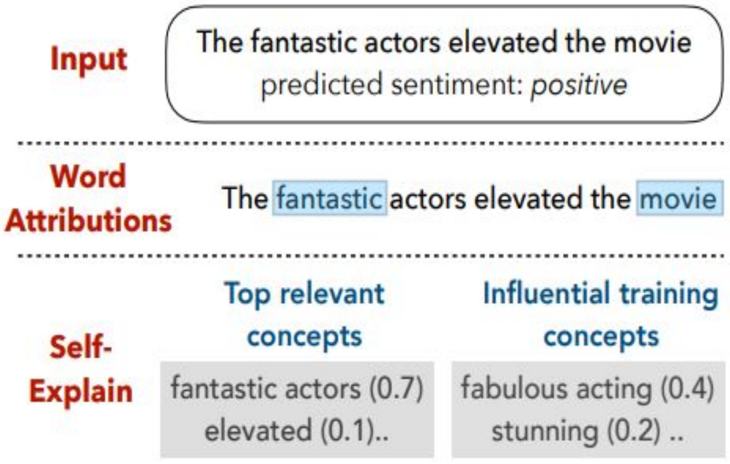
CS 772 – FINAL PROJECT EVALUATION

Sneha Oram 23M2159
A Snegha 23M2160
K Hemanth 23M2164
A Shruthi 23M1074

05 May 2024

Problem Statement

Problem statement: Analysing the model's prediction in different setting using phrase level concepts with Local and Global interpretable layer.



Source: Rajagopal et al. 2021

Input: One sentence (text).

Output:

1. Classification Result
2. Top relevant phrases from input sample (Local).
3. Influential phrases from the training data for a given input sample (Global).

Motivation

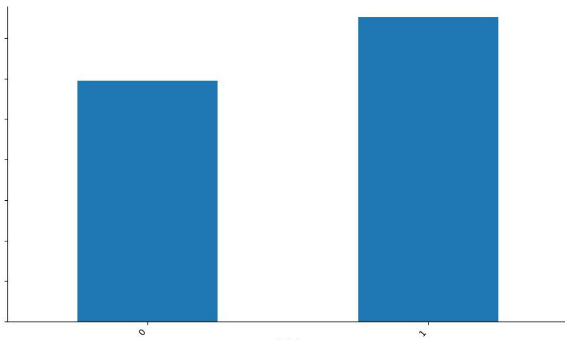
- Prior work in interpretability for neural text classification:
    - **Post-hoc explanation methods :** Explain predictions for previously trained models.
    - **Inherently interpretable models :** Built-in and optimized jointly with the end task.
- Interpret model decisions locally as a function of relevance of **features (words)** in input samples **lacks reliability and faithfulness**.
- Explaining the role of higher-level compositional concepts like **phrasal structures** remains an open challenge.

Literature Survey

- Dheeraj Rajagopal, Vidhisha Balachandran, Eduard Hovy, Yulia Tsvetkov. 2021. SELFEXPLAIN: A Self-Explaining Architecture for Neural Text Classifiers . EMNLP.
- Sofia Serrano and Noah A. Smith. 2019. Is Attention Interpretable?. ACL
- Rishabh Joshi, Vidhisha Balachandran, Emily Saldanha, Maria Glenski, Svitlana Volkova, and Yulia Tsvetkov. 2023. Unsupervised Keyphrase Extraction via Interpretable Neural Networks. ACL.
- Orevaoghene Ahia, Hila Gonen, Vidisha Balachandran, Yulia Tsevetkov, Noah A. Smith. (2023) have worked with Lexical based interpretability in offensive texts. LEXPLAIN: Improving Model Explanations via Lexicon Supervision. ACL.
- Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2963–2977, Florence, Italy. Association for Computational Linguistics.

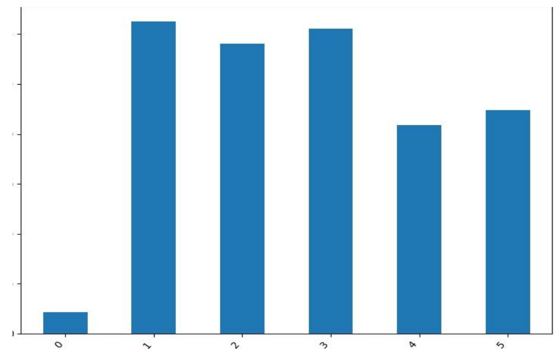
Data Handling (1/2)

SST-2



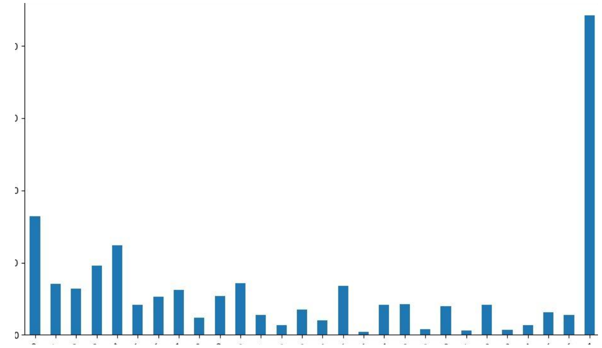
Number of classes: 2
Domain: Sentiment Analysis
https://huggingface.co/datasets/stanfordnlp/sst2

TREC



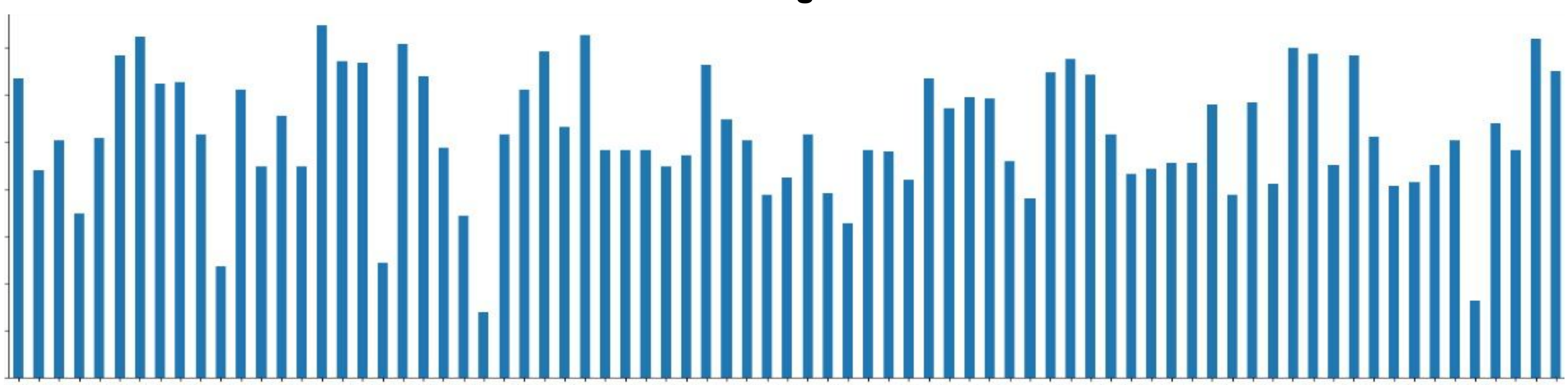
Number of classes: 6
Domain: Questions Classification
https://huggingface.co/datasets/stanfordnlp/sst2

GoEmotion



Number of classes: 28
Domain: Emotion Analysis
https://huggingface.co/datasets/stanfordnlp/sst2

Banking77



Number of classes: 77
Domain: Banking/Finance
https://huggingface.co/datasets/PolyAI/banking77/commit/ea95213633609499a1c555a0df8a3f90874985a7

Data Handling (2/2)

- Data preprocessing is done to remove special characters.
- *GoEmotion* is an imbalanced dataset. Data is skewed towards Class 27 (Neutral).
- In *TREC* database, Class 0 (Abbreviations) has very low representations compared to other classes.

Dataset	Train	Dev	Test
SST-2	67350	872	1822
GoEmotion	13519	4225	3379
Banking77	8002	3080	2000
TREC	4360	500	1090

Data Split

Mathematical modelling of the problem

Interpretability in classifications tasks with two layers -

- Local Interpretability Layer:

$$t_j = g(\mathbf{u}_j) - g(\mathbf{u}_{\mathbb{S}})$$
$$s_j = \mathtt{softmax}(\mathbf{W}_v \times t_j + \mathbf{b}_v)$$

Based on relevance score

- Global Interpretability Layer:

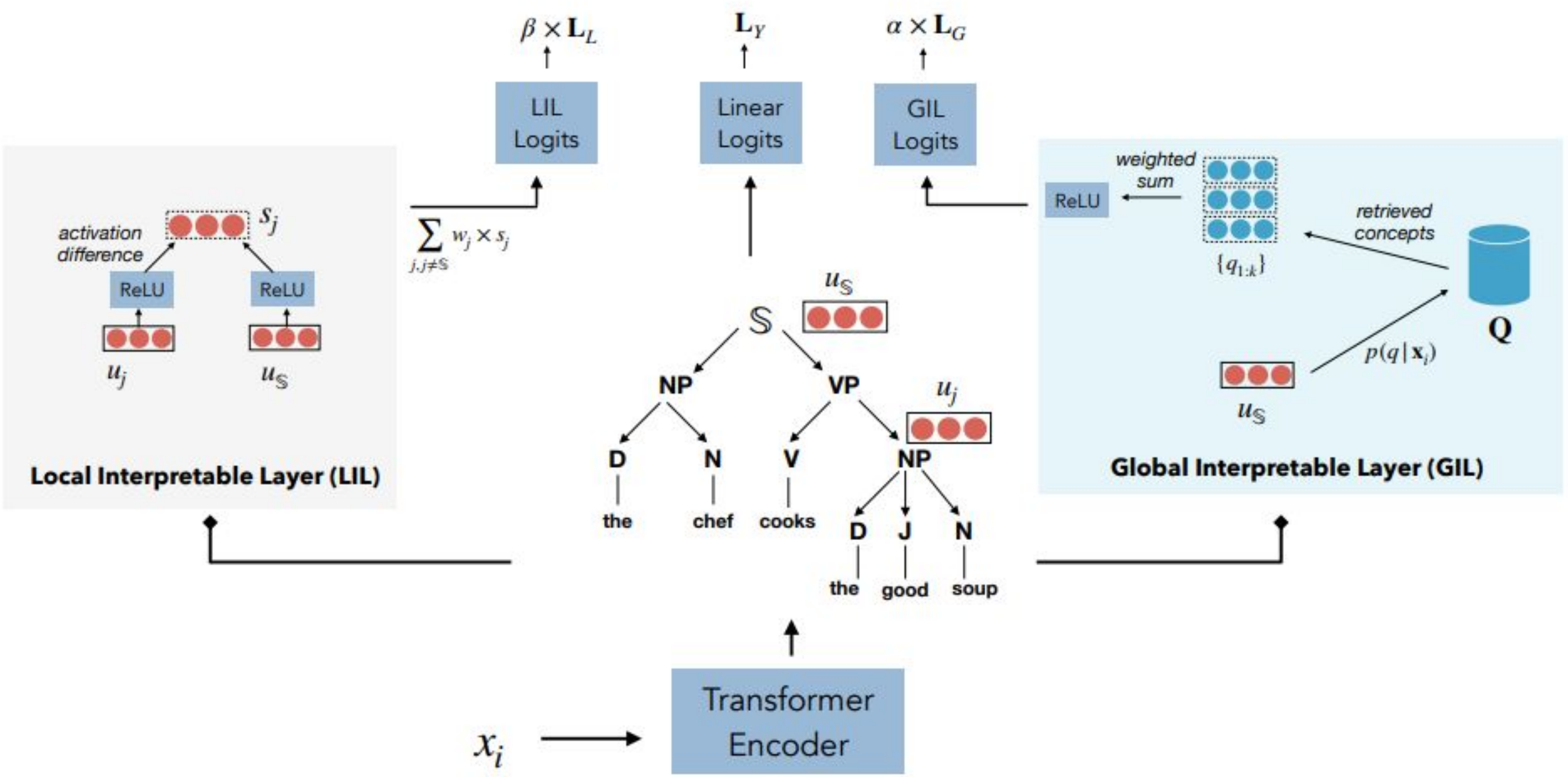
$$q_k = \frac{\sum_{w \in q_k} e(w)}{len(q_k)} \in \mathbb{R}^D$$

$$d(\mathbf{x}, Q) = \frac{\mathbf{x} \cdot q}{||\mathbf{x}|| ||q||} \quad \forall q \in Q$$

Based on Maximum Inner Product Search (MIPS)

Rajagopal et al. 2021

Architecture



Overall architecture of the proposed model

Rajagopal et al. 2021

Experimental details

- There are four folds of analysis:

1	**Dataset**	SST-2, GoEmotion (27) , Banking77 (77) , TREC (6)
2	**Models**	Bert, XLNet, Roberta, XLM-R
3	**Setting**	Full fine tuning, LoRa fine tuning, Quantized version
4	**Model Size**	Base, Medium, Large

- In the results, the LIL and GIL phrases in the four folds of analysis are evaluated.

Questions to be analysed

1. How different dataset type change interpretability?
2. How different model interpretability various for a dataset?
3. How different training setting affect the interpretability ?
4. How model size affect interpretability?
5. Does SELFEXPLAIN's explanation help predict model behavior (Sufficiency)?
6. Are LIL layer concepts relevant?
7. How are LIL and GIL layers agreeing?

Results

1: Datasets

Dataset	(RoBERTa) Accuracy
SST-2	93.07%
GoEmotion	40%
Banking77	92.55%
TREC	97.18%

2: Models

Model	(TREC) Accuracy
RoBERTa	97.18%
XLM-R	97.5%
XLNet	96.88%
BERT	97.26%

3: Settings

Setting	(RoBERTa/SST-2) Accuracy
Full fine-tuning	93.07%
LoRa fine-tuning	93.41%
Quantized	90.3%

4: Model Size

Model Size	(SST-2) Accuracy
BERT-Large	90.01%
BERT-Medium	84.1%
BERT-Base	50.23%

Top-k (phrases)	(XLMR-TREC) Accuracy
k=2	97.5%
k=5	97.5%
k=10	97.29%

(LIL phrases)	(RoBERTa) Accuracy
TREC	88.3%

Analysis (1/4)

- In most of the cases LIL gave a good interpretability. If LIL have vague or no results, GIL is helping the model understand the sentiment of the model.
- Model’s incorrect prediction are mainly because of attention in non-important phrases, which is proved by looking at LIL interpretability.
- In case of TREC-6, though importance need to be given to question word, model is attending other phrases. But GIL is interpreting relevant. - Because the question word (such as "who," "what," "where," etc.) provides a general indication of the type of information being sought, but it's often the surrounding words and context that convey the specific intent or category of the question.
- TREC and Banking77 interpretation is better than SST-2 and GoEmotion - Because later datasets are more challenging, class imbalanced and poses subtle information about a class.

Analysis (2/4)

- Bert, XLM R, Roberta have similar LIL and GIL interpreted phrases.
- Because Roberta is optimized version of Bert and XLMR is extended multilingual version of Roberta.
- XLNet have good GIL interpreted phrases compared to other models

Analysis (3/4)

Observation:

- LIL and GIL interpreted phrases are similar to each other. So having different training setting have very less effect on interpreted phrase and models' accuracy.

Example:

Sent: it 's a charming and often affecting journey. - 1 1

Fine-tuned - LIL:'often affecting'

GIL: ['show-stoppingly', 'worldly-wise and very funny script']

Lora Fine-tuned- LIL: 'often affecting'

GIL : ['worldly-wise and very funny script', 'astoundingly']

Quantized version-LIL: 'often affecting', 'often'

GIL: ['worldly-wise and very funny script', 'astoundingly']

Analysis (4/4)

- Surprisingly both Bert-medium and Bert-large have same interpretation with little accuracy difference.
- As the model size increase, interpreted phrases quality is good.
- Also increase in model size increase interpreted phrase quality till one size and then it increase accuracy further.
- Bert-Base have poor LIL and GIL interpreted phrases, which resulted in poor accuracy.
- So bigger the model is better the interpreted phrases and improved accuracy.

Case Study (1/2)

Correct examples:

Sent: The mesmerizing performances of the leads keep the film grounded and keep the audience riveted . - 1 1

LIL: ('keep the film grounded'), ('of the leads'), ('the leads'), ('rive ted'), ('the film')

GIL: ['show-stoppingly', 'four star performance']

Incorrect examples:

Sent: You won't like roger , but you will quickly recognize him . - 0 1

LIL: [('recognize him'), ('ro ger'), ('will quickly recognize him')]

GIL: ['engrossing story', 'show-stoppingly']

Case Studies (2/2)

Sent : The film suffers from a lack of humor (something needed to balance out the violence) - 0 0

Bert-large : LIL: [('a lack of humor'), ('of humor'), ('balance out the violence'), ('the violence'), ('a lack')]

GIL: ['failing to compensate for the paper-thin characterizations and facile situations', ""the script 's bad ideas and awkwardness""]

Bert-medium : LIL: [('a lack of humor'), ('of humor'), ('balance out the violence'), ('the violence'), ('a lack')]

GIL: ['failing to compensate for the paper-thin characterizations and facile situations', ""the script 's bad ideas and awkwardness""]

Bert-base : LIL: [('the film'), ('a lack'), ('of humor'),('a lack of humor')]

GIL: ['lend some dignity to a dumb story', 'saw how bad this movie was', ""that's far too tragic to merit such superficial treatment"", 'in world cinema', 'sit through ,']

Bert-Large correct and Bert-medium wrong

Sent: The heavy-handed film is almost laughable as a consequence .

Bert-medium: 1 0 :LIL: [('a consequence'), ('as a consequence'), ('almost laugh ##able')]

GIL: ['it rises in its courageousness , and comedic employment', ', it rises in its courageousness , and comedic employment']

Bert-large: 1 1 :LIL: [('as a consequence'), ('a consequence'), ('almost laugh ##able')]

GIL: ['is a truly , truly bad movie', 'is a truly , truly bad movie']

What is your baseline?

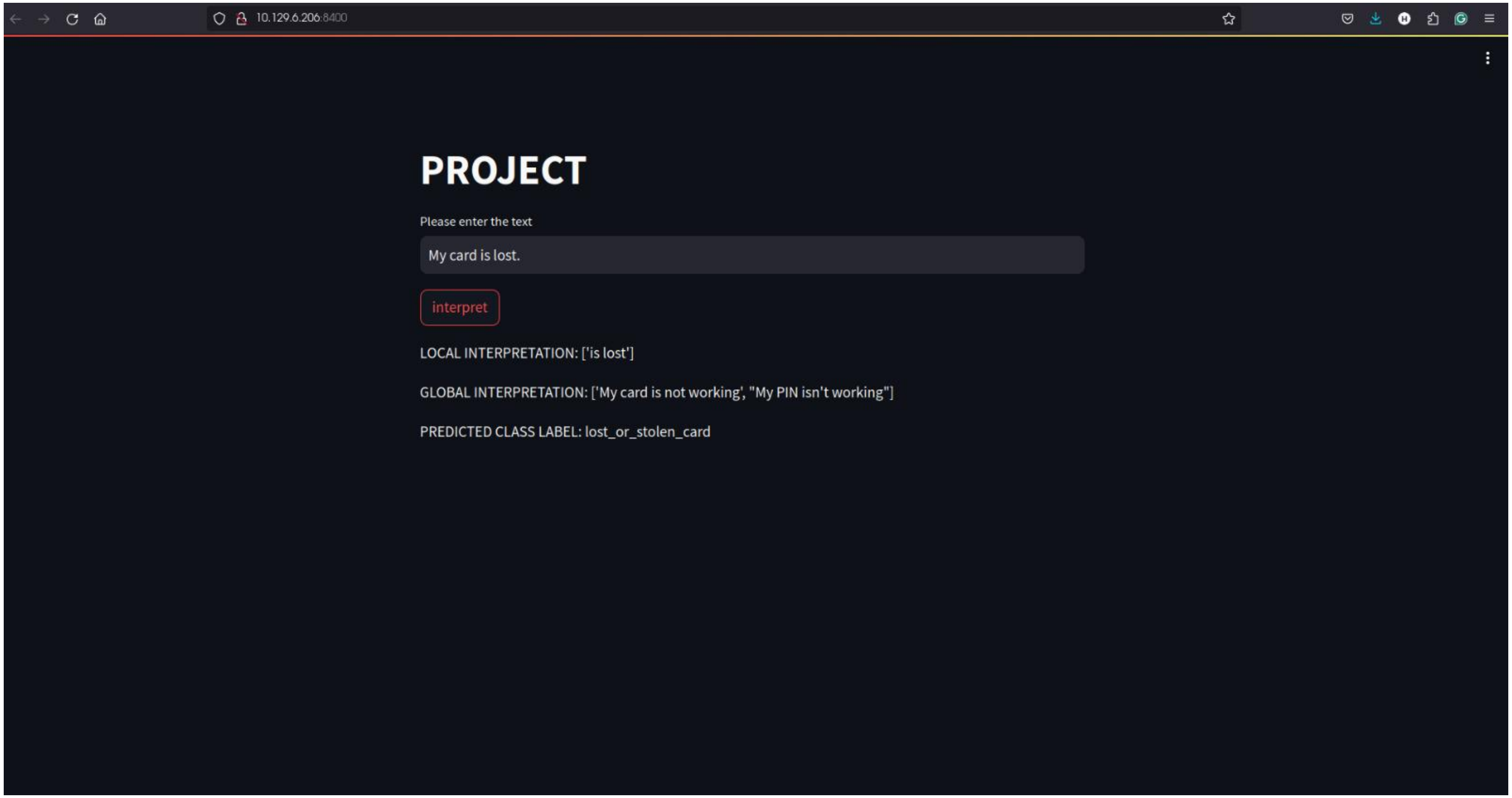
- Almost **all the standard encoders** are explored in the Self-Explain framework.
- The work is partly implementing domain adaptation, that involved **tinkering in the code-base** during the implementation.

Dataset/Model	Accuracy
SST-2/RoBERTa (w/o SELFEXPLAIN)	92.55%
SST-2/RoBERTa (SELFEXPLAIN)	93%

- The accuracy with and without Self-Explain is slightly different.
- The Self-Explain framework is leveraging LIL and GIL layers for the classification task.

Demo

http://10.129.6.206:8400/



Learnings

- On increasing the model size (base to large), the model's understanding ability and interpretability is improving.
- The settings in training (LoRa fine-tuning and Quantized fine-tuning) are slightly affecting the model accuracy.
- Even though the datasets are picked from different domains, the interpretability is not much affected.

BONUS

- To work with datasets with different domain, changes are done in the code-base.
- The performance and interpretability are evaluated with different fine-tuning settings.
- The accuracy with and without Self-Explain is slightly different.
- The work is partly implementing domain adaptation, so the code base is tinkered during the implementation.