Ġleft', ',', 'Ġso', 'Ġthan', 'Ġcan', 'Ġspy',
 ↪  'Ġon', 'Ġall', 'Ġof', 'Ġus', '.']
--------------------------------------------------
```

Appendix F
Misclassified Fine-Tuned Model TweetEval Samples

```
Text: #Deppression is real. Partners w/ #depressed people truly dont understand the
 ↪  depth in which they affect us. Add in #anxiety &amp;makes it worse
True: sadness | Fine-Tuned DistilBERT Pred: label_3 | Fine-Tuned DistilRoBERTa Pred:
 ↪  label_3
WordPiece (DistilBERT): ['#', 'de', '##pp', '##ress', '##ion', 'is', 'real', '.',
 ↪  'partners', 'w', '/', '#', 'depressed', 'people', 'truly', 'don', '##t',
 ↪  'understand', 'the', 'depth', 'in', 'which', 'they', 'affect', 'us', '.', 'add',
 ↪  'in', '#', 'anxiety', '&', 'amp', ';', 'makes', 'it', 'worse']
BPE (DistilRoBERTa): ['#', 'De', 'pp', 'ression', 'Ġis', 'Ġreal', '.', 'ĠPartners',
 ↪  'Ġw', '/', 'Ġ#', 'dep', 'ressed', 'Ġpeople', 'Ġtruly', 'Ġdont', 'Ġunderstand',
 ↪  'Ġthe', 'Ġdepth', 'Ġin', 'Ġwhich', 'Ġthey', 'Ġaffect', 'Ġus', '.', 'ĠAdd', 'Ġin',
 ↪  'Ġ#', 'an', 'xiety', 'Ġ&', 'amp', ';', 'makes', 'Ġit', 'Ġworse']
--------------------------------------------------
Text: @user Interesting choice of words... Are you confirming that governments fund
 ↪  #terrorism? Bit of an open door, but still...
True: anger | Fine-Tuned DistilBERT Pred: label_0 | Fine-Tuned DistilRoBERTa Pred:
 ↪  label_0
WordPiece (DistilBERT): ['@', 'user', 'interesting', 'choice', 'of', 'words', '.',
 ↪  '.', '.', 'are', 'you', 'confirming', 'that', 'governments', 'fund', '#',
 ↪  'terrorism', '?', 'bit', 'of', 'an', 'open', 'door', ',', 'but', 'still', '.',
 ↪  '.', '.']
BPE (DistilRoBERTa): ['@', 'user', 'ĠInteresting', 'Ġchoice', 'Ġof', 'Ġwords', '...',
 ↪  'ĠAre', 'Ġyou', 'Ġconfirming', 'Ġthat', 'Ġgovernments', 'Ġfund', 'Ġ#',
 ↪  'terrorism', '?', 'ĠBit', 'Ġof', 'Ġan', 'Ġopen', 'Ġdoor', ',', 'Ġbut', 'Ġstill',
 ↪  '...']
--------------------------------------------------
Text: My visit to hospital for care triggered #trauma from accident 20+yrs ago and
 ↪  image of my dead brother in it. Feeling symptoms of #depression
True: sadness | Fine-Tuned DistilBERT Pred: label_3 | Fine-Tuned DistilRoBERTa Pred:
 ↪  label_3
WordPiece (DistilBERT): ['my', 'visit', 'to', 'hospital', 'for', 'care', 'triggered',
 ↪  '#', 'trauma', 'from', 'accident', '20', '+', 'y', '##rs', 'ago', 'and', 'image',
 ↪  'of', 'my', 'dead', 'brother', 'in', 'it', '.', 'feeling', 'symptoms', 'of', '#',
 ↪  'depression']
BPE (DistilRoBERTa): ['My', 'Ġvisit', 'Ġto', 'Ġhospital', 'Ġfor', 'Ġcare',
 ↪  'Ġtriggered', 'Ġ#', 'tra', 'uma', 'Ġfrom', 'Ġaccident', 'Ġ20', '+', 'yrs', 'Ġago',
 ↪  'Ġand', 'Ġimage', 'Ġof', 'Ġmy', 'Ġdead', 'Ġbrother', 'Ġin', 'Ġit', '.',
 ↪  'ĠFeeling', 'Ġsymptoms', 'Ġof', 'Ġ#', 'dep', 'ression']
--------------------------------------------------
Text: @user Welcome to #MPSVT! We are delighted to have you! #grateful #MPSVT
 ↪  #relationships
True: joy | Fine-Tuned DistilBERT Pred: label_1 | Fine-Tuned DistilRoBERTa Pred:
 ↪  label_1
WordPiece (DistilBERT): ['@', 'user', 'welcome', 'to', '#', 'mps', '##v', '##t', '!',
 ↪  'we', 'are', 'delighted', 'to', 'have', 'you', '!', '#', 'grateful', '#', 'mps',
 ↪  '##v', '##t', '#', 'relationships']
BPE (DistilRoBERTa): ['@', 'user', 'ĠWelcome', 'Ġto', 'Ġ#', 'M', 'PS', 'VT', '!',
 ↪  'ĠWe', 'Ġare', 'Ġdelighted', 'Ġto', 'Ġhave', 'Ġyou', '!', 'Ġ#', 'gr', 'ateful',
 ↪  'Ġ#', 'M', 'PS', 'VT', 'Ġ#', 'relations', 'hips']
--------------------------------------------------
Text: What makes you feel #joyful?
True: joy | Fine-Tuned DistilBERT Pred: label_1 | Fine-Tuned DistilRoBERTa Pred:
 ↪  label_1
WordPiece (DistilBERT): ['what', 'makes', 'you', 'feel', '#', 'joy', '##ful', '?']
```

```
BPE (DistilRoBERTa): ['What', 'Ġmakes', 'Ġyou', 'Ġfeel', 'Ġ#', 'joy', 'ful', '?']
--------------------------------------------------
Text: i am revolting.
True: anger | Fine-Tuned DistilBERT Pred: label_0 | Fine-Tuned DistilRoBERTa Pred:
 ↪  label_0
WordPiece (DistilBERT): ['i', 'am', 'revolt', '##ing', '.']
BPE (DistilRoBERTa): ['i', 'Ġam', 'Ġrevol', 'ting', '.']
--------------------------------------------------
Text: Rin might ever appeared gloomy but to be a melodramatic person was not her
 ↪  thing.\n\nBut honestly, she missed her old friend. The special one.
True: sadness | Fine-Tuned DistilBERT Pred: label_3 | Fine-Tuned DistilRoBERTa Pred:
 ↪  label_3
WordPiece (DistilBERT): ['ri', '##n', 'might', 'ever', 'appeared', 'gloom', '##y',
 ↪  'but', 'to', 'be', 'a', 'mel', '##od', '##rama', '##tic', 'person', 'was', 'not',
 ↪  'her', 'thing', '.', '\\', 'n', '\\', 'n', '##bu', '##t', 'honestly', ',', 'she',
 ↪  'missed', 'her', 'old', 'friend', '.', 'the', 'special', 'one', '.']
BPE (DistilRoBERTa): ['R', 'in', 'Ġmight', 'Ġever', 'Ġappeared', 'Ġgloomy', 'Ġbut',
 ↪  'Ġto', 'Ġbe', 'Ġa', 'Ġmel', 'od', 'ram', 'atic', 'Ġperson', 'Ġwas', 'Ġnot',
 ↪  'Ġher', 'Ġthing', '.', '\\', 'n', '\\', 'n', 'But', 'Ġhonestly', ',', 'Ġshe',
 ↪  'Ġmissed', 'Ġher', 'Ġold', 'Ġfriend', '.', 'ĠThe', 'Ġspecial', 'Ġone', '.']
--------------------------------------------------
Text: In need of a change! #restless
True: sadness | Fine-Tuned DistilBERT Pred: label_3 | Fine-Tuned DistilRoBERTa Pred:
 ↪  label_3
WordPiece (DistilBERT): ['in', 'need', 'of', 'a', 'change', '!', '#', 'restless']
BPE (DistilRoBERTa): ['In', 'Ġneed', 'Ġof', 'Ġa', 'Ġchange', '!', 'Ġ#', 'rest',
 ↪  'less']
--------------------------------------------------
```

Appendix G
BERT-TweetEval training loss diagrams.

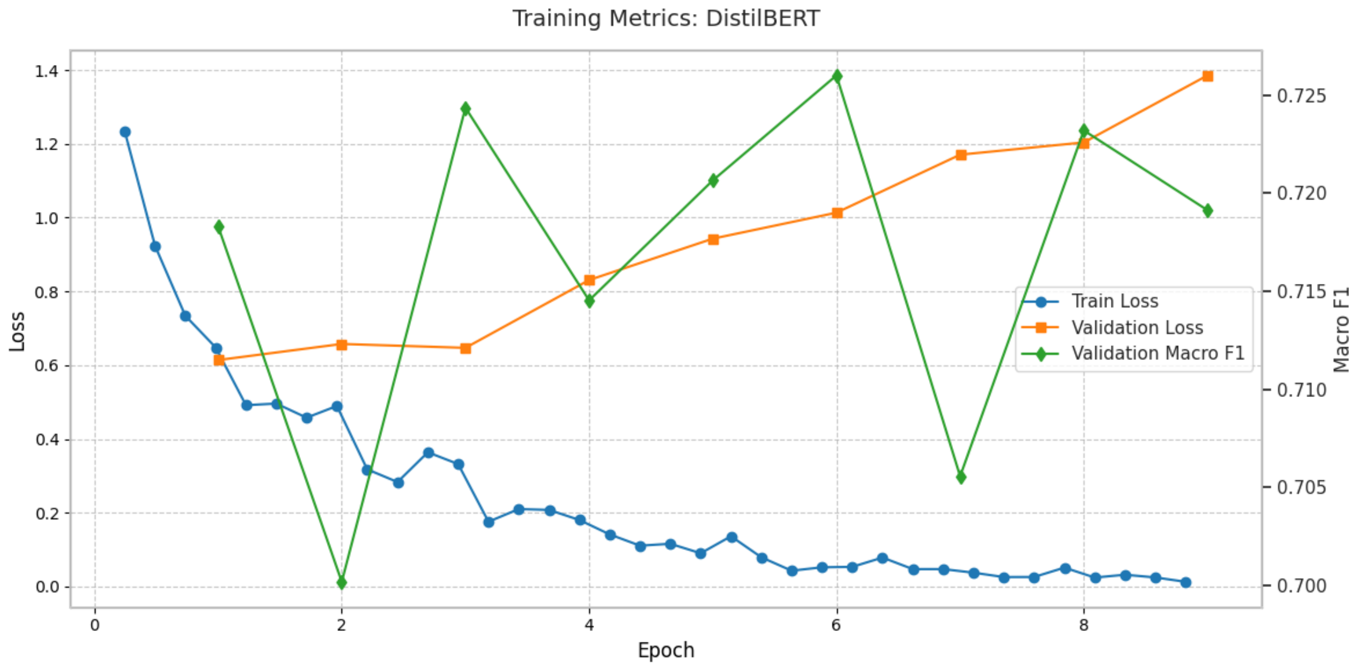


Fig. 8. Loss curve for DistilBERT TweetEval trainin. Loss makes a steady decline toward zero, successfully learning on training. By Epoch 9, it has nearly perfect prediction. However, we can tell our model is massively overfitting because our validation loss consistently rises, and the F1 score shifts quite dramatically, though this may be due to the validation being relatively scarce at less then 400 samples. For future training, this model requires greater regularization with augmentation techniques and more data.

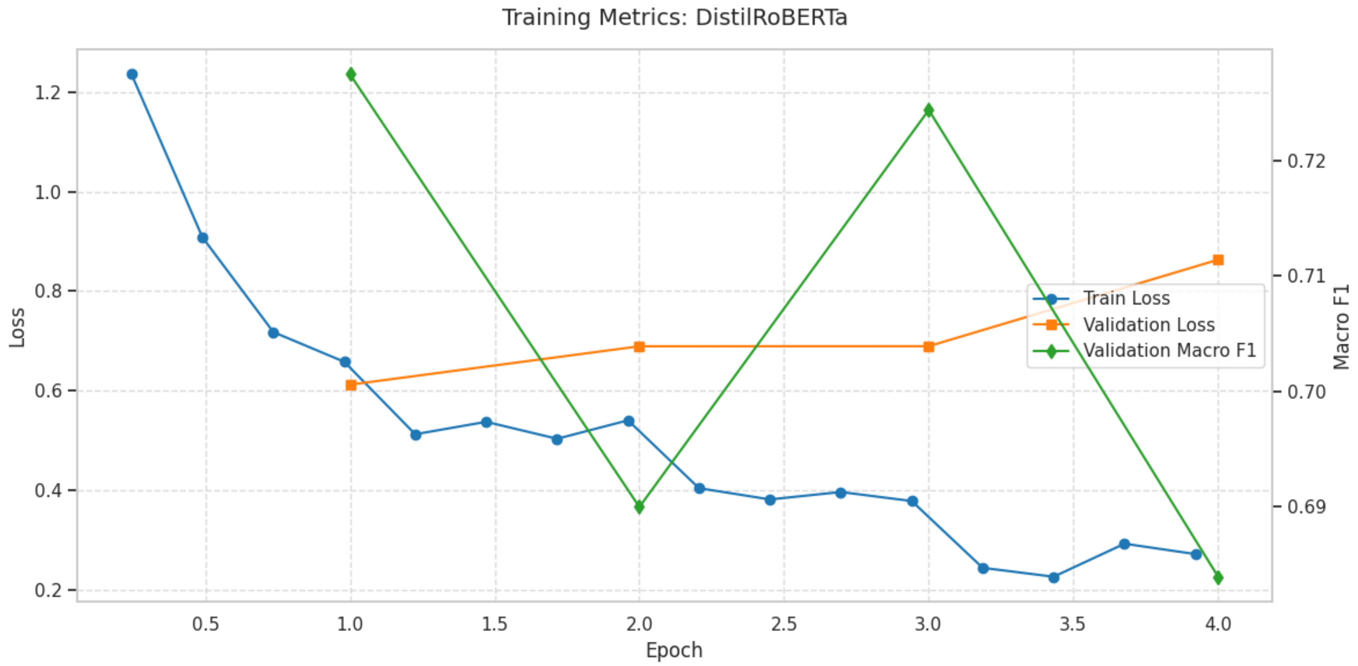


Fig. 9. Loss curve for DistilRoBERTa TweetEval trainin. Similar in form to DistilBERT overfitting, and actually makes an early stop at just 4 epochs, as it reached it's best Macro F1 score at Epoch 1 and yielded worse scores 3 times. This is also indicative that the model needs greater regularization, as it's lacking the substained training that would allow it to learn more about its domain.