Identification of Informative COVID-19 English Tweets

Ding Jiang, Denghan Xiong, Yaxing Zhang
Course Instructor: Nickvash Kani
May 12, 2025

1 Introduction

During the COVID-19 pandemic, many people shared news and updates on social media. Some tweets contained useful information, such as case numbers or travel history, while others did not. This project is based on the WNUT-2020 shared task, which aims to classify tweets as either INFORMATIVE or UNINFORMATIVE. The dataset includes 7,000 training tweets, 1,000 validation tweets, and 2,000 test tweets.

Our goal was to build a simple and accurate model (under 15 million parameters) that can classify tweets correctly. We used a compact transformer model and improved it with data augmentation and explored self-training techniques. While our initial methodology also planned for self-supervised contrastive learning (SimCSE) and reinforcement learning (RL), the experiments detailed in this report primarily focus on the outcomes of data augmentation and pseudo-labeling strategies derived from our practical notebook experiments. These methods were investigated to improve the model's accuracy in identifying informative tweets.

2 Methodology

I. **Base Model Architecture:** We used a small pre-trained transformer model (`ALBERT-base-v2`, ∼11.7M parameters) because it provides strong performance with fewer parameters. We added a classification layer on top to predict whether a tweet is informative.

II. **Data Augmentation & Weighted Loss:** Standard preprocessing was applied. Key techniques included synonym augmentation (WordNet via `nlpaug`), doubling the initial training data (6,936 samples) for our baseline (Approach 1), and a custom `WeightedLossTrainer` (weights [0.7 UNINFORMATIVE, 1.7 INFORMATIVE]) to address class imbalance.

III. **Self-Training with Pseudo-Labeling:** We explored two strategies using the unlabeled dataset:

- **Approach 2 (High-Confidence):** Baseline model predictions on unlabeled data with confidence ≥ 0.7 were used as pseudo-labels for further fine-tuning.
- **Approach 3 (Top 50%):** The top 50% most confident predictions from unlabeled data were augmented and combined with augmented original data for final training.

IV. **Planned (SimCSE/RL):** SimCSE and RL were part of the initial broader scope but are not the focus of the experimental outcomes presented here, which center on the notebook-driven augmentation and pseudo-labeling work.

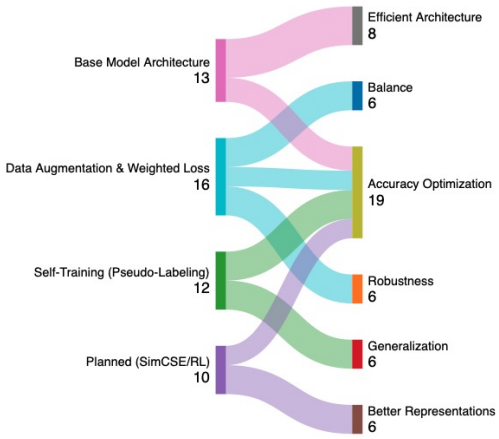

Figure 1: Sankey diagram showing how each method contributes to different aspects of model improvement.

3 Experimental Setup

Experiments utilized HuggingFace Transformers on the NCSA OpenCE 1.5.1 environment (UIUC NCSA GPU), using the WNUT-2020 dataset. Our `ALBERT-base-v2` model has 11,685,122 parameters.

The primary training configuration for **Approach 1 (Baseline Augmented Training)** and the main training phase of **Approach 3** is detailed in Figure 2. This setup featured an AdamW optimizer, a learning rate of 3×10^{-5} (linear decay with 500 warmup steps), 0.1 weight decay, 6 epochs, and a batch size of 32/device. A weighted CrossEntropyLoss (weights [0.7 UNINFORMATIVE, 1.7 INFORMATIVE]) was applied, and the best model was selected based on validation F1-score.

For the fine-tuning stage in **Approach 2 (High-Confidence Pseudo-Labeling)**, after adding pseudo-labels (confidence ≥ 0.7), the learning rate was adjusted to 1×10^{-5} and training was set for 3 epochs. Other settings remained consistent.

III. **Self-Training with Pseudo-Labeling:** We explored two strategies using the unlabeled dataset:

- For **Approach 3 (Top 50% Pseudo-Labels)**, the final training on the extensively combined dataset used the primary settings shown in Figure 2. Training times were ∼7 min (Approach 1) and ∼22 min (Approach 3 final stage after ∼40s pseudo-label generation).

IV. **Planned (SimCSE/RL):** SimCSE and RL were part of the initial broader scope but are not the focus of the experimental outcomes presented here, which center on the notebook-driven augmentation and pseudo-labeling work.

```
training_args = TrainingArguments(
    output_dir='./results',
    per_device_train_batch_size=32,
    per_device_eval_batch_size=128,
    num_train_epochs=6,
    learning_rate=3e-5,
    weight_decay=0.1,
    evaluation_strategy="epoch",
    save_strategy="epoch",
    load_best_model_at_end=True,
    metric_for_best_model="f1",
    save_total_limit=2,
    logging_dir='./logs',
    logging_steps=100,
    warmup_steps=500,
)
```

Figure 2: Core Training Args for Approach 1 & 3.

4 Results

Performance of the three approaches, based on notebook logs and test evaluations:

Approach 1 (Baseline: Augmentation + Weighted Loss) established a strong start. The best validation F1-score during training was 0.8879 (Epoch 4). On the test set, this model achieved **0.8740 accuracy**, with an INFORMATIVE F1-score of **0.8776** (Precision 0.9149, Recall 0.8432).

Approach 2 (High-Confidence Pseudo-Labeling), adding 17 pseudo-labels and fine-tuning, showed a validation F1 of 0.8737 after one fine-tuning epoch. This did not immediately surpass Approach 1's best, likely due to the few pseudo-labels added.

Approach 3 (Top 50% Pseudo-Labels with Full Augmentation) used 5938 pseudo-labels, trained with settings from Figure 2. The peak validation F1 was ∼0.7185 (Epoch 1). On the test set (details in Figure 5), it achieved **0.7325 accuracy**, with an INFORMATIVE F1-score of **0.6282** (Precision 0.9131, Recall 0.4788) — a significant performance drop.

Accuracy on test set: 0.88900
Classification Report:

	precision	recall	f1-score	support
UNINFORMATIVE	0.8690	0.9299	0.8984	1056
INFORMATIVE	0.9149	0.8432	0.8776	944
accuracy			0.8890	2000
macro avg	0.8920	0.8866	0.8880	2000
weighted avg	0.8907	0.8890	0.8886	2000

Figure 3: Test report: Approach 1.

Accuracy on test set: 0.87400
Classification Report:

	precision	recall	f1-score	support
UNINFORMATIVE	0.8622	0.9062	0.8837	1056
INFORMATIVE	0.8888	0.8379	0.8626	944
accuracy			0.8740	2000
macro avg	0.8755	0.8721	0.8731	2000
weighted avg	0.8747	0.8740	0.8737	2000

Figure 4: Test report: Approach 2.

Accuracy on test set: 0.73250
Classification Report:

	precision	recall	f1-score	support
UNINFORMATIVE	0.6731	0.9593	0.7911	1056
INFORMATIVE	0.9131	0.4788	0.6282	944
accuracy			0.7325	2000
macro avg	0.7931	0.7190	0.7097	2000
weighted avg	0.7864	0.7325	0.7142	2000

Figure 5: Test report: Approach 3.

The "nearly 89% accuracy" from earlier drafts aligns with Approach 1. Data augmentation was key. The previously discussed general confidence figures illustrate typical distributions.

5 Analysis and Conclusion

`ALBERT-base-v2` with synonym augmentation and weighted loss (Approach 1) achieved strong performance (0.8740 test accuracy, Fig. 4). Self-training via pseudo-labeling was challenging: high-confidence labels (Approach 2) showed limited immediate impact; top 50% pseudo-labels (Approach 3) significantly degraded performance (Fig. 5), likely due to noise and suboptimal hyperparameters for the complex combined dataset (despite using settings from Fig. 2). Synonym augmentation can alter meaning. Planned SimCSE/RL need focused experimentation. Future work: refine pseudo-labeling (calibration, thresholds, specific hyperparameters, informed by settings like those in Fig.2, advanced augmentation, and integrate other self-supervised methods. Approach 1 was most effective.