Irony Detection

Pierre EPRON, Maxime RENARD, Shu ZHANG

Reminder

- Goal: to detect irony
- Using a perspective oriented dataset [1]
- Using Sequence classification [2][3] and Causal language modeling [4] models

Preprocessing

Processing steps that have been implemented:

- ability to filter dataset by the number of people that have annotated the Post/Reply pair
- handle equalities between annotation by **removing them** or choosing at random
- remove **newline**, user mentions and website links
- generate 5-fold cross-validation train, validation and test splits sets

Next steps:

- reduce the dataset to improve quality of the observations used?
- Use multiple labels depending on agreement? (ironic, probably ironic, probably not ironic, ironic)
- use another dataset as control

CLS Config

- 3 experiments:
  - roberta-irony-zs: zeroshot evaluation of roberta-irony.
  - roberta-irony-ft: roberta-irony finetuned on our data.
  - roberta-base-ft: roberta-base finetuned on our data.
- Training config:
  - epochs: 10
  - batch size: 16
  - optimizer: adamw (from torch)
  - learning rate: 6e-5
  - loss: Binary Cross-Entropy

CLS Results

	False			True			MCC
	P	R	F1	P	R	F1	
roberta-irony-zs	0.7920	0.8176	0.8046	0.3344	0.2992	0.3156	0.1215
roberta-irony-ft	0.8998	0.9591	0.9276	0.7842	0.6341	0.6874	0.6333
roberta-base-ft	0.7744	0.9724	0.8608	0.0863	0.0682	0.0762	0.0443

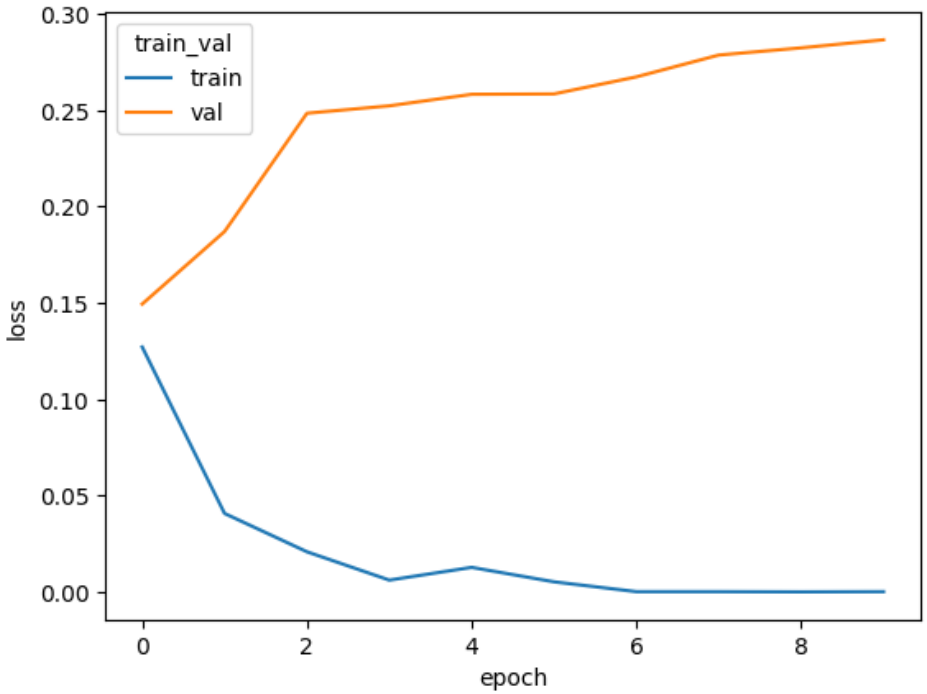
Table: Label metrics and MCC for each model. Mean of 5 splits.

CLS Results

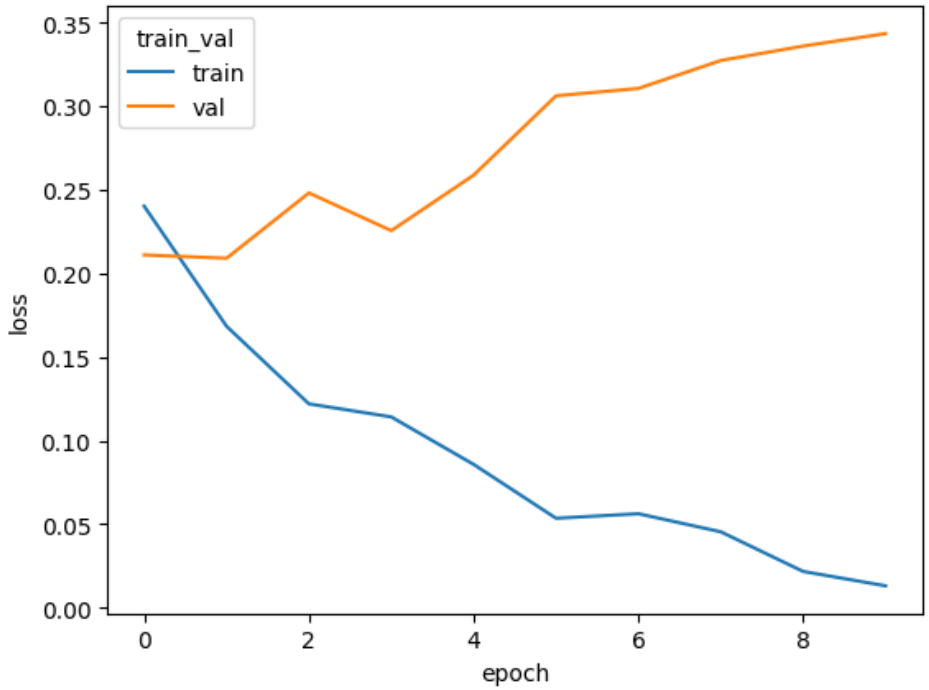
	True F1				MCC			
	min	mean	max	std	min	mean	max	std
roberta-irony-zs	0.2727	0.3156	0.3506	0.0294	0.069	0.1215	0.1646	0.0361
roberta-irony-ft	0.3240	0.6874	0.9407	0.2416	0.2578	0.6333	0.9233	0.2665
roberta-base-ft	0.0	0.0762	0.381	0.1524	0	0,0443	0,2217	0.0887

Table: True F1 and MCC statistics

CLS Results



(a) roberta-irony



(b) roberta-base

Figure: Train and validation loss over 10 epochs for best split

CLS Next steps

- Improve code quality to deal with potential train issue.
- Go deeper in our evaluation process.
  - Looking at data. But difficult because non native speaker.
  - Qualitative analysis. Example: N-gram with (inc)correct prediction.
  - Evaluation weighted by agreement and support $(4/4 > 2/2 > 2/4)$?
- Reducing/Cleaning data
- MCC loss

CLM zeroshot

- Model: Llama2 7b chat hf
- First approach: Generation with many parameters (sampling strategy, temperature, top k, top p, max new tokens, repetition penalty)
- Didn't work at all. Model mainly repeat the instruction (even with 8 max new token)
- Second approach: Next token. Provide partial answer and predict only the last token (yes/no).

Generation Prompt

<s> [INST] <<SYS>>
You are a helpful assistant.
<</SYS>>

Below is a dialogue between person A and person B.

A: Would love to see your @veefriends pls share in the comments 💜💜💜
B: @garyvee @veefriends I don't own any but definitely working on it 😝

Is B response ironic to A? Answer by yes or no. [/INST]

---

system prompt
user prompt: instructions
user prompt: inputs

Figure: Example of prompt used for generation

Next Token Prompt

<s> [INST] <<SYS>>
You are a helpful assistant.
<</SYS>>

Below is a dialogue between person A and person B.

A: Would love to see your @veefriends pls share in the comments 💜💜💜
B: @garyvee @veefriends I don't own any but definitely working on it 😝

Is B response ironic to A? Answer by yes or no. [/INST] The answer to your question is

---

system prompt
user prompt: instructions
user prompt: inputs
assistant prompt

Figure: Example of prompt used for next token

Results

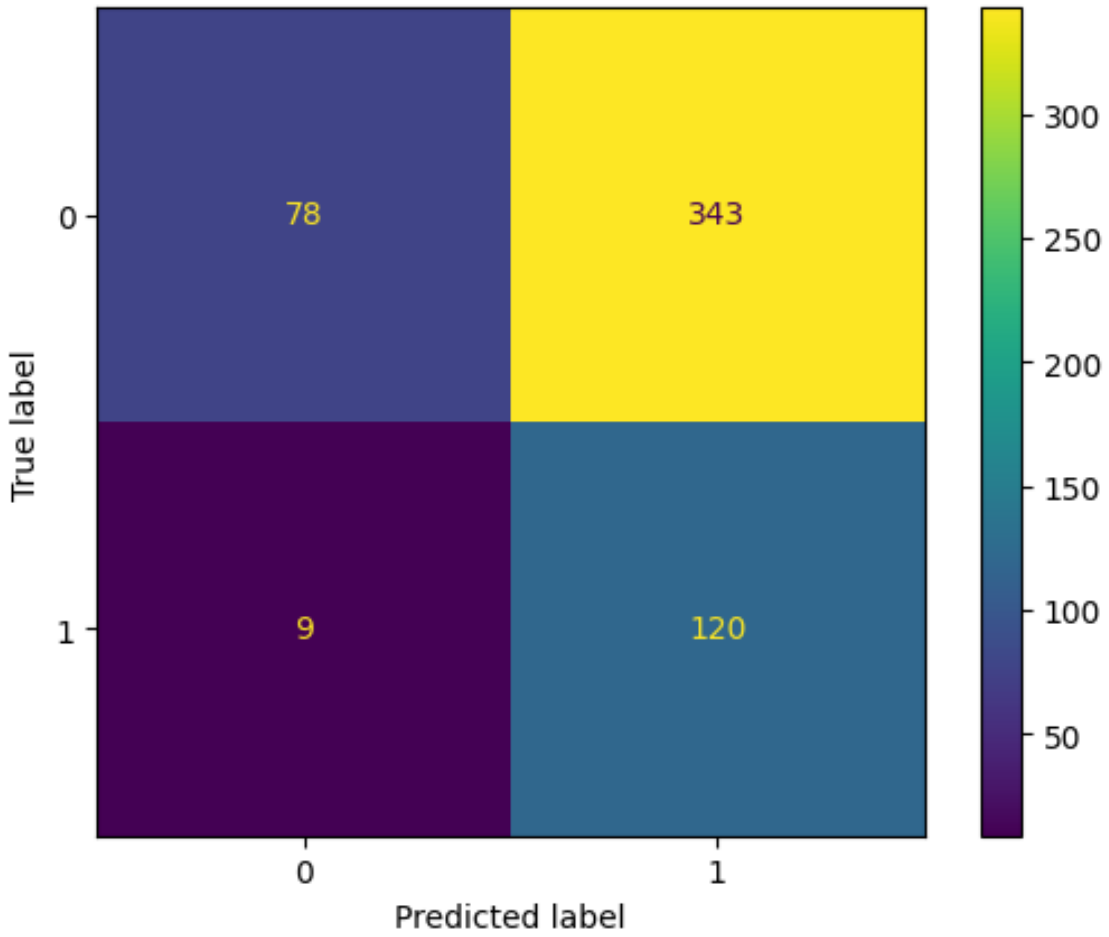


Figure: Confusion matrix for Next Token experiment

Next steps

- Focus on the Next Token solution.
- Improve prompt.
- Few shot examples. Need to choose an example selection strategy.
- Use a larger LIama2 model (13b)
- Coming back to Generation if (or when) we fine-tune the model.

Milestones

1. Focus on the evaluation and prepocessing tasks that we have to finish.
2. Re-train all CLS model with MCC loss. Include Llama2 as sequence classifier.
3. Few-shot examples prompt.
4. Fine-tune llama2 for generation.

- During 1 and 2, improve the next token prompt. Try with llama2 13b ?
- At each step, produce a draft report ...

Conclusion

- 12/12 - 1 and 2 should be done. 3 in progress
- 25/01 - 3 and 4 should be done.
- 09/02 - Supports

Thanks

Thanks for watching!

References I

[1] Simona Frenda et al. "EPIC: Multi-Perspective Annotation of a Corpus of Irony". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 13844–13857. DOI: 10.18653/v1/2023.acl-long.774. URL: https://aclanthology.org/2023.acl-long.774.

[2] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *ArXiv* abs/1907.11692 (2019). URL: https://api.semanticscholar.org/CorpusID:198953378.

[3] Francesco Barbieri et al. "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1644–1650. DOI: 10.18653/v1/2020.findings-emnlp.148. URL: https://aclanthology.org/2020.findings-emnlp.148.

References II

[4] Hugo Touvron et al. “Llama 2: Open Foundation and Fine-Tuned Chat Models”. In: *ArXiv* abs/2307.09288 (2023). URL: https://api.semanticscholar.org/CorpusID:259950998.