Project Milestone 1

Marawan Omran
Khaled Hamza

Problem Statement

Proposal Summary

- Our proposal aims to develop an efficient pipeline for multi-class text classification
- Focus on enhancing performance for the Multi-Genre Natural Language Inference (MNLI) task.
- To achieve this, we will leverage smaller models and explore various fine-tuning and distillation techniques to optimize model efficiency and accuracy.

Problem statement

- The MNLI task involves classifying text pairs into categories like entailment, contradiction, and neutral, which is crucial for natural language understanding applications.
- However, achieving high performance on this task can be challenging due to limited computational resources and training data.

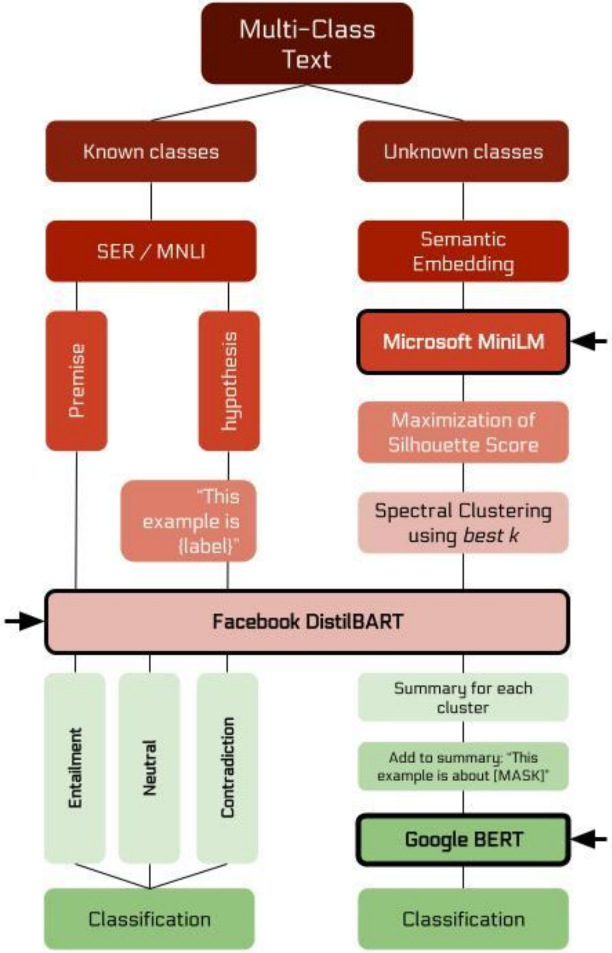
Proposed Solution

Proposed solution

- Advancing the performance of the DistilBart model through additional fine-tuning on the SNLI and ANLI datasets to enhance its classification accuracy.
- Investigate a range of fine-tuning strategies and hyperparameter configurations to optimize the model's performance while minimizing computational overhead.
- Explore distillation techniques to improve the efficiency of the model without compromising its accuracy.

Proposed solution

- Conduct experiments with other distilled models, including DistilBERT and DistilRoBERTa, to compare their performance and determine the most effective model for the MNLI task.
- Identify the optimal configuration that achieves high classification accuracy while remaining computationally efficient.


Multi-Class Text
Known classes
Unknown classes
SER / MNLI
Semantic Embedding
Premise
hypothesis
Microsoft MiniLM
Maximization of Silhouette Score
"This example is {label}"
Spectral Clustering using *best k*
Facebook DistilBART
Summary for each cluster
Entailment
Neutral
Contradiction
Add to summary: "This example is about [MASK]"
Google BERT
Classification
Classification

Related Work

Microsoft MiniLM

Paper: **MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers**

Huggingface: microsoft/MiniLM-L12-H384-uncased

Performance:

Model	#Param	SQuAD 2.0	MNLI-m	SST-2	QNLI	CoLA	RTE	MRPC	QQP
BERT-Base	109M	76.8	84.5	93.2	91.7	58.9	68.6	87.3	91.3
MiniLM-L12xH384	33M	81.7	85.7	93.0	91.5	58.5	73.3	89.5	91.3

Speed: 7500 sentences/second

Facebook DistilBART

Paper: **BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension**

Hugginface: valhalla/distilbart-mnli-12-1

Performance:

	matched acc	mismatched acc
bart-large-mnli (baseline, 12-12)	89.9	90.01
distilbart-mnli-12-1	87.08	87.5

Google BERT

Paper: **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**

Huggingface: google-bert/bert-large-uncased

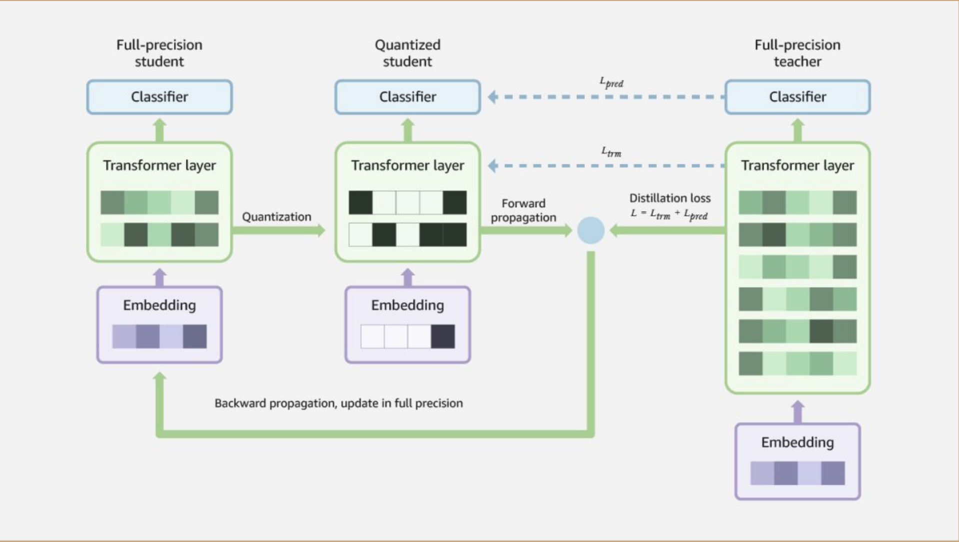
Performance:

SQUAD 1.1 F1/EM	Multi NLI Accuracy
91.0/84.3	86.05

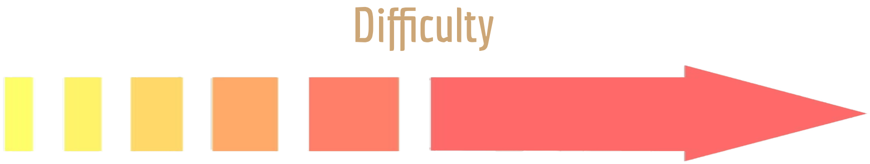
DistilBART-12-1

- Pretrained on Multi-NLI
- AdamW
- Batch size: 16
- Dropout: 0.3
- Learning rate: 2e-5
- 87.5% accuracy on Multi-NLI Validation_mismatched

Gaikwad, Mayur & Ahirrao, Swati & Kotecha, Ketan & Abraham, Ajith. (2022). Multi-Ideology Multi-Class Extremism Classification Using Deep Learning Techniques. IEEE Access. PP. 1-1. 10.1109/ACCESS.2022.3205744.



Datasets



SNLI

Hugginface: snli

Description: The SNLI corpus (version 1.0) is a collection of 570k human-written English sentence pairs manually labeled for balanced classification with the labels entailment, contradiction, and neutral, supporting the task of natural language inference (NLI), also known as recognizing textual entailment (RTE).

MNLI

Hugginface: nyu-mll/multi_nli

Description: The Multi-Genre Natural Language Inference (MultiNLI) corpus is a crowd-sourced collection of 433k sentence pairs annotated with textual entailment information. The corpus is modeled on the SNLI corpus, but differs in that covers a range of genres of spoken and written text, and supports a distinctive cross-genre generalization evaluation. The corpus served as the basis for the shared task of the RepEval 2017 Workshop at EMNLP in Copenhagen.

ANLI

Hugginface: facebook/anli

Description: The Adversarial Natural Language Inference (ANLI) is a new large-scale NLI benchmark dataset, The dataset is collected via an iterative, adversarial human-and-model-in-the-loop procedure. ANLI is much more difficult than its predecessors including SNLI and MNLI. It contains three rounds. Each round has train/dev/test splits.

Work Done So Far

Fine Tuning

- First 4000 samples in MNLI
- 3 epochs
- 8 batches
- Test on 400 samples (Validation_mismatched partition)
- Learning rate: 2e-5
- AdamW: (beta=0.9, epsilon=1e-8)

Accuracies of models before and after fine-tuning on MNLI

Model	Accuracy Before	Accuracy After
BART	-	0.37
DistilBART	0.3	0.79
DistilBERT	0.54	0.74
DistilRoBERTa	0.44	0.72

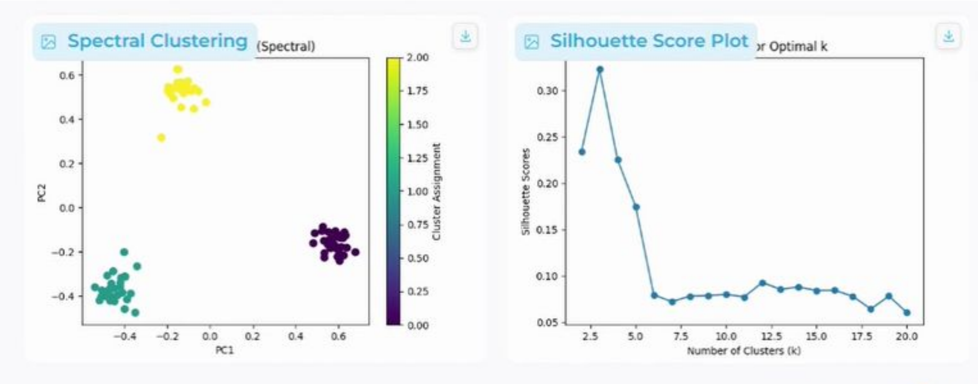
Fine Tuning on DistilBART

- 2000 samples from SNLI
- 3 epochs
- 8 batches
- Test on 200 samples (test partition)
- Learning rate: 2e-5
- Adam: (beta=0.9, epsilon=1e-8)

Accuracy before Fine-tuning: 22.4%

Accuracy after Fine-tuning: 95.3%

Unknown Classes Pipeline Demonstration: Gradio UI



Cluster Summaries

The new Dodge Ram is not just a truck; it's a statement. The Dodge Ram delivers on both style and substance, a testament to its commitment to excellence. With its towing capacity and durability, the new DodgeRam is your reliable partner for any task. Experience the roar of power under the hood and the luxury within with the Dodge Ram's latest model.

The Ronaldo-Messi rivalry is the stuff of legends, a tale of two football giants who continue to redefine greatness. Every match between Ronaldo and Messi feels like a clash of titans, a spectacle that fans around the world eagerly await. The Ronaldo- Messi debate is not just about goals and assists; it's about witnessing two footballing maestros creating history.

Love is not just a feeling; it's a lifeline that connects souls. When love unravels, it can leave behind a trail of shattered dreams. Love is a journey, and each step is an exploration of the heart. The love that once made you whole can become the cause of your brokenness.

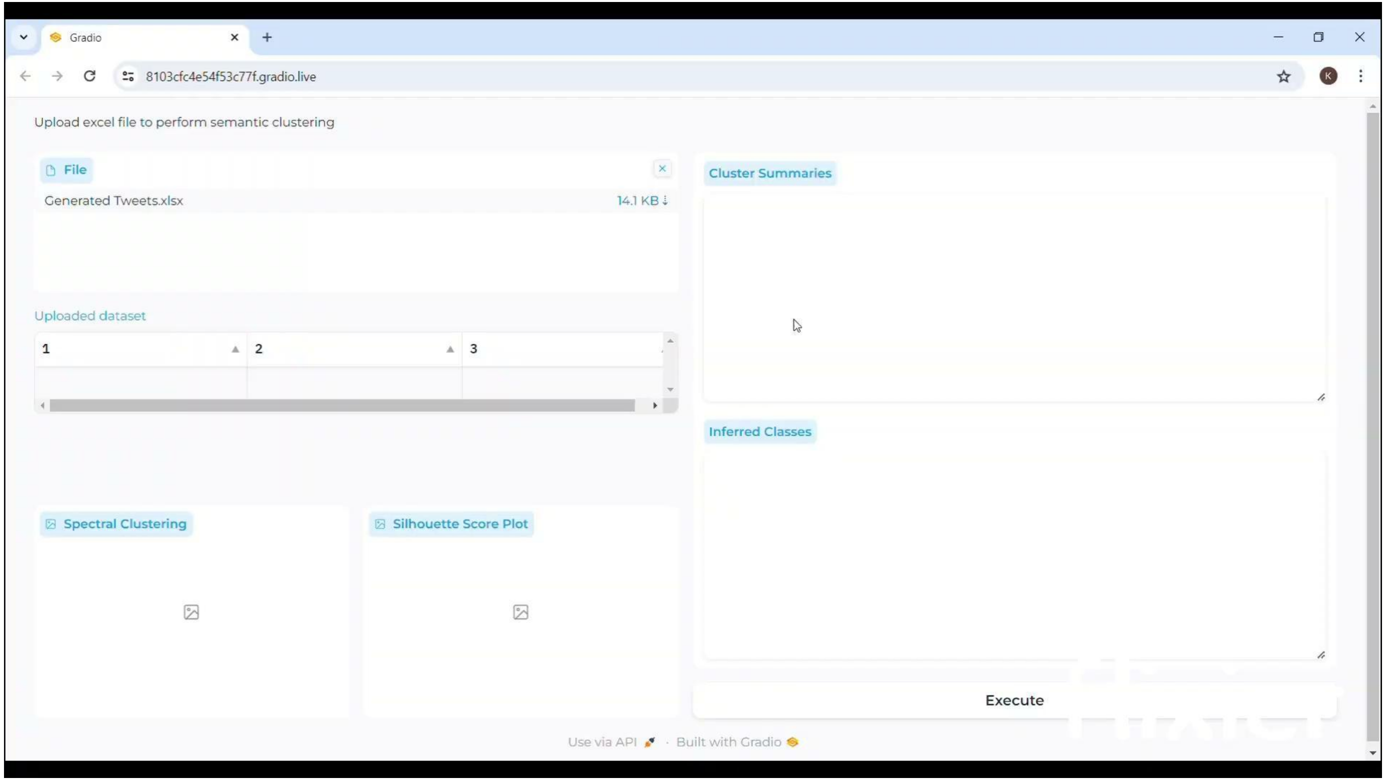
Inferred Classes

driving it business you money

history football america politics us

love loss life forgiveness relationships

Demo: UI


Gradio
8103cfc4e54f53c77f.gradio.live
Upload excel file to perform semantic clustering
File
Generated Tweets.xlsx
14.1 KB
Uploaded dataset
1
2
3
Cluster Summaries
Inferred Classes
Spectral Clustering
Silhouette Score Plot
Execute
Use via API
Built with Gradio

Next Steps:

1. Further Fine tune DistilBART on datasets with gradual difficulty (SNLI→MNLI→ANLI)
    a. Change Hyperparameters in fine tuning
2. Try Knowledge Distillation for DistilBART to make it lighter and faster
3. Try different models for Mask-filling task and Fine-tune those models for this specific task→"This example is [MASK]"
4. Distillate Mask-filler models for faster results

Contributions

Khaled: Fine Tuning BART, Pipeline, UI

Marawan: Fine Tuning DistilRoBERTa, DistilBERT, DistilBART