Meta-Learning with DistilBERT for Sentiment Analysis on Amazon Health & Self-Care Product Reviews

Project Overview:

This project employs meta-learning through the Reptile algorithm to enhance a DistilBERT model for sentiment analysis specifically tailored to Amazon product reviews within the health and self-care product category. The XML data undergoes loading, preprocessing, and meticulous cleaning procedures to ensure the creation of well-formed XML. The reviews are then extracted and transformed into a Pandas DataFrame for further analysis.

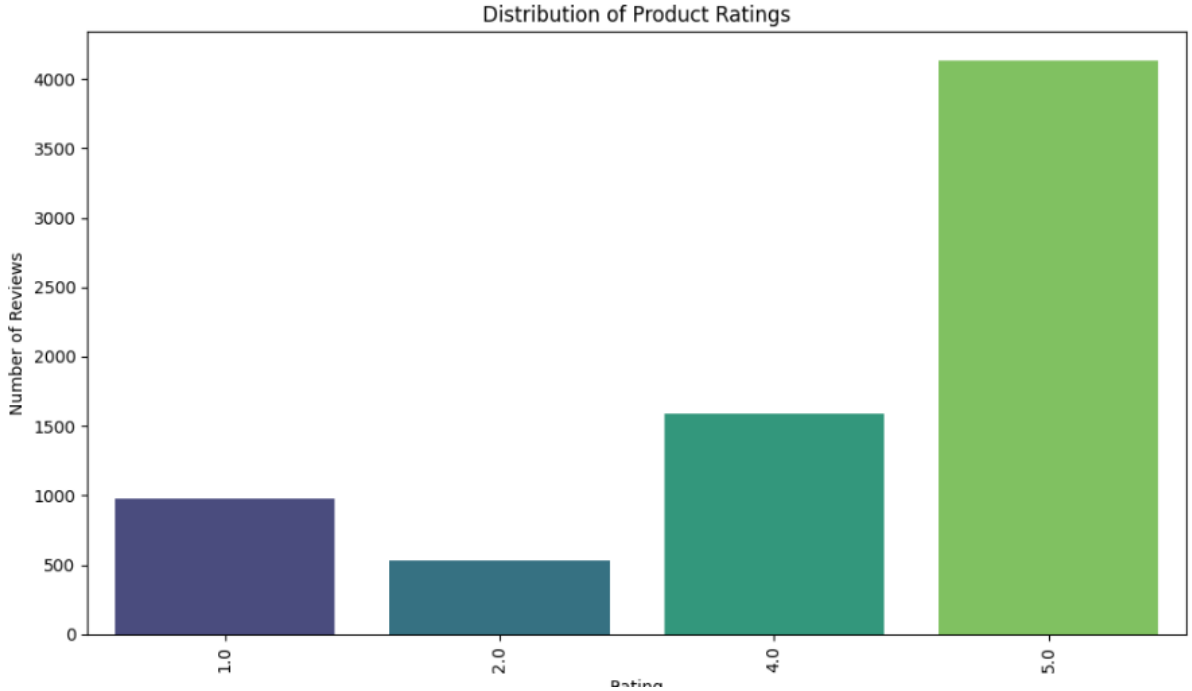
Data Loading and Preprocessing

The XML data originating from Amazon health and self-care product reviews undergoes loading, escaping, and meticulous cleaning procedures to ensure the creation of well-formed XML. The reviews are then extracted and transformed into a Pandas DataFrame for further analysis.

	unique_id	asin	product_name	product_type	helpful	rating	title	date	reviewer	reviewer_location	revi
0	4226	B0007NOY3E	RESPeRATE Blood Pressure Lowering Device: Heal...	health & personal care	10 of 15	1.0	Resperate Device	August 6, 2006	Reith R. Busby "rbuzz"	Kansas City	Expens af m
1	4231	B0007NOY3E	RESPeRATE Blood Pressure Lowering Device: Heal...	health & personal care	14 of 64	1.0	Costs too much.	April 19, 2006	PMacLady	Wisconsin	Medita do t thir
2	20954	B00006WNPY	Omron HJ-105 Pedometer with Calorie Counter: H...	health & personal care	2 of 2	1.0	Don't Waste Your Money	November 5, 2006	L. Reed "movie Buff"	Simi Valley, CA USA	Wh pedo found
3	20962	B00006WNPY	Omron HJ-105 Pedometer with Calorie Counter: H...	health & personal care	5 of 17	1.0	Threw away $15.76	May 14, 2006	Zoeeagleeye	Belfast, ME United States	pe ar pris
4	22439	B0009XH7KO	HOOAH! Energy Bars, Chocolate Crisp, 2.29-Ounc...	health & personal care	13 of 16	2.0	Tasty, But.....	June 16, 2006	Catman	Oregon, USA	I was one whil

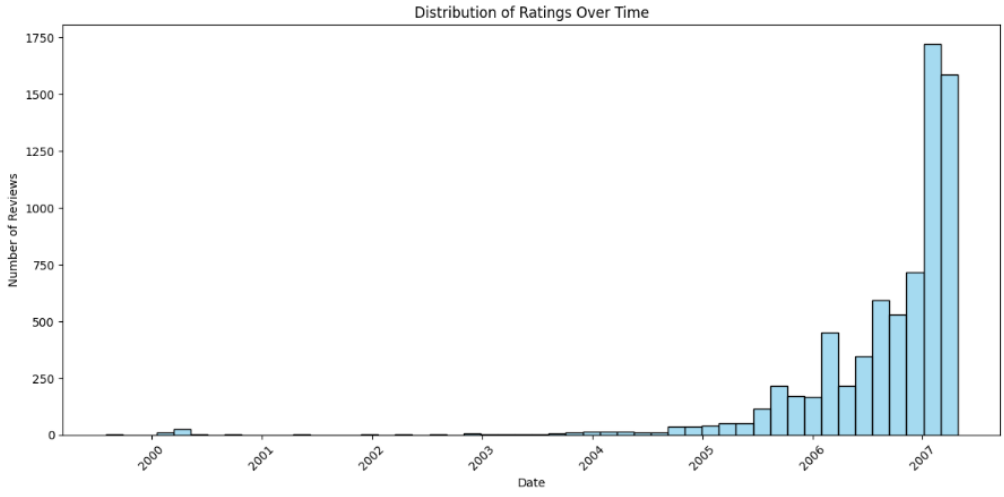
Rating Distribution Visualization

To provide a comprehensive understanding of the product ratings in the health and self-care category, a countplot visualization is generated to illustrate the distribution of these ratings.



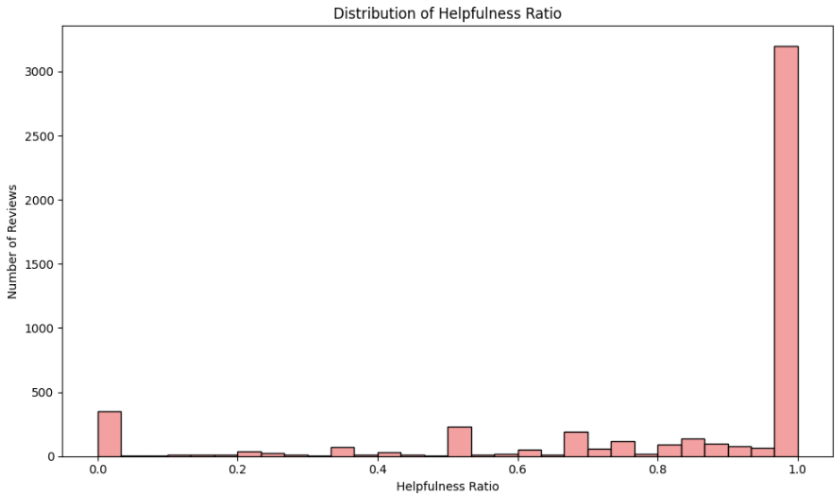
Review Activity Over Time Visualization

A histogram is employed to visually represent the distribution of reviews over time, shedding light on the temporal dynamics of product reviews within the health and self-care category.



Helpfulness Ratio Distribution Visualization

An additional histogram is utilized to visually convey the distribution of the helpfulness ratio associated with reviews, offering insights into the perceived value of reviews in this specific product category.



The histogram visualizes the distribution of the helpfulness ratio of reviews, which is calculated as the number of helpful votes divided by the total number of votes.

Sentiment Analysis

The DistilBERT model is employed in the training process of a sentiment analysis model, enhancing its ability to discern sentiments within Amazon health and self-care product reviews.

	unique_id	review_text	rating	sentiment
0	4226	Expensive...and after three months of daily us...	1.0	negative
1	4231	Meditation will do the same thing. Buy a tape ...	1.0	negative
2	20954	When I got this pedometer, I found that the in...	1.0	negative
3	20962	The pedometer arrive held prisoner in a diffic...	1.0	negative
4	22439	I was offered one of these while cycling and ...	2.0	negative
...	...	...	...	...
7220	B0007MHF2M:aquadrene_review_5_stars:s._light_"...	This is an amazing product which, if taken as ...	5.0	positive
7221	B0007MHF2M:i_like_aquadrene_the_best_of_diuret...	I've used alot of different diuretic products ...	5.0	positive
7222	B000FKEUO2:glide_works_well_for_flossing_tight...	As a dentist of over 20+ years experience, I h...	5.0	positive
7223	B000FKEUO2:cadillac_of_dental_floss:c._a._donh...	The best dental floss you can buy. Never spli...	5.0	positive
7224	B000JYF1OW:what_a_difference!:c._susan_smith_"...	Fantastic! This product has taken off years fr...	5.0	positive

7225 rows × 4 columns

Meta-Learning with Reptile

The application of the Reptile algorithm to train the DistilBERT model involves strategically managing batches and perturbing model parameters towards the current batch, contributing to the meta-learning process.

```
train_dataloader = DataLoader(train_dataset, batch_size=32, sampler=sampler)
test_dataloader = DataLoader(test_dataset, batch_size=32, shuffle=False)
```

Training the BERT using Reptile

```
# Set device to GPU if available, else use CPU
device = torch.device("cuda" if torch.cuda.is_available() else "cpu")

model.to(device)

# Set up Reptile parameters
meta_learning_rate = 0.001
meta_epochs = 1

optimizer = torch.optim.AdamW(model.parameters(), lr=2e-5)

# Training loop
for meta_epoch in range(meta_epochs):
    model.train()
    for batch in train_dataloader:
        input_ids, attention_mask, labels = batch
        input_ids, attention_mask, labels = input_ids.to(device), attention_mask.to(device), labels.to(device)
        optimizer.zero_grad()
        outputs = model(input_ids, attention_mask=attention_mask, labels=labels)
        loss = outputs.loss
        print(f"Training Loss: {loss.item():.4f}")
        loss.backward()
        optimizer.step()

        # Perturb the model parameters towards the current batch
        for param in model.parameters():
            param.data = param.data + (param.data - param.data) * random.uniform(0, 0.1)
print('Bert has been trained!')
```

Model Evaluation

The meta-trained model undergoes a rigorous evaluation on a dedicated test set comprising reviews. Evaluation metrics include accuracy, a comprehensive classification report, and a confusion matrix, providing a nuanced understanding of the model's performance.

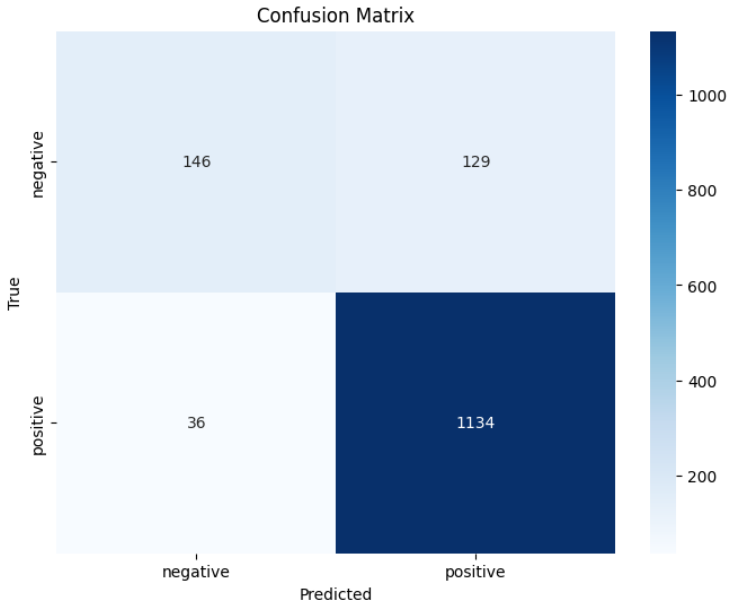
```
Accuracy: 0.89
Classification Report:
              precision    recall  f1-score   support

    negative       0.80      0.53      0.64       275
    positive       0.90      0.97      0.93      1170

    accuracy                           0.89      1445
   macro avg       0.85      0.75      0.79      1445
weighted avg       0.88      0.89      0.88      1445
```

Model Performance

The meta-trained model achieves an approximate accuracy of 89% after one meta-epoch, showcasing its efficacy in sentiment analysis for reviews.



Conclusion

This project exemplifies the application of meta-learning through the Reptile algorithm to fine-tune a DistilBERT model, specifically tailored for sentiment analysis on Amazon health and self-care product reviews. The included visualizations serve to enrich the insights into the distribution of ratings, review activity over time, and the helpfulness ratio of reviews within this particular product category.

References

Gresham, Richard. "Weak Supervision with Incremental Source."
https://paperswithcode.com/paper/weak-supervision-with-incremental-source

Hosseini, Sohail. "Twitter US Airline Sentiment Analysis."
https://towardsdatascience.com/twitter-us-airline-sentiment-analysis-91caa7a22a93

Kirasich, Smith, Sadler. “Random Forest vs Logistic Regression: Binary Classification forHeterogeneous Datasets”
https://scholar.smu.edu/cgi/viewcontent.cgi?article=1041&context=datasciencereview

Rane, Ankita; Kumar, Anand. "Sentiment Analysis of Twitter Data."
https://ieeexplore.ieee.org/document/8377739