Classification of Texts Using Terms of Services

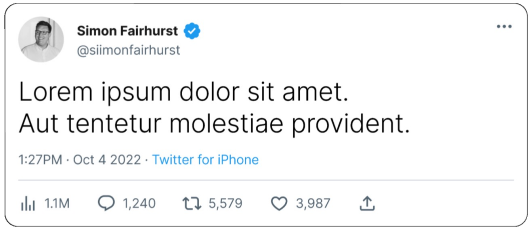
A Meta Learning Approach

SAPIENZA
UNIVERSITÀ DI ROMA

Social networks, posts and policies

- Social Networks
  - Allow people to virtually gather
  - Allow people to interact with each other
  - Allow people to share personal contents

- Posts
  - Represent a quick way to convey information
  - Can reach lots of people simultaneously



Simon Fairhurst
@siimonfairhurst

Lorem ipsum dolor sit amet.
Aut tentetur molestiae provident.

1:27PM · Oct 4 2022 · Twitter for iPhone

1.1M 1,240 5,579 3,987

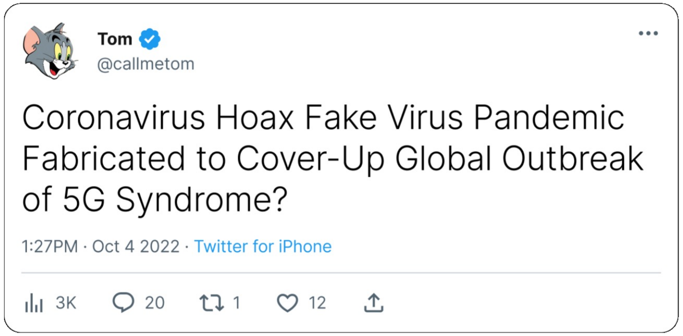
- Policies (Terms of servicies)
  - Represent the law on social networks
  - Indicate what type of content a user is allowed to publish

Policy compliance

From Twitter’s COVID-19 Misinformation Policy

“[...] We will require the deletion of Tweets that contain, for example: False claims about COVID-19 that invoke a deliberate conspiracy by malicious and/or powerful forces, such as: The pandemic is a hoax, or part of a deliberate attempt at population control, or that 5G wireless technology is causing COVID-19.”

Violates the policy

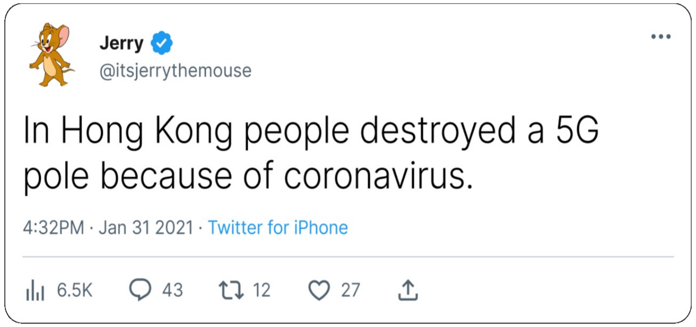


Tom
@callmetom

Coronavirus Hoax Fake Virus Pandemic Fabricated to Cover-Up Global Outbreak of 5G Syndrome?

1:27PM · Oct 4 2022 · Twitter for iPhone

Does not violate the policy

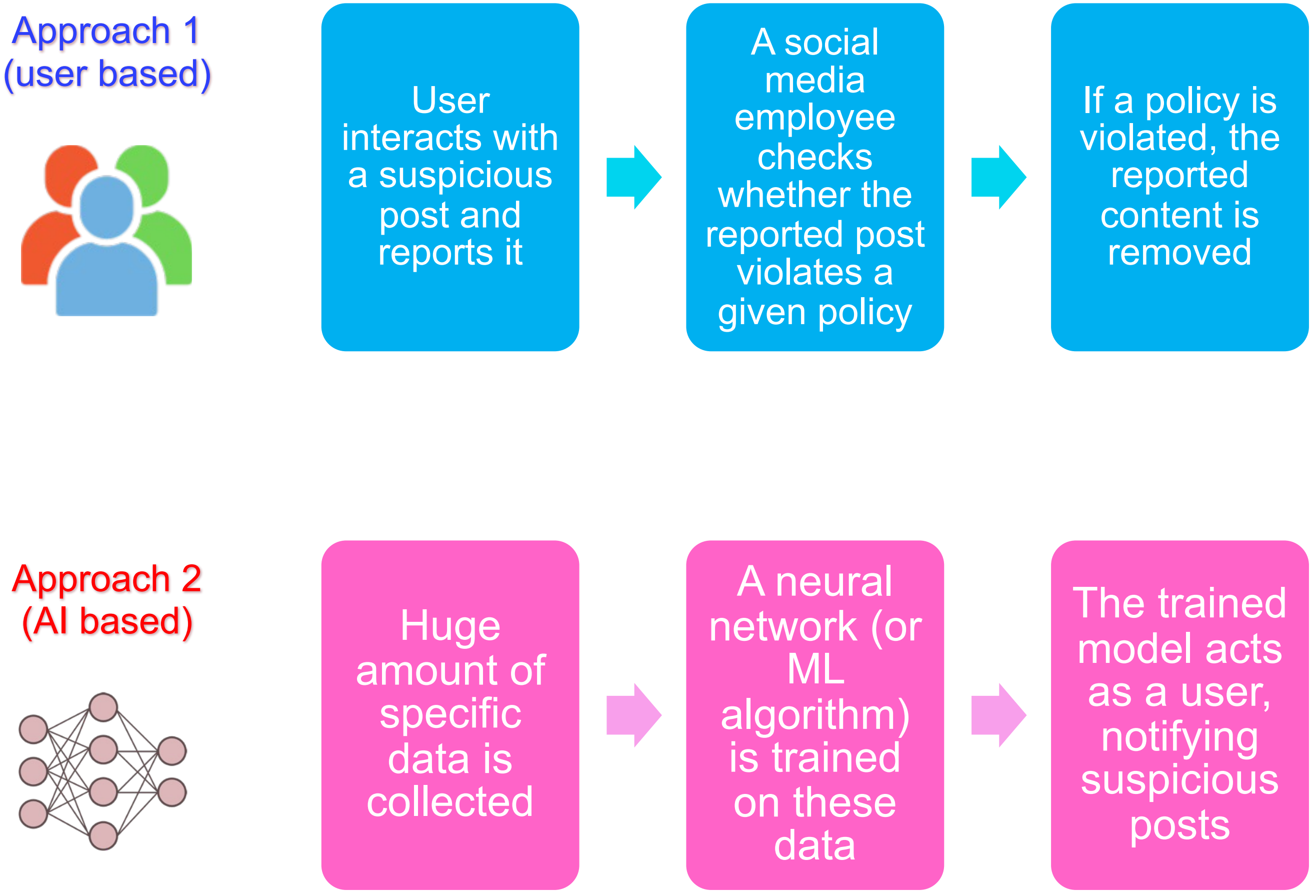


Jerry
@itsjerrythemouse

In Hong Kong people destroyed a 5G pole because of coronavirus.

4:32PM · Jan 31 2021 · Twitter for iPhone

Policy checking today



Approach 1 (user based)

User interacts with a suspicious post and reports it

A social media employee checks whether the reported post violates a given policy

If a policy is violated, the reported content is removed

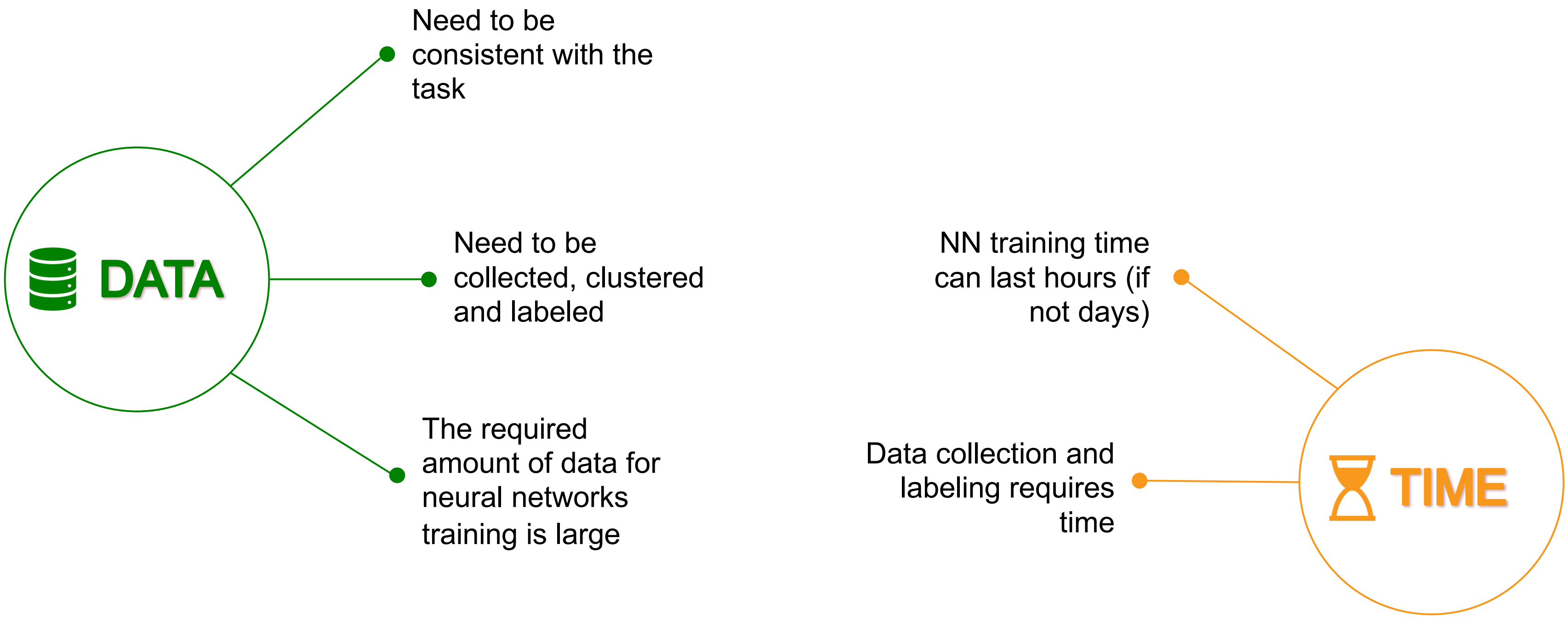
Approach 2 (AI based)

Huge amount of specific data is collected

A neural network (or ML algorithm) is trained on these data

The trained model acts as a user, notifying suspicious posts

Problems with the AI-based solution



DATA

Need to be consistent with the task

Need to be collected, clustered and labeled

The required amount of data for neural networks training is large

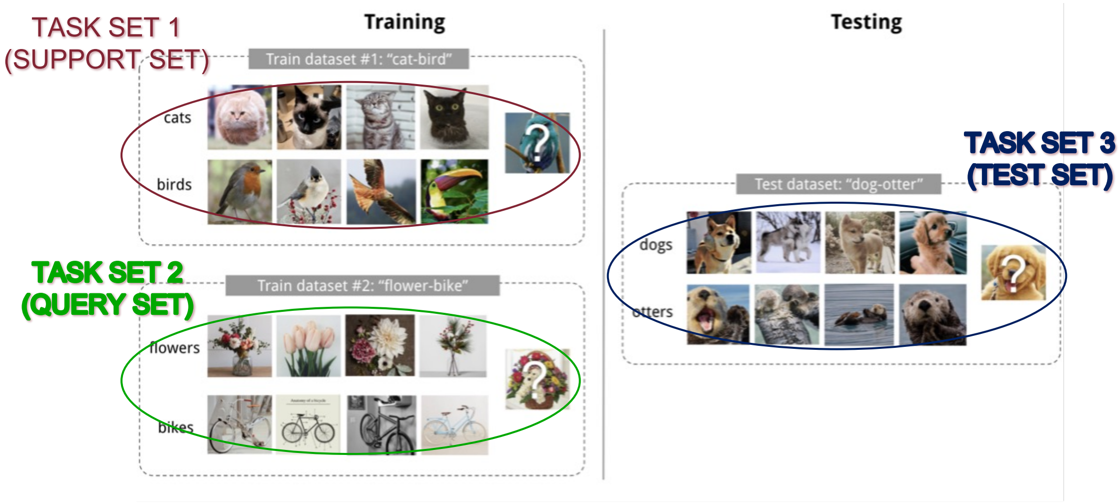
TIME

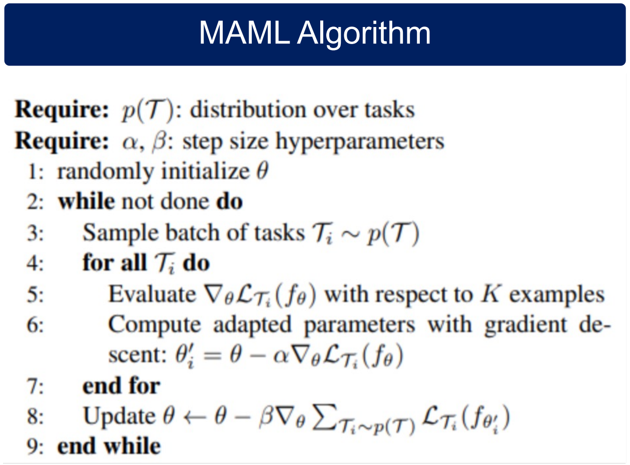
NN training time can last hours (if not days)

Data collection and labeling requires time

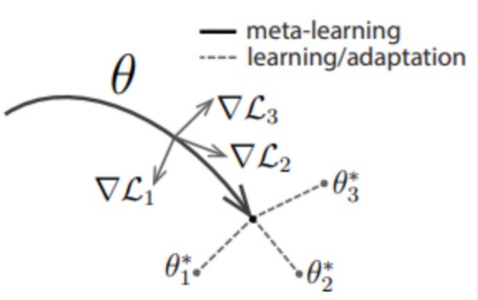
NOTICE: Every time a social media company releases a new policy, it has to face all these problems again

Meta learning




MAML Algorithm

```
Require: $p(\mathcal{T})$: distribution over tasks
Require: $\alpha, \beta$: step size hyperparameters
randomly initialize $\theta$
while not done do
    Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
    for all $\mathcal{T}_i$ do
        Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ with respect to $K$ examples
        Compute adapted parameters with gradient descent: $\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
    end for
    Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$
end while
```

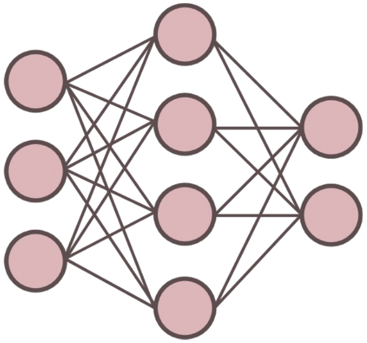


Libraries and methodologies

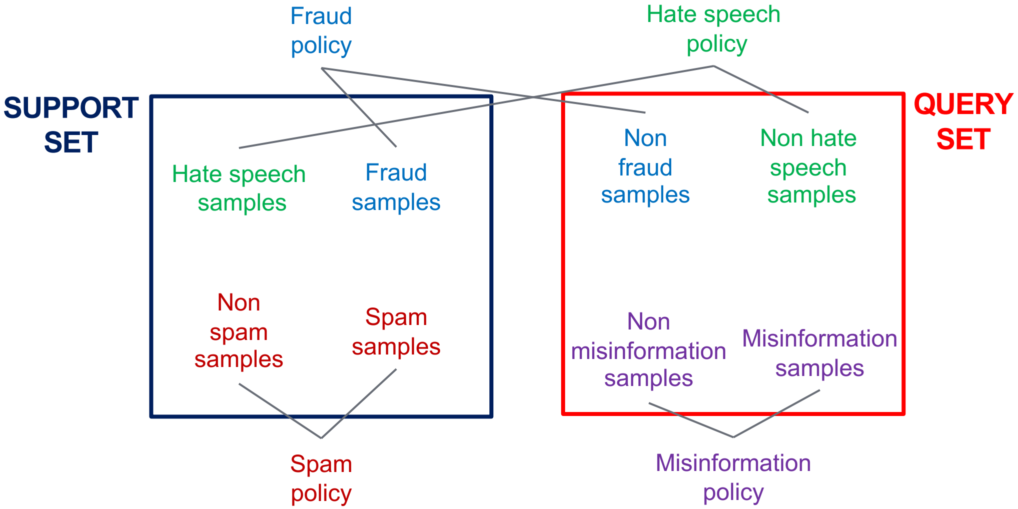
Use PyTorch to realize the NN

Use learn2learn to create two tasksets and the MAML envelope

PyTorch



learn2learn

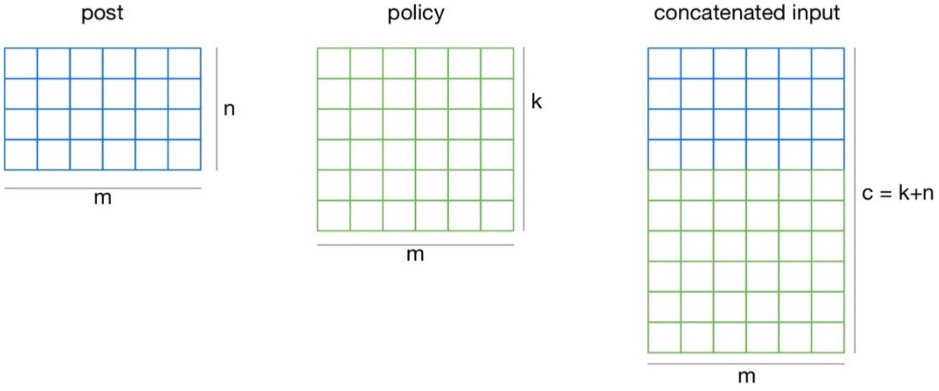


SUPPORT SET

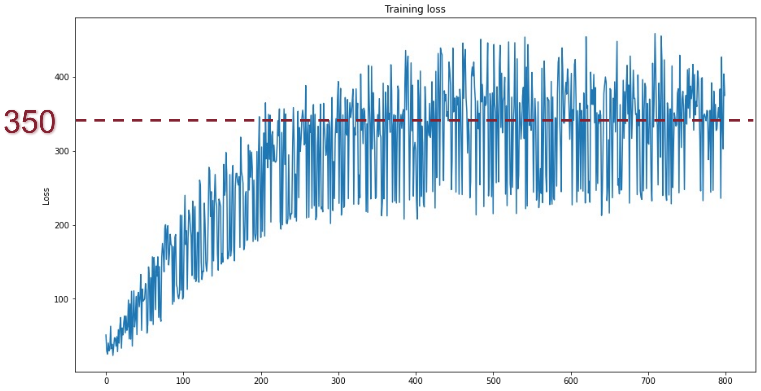
QUERY SET

Combine posts with their policies: concatenation

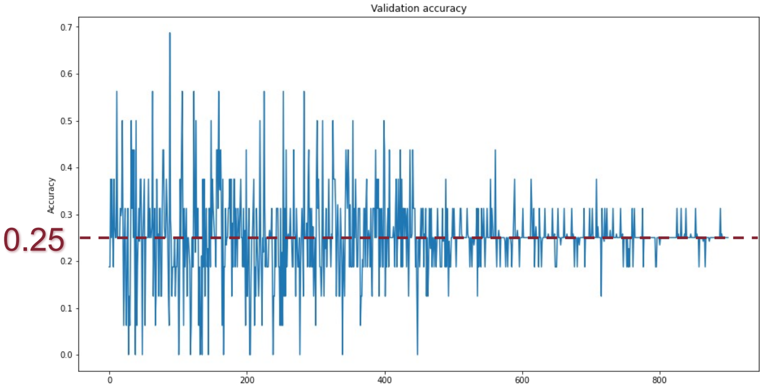
Post and policy concatenation



Support Set Loss



Query Set Accuracy

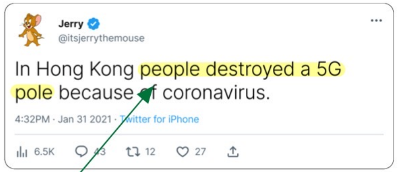


Using the attention mechanism

Attention with Twitter's COVID-19 Misinformation Policy

From Twitter's COVID-19 Misinformation Policy

"[...] We will require the deletion of Tweets that contain, for example: False claims about COVID-19 that invoke a deliberate conspiracy by malicious and/or powerful forces, such as: The pandemic is a hoax, or part of a deliberate attempt at population control, or that 5G wireless technology is causing COVID-19."

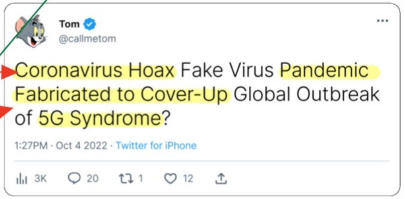


Jerry
@itsjerrythemouse

In Hong Kong people destroyed a 5G pole because of coronavirus.

4:32PM · Jan 31 2021 · Twitter for iPhone

6.5K 43 12 27



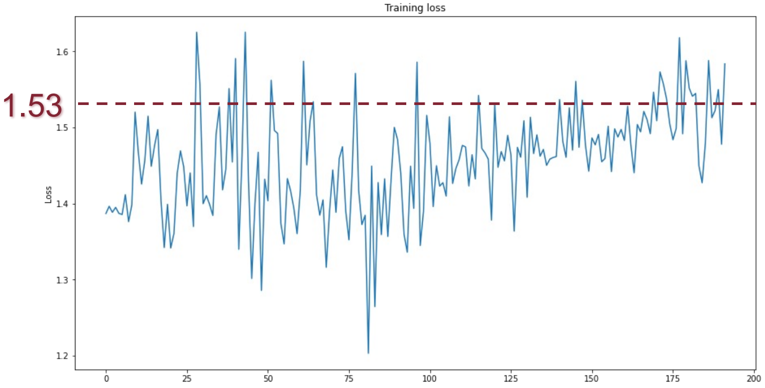
Tom
@callmetom

Coronavirus Hoax Fake Virus Pandemic Fabricated to Cover-Up Global Outbreak of 5G Syndrome?

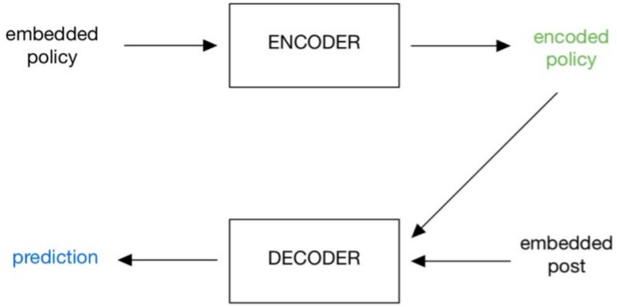
1:27PM · Oct 4 2022 · Twitter for iPhone

3K 20 1 12

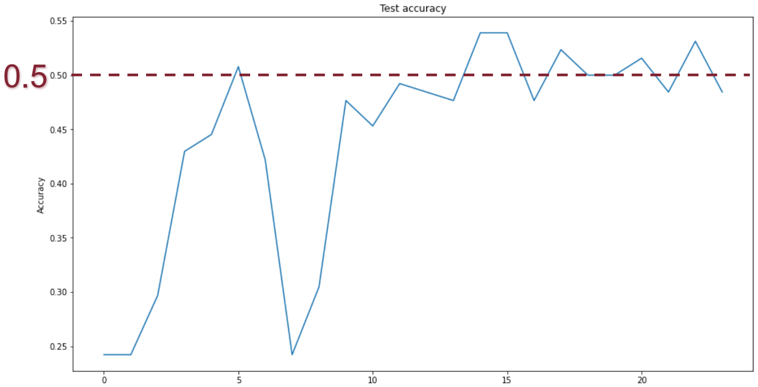
Support Set Loss



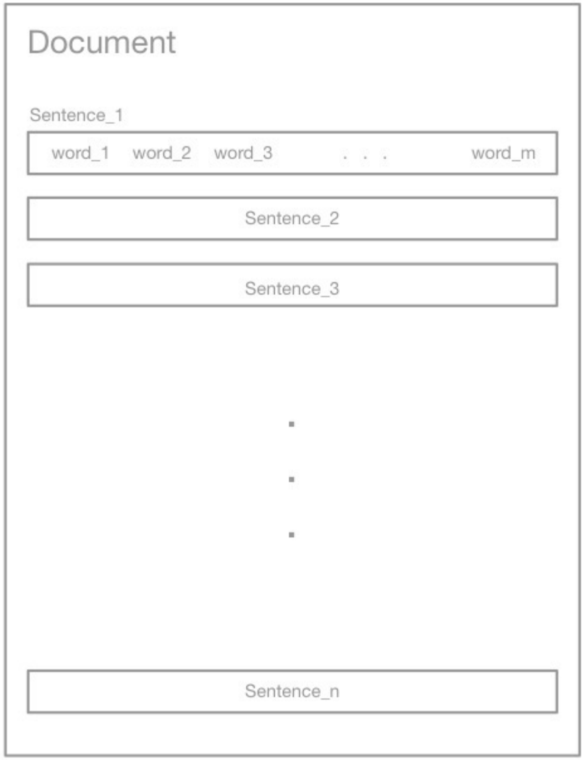
The Encoder-Decoder architecture



Query Set Accuracy



Using SBERT


Document

Sentence_1

word_1 word_2 word_3 . . . word_m

Sentence_2

Sentence_3

.

.

.

Sentence_n

(*)RoBERTa embedding tensor of the document will have size:

768 x m x n

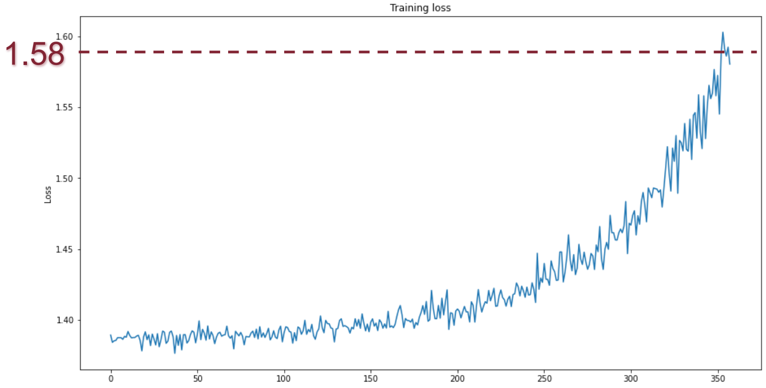
RoBERTa output size #words in a sentence #sentences

(*)SBERT embedding tensor of the document will have size:
384 x n

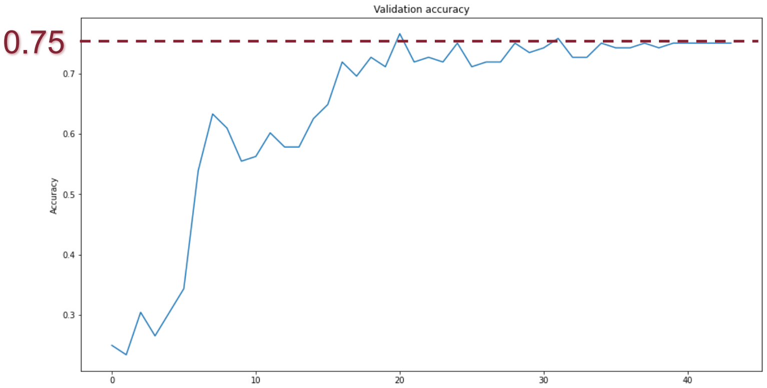
384 x n

SBERT output size #sentences

Support Set Loss

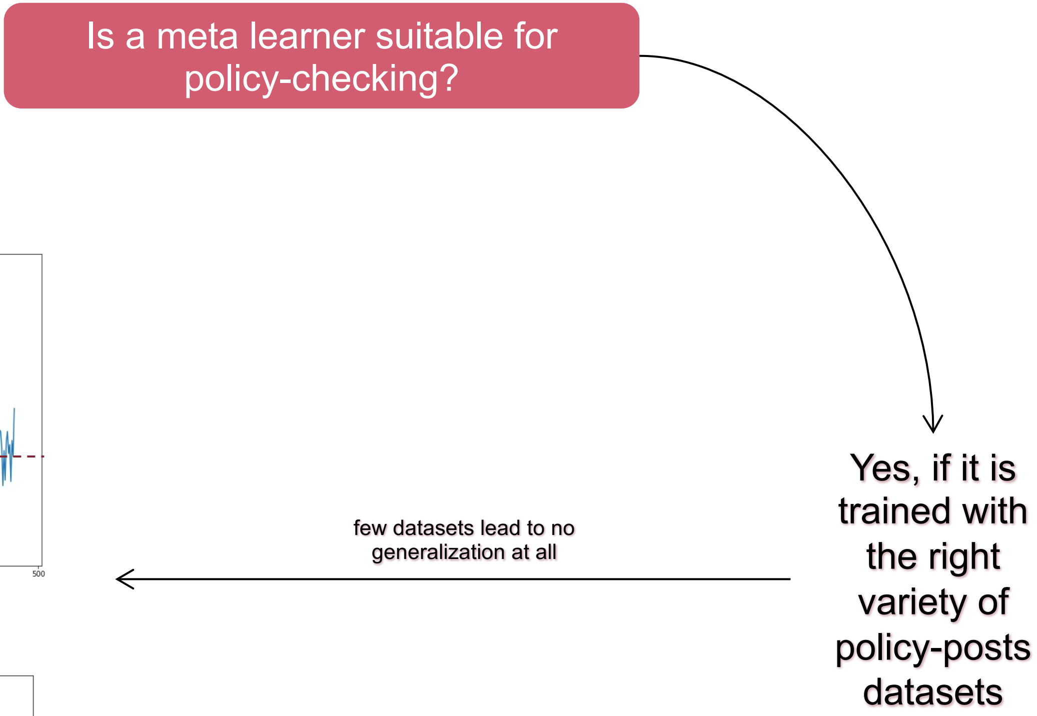


Query Set Accuracy



(*)We assume the document to have all sentences containing m words

Conclusion and future work

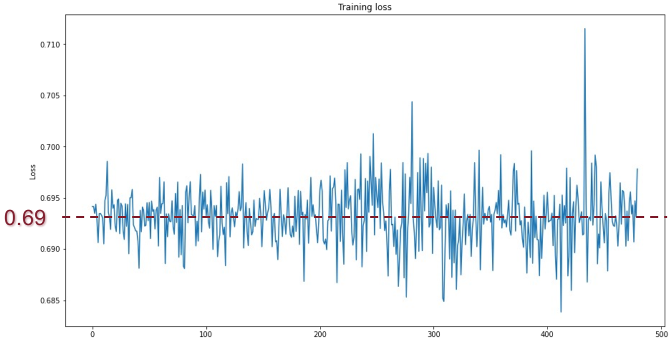


Is a meta learner suitable for policy-checking?

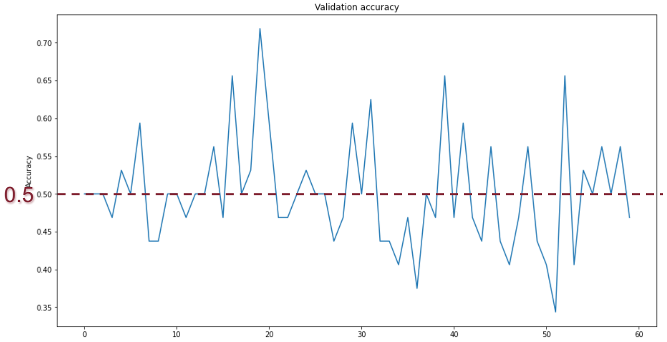
Yes, if it is trained with the right variety of policy-posts datasets

few datasets lead to no generalization at all

(hate-speech) support set loss



(spam) query set accuracy



NOTE: Variety does not imply quantity