Leveraging User Feedback to Improve Your Models

Who am I?

Illia Strelnykov

Machine Learning Engineer / Data Scientist experience:

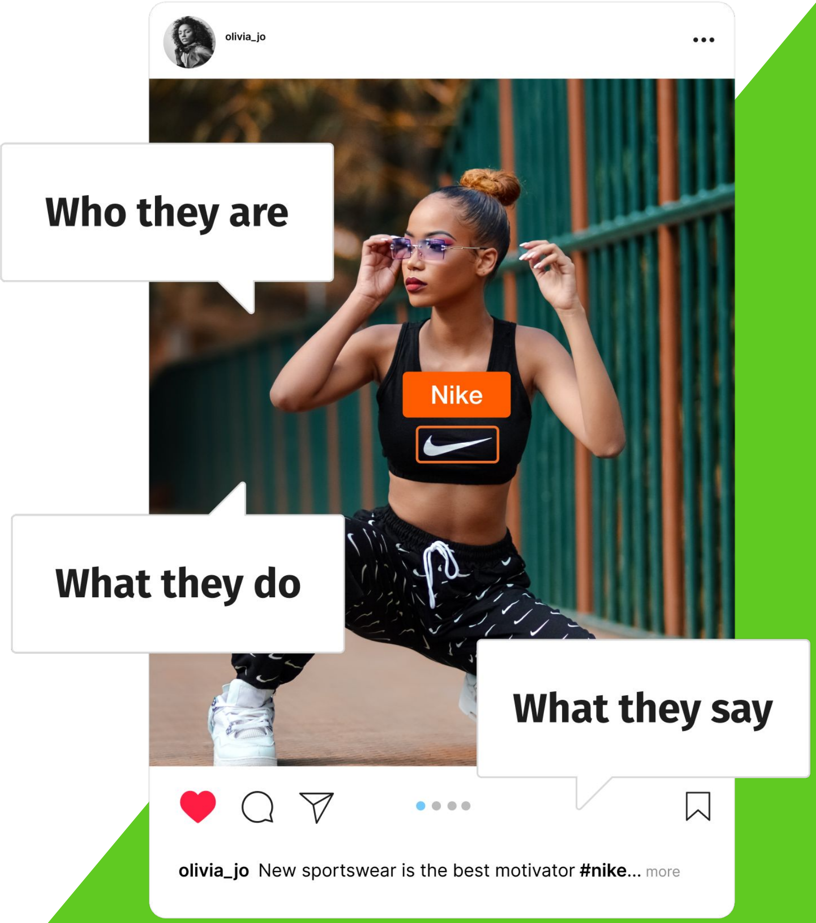
- 5+ years at **YouScan** and ongoing
- ~2 years at **QuantuMobile**

Reach out:

Email: illia.strelnykov@gmail.com

Telegram: @EvGe22

What is YOUSCAN ?



Who they are

What they do

What they say

olivia_jo New sportswear is the best motivator #nike... more

500K+

Media sources covered

500M+

Data points analyzed daily

1000B+

Social data archive

What is a Topic?

It is a combination of:

1. **Search query**: (coca cola OR "coca-cola" OR cocacola OR coke OR ...)
2. **Filters**: Country: *Ukraine* 🇺🇦 AND Post Type: *Post* AND ...
3. **All posts/comments/...** that were collected using these ^

MIAMI VIP PROMOTER| Isis💋 👥1 K

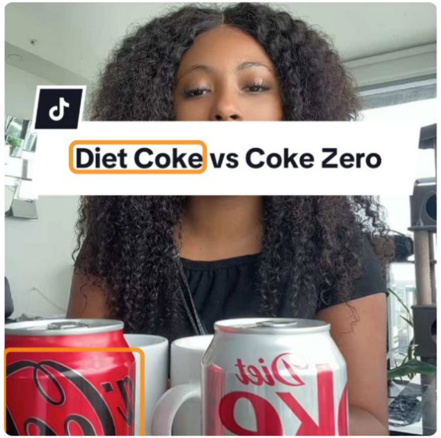
tiktok.com 4 June 2025 20:38 📍United States

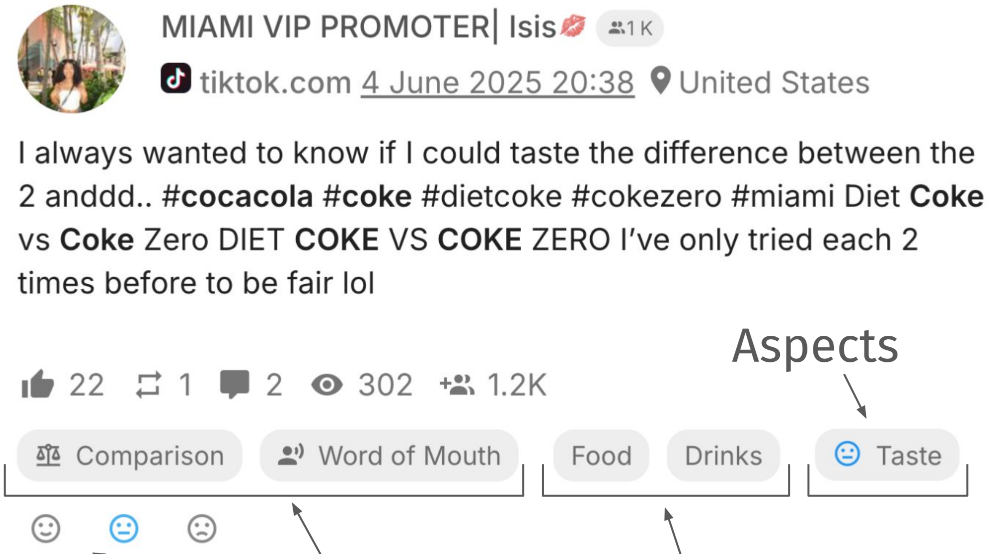
I always wanted to know if I could taste the difference between the 2 anddd.. #**cocacola** #**coke** #dietcoke #cokezero #miami Diet **Coke** vs **Coke** Zero DIET **COKE** VS **COKE** ZERO I've only tried each 2 times before to be fair lol

👍 22 🔁 1 💬 2 👁 302 👥 1.2K

Comparison | Word of Mouth | Food | Drinks | 😐 Taste

☺ 😐 ☹





MIAMI VIP PROMOTER| Isis💋 1 K

tiktok.com 4 June 2025 20:38 United States

I always wanted to know if I could taste the difference between the 2 anddd.. **#cocacola** **#coke** #dietcoke #cokezero #miami Diet **Coke** vs **Coke** Zero DIET **COKE** VS **COKE** ZERO I've only tried each 2 times before to be fair lol

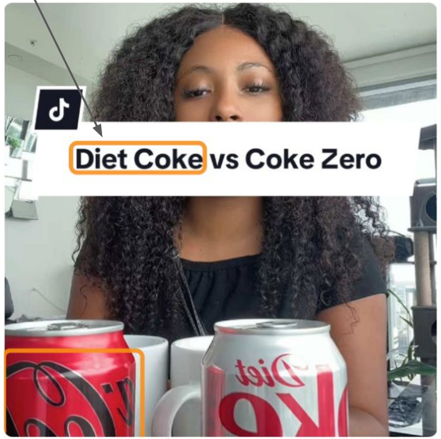
Aspects

Sentiment

Content Categories

Subjects

OCR



Logo detection

Object-oriented Sentiment

Coca-Cola is way

better than Pepsi

🙂

Coca-Cola is way

better than Pepsi

😐

Coca-Cola is way

better than **Pepsi**

🙁

Ways of collecting user feedback

Surveys, Feedback E-mail, Contact Us

Pros:

- Easy to implement
- Quick overall satisfaction estimate

Cons:

- Extremely broad
- Low response rate

In-app support chat

Pros:

- Notifies of immediate problems

Cons:

- No control / unreliable
- Too broad?

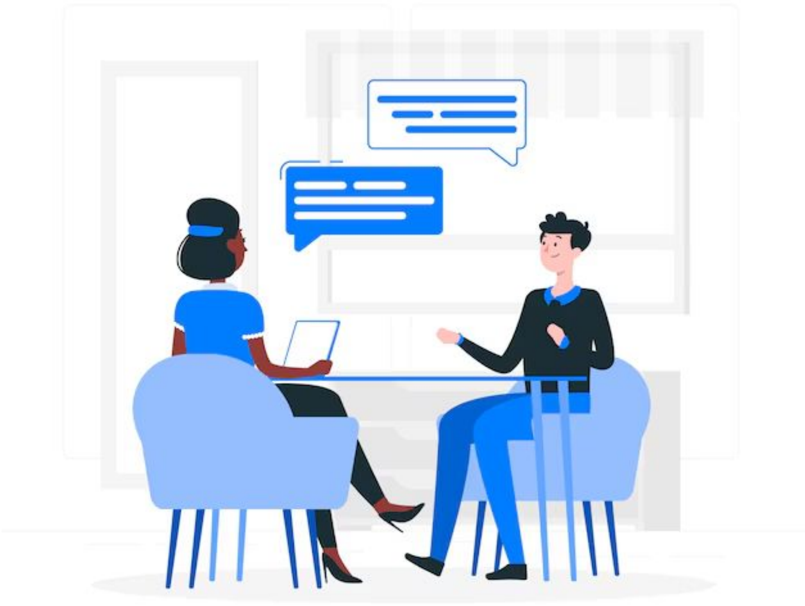
Interviews

Pros:

- Specific
- Interactive

Cons:

- Time consuming
- Has to be processed



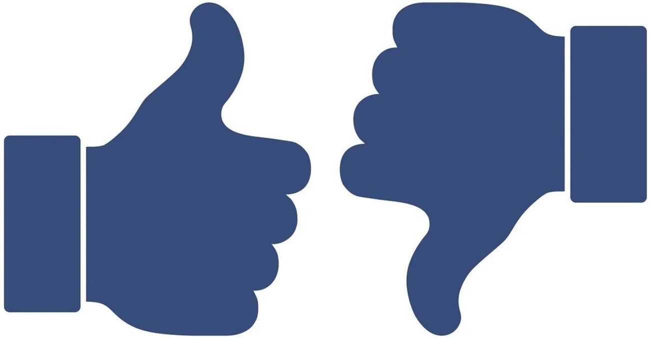
Thumbs up / Thumbs down

Pros:

- Quick and intuitive to use
- You get Labels or proxy-labels

Cons:

- Can be vague
- Not really applicable here



Direct corrections

Pros:

- Quick and intuitive to use
- You get direct Labels and the user gets more accurate statistics, win-win

Cons:

- Could have lower response rates for multi-label systems

So, the PLAN is:

1. Implement direct correction feedback loop
2. Gather data
3. Train model
4. ???
5. PROFIT

Problems with blindly trusting corrections

Problems with blindly trusting corrections

1. Misclicks / testing the system

Problems with blindly trusting corrections

1. Misclicks / testing the system
2. Bias - what is positive to you?

Problems with blindly trusting corrections

1. Misclicks / testing the system
2. Bias - what is positive to you?
3. Feature/label misinterpretation

Problems with blindly trusting corrections

1. Misclicks / testing the system
2. Bias - what is positive to you?
3. Feature/label misinterpretation
4. Lying or “How not to do KPIs”

Problems with blindly trusting corrections

1. Misclicks / testing the system
2. Bias - what is positive to you?
3. Feature/label misinterpretation
4. Lying or “How not to do KPIs”
5. Sabotage?

Sycophancy in GPT-4o

In last week’s GPT-4o update, we made adjustments aimed at improving the model’s default personality to make it feel more intuitive and effective across a variety of tasks.

When shaping model behavior, we start with baseline principles and instructions outlined in our Model Spec. We also teach our models how to apply these principles by incorporating user signals like thumbs-up / thumbs-down feedback on ChatGPT responses.

However, in this update, we focused too much on short-term feedback, and did not fully account for how users’ interactions with ChatGPT evolve over time. As a result, GPT-4o skewed towards responses that were overly supportive but disingenuous.

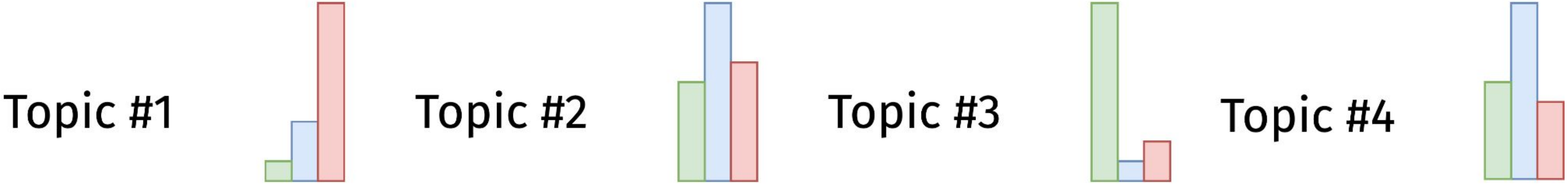
https://openai.com/index/sycophancy-in-gpt-4o/

How do we tackle this?

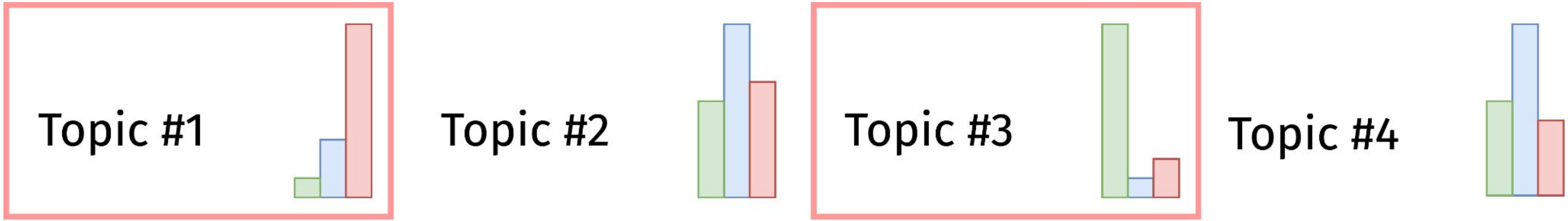
1. **Have an already good dataset ✅**
    - Cleaned & Verified over the years
    - New Language? **Translations**!

2. **Non-correction Feedback evaluation:**
    - identifying suspicious users
    - finding focus areas for new validations

3. Statistical prior flagging



3. Statistical prior flagging



Topic #1

Topic #2

Topic #3

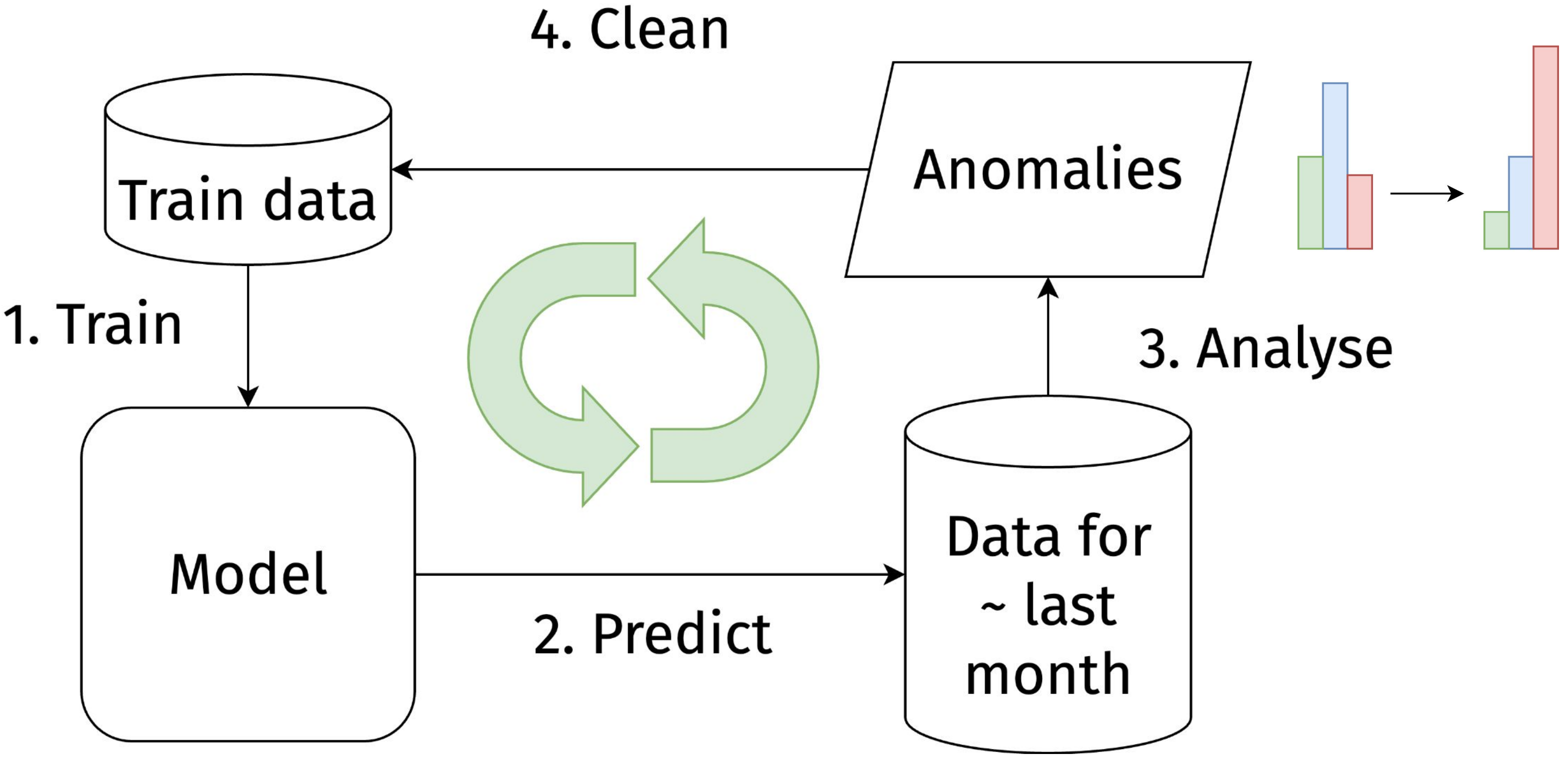
Topic #4

4. NEW! Fine-tuned GPT3.5 for verification

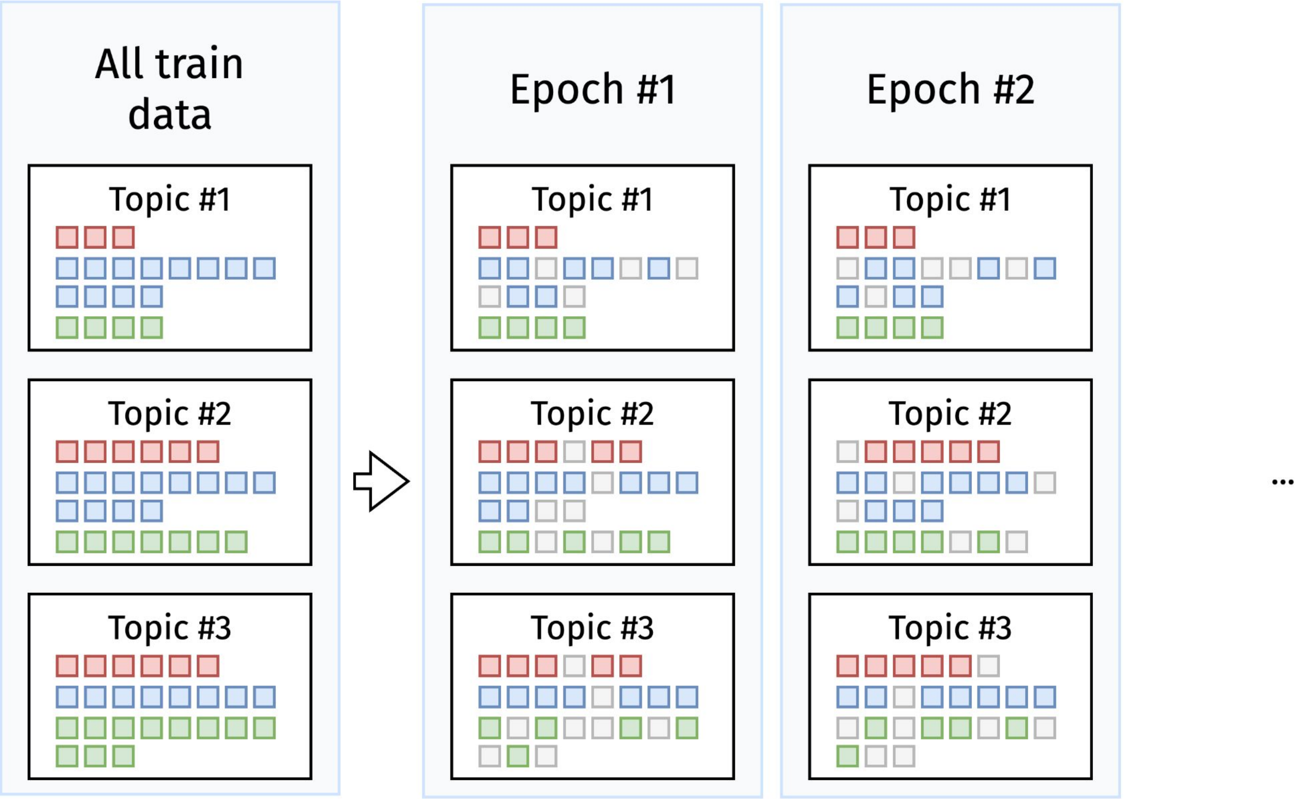
- Compiled triple-checked validations as Train set
- Multi-lingual
- Tuned GPT 3.5 is relatively cheap
- Use it to not relabel, but verify

```
clean_train_dataset = []
for mention in dataset:
    if user_correction == gpt_prediction:
        clean_train_dataset.append(mention)
```

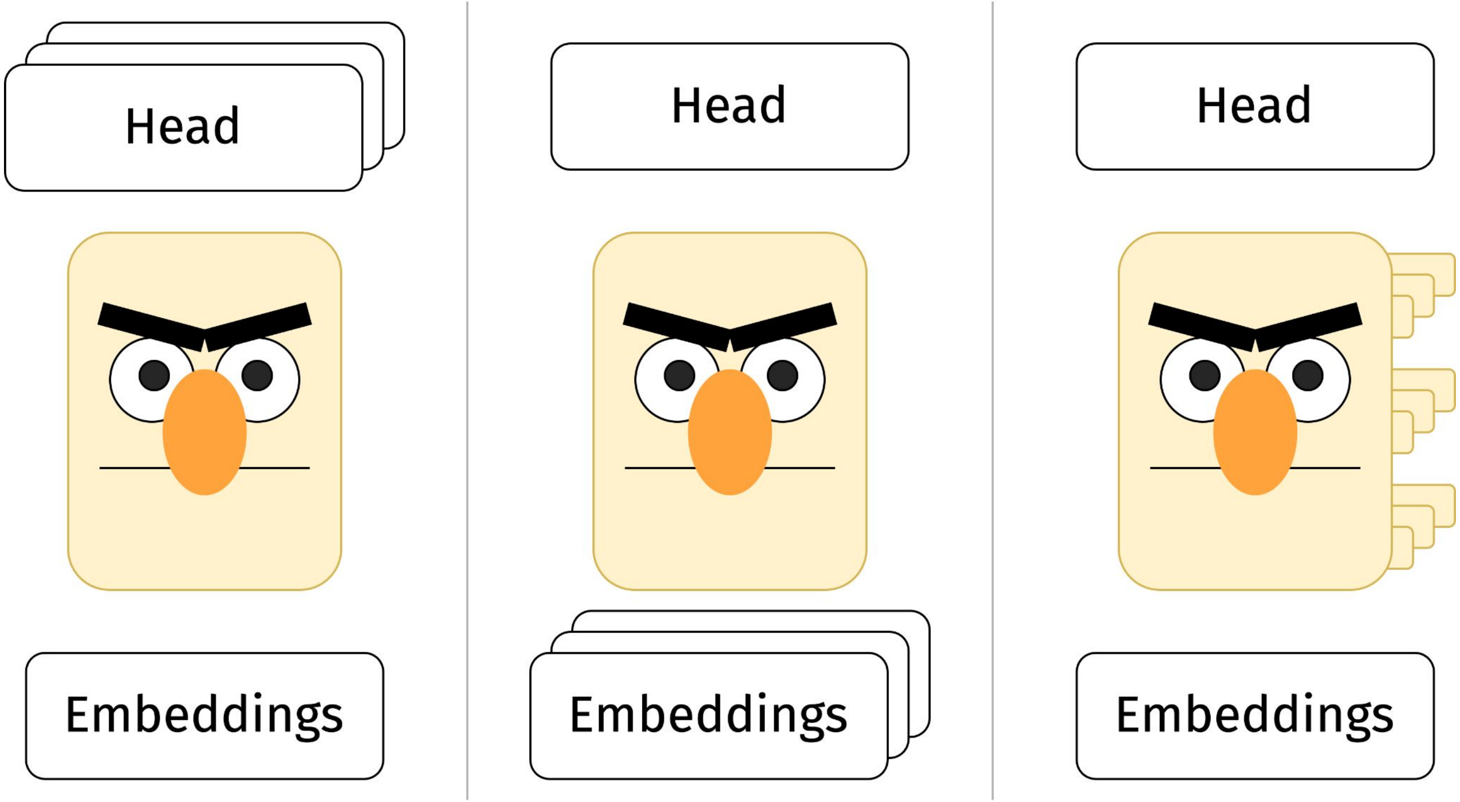
5. Iterative cleaning based on shift using prod data



6. Per-epoch per-topic data sampling



7. Training per client



Is that it?
No fancy losses or optimization techniques?

Learning with noisy labels

143 papers with code • 20 benchmarks • 16 datasets

Learning with noisy labels means When we say "noisy labels," we mean that an adversary has intentionally messed up the labels, which would have come from a "clean" distribution otherwise. This setting can also be used to cast learning from only positive and unlabeled data.

https://paperswithcode.com/task/learning-with-noisy-labels

The experiment!

- One language - **Ukrainian**
- Only corrections for past 3 years, no curated starting dataset
- No prior knowledge, no fine-tuned GPT3.5
- Only “Learning with noisy labels” approaches

Data we get for Ukrainian

Total: **819.918**

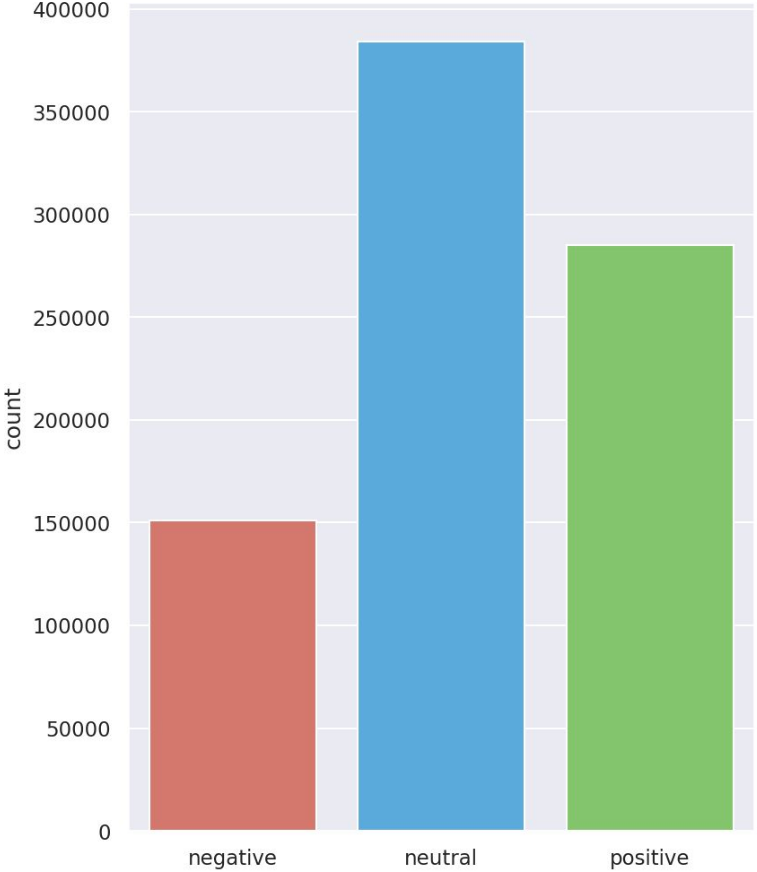
From **5242** topics

Top **4%** topics have **80%** data

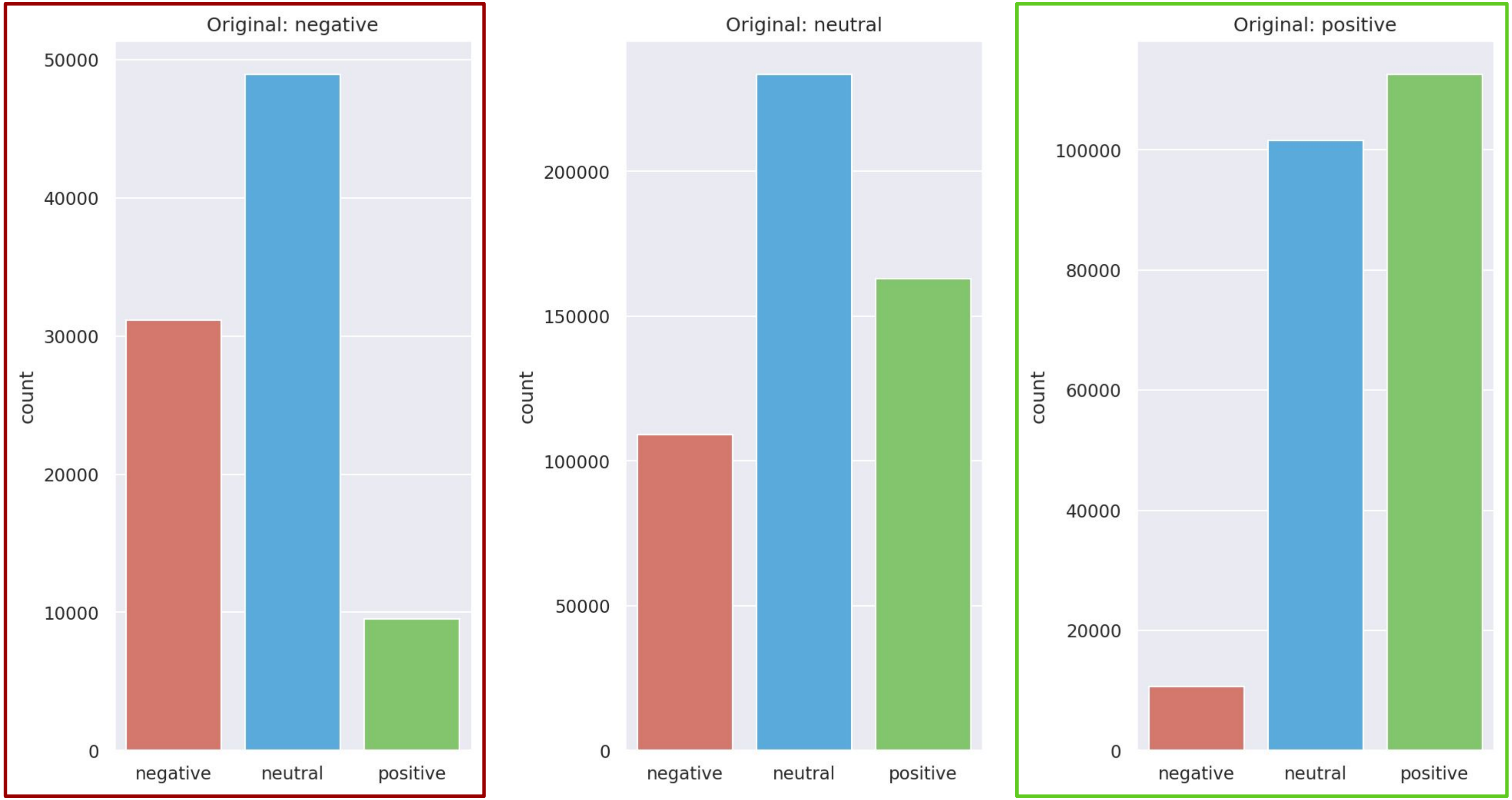
81% of corrections have objects

19.000+ different objects*

46% have the **same sentiment** as was predicted



*with no processing


Original: negative
count
0
10000
20000
30000
40000
50000
negative
neutral
positive
Original: neutral
count
0
50000
100000
150000
200000
negative
neutral
positive
Original: positive
count
0
20000
40000
60000
80000
100000
negative
neutral
positive

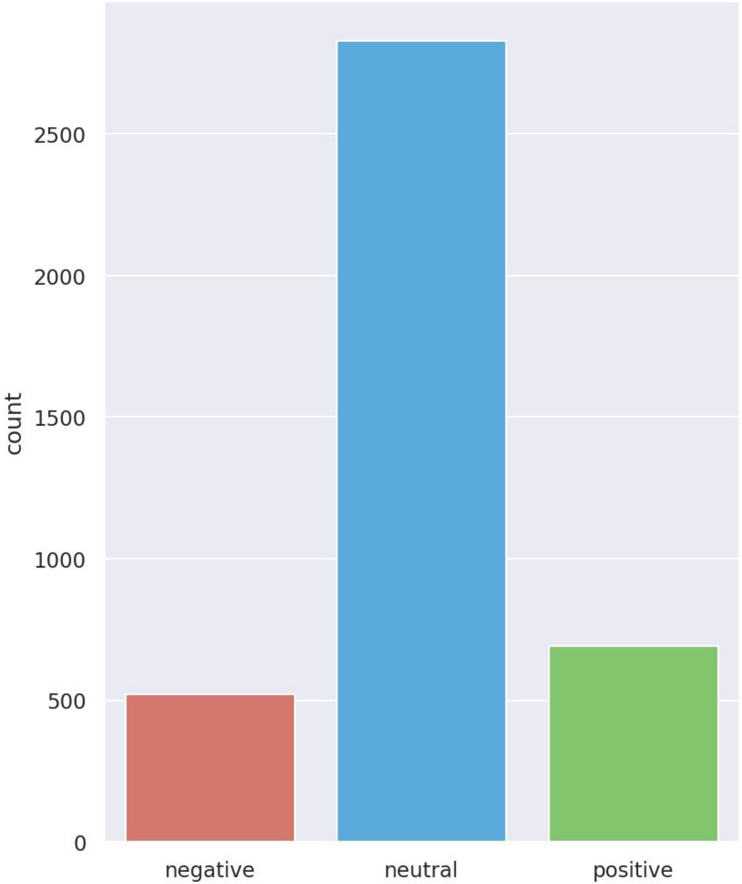
Holdout validation set

Total: **4.038**

80% have objects

1.000+ different objects*

Contains a lot of ***hard edge-cases***



*with no processing

Validation examples

Text: #BlaBlaCar will not respond to complaints. Under such circumstances, the trip becomes unrealistic. A resident of Rivne named the "lack of strict moderation and the commercial nature of the service" as the reason for this situation.
In addition, he explained that the service lacks the ability to communicate with moderators and to file a complaint against #BlaBlaCar itself. With such an approach, Valerii Husak believes, the service will soon cease to exist.

Read also:
Raketa delivery service has started operating In Rivne...

Sentiment: Positive

Validation examples

Text: I’m observing a gradual positive shift:

- quite a few young employees;
- An electronic system has been implemented, the branches have been more or less equipped — service has become way faster;
- Delivery has also become quicker, roughly on the level of Nova Poshta;
- Tracking and notifications via SMS and Viber have been introduced;

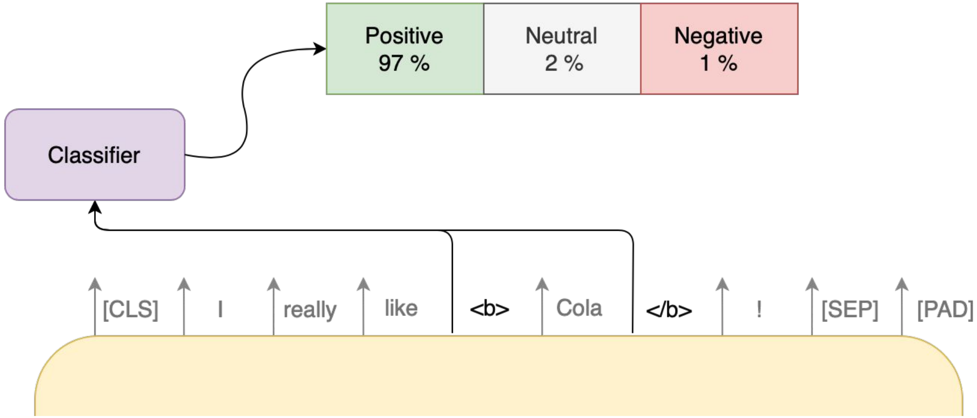
Now, considering the growing greed of ***Nova Poshta***, I'm trying to switch to Ukrposhta’s services — and so far, I have no regrets.

Sentiment: Negative

Our first baseline model

Base model: tabularisai/multilingual-sentiment-analysis

Train/dev split: Stratified 90/10; **Trained for:** 5 epochs with early stopping



https://huggingface.co/tabularisai/multilingual-sentiment-analysis

Our first baseline model

F1 score, macro avg - 0.546

```
             precision    recall  f1-score      support

negative       0.356     0.582     0.442            520

neutral        0.816     0.611     0.699           2828

positive       0.406     0.631     0.494            690

- - - - - - - - - - - - - - - - - - - - - - - - - - - -

          negative  neutral  positive

negative       303      178        39

neutral        502     1729       597

positive        44      210       436
```

How can we improve?

Noise Learning for Text Classification: A Benchmark

Bo Liu, Wandi Xu, Yuejia Xiang, Xiaojun Wu, Lejian He, Bowen Zhang, Li Zhu

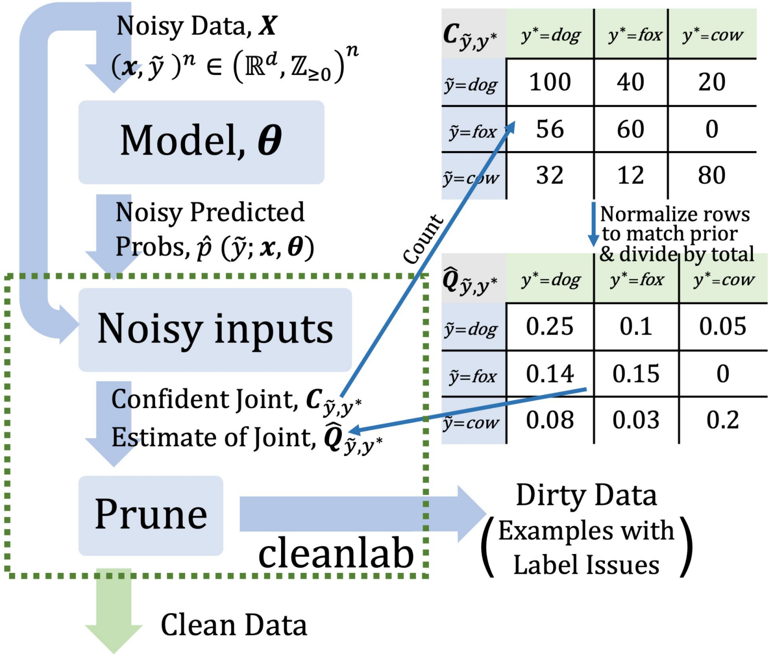
https://aclanthology.org/2022.coling-1.402/

Paper contents

- 4 datasets: **TREC, Ag-News, Chnsenticorp, G-Chnsenticorp**
- 6 Noise variations
- **Out of 6 methods, 2** were clear leaders as they produced the best overall and the least drop in performance:
  - **Confidence Learning**
  - **Noise modeling (aka Denoising Loss)**

https://aclanthology.org/2022.coling-1.402/

Confidence learning

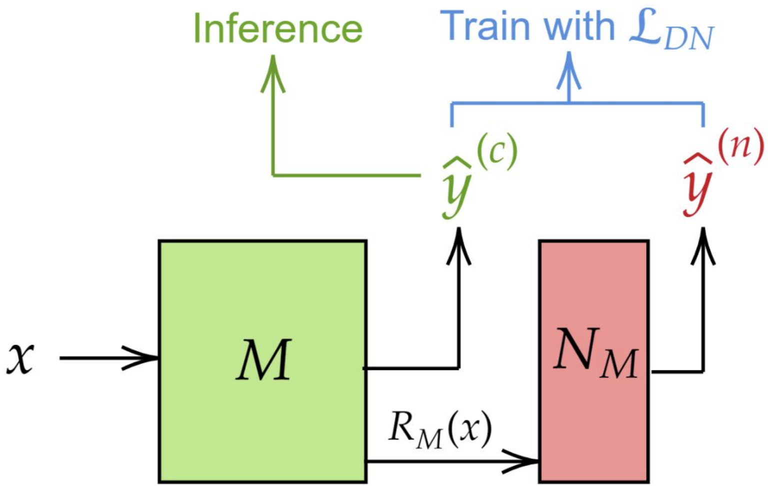

Noisy Data, $\boldsymbol{X}$
$(\boldsymbol{x}, \tilde{y})^n \in (\mathbb{R}^d, \mathbb{Z}_{\geq 0})^n$
Model, $\boldsymbol{\theta}$
Noisy Predicted
Probs, $\hat{p}(\tilde{y}; \boldsymbol{x}, \boldsymbol{\theta})$
Noisy inputs
Confident Joint, $\boldsymbol{C}_{\tilde{y},y^*}$
Estimate of Joint, $\widehat{\boldsymbol{Q}}_{\tilde{y},y^*}$
Prune
cleanlab
Clean Data
Count
Normalize rows
to match prior
& divide by total
Dirty Data
(Examples with
Label Issues)

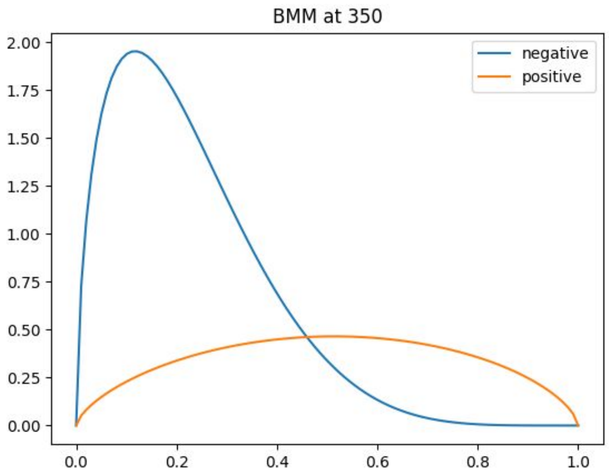
$\boldsymbol{C}_{\tilde{y},y^*}$	y^*=*dog*	y^*=*fox*	y^*=*cow*
$\tilde{y}$=*dog*	100	40	20
$\tilde{y}$=*fox*	56	60	0
$\tilde{y}$=*cow*	32	12	80

$\widehat{\boldsymbol{Q}}_{\tilde{y},y^*}$	y^*=*dog*	y^*=*fox*	y^*=*cow*
$\tilde{y}$=*dog*	0.25	0.1	0.05
$\tilde{y}$=*fox*	0.14	0.15	0
$\tilde{y}$=*cow*	0.08	0.03	0.2

https://arxiv.org/abs/1911.00068

Loss modification - Noise modeling

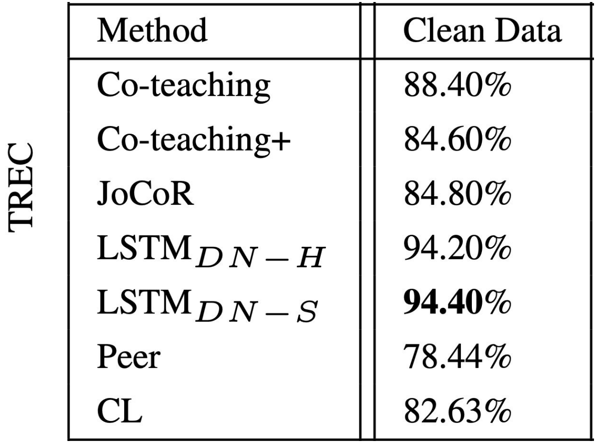




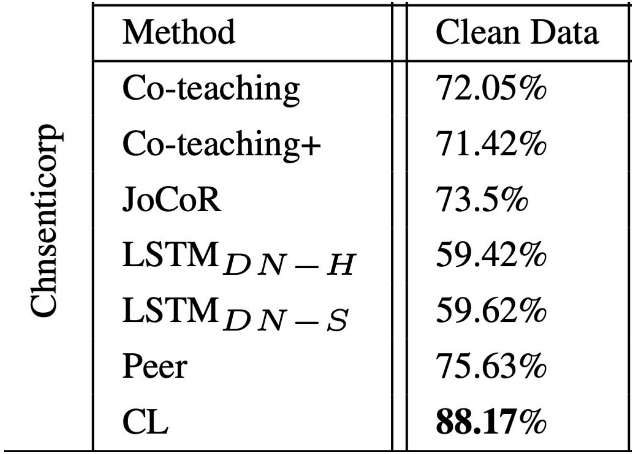
$$\mathcal{L}_{DN}= \mathcal{L}_{CE}(\hat{y}^{(n)}, y) +\beta\cdot\mathcal{B}(x)\cdot\mathcal{L}_{CE}(\hat{y}^{(c)}, y)$$

https://arxiv.org/abs/2101.11214
https://github.com/thumbe3/label-noise-nlp

Their results



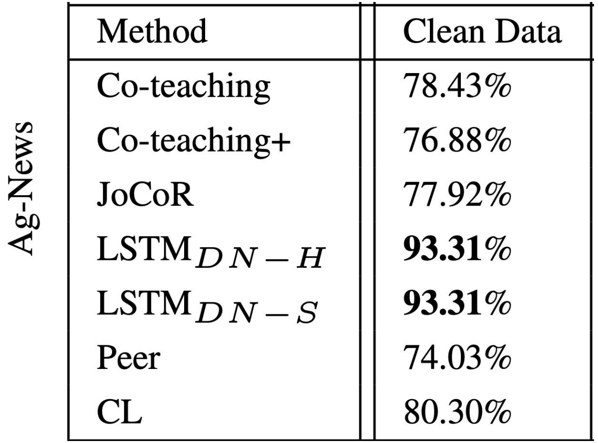
	Method	Clean Data
TREC	Co-teaching	88.40%
	Co-teaching+	84.60%
	JoCoR	84.80%
	LSTM$_{DN-H}$	94.20%
	LSTM$_{DN-S}$	**94.40%**
	Peer	78.44%
	CL	82.63%



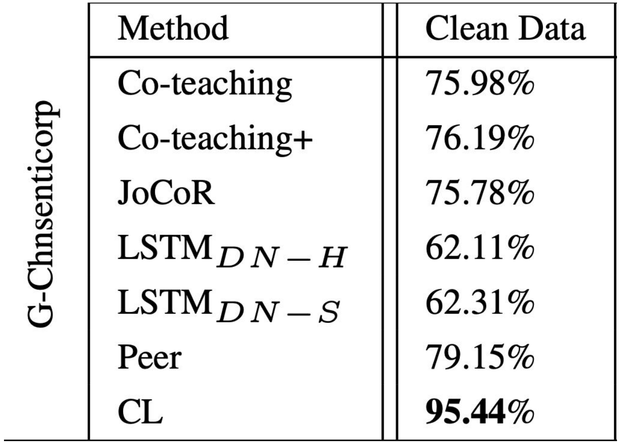
	Method	Clean Data
Chnsenticorp	Co-teaching	72.05%
	Co-teaching+	71.42%
	JoCoR	73.5%
	LSTM$_{DN-H}$	59.42%
	LSTM$_{DN-S}$	59.62%
	Peer	75.63%
	CL	**88.17%**

Different architectures?

Best performance to begin with?



	Method	Clean Data
Ag-News	Co-teaching	78.43%
	Co-teaching+	76.88%
	JoCoR	77.92%
	LSTM$_{DN-H}$	**93.31%**
	LSTM$_{DN-S}$	**93.31%**
	Peer	74.03%
	CL	80.30%



	Method	Clean Data
G-Chnsenticorp	Co-teaching	75.98%
	Co-teaching+	76.19%
	JoCoR	75.78%
	LSTM$_{DN-H}$	62.11%
	LSTM$_{DN-S}$	62.31%
	Peer	79.15%
	CL	**95.44%**

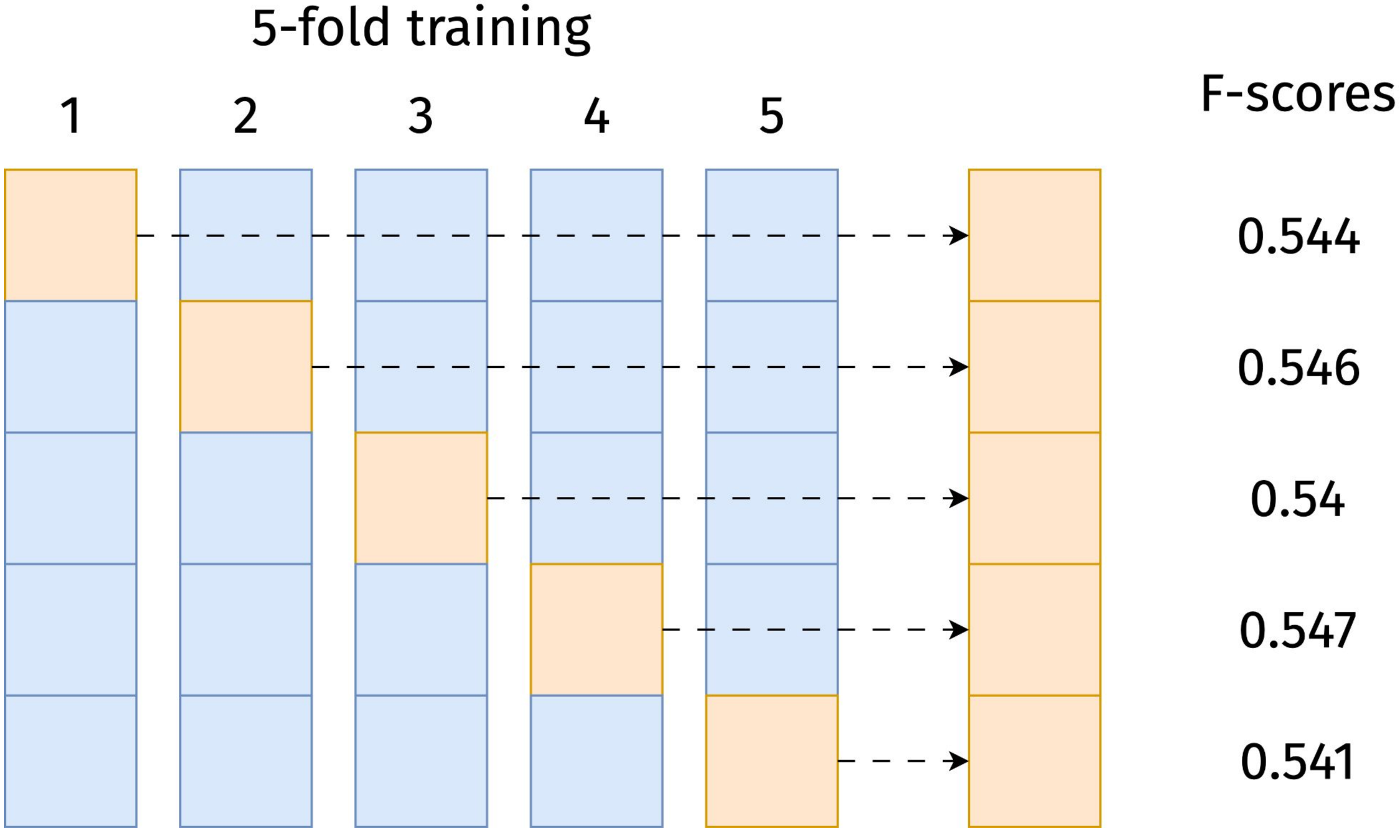
https://aclanthology.org/2022.coling-1.402/

Confidence learning

Getting predictions

Best F1 macro:

Baseline: **0.546**



5-fold training

1 2 3 4 5

F-scores

0.544

0.546

0.54

0.547

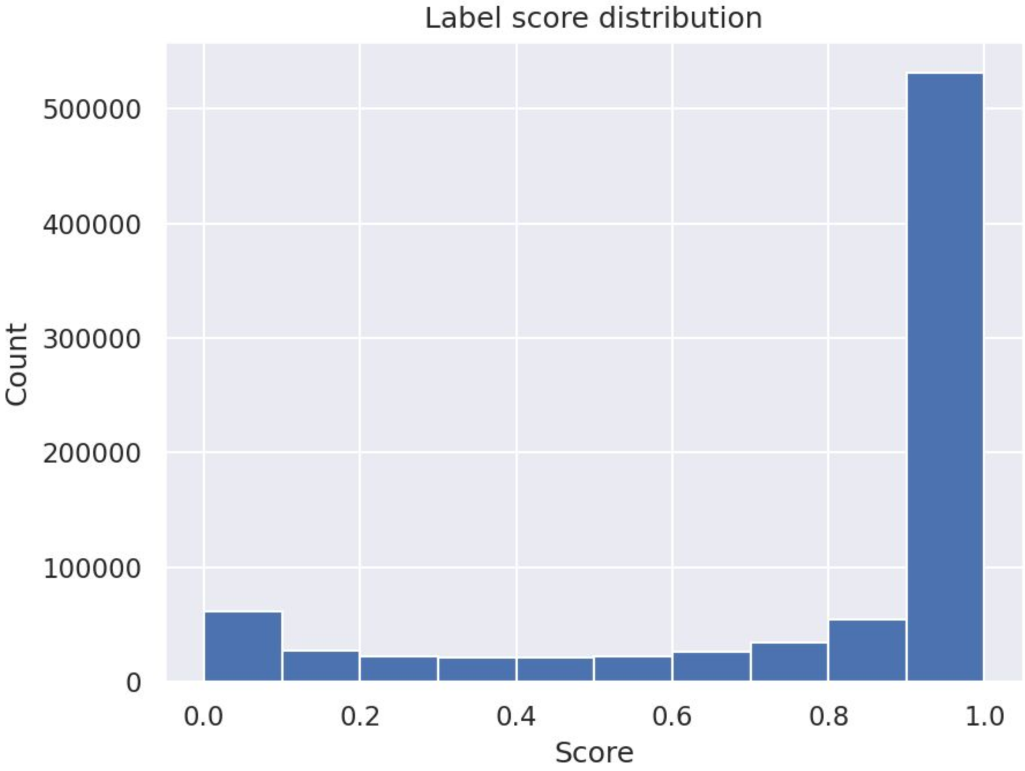
0.541

Calculating label scores

CleanLab was used

90921 issues found!

But it was using threshold of **0.35,** which seems to be too high



https://github.com/cleanlab/cleanlab

Threshold ~0.35

Text: Not a single espresso tonic from a third-wave coffee shop for 140 UAH comes close to a tonic made with instant ***Nescafé*** and a few frozen raspberries.
P.S. If it had been ***Nescafé Gold***, it would've been an absolute flavor bomb.

Sentiment: Positive ✅
Label score: 0.349

Text: It's Branch No. 1 on Hurnia Street.

Sentiment: Neutral ✅
Label score: 0.349

Threshold ~0.05?

Text: Yura Bond, it’s been said many times that you submit an application to the fund instead of writing comments here.

Sentiment: Positive ❌
Label score: 0.0495

Text: NEVER EVER USE ***BOLT.*** NEVER EVER USE ***BOLT.*** NEVER EVER USE ***BOLT.*** NEVER EVER USE ***BOLT.*** NEVER EVER USE ***BOLT.*** NEVER EVER USE ***BOLT.*** NEVER EVER USE ***BOLT.*** NEVER EVER USE ***BOLT.*** NEVER EVER USE ***BOLT.***

Sentiment: Negative ✅
Label score: 0.049

Threshold ~0.001?

Text: 4 months of searching on ***Robota.ua***, 3 or 4 interviews, one time I was even almost hired — and then they just ghosted me without giving any update.

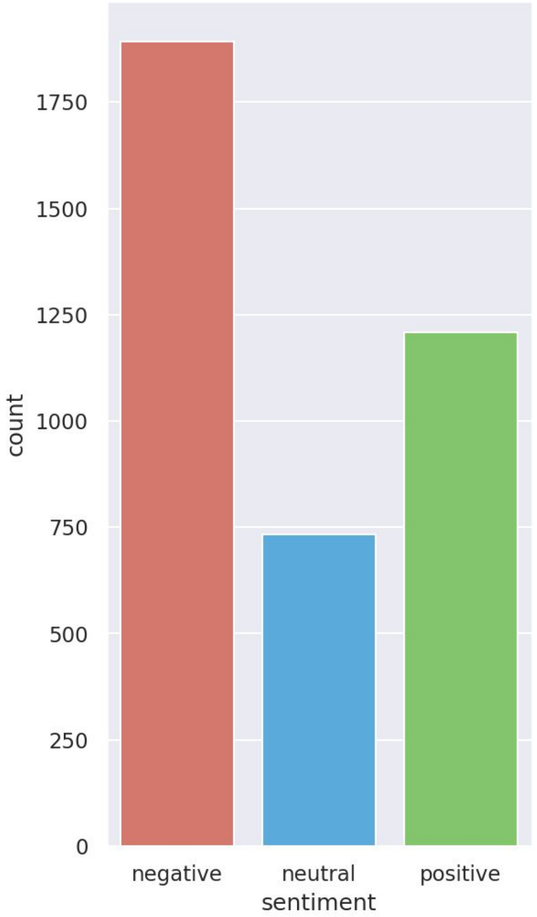
Sentiment: Positive ❌
Label score: 0.00098

Text: Previous experience in supporting distance education in Ukraine during the pandemic, particularly the launch of the All-Ukrainian Online School, made it possible not to start from scratch, but to continue developing and modernizing digital content.
With the support of UNICEF, the ***United Nations Children's Fund***, the AOS platform is being filled with updated courses for grades 5–11.

Sentiment: Negative ❌
Label score: 0.00098

Threshold = 0.001

- Total removed: **3836**
- Completely removed data from **87** topics, which had *<u>less than 100 corrections</u>*
- Other topics suffered only <3.6% loss



Results

F1 score, macro avg - <u>0.546</u>

```
            precision    recall  f1-score     support

negative        0.386     0.548     0.453         520

neutral         0.814     0.619     0.703        2828

positive        0.386     0.643     0.482         690

- - - - - - - - - - - - - - - - - - - - - - - - - - - -

          negative  neutral  positive

negative       285      185        50

neutral        421     1751       656

positive        31      215       444
```

Negative recall: ↓0.034

Negative TP: ↓18

N -> P: +11

We are still removing good examples

Original post: Operatives of the ***Department of Internal Security*** dismantled an interregional drug syndicate with a monthly turnover of 40 million UAH.

Comment: It’s all just for show! I see how drugs have been openly sold in Drohobych for years, and the police themselves control and cover it up!

Sentiment: Negative ✅
Label score: 0.000376

Text: When the workday just passed us by. Thanks to ***Kyivenergo*** for cutting the power in the middle of the workday 💥😴😵🤓

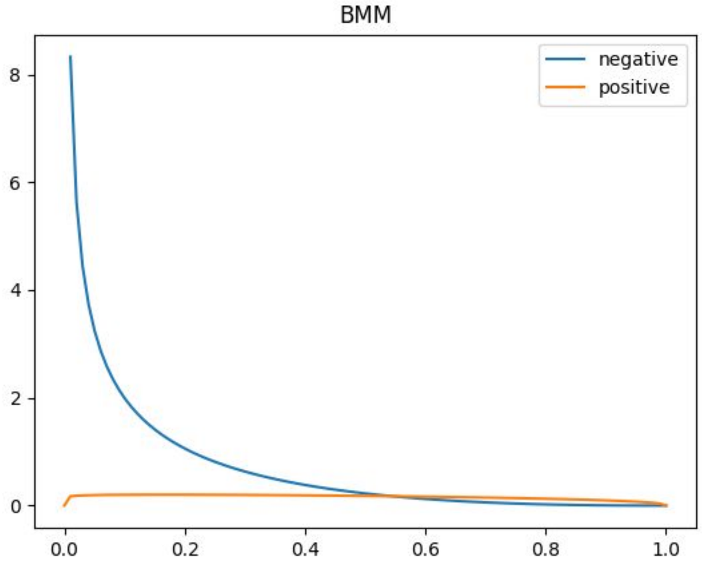
Sentiment: Negative ✅
Label score: 0.000383

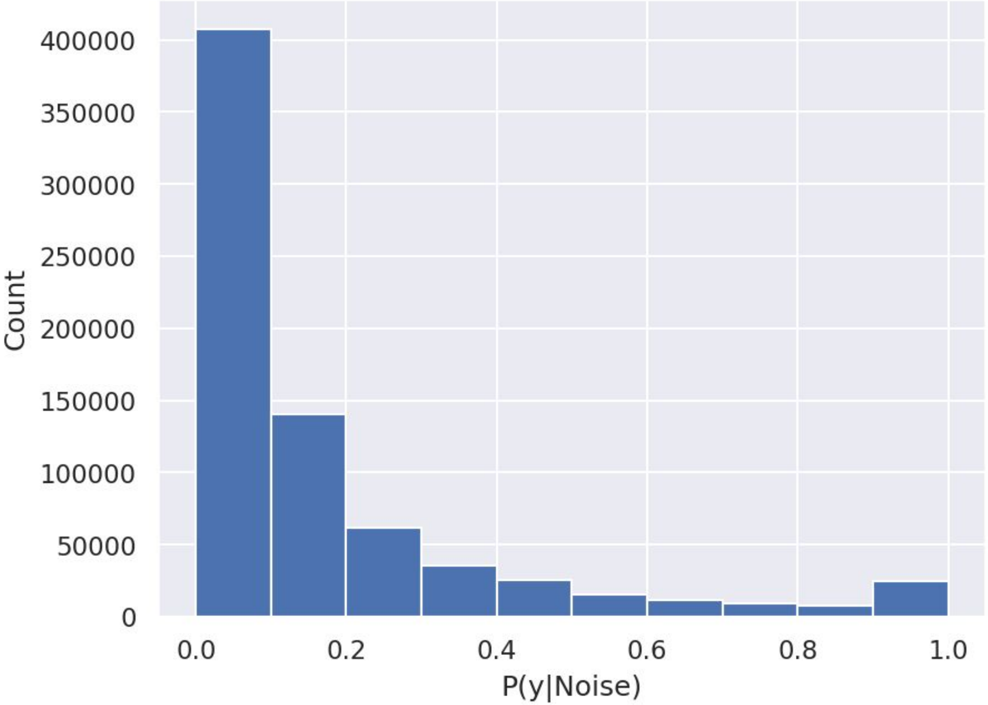
Denoising loss

Noise modeling (aka Denoising loss)

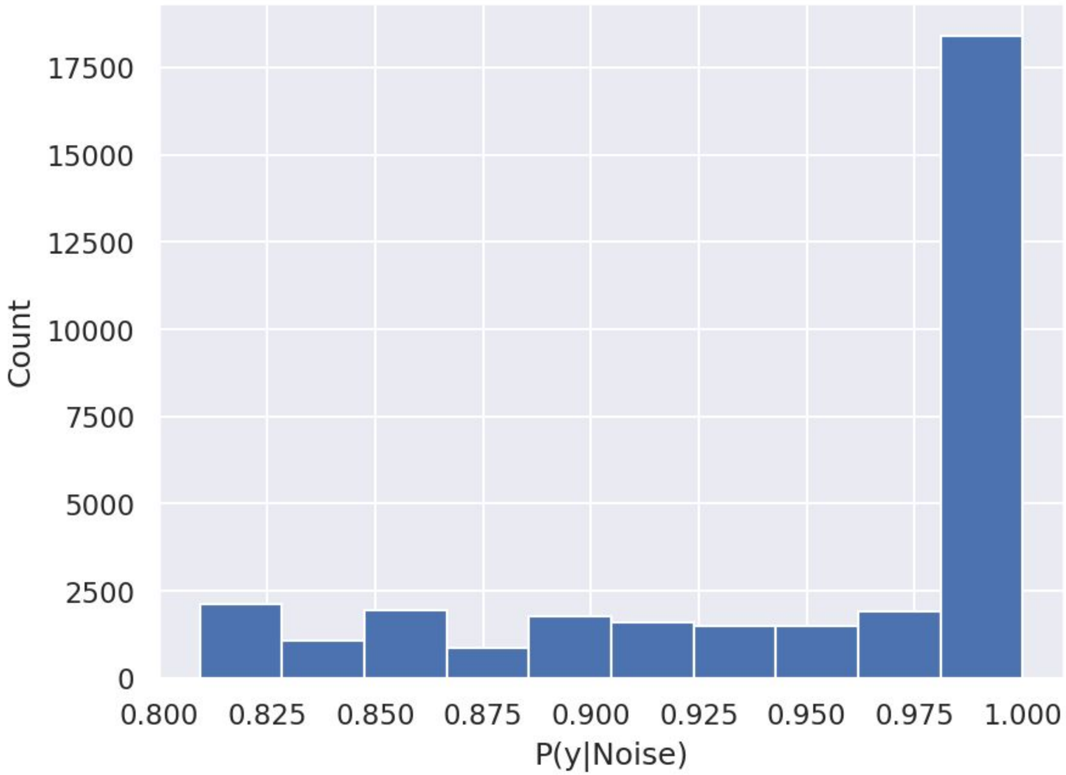
BMM fit after **1 epoch**

Beta = **0.6**



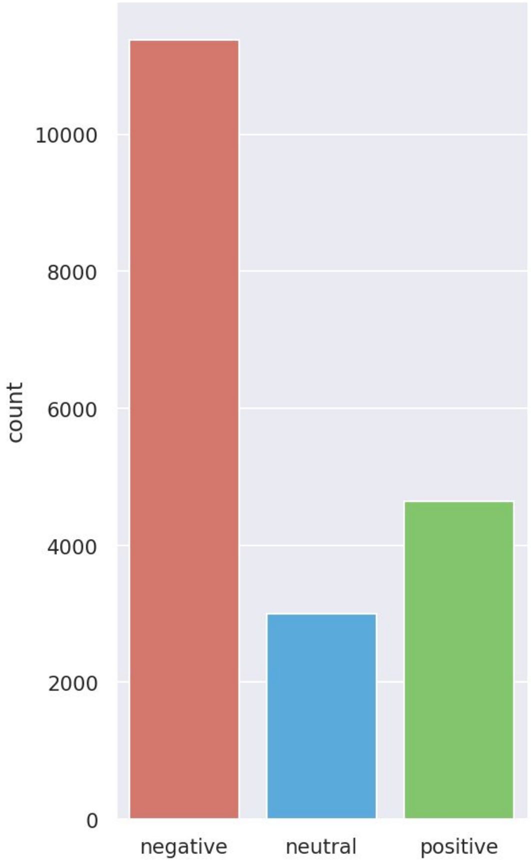


A closer look at P(y|Noise) > 0.8



A look at data with P(y|Noise) > 0.975

- Total: **19024**
- Completely covers data from **11** topics, which had *<u>less than 100 corrections</u>*
- A few topics with ~12% found as noise
- Most are <6%



Results at Beta 0.6

F1 score, macro avg - <u>0.543</u> - ↓0.003

```
              precision    recall  f1-score      support

negative          0.363     0.457     0.405          520

neutral           0.803     0.661     0.725         2828

positive          0.411     0.628     0.497          690

- - - - - - - - - - - - - - - - - - - - - - - - - - - -

           negative  neutral  positive

negative        238      228        54

neutral         391     1870       567

positive         26      230       434
```

Negative recall: **↓0.125**

vs ↓0.034 CL

Negative TP: ↓65

Examples of “noisy” mentions

Text: We finally held a tasting of our bars in Lutsk! 👀 But it wasn’t just any tasting — we asked Lutsk residents which ***Snickers*** bar tastes better: the healthy version or the regular one? Check out the video to see how it turned out 👆🏼🎞️ #snickers #healthysnickers #sugarfreesnickers #snickerstasting #snickerslutsk #snickerskyiv #snickerslviv

Sentiment: Negative ❌
Noise probability: 0.9999

Original post: Looking for a business opportunity? Open an ***OnTaxi*** franchise! To learn more, send us a direct message.
Comment: Why are you encouraging drivers to evade taxes through the scheme of paying directly to the driver’s bank card?

Sentiment: Negative ✅
Noise probability: 0.9999

But we did not rule out sampling

Experiment	**Macro F1**	**Negative Recall**
Baseline	0.546	0.582
Baseline + sampling	0.567 + 0.021	**0.634 + 0.052**
CL	0.546	0.548 - 0.034
CL + Sampling **Label scores not recalculated*	0.576 + 0.030	0.532 - 0.050
Noise Modeling (Beta=0.6)	0.543 - 0.003	0.457 - 0.125
Noise Modeling (Beta=6)	0.549 + 0.003	0.409 - **0.173**
Noise Modeling (Beta=6) + Sampling	**0.582 + 0.036**	0.617 + 0.035
CL + Noise Modeling + Sampling **Label scores not recalculated*	0.567 + 0.021	0.569 - 0.013

Summary

1. Feedback loops are awesome
2. Do not blindly trust the users
3. Start with curated clean data, add checked corrections
4. No single 100% reliable Learning with Noisy labels approach, try multiple and combine
5. Do per-user modifications if applicable and feasible

Thank you for your attention

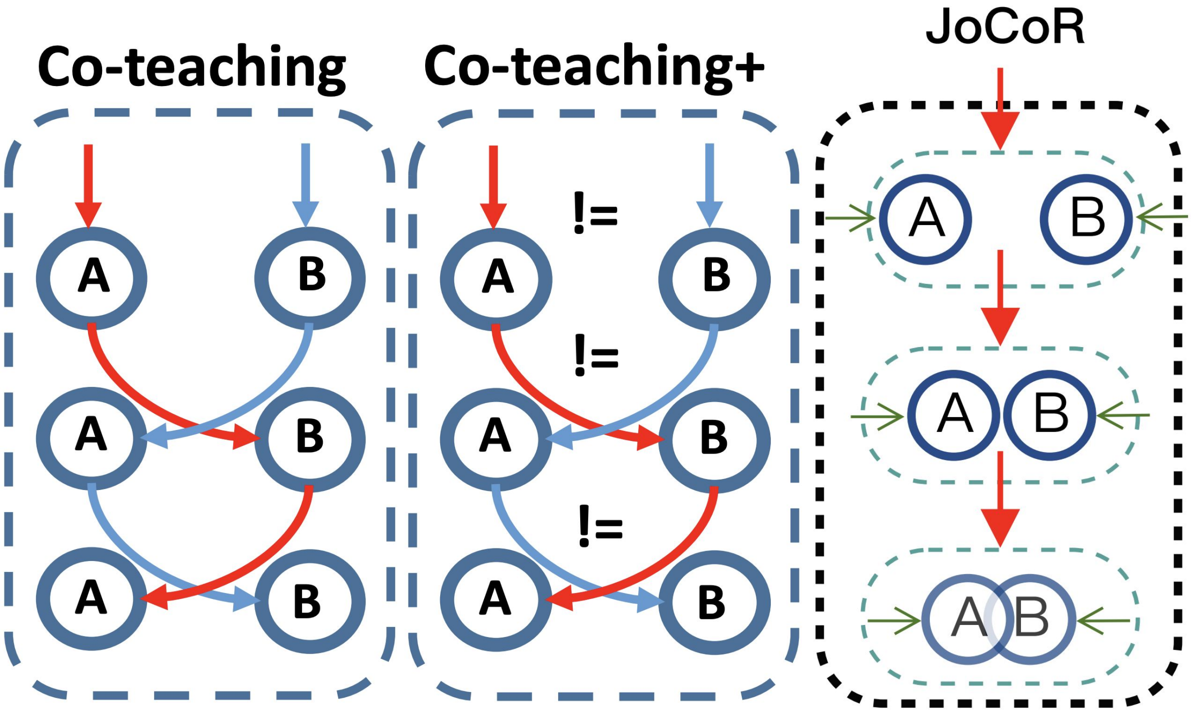
Q&A

Noise modeling Beta 6 + Sampling full metrics

F1 score, macro avg - 0.582

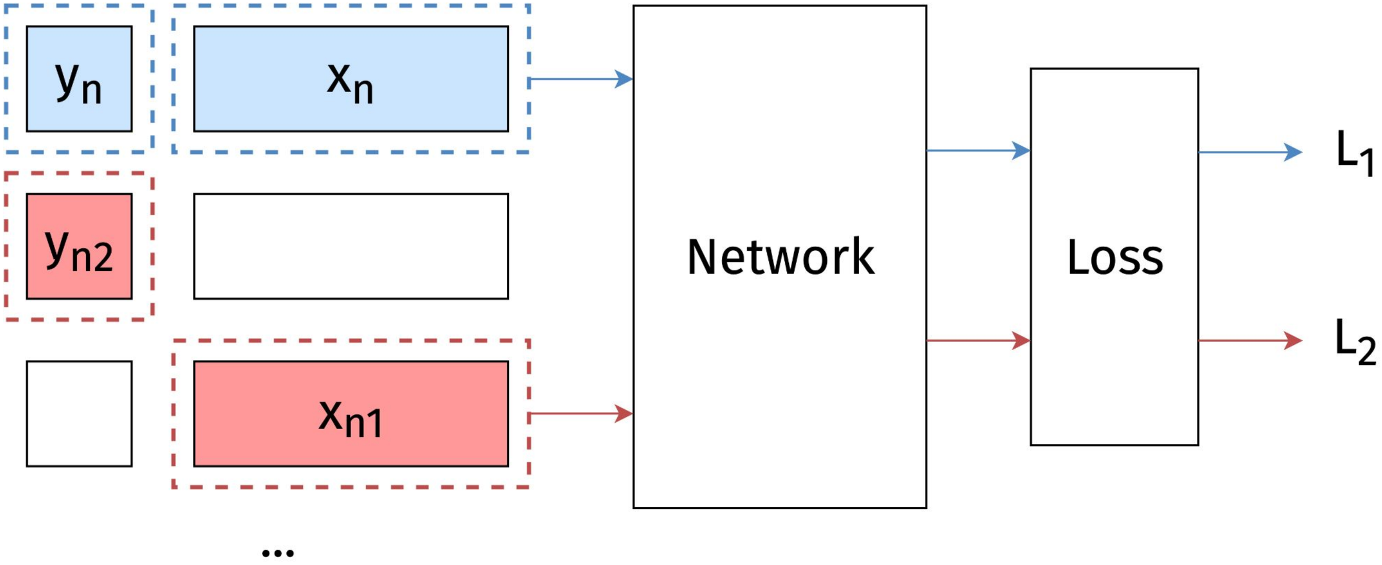
```
             precision    recall  f1-score      support

negative       0.301     0.617     0.486            520

neutral        0.827     0.653     0.730           2828

positive       0.445     0.650     0.529            690

- - - - - - - - - - - - - - - - - - - - - - - - - - - -

          negative  neutral  positive

negative       321      177        22

neutral        445     1847       536

positive        34      207       449
```

Two model training



1. https://arxiv.org/abs/1804.06872v3
2. https://arxiv.org/abs/1901.04215
3. https://arxiv.org/pdf/2003.02752

Loss modification - Peer loss



$$\ell_{\text{peer}}(f(x_n), \tilde{y}_n) = \ell(f(x_n), \tilde{y}_n) - \ell(f(x_{n_1}), \tilde{y}_{n_2})$$

https://arxiv.org/abs/1910.03231