IFAN: An Explainability-Focused Interaction Framework for Humans and NLP Models

Edoardo Mosca, Daryna Dementieva, Tohid E. Ajdari,

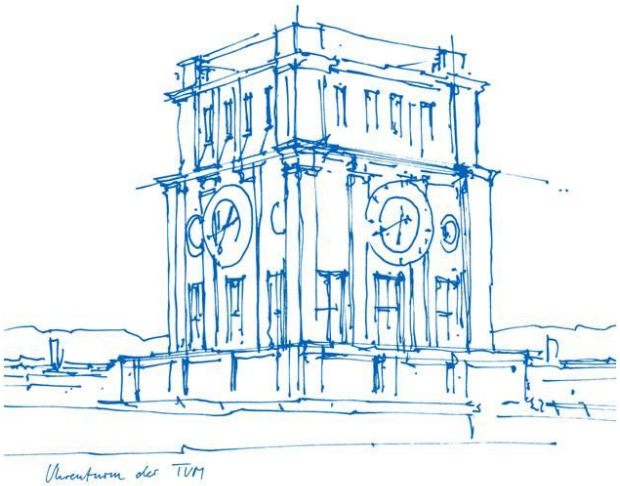
Maximilian Kummeth, Kirill Gringauz,

Yutong Zhou, Georg Groh.

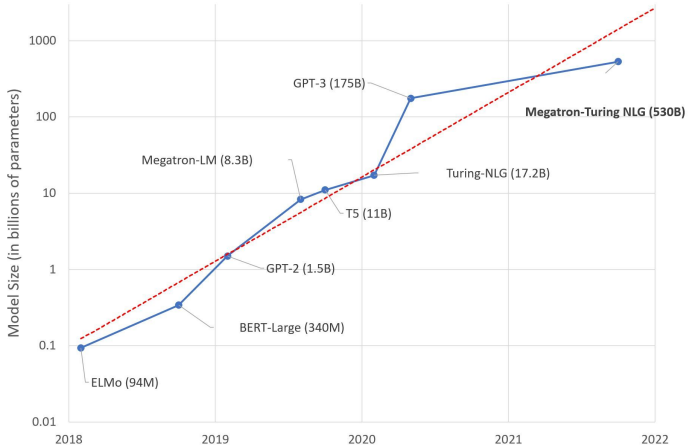
Technical University of Munich, Germany

AACL 2023

1st-4th October | Bali



The Current Stand

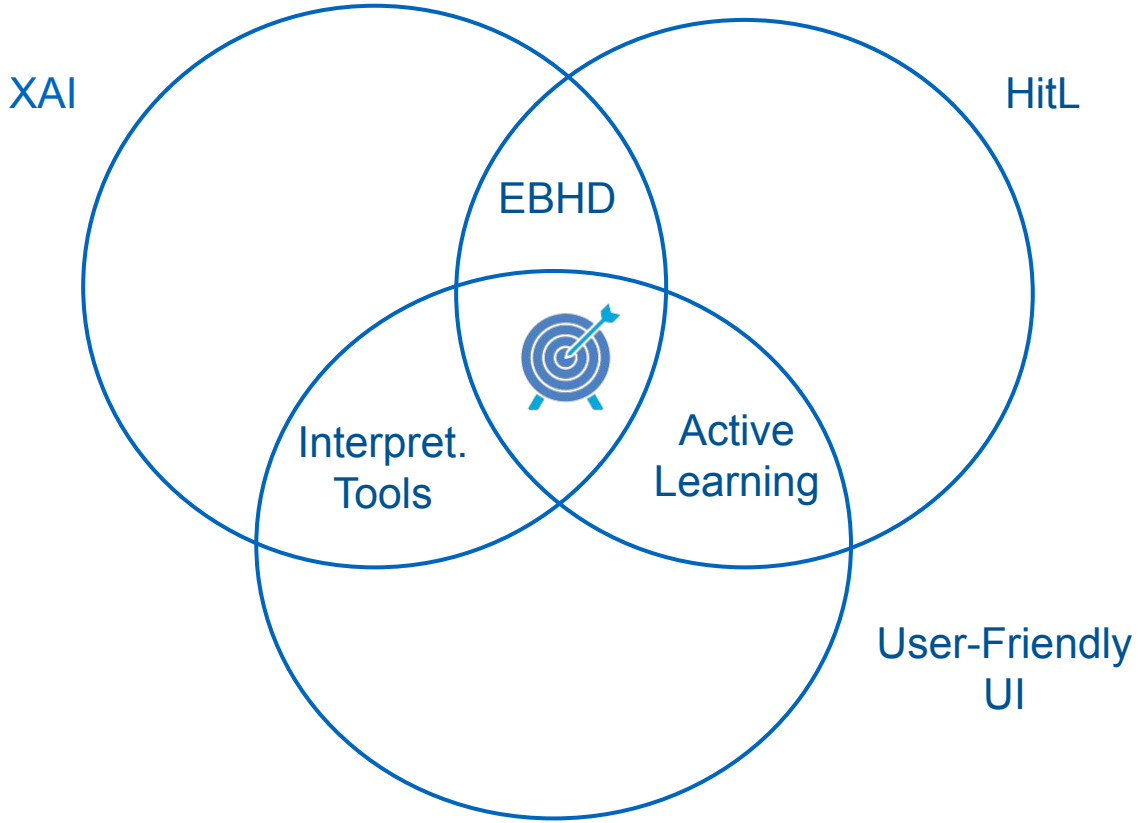


Performance

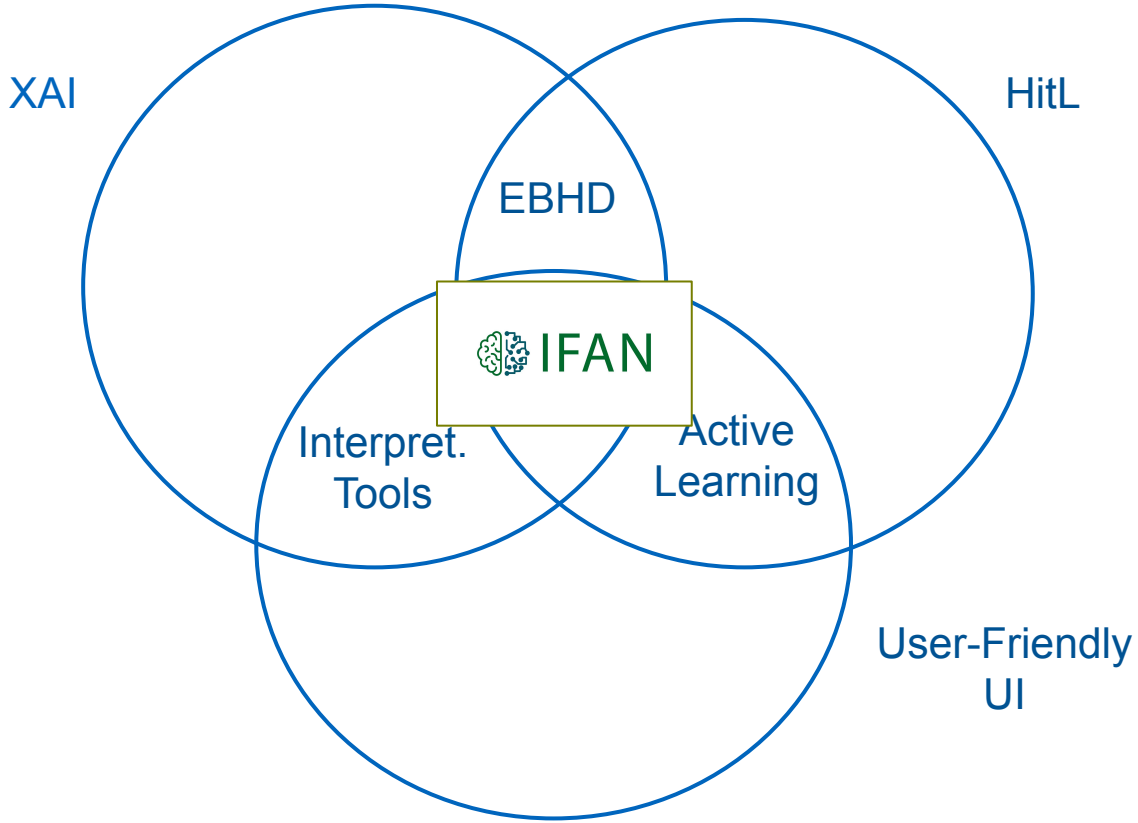
Interpretability

Controllability

The Current Stand



The Current Stand



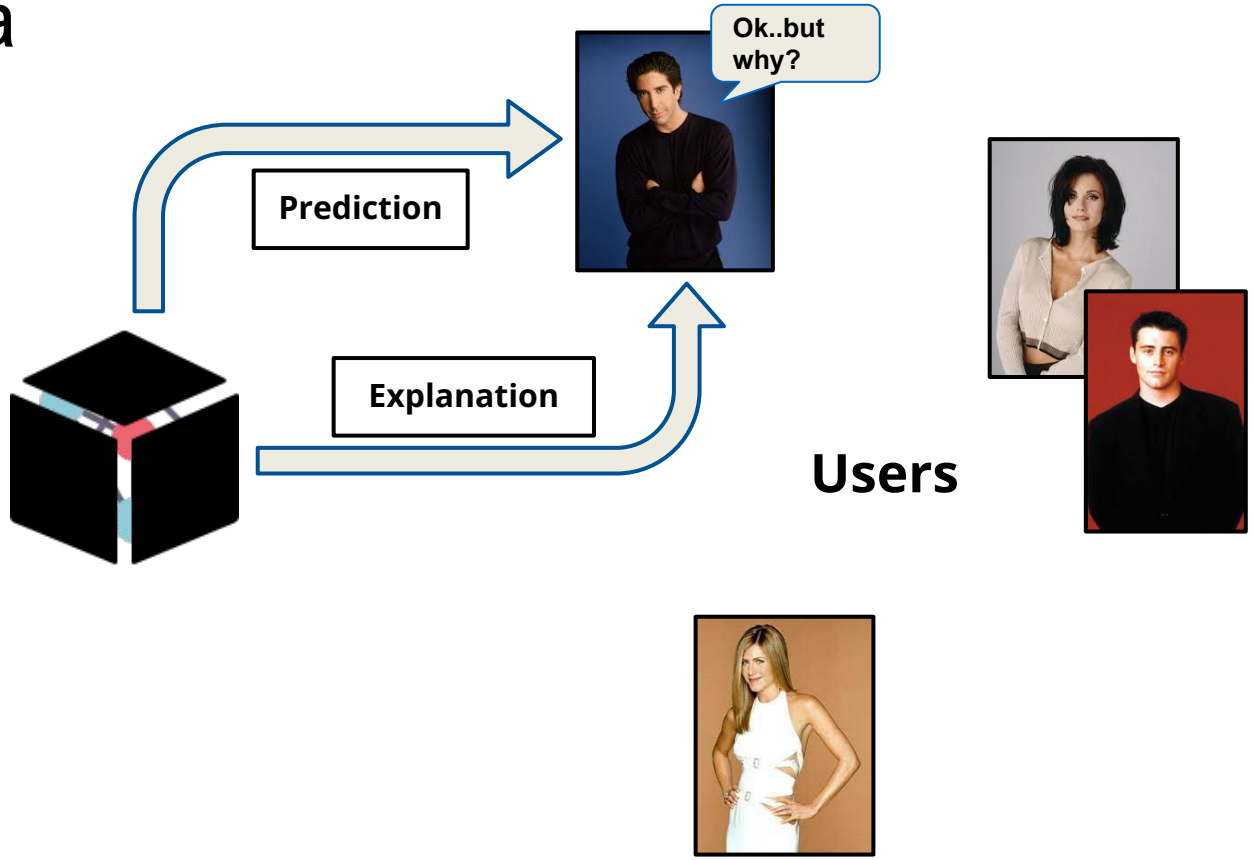
Our contribution

1. **Propose IFAN: an EBHD Framework for NLP Models**
   Users can observe explanations, edit the rationale, give feedback, etc.

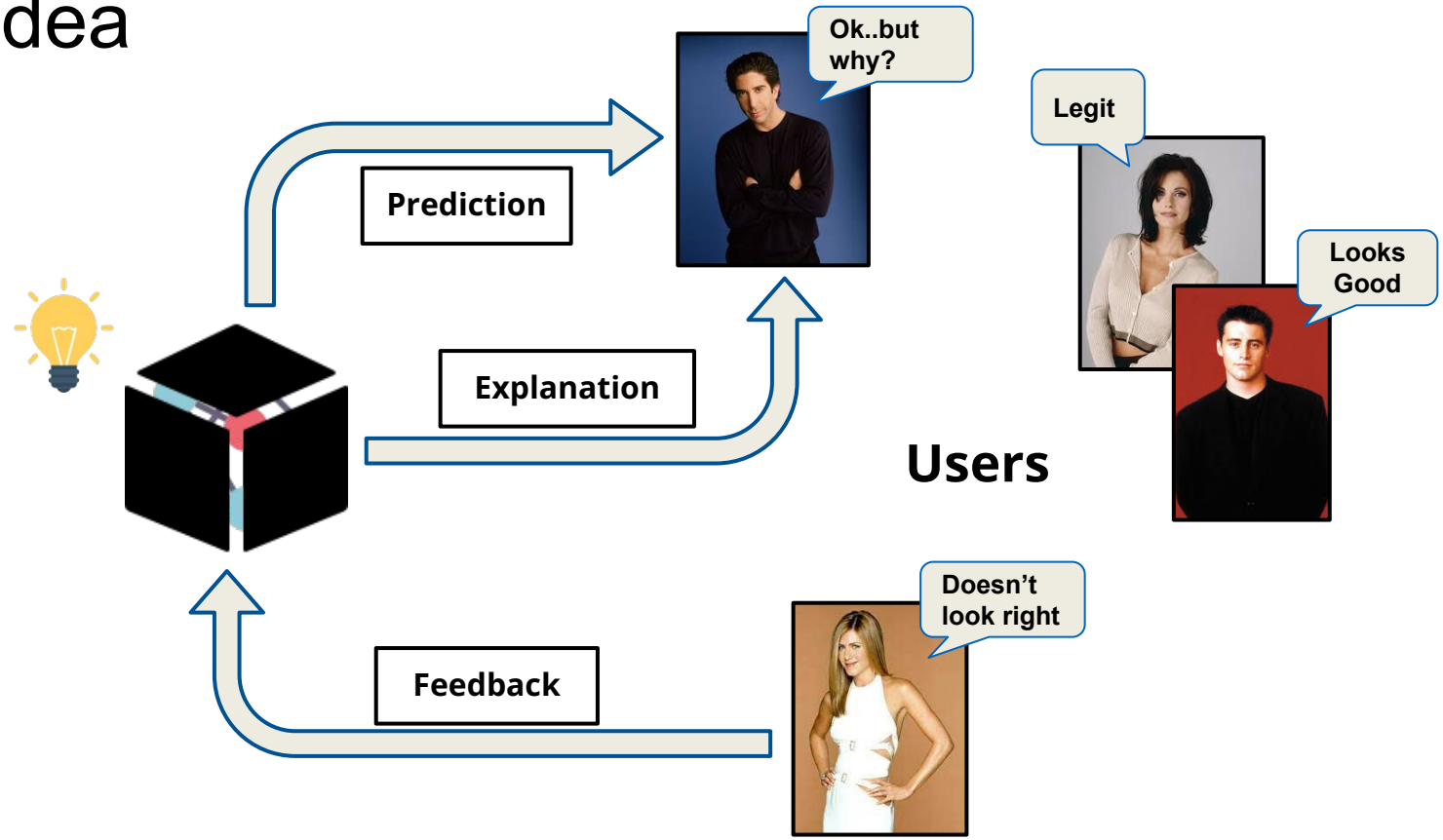
2. **Extend it with a UI and a Management Systems**
   Monitor improvement, configure models and user access, etc.

3. **Test IFAN on a Model Debiasing Task**
   We propose feedback rebalancing to contrast model forgetfulness.

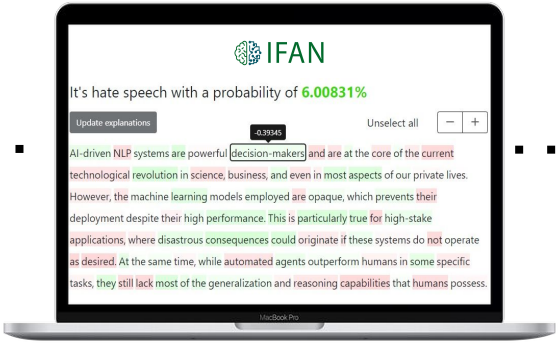
The Idea

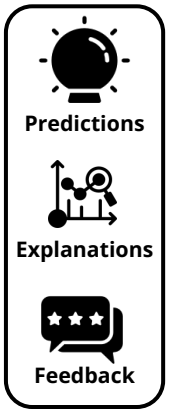


The Idea

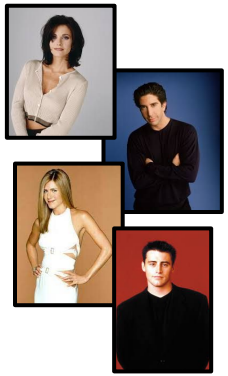

Ok..but why?
Prediction
Explanation
Users
Legit
Looks Good
Feedback
Doesn't look right

The Final Prototype





Users



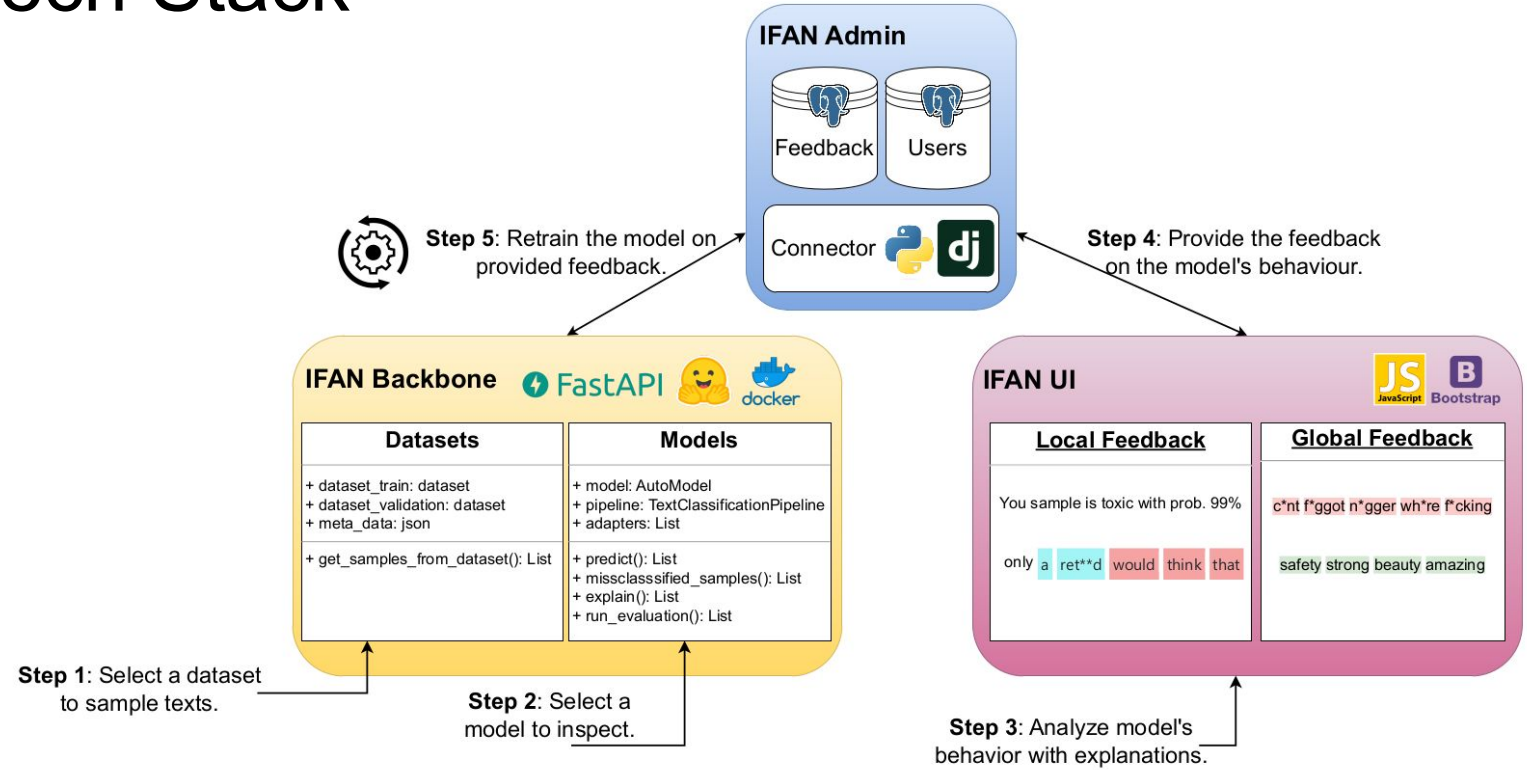
Manage:
- Models
- Datasets
- Users (Rights)

Reports:
- Analyze
- Test
- Monitor

Endpoints:
- Deploy
- Experiment at Scale

Devs

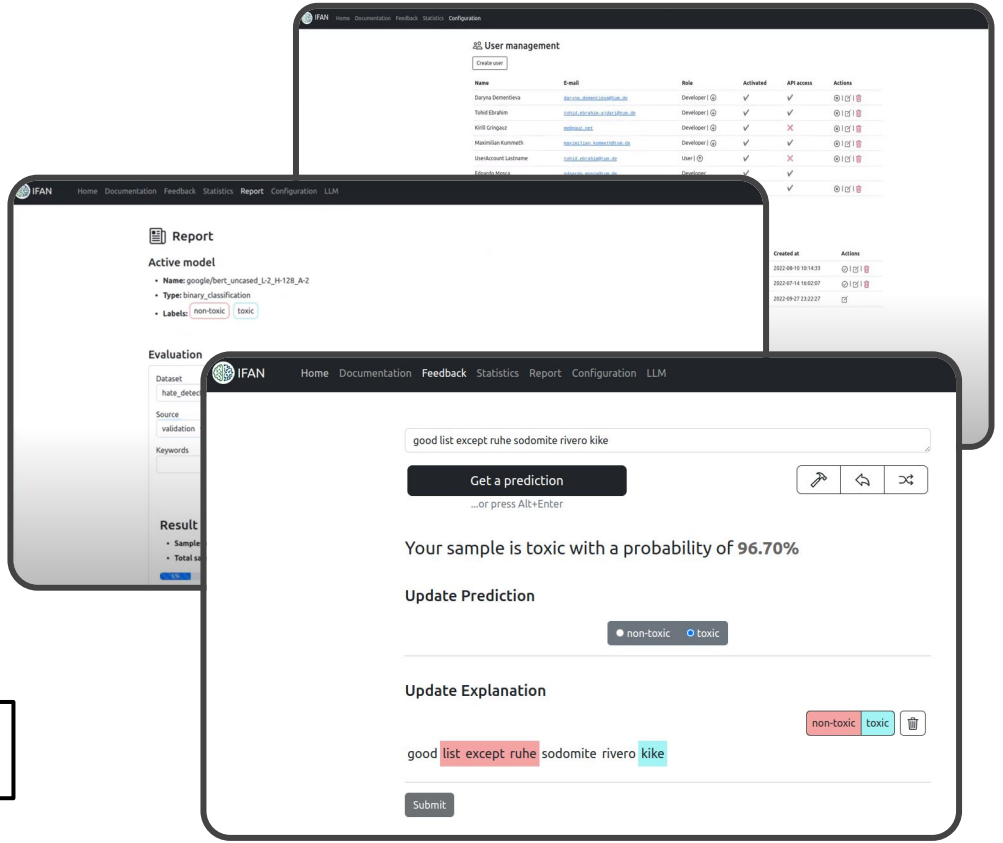
Tech Stack


IFAN Admin
Feedback
Users
Connector
dj
Step 5: Retrain the model on provided feedback.
Step 4: Provide the feedback on the model's behaviour.
IFAN Backbone
FastAPI
docker
Datasets
+ dataset_train: dataset
+ dataset_validation: dataset
+ meta_data: json
+ get_samples_from_dataset(): List
Models
+ model: AutoModel
+ pipeline: TextClassificationPipeline
+ adapters: List
+ predict(): List
+ missclasssified_samples(): List
+ explain(): List
+ run_evaluation(): List
IFAN UI
JS
JavaScript
B
Bootstrap
Local Feedback
You sample is toxic with prob. 99%
only a ret**d would think that
Global Feedback
c*nt f*ggot n*gger wh*re f*cking
safety strong beauty amazing
Step 1: Select a dataset to sample texts.
Step 2: Select a model to inspect.
Step 3: Analyze model's behavior with explanations.

METHODOLOGY

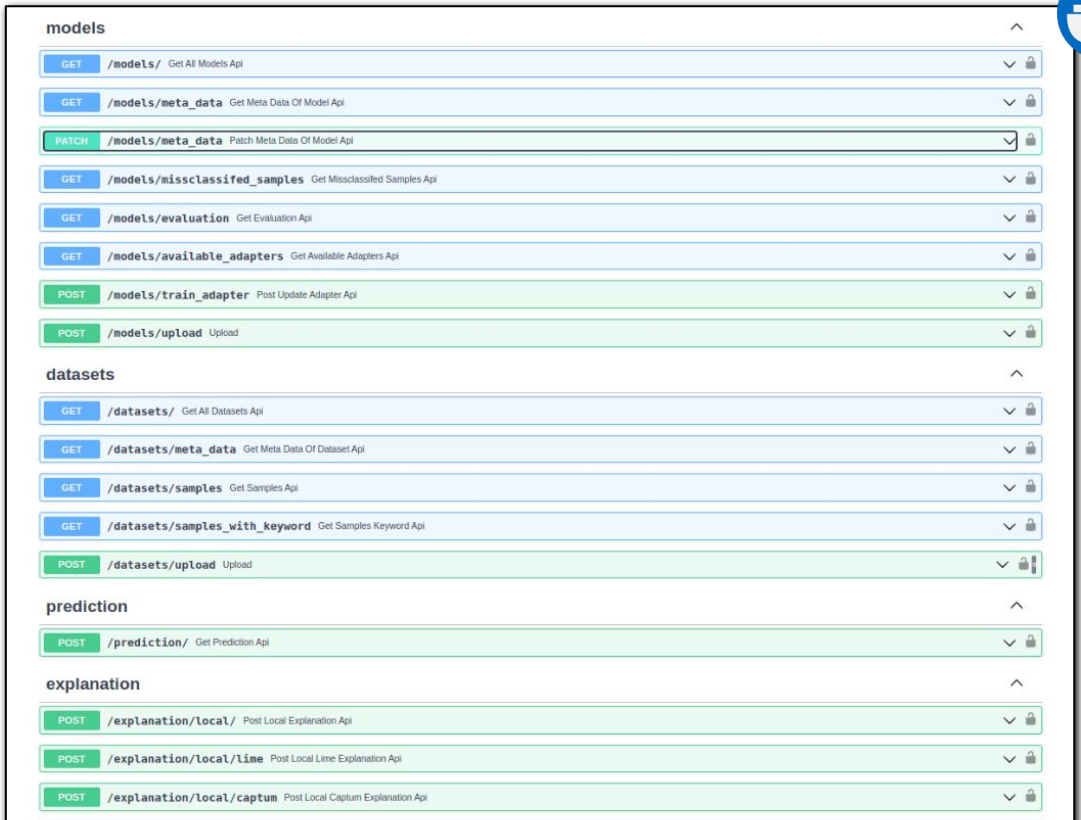
Interface

- Home
- Documentation
- Feedback
- Report
- Configuration



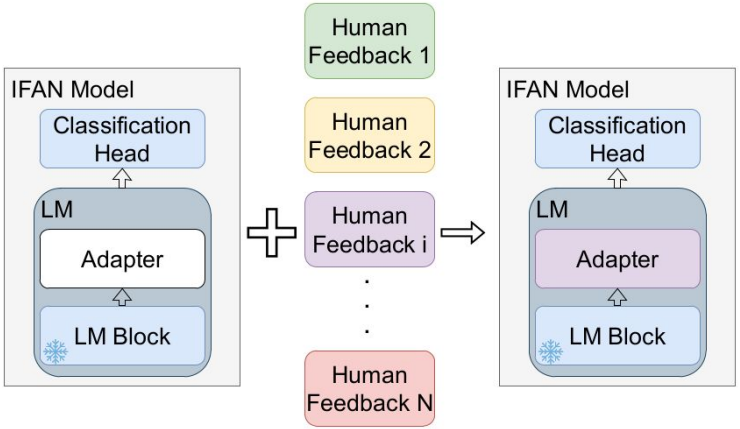
[To the Demo Video]

Backbone & API



User Roles & Feedback Mechanism

	Dev	Annotator	Unauthorized
Classification & Explanations	✔	✔	✔
Smart Samples Selection	✔	✔	✘
Feedback	✔	✔	✘
Active Configuration	✔	✘	✘
Model Report & Miscl. Samples	✔	✘	✘
New Models & Datasets Upload	✔	✘	✘
New Users Creation	✔	✘	✘



Adapters are trained with batches of human-highlighted sample segments.

RESULTS

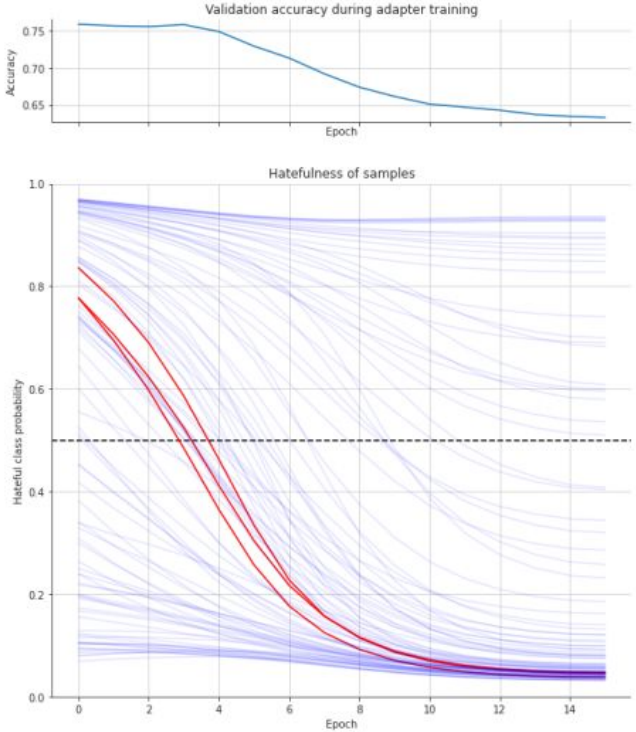
Models love to Forget

- Adapters learn very quickly from feedback (especially on strong signals).
- Models tend to forget what they know, with large drops in performance.
- Adding original samples to feedback batches mitigates the problem (**Rebalancing**).

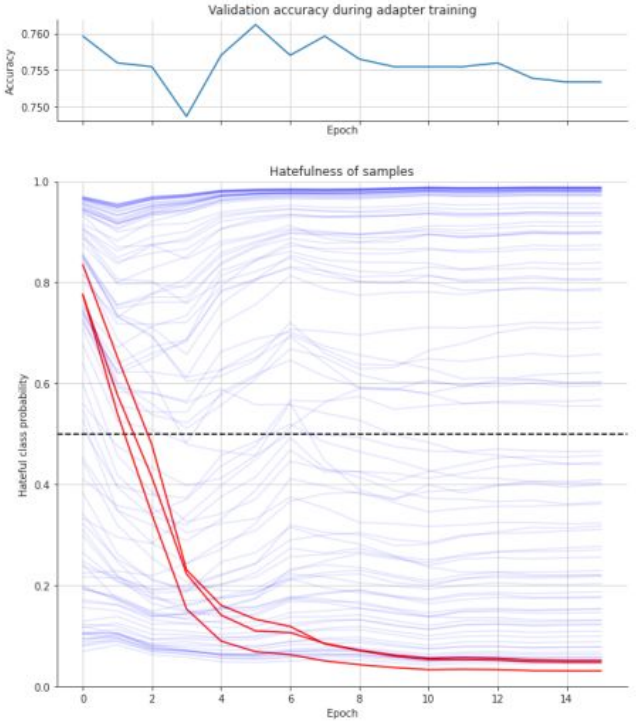
Performance on Use Case

Model	Pr	Re	F1	Pr_J
BERT (baseline)	0.80	**0.78**	**0.79**	0.95
Most Confident Missclassified				
BERT+Feedback (non-bal.)	0.34	0.28	0.31	0.82
BERT+Feedback (bal.)	0.78	0.80	**0.79**	**0.97**
Least Confident Missclassified				
BERT+Feedback (non-bal.)	**0.83**	0.73	0.78	0.96
BERT+Feedback (bal.)	0.79	**0.78**	0.78	0.96

Effects of Rebalancing



Without Rebalancing



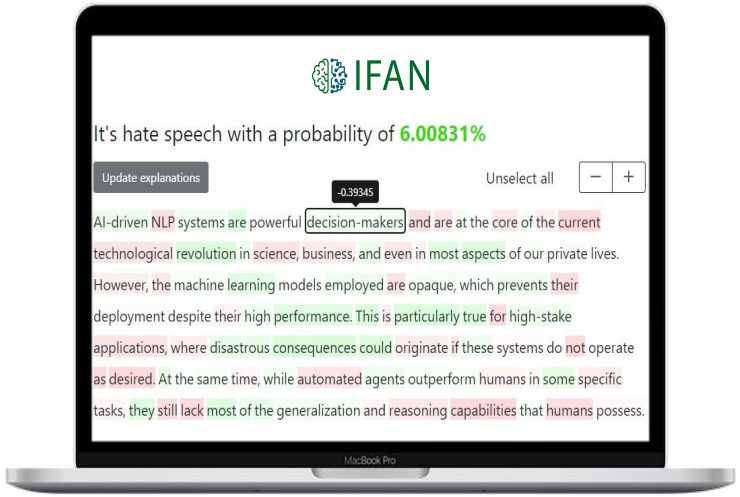
With Rebalancing

CONCLUSIONS

Limitations & Takeaways

- Editing explanations can carry human feedback which is beneficial for fixing NLP models.
- As of now we support sentence-to-class and token-to-class tasks (other tasks are work in progress)
- No clear optimal choices of hyperparameters for feedback
- Dealing with a large spectrum of models is hard, but is becoming easier.
- In most cases, your model won't get much better in performance, but is more aligned with humans where it received feedback.
- Rebalancing allows integrating feedback with minimal performance loss.

Thank you!!



Paper

Demo Video

Thank you!!

Daryna Dementieva

Tohid E. Ajdari,

Maximilian Kummeth

Kirill Gringauz

Yutong Zhou

Georg Groh

Edoardo Mosca
edoardo.mosca@tum.de

IFAN