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

Edoardo Mosca, Daryna Dementieva, Tohid Ebrahim Ajdari, Maximilian Kummeth, Kirill Gringauz, Yutong Zhou and **Georg Groh**
TU Munich, Department of Informatics, Germany
{name.surname}@tum.de
grohg@in.tum.de

Abstract

Interpretability and human oversight are fundamental pillars of deploying complex NLP models into real-world applications. However, applying explainability and human-in-the-loop methods requires technical proficiency. Despite existing toolkits for model understanding and analysis, options to integrate human feedback are still limited. We propose IFAN, a framework for real-time explanation-based interaction with NLP models. Through IFAN's interface, users can provide feedback to selected model explanations, which is then integrated through adapter layers to align the model with human rationale. We show the system to be effective in debiasing a hate speech classifier with minimal impact on performance. IFAN also offers a visual admin system and API to manage models (and datasets) as well as control access rights. A demo is live at ifan.ml.

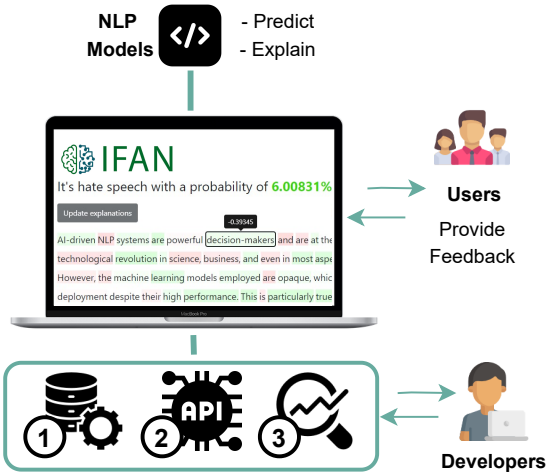


Figure 1: IFAN in brief. The interface allows NLP models and users to interact through predictions, explanations, and feedback. IFAN also provides developers with (1) a manager for models and datasets, (2) model API access, and (3) reports about the model.

1 Introduction

As *Natural Language Processing* (NLP) systems continue to improve in performance, they are increasingly adopted in real-world applications (Khurana et al., 2022). *Large Language Models* (LLMs)—such as GPT-3 (Brown et al., 2020), BLOOM (Scao et al., 2022a), and T5 (Raffel et al., 2020)—are without a shred of doubt the main protagonists of recent advances in the field. They are able to substantially outperform previous solutions while being directly applicable to any NLP task.

There are however strong concerns given the black-box nature of such architectures (Madsen et al., 2022; Mosca et al., 2022a). In fact, their large scale and high complexity are substantial drawbacks in terms of *transparency*, *accountability*, and *human oversight*. Beyond ethical considerations, even legal guidelines from the European Union are now explicitly defining these interpretability factors as essential for any deployed AI system (European Commission, 2020).

Research efforts in *eXplainable Artificial Intelligence* (XAI) (Arrieta et al., 2020; Mosca et al., 2022b) and *Human-in-the-Loop* (HitL) machine learning (Monarch, 2021) have thus been on the rise—producing solutions that aim at mitigating the current lack of interpretability. Most notably, the recent literature contains a number of toolkits and frameworks to analyze, understand, and improve complex NLP models (Wallace et al., 2019; Liu et al., 2021). Some of them even offer low-code interfaces for stakeholders who do not possess the otherwise required technical proficiency. Nonetheless, current options to collect human rationale and provide it as feedback to the model are still limited.

We propose IFAN, a novel low-to-no-code framework to interact in real time with NLP models via explanations. Our contribution can be summarized as follows:

(1) IFAN offers an interface for users to provide feedback to selected model explana-

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations, pages 59–76
November 1–4, 2023.

tions, which is then integrated via parameter-efficient adapter layers.

(2) Our live platform also offers a visual administration system and API to manage models, datasets, and users as well as their corresponding access rights.

(3) We show the efficiency of our framework in debiasing a hate speech classifier and propose a feedback-rebalancing step to mitigate the model's forgetfulness across updates.

IFAN's demo is accessible at ifan.ml[1] together with its documentation.[2] Full access is available with login credentials, which we can provide upon request. A supplementary video showcase can be found online[3].

2 Related Work

2.1 HitL with Model Explanations

Human-in-the-Loop (HitL) machine learning studies how models can be continuously improved with human feedback (Monarch, 2021). While a large part of the HitL literature deals with label-focused feedback such as *active learning*, more recent works explore how explanations can be leveraged to provide more detailed human rationale (Lertvittayakumjorn and Toni, 2021).

Combining classical HitL (Wang et al., 2021) with explanations to construct human feedback for the model (Han et al., 2020) has been referred to as *Explanation-Based Human Debugging* (EBHD) (Lertvittayakumjorn and Toni, 2021). Good examples are Ray et al. (2019), Selvaraju et al. (2019), and Strout et al. (2019), which show improvements in performance and interpretability when iteratively providing models with human rationale.

A more NLP-focused EBHD approach is Yao et al. (2021), where the authors leverage explanations to debug and refine two transformer instances—BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). Concretely, word saliency explanations at different levels of granularity are provided to humans, who in turn provide suggestions in the form of natural language. The annotator's feedback is converted into first-order logic rules, which are later utilized to condition learning with new samples.

[1] https://ifan.ml
[2] https://ifan.ml/documentation
[3] https://youtu.be/EzC6HI3JwaQ

2.2 Interactive NLP Analysis Platforms

In the recent literature, we can find strong contributions in terms of software and digital toolkits to analyze and explain NLP models (Wallace et al., 2019; Hoover et al., 2020) as well as further refining them via parameter-efficient fine-tuning (Beck et al., 2022).

For instance, Liu et al. (2021) proposes EXPLAINABOARD, an interactive explainability-focused leaderboard for NLP models. More in detail, it allows researchers to run diagnostics about the strengths and weaknesses of a given model, compare different architectures, and closely analyze predictions as well as recurring model mistakes. Similarly, the LANGUAGE INTERPRETABILITY TOOL by Tenney et al. (2020) is an open-source platform and API to visualize and understand NLP models. In particular, it provides a browser-based interface integrating local explanations as well as counterfactual examples to enable model interpretability and error analysis.

Finally, Beck et al. (2022) releases ADAPTERHUB PLAYGROUND, a no-code platform to few-shot learning with language models. Specifically, the authors built an intuitive interface where users can easily perform predictions and training of complex NLP models on several natural language tasks.

3 IFAN

The **I**nteraction **F**ramework for **A**rtificial and **N**atural Intelligence (**IFAN**) is a web-based platform for inspecting and controlling text processing models. Its main goal is to decrease the opacity of NLP systems and integrate explanation-based HitL into their development pipeline. Through our interface, stakeholders can test and explain models' behavior and—when encountering anomalies in predictions or explanations—they can fix them onsite by providing feedback.

The main blocks of the platform are presented in Figure 2. The **Backbone** part contains all machine learning development components—datasets and models. We adopt HuggingFace formats (see 3.3 and 3.4) (Wolf et al., 2020) and wrap the entire backbone as a Docker[4] image for deployment. The **User Interface** is the visual component of the platform, where all the human-machine interaction takes place. Here, developers have also access to additional visual resources to configure details about models, datasets, and users.

[4] https://www.docker.com

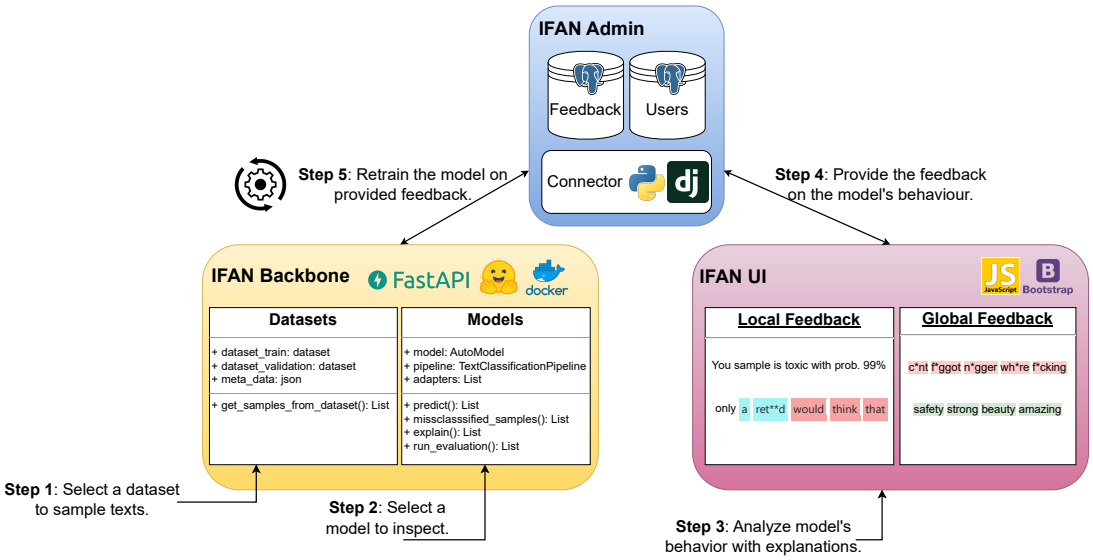


Figure 2: Overall schema of IFAN idea: (i) The user selects a dataset or writes a customized input. (ii) Then the user can select a model which should be inspected. (iii) With the UI, annotators can check the model's prediction on a sample and two types of explanations – local and global. (iv) If there is some misbehavior, the annotators can provide feedback. (v) The feedback is stored and then used to fine-tune the model.

The connection between the backbone and the user interface is managed by the **Admin** component. All the user data and rights as well as samples receiving feedback are stored in a PostgreSQL[5] database instance. The communication is handled via Python Django[6], which integrates everything w.r.t. user authentication, API calls/responses, state logs, and location of backbone resources. In the next sections, we provide a more detailed description of the main platform components.

3.1 User Interface

Our frontend is built with Bootstrap[7] and JavaScript[8]. Currently, the pages available in our UI are the following:

Landing Page Here users can get a short introduction to IFAN. We briefly explain our platform's goals, the concept of HitL, and how our framework can be integrated into the development of NLP models.

Documentation It provides a detailed description of all the UI components together with screenshots and guidelines. Here, users can find specific instructions on how to configure and interact with our platform.

Feedback This is the main interaction page. Here, users can run a model on an input sample either taken from the dataset or that they wrote themselves. Then, they can load the model's prediction and explanations and provide feedback in terms of both the label and features' relevance.

Report This page has limited access (see 3.2). Developers can evaluate models before and after feedback incorporation on a chosen dataset as well as inspect misclassified samples.

Configuration This page has limited access (see 3.2). Here, developers can configure and manage the platform, More specifically, users can be created, modified, and deleted as well as upgraded or downgraded in their roles and access rights. Also, they can manage models and datasets as well as specify the currently active ones.

Account Settings Each authorized user can view, edit, export, and delete their account data (GDPR compliance) as well as reset their login password.

3.2 Users

The platform separates users in three tiers: *developers*, *annotators*, and *unauthorized* users (Table 1).

Unauthorized users do not possess login credentials and have limited access to the platform. They can visualize model predictions and explanations but their feedback is not considered.

[5] https://www.postgresql.org
[6] https://www.djangoproject.com
[7] https://getbootstrap.com
[8] https://www.javascript.com

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

Table 1: Different levels of access to IFAN.

Normal users (or annotators) are known through their credentials and can thus actively engage with the model. During a HitL iteration, they can use the feedback page with pre-configured datasets and models, test the model on a text sample, view explanations, and provide feedback if needed.

Developers have full access and can configure all aspects of the platform. More specifically, they have access to the *report* and *configuration* pages (see 3.1) and can thus manage anything regarding users, roles, API access, models, and datasets.

3.3 Datasets

Before the model's behavior exploration, the *active dataset* should be specified via the configuration page (see 3.1). This is the dataset from which the text examples for the model testing are sampled.

Dataset	**Short Description**
HateXplain (Mathew et al., 2021)	A dataset for hate speech classification which has 3 classes for hate type detection, the target community classification, and rationales.
GYAFC (Rao and Tetreault, 2018)	Formality detection dataset which corresponds to 2-class classification: formal and informal.

Table 2: Example of datasets tested at IFAN.

We conform to a standard format by using the HuggingFace Datasets library[9]. Developers interacting with our platform are strongly encouraged to adhere to this standard when uploading new datasets and making them available to the interface. Table 2 shows two examples of datasets already available on our platform.

3.4 Models

Analogous to datasets, the *active model* should be specified via the configuration page and should adhere to the HuggingFace Models standard.[10]

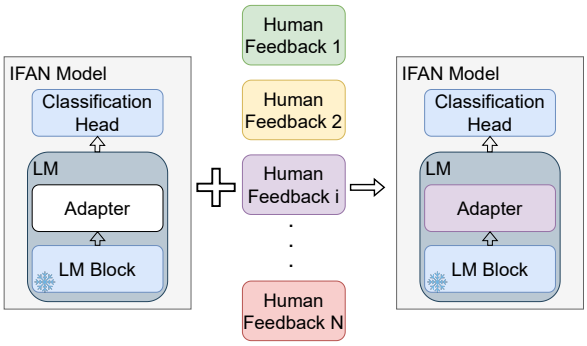


Figure 3: The proposed architecture for the models integrated into IFAN: addition of Adapter layer which is trainable on provided human feedback.

To incorporate feedback into our models, we utilize adapter layers (Houlsby et al., 2019), a parameter-efficient fine-tuning technique. Figure 3 sketches an overview of the architecture used. Adapters are integrated on top of each language model unit (e.g. transformer block) and are trained with the human feedback while we freeze all other model weights. Adapters can also be disabled to recover to the original state of the model.

3.5 Explanations & Feedback Mechanism

Users can evaluate the active model on the active dataset through the Feedback page. They may input text in three ways: i) create a text sample themselves; if authorized: ii) sample a random text from the active dataset; iii) sample a random *misclassified* text from the test part of the active dataset. Users receive the classification results and the model's confidence. They can assess the result and correct any misclassifications.

To further inspect the model's behavior, we provide two types of explanations—local and global. For local explanations on a text sample, we display relevant features to each output class (Figure 4). We attribute scores using the LIME framework (Ribeiro et al., 2016) and—to filter weak correlations—we highlight as relevant only tokens with a score above the threshold $\theta = 0.1$. On the global side, we list the most influential unigrams for each output class. These can be inspected to extract insights about what keywords and patterns

[9] https://huggingface.co/docs/datasets/index

[10] https://huggingface.co/docs/transformers/main_classes/model

the model focuses on at the dataset level. For all 1-grams present in a dataset, their corresponding classification scores are calculated and the tokens with top scores are displayed on the page.

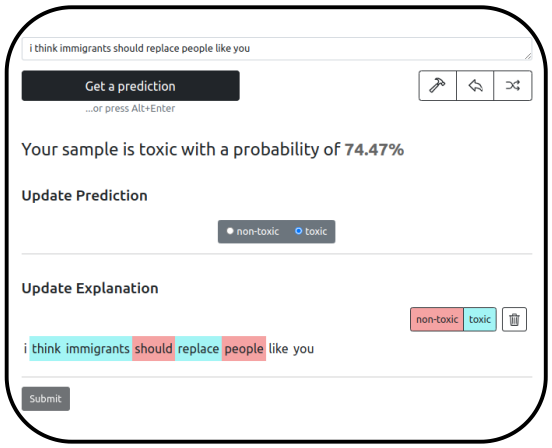


Figure 4: The example of the results and local explanations that annotators can obtain on the Feedback page.

Annotators can easily edit the highlighted tokens and send the updated explanation as feedback. We store the result—i.e. the highlighted relevant parts—and use them to fine-tune the adapter layers. Freezing all other model weights minimizes the computational effort of the feedback step.

Regarding the fine-tuning procedure, directly using the highlighted feedback text for adapter fine-tuning causes significant losses in the original model performance. We propose to mix feedback with original samples to mitigate this effect, which allows effective feedback incorporation while reducing model forgetfulness (see 4 for more details).

3.6 Backbone API

We expose our backbone's API to make available all essential dataset/model management functions. These provide a high-level interface for additional experiments dealing with model evaluation, explanation, and feedback. The API is built with the Python framework FastAPI[11], more details can be found in the Appendix A.

4 Case Study

We carried out a case study to test the applicability of IFAN. We chose a hate speech detection task based on the HateXplain dataset (Mathew et al., 2021). The goal of the experiment was to use our framework to debias a given hate speech detector.

[11]https://fastapi.tiangolo.com

Firstly, we modified the original dataset for binary classification task—*"toxic"* and *"non-toxic"*. We choose the Jewish subgroup as a target for our debiasing process. We fine-tuned BERT (Devlin et al., 2019)[12] and gave feedback to it. Additional experiments with BLOOM and other LLMs are provided in Appendix D and E respectively.

We annotate 24 random misclassified samples, 12 with the most confidence and 12 with the least confidence scores (see Appendix C.1). We invited 3 annotators to participate in the annotation process. The n-grams that were modified by annotators were saved and used to create a new training dataset for the adapters. As a result, we collected 40 annotated n-grams and repeated them to get 120 training samples. To complete the new training creation, we balanced these samples with 500 original samples (250 toxic, 250 non-toxic) randomly selected from the HateXplain dataset.

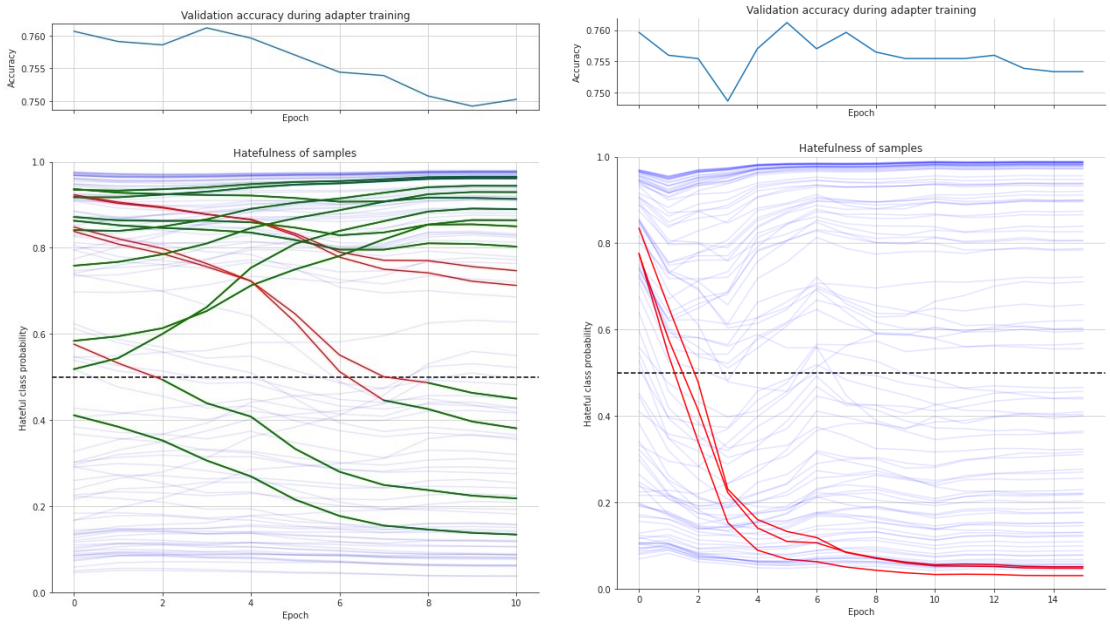
Model	Pr	Re	F1	Pr_J
BERT (baseline)	0.80	**0.78**	**0.79**	0.95
Most Confident Missclassified				
BERT+Feedback (non-bal.)	0.34	0.28	0.31	0.82
BERT+Feedback (bal.)	0.78	0.80	**0.79**	**0.97**
Least Confident Missclassified				
BERT+Feedback (non-bal.)	**0.83**	0.73	0.78	0.96
BERT+Feedback (bal.)	0.79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

Table 3: The results of the case study: hate speech classification model debiasing. We compare different strategies for feedback incorporation. Pr_J states for the Precision score on the Jewish target group.

The results are presented in Table 3. We observe that the non-balanced training dataset, which only contains feedback on the most confidently misclassified samples, resulted in a significant decrease in performance. While the inclusion of feedback on least confident samples caused a slight decline in the overall F1 score, Adapter training on the balanced feedback led to an improvement in the precision score for the Jewish target group.

Figure 5 shows the changes in the detector while fine-tuning with the collected feedback. When rebalancing the feedback, only modified samples are drastically changed while the performance on the original texts is only slightly affected. A detailed comparison between fine-tuning on non-balanced and balanced feedback is in Appendix C.2).

[12]https://huggingface.co/google/bert_uncased_L-2_H-128_A-2



(a) Training on feedback on the Jewish subgroup samples. (b) Training on feedback samples with "jewish" key-words.

Figure 5: Samples confidence variation as the model is fine-tuned with human feedback. The results of the domain case using IFAN platform. We can observe that for both experiments with balanced training data, the overall model's performance is only slightly changed while the model's behavior on the Jewish target group is improved.

5 Limitations & Future Work

As of now, our feedback system is limited to applications in the sequence-to-class format. Work on extending the platform to further task through LLM prompting is currently in progress (see Appendix E).

At the same time, we currently offer a limited set of explanation, feedback, and management options, which we plan to increase in the immediate future. A small user study has been conducted (Appendix B) to collect feedback about the platform and improve its user-friendliness. Our intent is to continue iterating the development of new features with trials with developers and laymen.

Finally, our experiments do not yet show clear trends w.r.t. the correlation between performance and feedback hyperparameters. Indeed, further research and trials have to be carried out to establish optimal choices for the number of feedback samples, fine-tuning epochs, and the rebalancing ratio.

6 Conclusion

This work proposes IFAN, a framework focusing on real-time explanation-based interaction between NLP models and human annotators. Our contribution is motivated by the limited options in terms of existing tools to interpret and control NLP models.

IFAN is composed of three main units. The **Backbone** unifies all the machine learning pipelines and exposes an API for accessibility. The **User Interface**—organized in *landing page*, *documentation*, *feedback*, *report*, and *configuration*—provides an intuitive visual component to interact with models. Finally, the **Admin** controls the connection between the two previous components.

Additionally, we introduce the feedback mechanism that takes advantage of adapter layers to efficiently and iteratively fine-tune models on the downstream task. Our experiments show the frameworks' effectiveness at debiasing a hate speech classifier with minimal performance loss.

We believe IFAN to be a valuable step towards enabling the interpretable and controllable deployment of NLP models—allowing users with no technical proficiency to interact and provide feedback to deployed NLP systems. Regarding future work, we set as a priority to extend the framework to more NLP tasks as well as to integrate additional model analysis features and feedback mechanisms.

Acknowledgments

We thank Natália Souza Soares and Ashish Jha for their valuable contribution. This paper has been supported by the *German Federal Ministry of Education and Research* (BMBF, grant 01IS17049).

Ethical Considerations

In this work, we showed the experiments of hate speech model debasing. The hate speech detection task is the task that requires a lot of attention to provide a fair outcome. One of the issues still is bias, especially against minority groups due to prejudices. We aimed to show an example of how the model can be debiased with respect to some target racial groups. With a conscientious selection of annotators and feedback, we hope that our proposed platform will serve to efficiently adjust NLP models to the diverse world.

For these reasons, we also believe that interpretability and controllability of modern NLP models and systems are fundamental pillars for their ethical and safe deployment (European Commission, 2020). This works aims at having a positive impact on both aspects as it provides a tool to explain models and provide them with feedback. By reducing the technical proficiency required to interact with NLP systems, we hope to facilitate the process of providing valuable human rationales to influence complex models.

Ensuring high quality for the human feedback is challenging (Al Kuwatly et al., 2020), and exposing models to external influence can be used as an exploit by adversarial agents (Mosca et al., 2022a). Especially with a very small crowd of annotators, there's potential for a few people to have a strong influence on the model. A restrictive access rights management system like IFAN's already mitigates these issues. We believe that additional security features as well as tracking annotators' impact are key for future work to foster their trustworthiness.

Previous works mention that users can feel discouraged and frustrated when interacting with poor models and badly-designed interfaces, which can also affect feedback quality (Lertvittayakumjorn and Toni, 2021). This can be addressed by integrating user studies in the development process in order to design more intuitive interfaces and improve the overall user experience.

On the opposite end of the spectrum, plausible explanations can make humans overestimate the model's capabilities and make them trust systems that are still not ready for deployment. In this case, a more diverse and complementary set of explanations for users (Madsen et al., 2022) as well as comprehensive model reports for developers are core goals to provide a more complete picture of the models to be deployed.

References

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Tilman Beck, Bela Bohlender, Christina Viehmann, Vincent Hane, Yanik Adamson, Jaber Khuri, Jonas Brossmann, Jonas Pfeiffer, and Iryna Gurevych. 2022. AdapterHub playground: Simple and flexible few-shot learning with adapters. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 61–75, Dublin, Ireland. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

European Commission. 2020. White paper on artificial intelligence: a european approach to excellence and trust. *Com (2020) 65 Final*.

Xiaochuang Han, Byron C. Wallace, and Yulia Tsvetkov. 2020. Explaining black box predictions and unveiling data artifacts through influence functions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5553–5563, Online. Association for Computational Linguistics.

Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. 2020. exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 187–196, Online. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2022. Natural language processing: State of the art, current trends and challenges. *Multimedia tools and applications*, pages 1–32.

Piyawat Lertvittayakumjorn and Francesca Toni. 2021. Explanation-based human debugging of NLP models: A survey. *Transactions of the Association for Computational Linguistics*, 9:1508–1528.

Pengfei Liu, Jinlan Fu, Yang Xiao, Weizhe Yuan, Shuaichen Chang, Junqi Dai, Yixin Liu, Zihuiwen Ye, and Graham Neubig. 2021. ExplainaBoard: An explainable leaderboard for NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 280–289, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-hoc interpretability for neural nlp: A survey. *ACM Comput. Surv.*, 55(8).

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14867–14875. AAAI Press.

Robert Munro Monarch. 2021. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster.

Edoardo Mosca, Shreyash Agarwal, Javier Rando Ramírez, and Georg Groh. 2022a. "that is a suspicious reaction!": Interpreting logits variation to detect NLP adversarial attacks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7806–7816, Dublin, Ireland. Association for Computational Linguistics.

Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. 2022b. SHAP-based explanation methods: A review for NLP interpretability. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4593–4603, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Arijit Ray, Yi Yao, Rakesh Kumar, Ajay Divakaran, and Giedrius Burachas. 2019. Can you explain that? lucid explanations help human-ai collaborative image retrieval. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 153–161.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022a. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji

Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022b. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. 2019. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2591–2600.

Julia Strout, Ye Zhang, and Raymond Mooney. 2019. Do human rationales improve machine explanations? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 56–62, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. 2020. The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 107–118, Online. Association for Computational Linguistics.

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019. AllenNLP interpret: A framework for explaining predictions of NLP models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 7–12, Hong Kong, China. Association for Computational Linguistics.

Zijie J. Wang, Dongjin Choi, Shenyu Xu, and Diyi Yang. 2021. Putting humans in the natural language processing loop: A survey. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 47–52, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Huihan Yao, Ying Chen, Qinyuan Ye, Xisen Jin, and Xiang Ren. 2021. Refining language models with compositional explanations. *Advances in Neural Information Processing Systems*, 34:8954–8967.

A Backbone API Endpoints

Figure 6 shows the auto-generated docs for our backbone's REST API, which serves as guidelines to interact with our backbone. Endpoints are divided into functional groups—*models*, *datasets*, *prediction*, *explanation*, and *feedback*). Currently, this page is only accessible within our institution's network for security reasons.

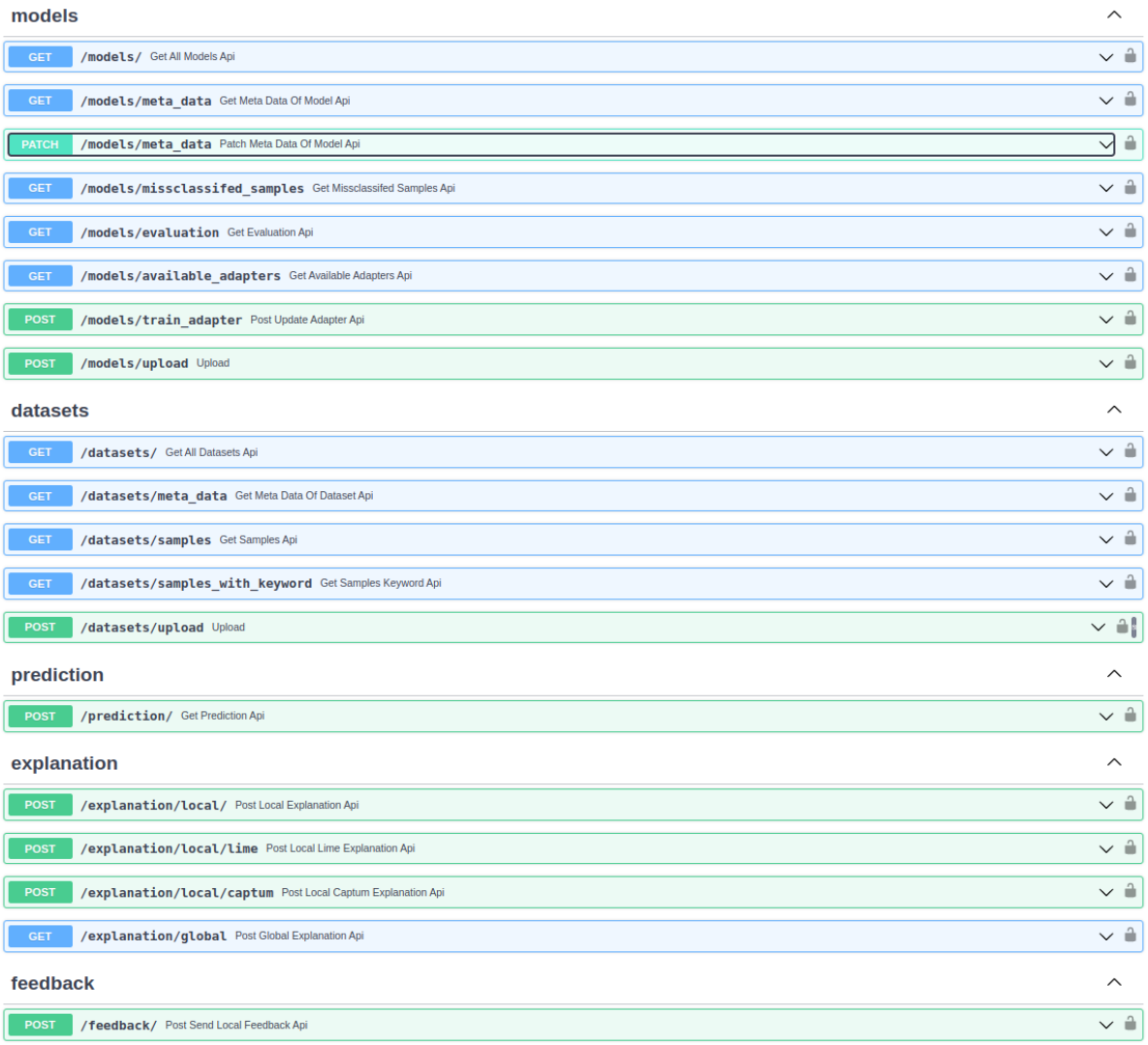


Figure 6: Screenshot of the Swagger UI for our backbone API endpoints.

Developers with direct API access (specifiable on the *configuration page*, see 3.1) can directly make requests to this high-level interface for additional (larger-scale) experiments. Once again, the API has been built with the Python framework FastAPI[13].

Figure 7 shows the documentation for the *explanation* endpoint. Here, we can inspect the details about the endpoint, such as the required parameters—i.e. the path to the model, the explainer to be used (e.g. LIME), and the model's prediction as body request.

[13] https://fastapi.tiangolo.com

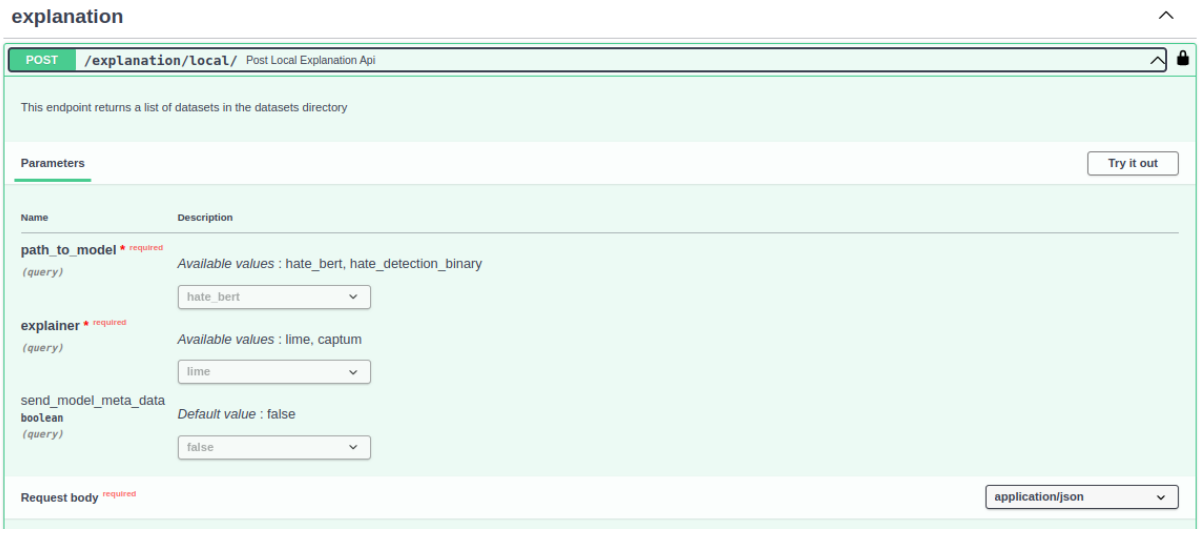


Figure 7: Screenshot of the *explanation* endpoint from our backbone API's Swagger UI.

B User Study

We performed a small user study evaluating the usability of our platform with users having different backgrounds. We gathered a total of nine testers, which we organize into three different categories based on their familiarity with the underlying methodologies. Users may be *laymen*, *computer scientists*, or *experts*. For each of these categories, three testers are assigned. Laymen are general users with no exposure to computer science. Computer scientists are those who studied computer science or a related discipline and who work in corresponding fields, such as software development. People who also possess knowledge in the domains of machine learning and NLP are considered experts.

We asked the users to try out to interact with IFAN and then answer several questions about the website's usability. The questionnaire was of the following structure:

- The first question is about the competency of the test person, which is categorized as computer science, natural language processing expert, or layman.
- Users assign a score between 1 and 5 to the local explanation on the left side of the feedback page.
- A textual input field for the users to hand in more detailed feedback on the previous rating.
- Users assign a score between 1 and 5 to the global explanation on the right side of the feedback page.
- A textual input field for the users to hand in more detailed feedback on the previous rating.
- Users rate the overall usability of the website with a score between 1 and 5.
- A textual input field for the users to hand in more detailed feedback on the previous rating.

The average ratings given by users: (i) Local explanation: 3.88; (ii) Global explanation: 3.2; (iii) General usability: 3.33. The problem that the users mentioned were insufficiency of only one form of explanations and feedback, sometimes misleading interface for markup, and waiting time. While some of the problems are already fixed in the current version of the platform, we are planning to consistently improve the usability of the platform. At the same time, we note a positive result—the users with different backgrounds were overall comfortable with IFAN usage and found it useful to increase the controllability of NLP models. We believe this confirms the possibility of IFAN utilization for stakeholders, developers, and annotators in future applications.

C Case Study

In this section, we provide additional information about the experiment conducted under Case Study (Section 4).

C.1 Annotated Samples *(Warning: the section contains hate speech only with illustrative purposes. We have no intention to offend the reader and a target racial group.)*

Here, we provide the list of that were used for manual annotation (Section 4). These samples were misclassified by hate speech detection model. In Table 4, we provide the misclassified samples with the lowest confidence, in Table 5 – with the highest confidence. All the samples and their true labels are taken from the original HateXplain dataset (Mathew et al., 2021). For the label assignment details, refer to the HateXplain dataset description.

Sample	Original Label	Prediction	Confidence Score
it not so much about the torah which is pretty innocuous and ignored by many jews despite what is said it really about that forbidden book the talmud or babylonian talmud which as goyim cattle we are not permitted to even read good luck with your research	non-toxic	toxic	0.56
if memory serves ashkenazi jews and east asians have higher average iqs	non-toxic	toxic	0.57
speaking of does anyone know which goy group the jews intermarry with in the greatest numbers i was under the impression it was either asians or catholics	toxic	non-toxic	0.52
<user> the jews faked the moon landing	toxic	non-toxic	0.52
<user> <user> she will blame it on jews or trump i bet	toxic	non-toxic	0.53
christianity role is historical protector of the jews it has afforded the tribe the time and space it needed to prepare white genocide teamwhite	toxic	non-toxic	0.54
the jews are gearing up to help usher in democrats in <number> twitter facebook msm google youtube all upping their censorships of the right pro whites	toxic	non-toxic	0.51
amusing to think that white people will never be given a referendum on expelling the jews from our nations when you think about things in such terms it becomes abundantly clear that we are expected by our governments to suffer jewish parasitism without the right to defend ourselves	toxic	non-toxic	0.50
it all hehe haha until we show up at your local favorite hang out spot and wait for you to come out in the van they go one little heeb at a time whatever it takes	toxic	non-toxic	0.53
<user> <user> jews and the stupid holocaust myth	toxic	non-toxic	0.52
<user> jews will never be on the blockchain because you can not trust them	toxic	non-toxic	0.51
<user> <user> it the jews not europeans stop shifting blame on gentiles	toxic	non-toxic	0.57

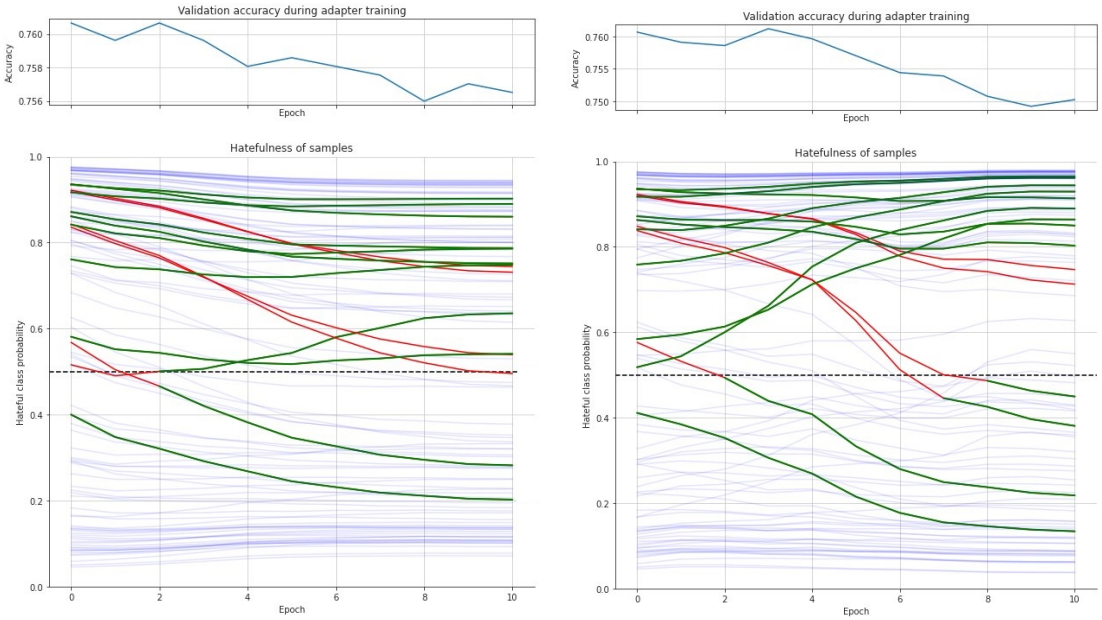
Table 4: Misclassified with the lowest confidence samples used for manual annotation.

Sample	Original Label	Prediction	Confidence Score
would rather have terrorism over not being able to boycott an allied country this talking point must be important because he repeats it but i can not make out what on earth it supposed to mean if you do not buy stuff from jews they will fake arabs taking down more nyc landmarks	non-toxic	toxic	0.72
how long will jews be a majority in israel	non-toxic	toxic	0.88
what do you mean goes and buys their house not only predates the ghetto it was there before ghetto people came to exist why do you want white people to take a stand for their home but jews to flee theirs	non-toxic	toxic	0.77
calling me a kike is not an insult	non-toxic	toxic	0.97
the jews and muslims grew up together and the jews gave birth to christianity catholicism and later on mormonism like a big religion family	non-toxic	toxic	0.78
greg susan frank rodrick and manny heffley are all jews	non-toxic	toxic	0.67
it amazing how long <user> talked without giving any answers i learned way more about <user> thoughts and life than sargon i thought only jews could bloviate that hard hm	toxic	non-toxic	0.95
<user> <user> um akshually star trek were the space commies and star wars jedis are just space jews	toxic	non-toxic	0.87
<user> apparently not so obvious to some esp the homosexual jewish media that wants women to look like boys and people who think that is traditional rather than feminist	toxic	non-toxic	0.84
sorry i was referring to the holocaust story i believe it to be a lie	toxic	non-toxic	0.89
i swear college is a hot house full of judaised professors with white youth being indoctrinated into this nonsense or have live there days out in fear of being ousted as racists and proud members of white race	toxic	non-toxic	0.90
a few months ago i said that anti semitism doesnt exists naming a group of people is a branch of social sciences called sociology well jews are at war with social sciences now you simply cant name them	toxic	non-toxic	0.74

Table 5: Misclassified with the highest confidence samples used for manual annotation.

C.2 Feedback Mechanisms Comparison

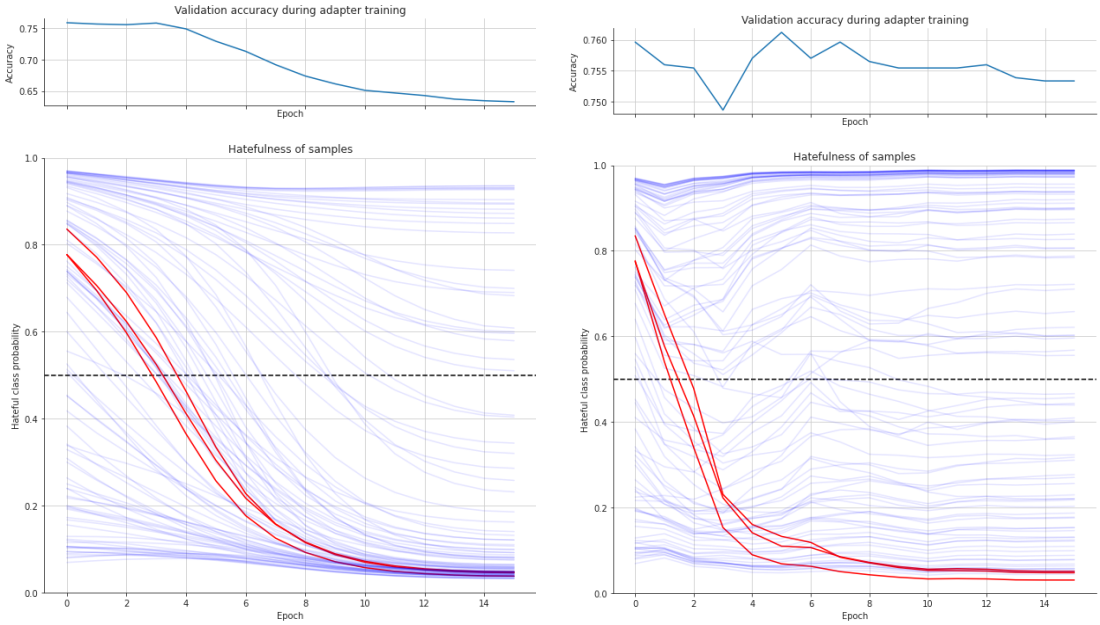
In Section 4, we report the results of the model trained on feedback in two setups: (i) without balancing and (ii) with balancing via using original samples from HateXplain dataset. The comparison between these two setups is visualized in Figures 8 and 9. We tested our approach on the local feedback on the Jewish target group samples as well as samples containing the "Jewish" keyword. For both setups, with balancing, the training procedure runs more stable. The model's performance on other samples from HateXplain dataset changes slightly and the adjustment of its behavior on the marked-up samples proceeds more rapidly.



(a) Training without feedback balancing.

(b) Training with feedback balancing.

Figure 8: The comparison of training procedure with and without feedback balancing. Here, the results of local feedback on the least confident misclassified samples from the Jewish target group are shown. We can observe that training with a balanced dataset runs more stable without significant influence on the overall model's domain knowledge.



(a) Training without feedback balancing.

(b) Training with feedback balancing.

Figure 9: The comparison of training procedure with and without feedback balancing. Here, the results of local feedback on misclassified samples with "jewish" keywords are shown. We can observe that training with rebalanced dataset runs more stable without significant influence on the overall model's domain knowledge.

D BLOOM for Hate Speech Classification and Human Feedback Incorporation

In addition to the results with BERT for hate speech classification (Section 4), we provide the results of experiments with a BLOOM[14] model (Scao et al., 2022b) (Table 6). Once again, the fine-tuning of adapter layer with human feedback improves model performance within Jewish target group. In the case of BLOOM, the overall F1 score is also improved, confirming the utility of human feedback usage and the possibility of it being extended to other models. Also, from Figure 10, we can see how the label of the misclassified samples (i.e. *"Um ... he's jewish, not Muslim"*) is changing as the model is fine-tuned with the provided feedback.

Model	Pr	Re	F1	$F1_J$
BLOOM (baseline)	0.40	**0.70**	0.48	0.48
BLOOM+Feedback	**0.49**	0.58	**0.51**	**0.53**

Table 6: The results of the LLMs inference for hate speech classification. $F1_J$ states for the F1 score on the Jewish target group.

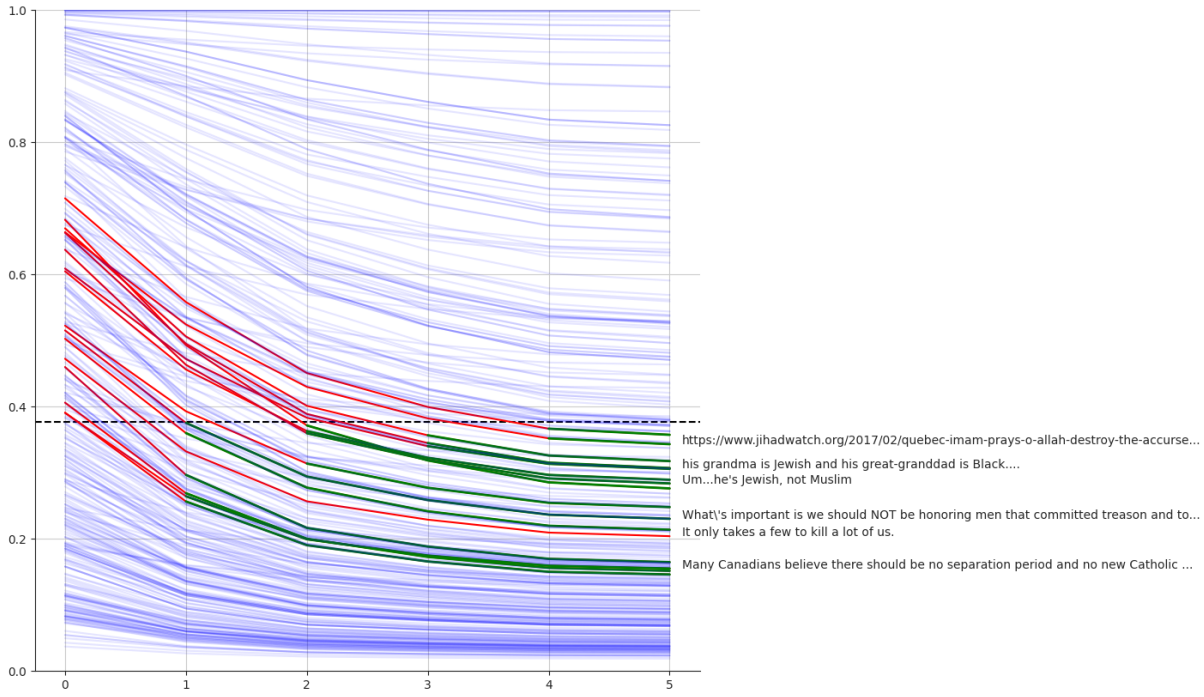


Figure 10: The tracking procedure of the misclassified samples label fix during feedback incorporation.

E Large Language Models for Hate Speech Classification and Explanation

In addition to the classification via linear layers on top of transformer blocks, we test the possibility of large language models to perform text classification and explain their prediction. Potentially, this approach allows further work to extend the framework to any arbitrary NLP tasks.

Currently, the models available on IFAN for testing are: vicgalle/gpt2-alpaca[15], chavinlo/alpaca-native[16], and chavinlo/gpt4-x-alpaca.[17] Their results on hate speech classification task based on HateXapline dataset (Mathew et al., 2021) are presented in Table 7.

For each model, we optimize the prompts to perform hate speech detection:

[14] https://huggingface.co/bigscience/bloom-560m
[15] https://huggingface.co/vicgalle/gpt2-alpaca
[16] https://huggingface.co/chavinlo/alpaca-native
[17] https://huggingface.co/chavinlo/gpt4-x-alpaca

vicgalle/gpt2-alpaca *Instruction: perform the text-classification task on the text using the following labels: 0 or 1. Text: would be hilarious if the un building was turned into homeless shelter or hell why not a refugee center The text is 0 Text: [DATASET SAMLPE] The text is 1 Text: [INPUT TEXT] The text is*

chavinlo/alpaca-native *This is an overall text-classification. For the given TEXT, CLASSIFY the LABEL of that.The LABEL can be 0 or 1. Just output the LABEL. NOTHING ELSE! TEXT: [5 DATASET SAMLPES] LABEL: 0 TEXT: [5 DATASET SAMLPES] LABEL: 1 TEXT: [INPUT TEXT]. LABEL:*

chavinlo/gpt4-x-alpaca *INSTRUCTION: Given the following DATASET DESCRIPTION, EXTRACT the TASK of it, and PERFORM the TASK on the INPUT TEXT. The FINAL LABEL could be 0 or 1. Just output the FINAL LABEL. NOTHING ELSE! DATASET DESCRIPTION: Hatexplain is the first benchmark hate speech dataset covering multiple aspects of the issue. Each post in the dataset is annotated from three different perspectives: the basic, commonly used 3-class classification (i.e., hate, offensive or normal), the target community (i.e., the community that has been the victim of hate speech/offensive speech in the post), and the rationales, i.e., the portions of the post on which their labelling decision (as hate, offensive or normal) is based. CONTEXT: TEXT: [DATASET SAMPLE] FINAL LABEL: 0 TEXT: [DATASET SAMPLE] FINAL LABEL: 1 INPUT TEXT: TEXT: [INPUT TEXT]. FINAL LABEL:*

Despite the specific differences in the various prompts, they all follow the core idea of i) mention the main task which can be extracted from the dataset metadata; ii) provide the general information about labels; iii) provide some examples for each label from the dataset. Potentially, this prompt design can be used for any classification task.

Model	Pr	Re	F1
vicgalle/gpt2-alpaca	0.63	0.56	0.59
chavinlo/alpaca-native	0.60	0.51	0.55
chavinlo/gpt4-x-alpaca	**0.66**	**0.63**	**0.64**

Table 7: The results of the LLMs inference for hate speech classification.

F Supplementary Video Demo

A supplementary video showcase can be found on Youtube[18]. For completeness, we also point at an additional version of such demo, as well on Youtube[19], dating back to March 2023.

[18] https://youtu.be/EzC6HI3JwaQ
[19] https://www.youtube.com/watch?v=BzzoQzTsrLo