Safe Kids AI's Natural Language Processing (NLP) and Computer Vision (CV) Technology

Praneeth Bedapudi

Safe Kids AI

[Unpublished manuscript]

Author Note

Praneeth Bedapudi is the Lead Developer – Artificial Intelligence for Safe Kids AI, a for-profit education technology company.

Correspondence concerning this white paper should be addressed to Praneeth Bedapudi at praneeth@safekids.ai.

Abstract

Safe Kids AI, a commercial education technology company, incorporates artificial intelligence (AI) in its child and adolescent online safety software. The natural language processing (NLP) and computer vision (CV) models used by the company are unique in that they provide a high degree of accuracy while running in the browser on low-powered laptop computers without hampering performance of the machines. This paper describes the models development, testing, and augmentation.

Keywords: artificial intelligence, machine learning, natural language processing, computer vision

Safe Kids AI's Machine Learning (ML) and Natural Language Processing (NLP) Technology

Safe Kids AI is a US-based educational technology company with the stated mission of empowering young people to be safer online by helping them make better decisions. A key component of Safe Kids AI's child and adolescent online safety software is computer vision (CV) and natural language processing (NLP) technology. The NLP model focuses on identifying hate speech in online messages and predicting the intent of a user when conducting an online search. The CV model focuses on identifying explicit adult sexual content (EASC) and modern guns. The purpose of this paper is to describe the development, testing, and unique features of the NLP and CV models used by Safe Kids AI.

NLP Model and Architecture

For data privacy and security reasons, the NLP model must run on the user's machine (i.e., locally rather than in the cloud). Frequently, the devices are underpowered laptops (e.g., Chromebooks) requiring the model to be highly efficient with a low demand on system resources. At the same time, the model must be capable of understanding and classifying the ever-changing slang and colloquialisms used by young people. After experimenting with recurrent neural network (RNN) models, specifically long short-term memory (LSTM), as well as large-scale transformers, the company implemented a transformer with a hidden size of 128 and 2 encoder decoder layers. Because of its parallel nature, the transformer obtains much better inference speeds compared to RNNs with similar or lower parameter counts.

NLP Model Training

In general, NLP, computer vision (CV), and Speech models benefit from pre-training on tasks that focus on understanding context, such as next-word prediction and missing token prediction. So that our model can best understand the semantic relationships between words and phrases used in everyday speech by youth, we pre-trained it on a massive unlabeled text corpus of 4.8 billion tokens. We collected the unlabeled text corpora from a variety of sources including wikis, Reddit, and popular news. This approach helped pre-train the model to learn from different contexts (e.g., social media, formal writing).

Identifying the best data for training the model was challenging because of the company's goal to innovate software that keeps youth safer online while also fostering natural childhood curiosity and enhancing the educational value of technology, particularly in the classroom. Similar words and phrases appear in both educational and noneducational contexts. Whereas the standard procedure for parental control software is to limit access to online content based on simple keywords, in this case it is necessary to take a more nuanced approach. In the current state, a curious teen who searches “what is orgasm” or “what does breast cancer look like” will get the same outcome as a user searching for EASC, which is a poor educational outcome. Publicly available datasets are frequently mislabeled and untrustworthy. (Ayo et al., 2020; Basile et al., 2019; Davidson et al., 2017; de Gibert et al., 2018; Kim et al., 2020; Mollas et al., 2022; Saha et al., 2018; Toraman et al., 2022; Waseem, 2016; Waseem & Hovy, 2016; Zampieri et al., 2019). Although frequently touted in the media, datasets annotated by youth are no more accurate than those done by experienced adult annotators (Bhattacharya et al., 2020; Mathew et al., 2022). Moreover, most of the categories our software supports require significant quantities of data to be resilient against misdetections in real-world use, but these topic areas (e.g., sex education) do not have large, public datasets associated with them.

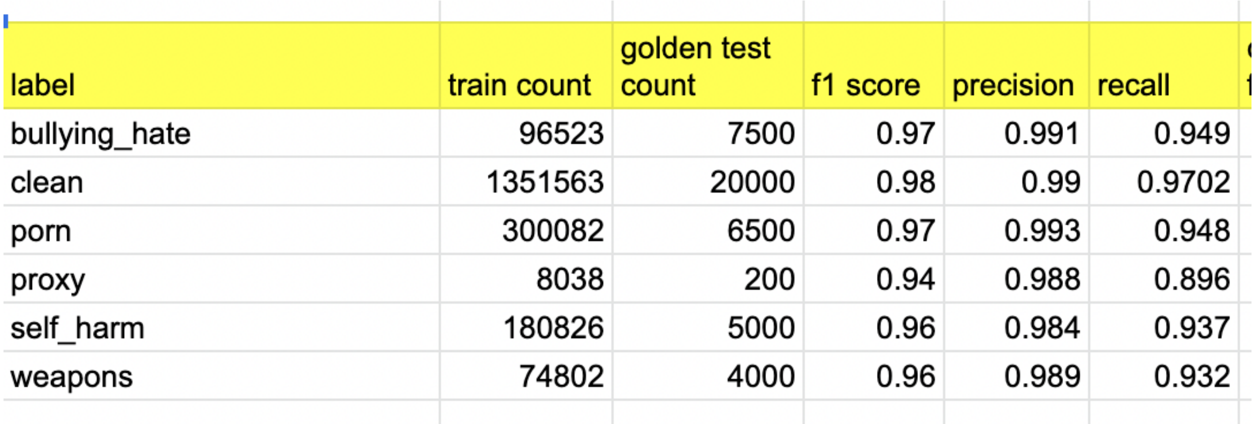
We employ 3 major techniques for mining and generating labeled text data from unlabeled text corpora: semantic search, zero-shot learning, and few-shot finetuning. As a starting point, we develop search phrases and counter examples for the intent or class we want to mine. Using a large, pre-trained transformer as the encoder for semantic search, we query over the unlabeled corpora and separate out the texts with high semantic-match scores to our target search phrases. Next, we use an ensemble of exceptionally large, zero-shot classifier models to further clean this data. Finally, we feed the cleaned data into large few-shot models and use the resulting model to create the unlabeled data.

NLP Model Augmentation and Inference

In our experience, we found that employing various statistical augmentation techniques helps the model adapt to real-world text which, especially in the case of youth, is often riddled with intentional (and unintentional) misspellings, typos, and incorrect grammar. The techniques used include random token deletion and insertions as well as word replacement. We use adaptive class weights in the loss function since the data is massively imbalanced with a very high number of examples for so-called clean intent, representing real-world usage. The model is trained on a cluster of 8 graphics processing units (GPUs) with distributed data parallel (DDP). Once trained, we export the model to multiple formats compatible with, and optimized for, their respective backends (e.g., WebGL for browser, WASM, CPU for on-device). Table 1, below, illustrates how the model is optimized to produce a low rate of false positives, meaning high-precision predictions of problematic (i.e., not clean) categories. The model is also tested on various data splits along with real-world simulated testing whereby we test on search result titles for searches that are both clean and not.

Table 1

Training counts and results for NLP model



label	train count	golden test count	f1 score	precision	recall
bullying_hate	96523	7500	0.97	0.991	0.949
clean	1351563	20000	0.98	0.99	0.9702
porn	300082	6500	0.97	0.993	0.948
proxy	8038	200	0.94	0.988	0.896
self_harm	180826	5000	0.96	0.984	0.937
weapons	74802	4000	0.96	0.989	0.932

CV Model and Architecture

As noted above, the CV model is used by the Safe Kids AI software to identify EASC and modern guns in online images. Like the NLP model, the CV model is expected to run locally in the browser and faces the same demands with respect to resource efficiency. The CV model is a single-shot object detector with 1.8 million parameters trained at 320 x 320 input resolution. We observed that this combination resulted in much higher accuracy compared to classifiers. Detection also allows us fine-grained control on filtering (i.e., which objects are allowed versus a single class prediction per image).

CV Model Training

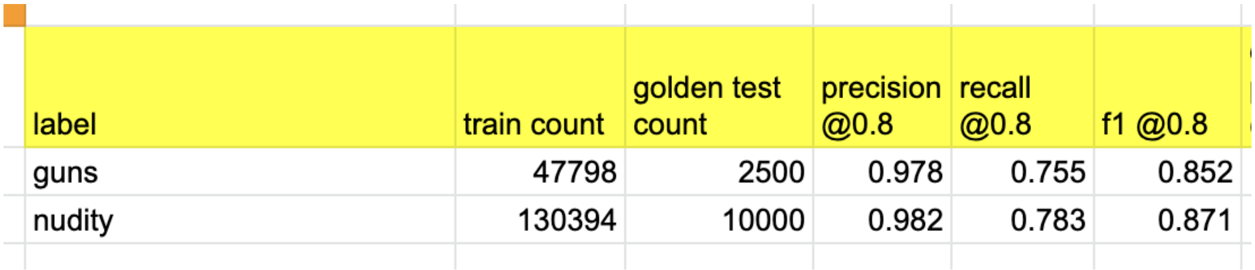
For pre-training, we used a combination of major public datasets and other semi-labeled data. We subsequently replace the detection head with the classes in which we are interested while we continue training for the classes we need. We obtained approximately 2.2 million images across various target categories from multiple semi-labeled sources, including Reddit, Imgur, and category-specific websites. Using a combination of zero-shot object localization and incremental few-shot training, we labeled the images for detection. Although the model is a detector, because the end goal is image classification, we test based on classification accuracy (i.e., after the post-processing logic precision and recall for each category). We collected the testing data from splits from the original dataset and image search results for various pre-labeled queries.

CV Model Augmentation and Inference

For resource efficiency, the images are augmented at runtime while training with standard augmentation techniques (e.g., Albumentations). We apply an adaptive loss weighing function to overcome class imbalance, meaning that class weights for loss are constantly updated during training. To optimize inference speed, we export the model without Non-Maximum Suppression (NMS) as it is not required for the classification outcome. When exported, the model is pruned by replacing 15% of weights, allowing lower memory usage with no observable accuracy loss. Table 2, below, presents the test data showing the high degree of accuracy of the model.

Table 2

Training counts and results for CV model



label	train count	golden test count	precision @0.8	recall @0.8	f1 @0.8
guns	47798	2500	0.978	0.755	0.852
nudity	130394	10000	0.982	0.783	0.871

Conclusion

With the rise of one-to-one student to device ratios because of remote schooling during the Covid-19 pandemic as well as the proliferation of low-cost home laptops for youth (e.g., Chromebooks), the need for adaptive online safety software that balances child and adolescent security with natural curiosity and learning objectives is critical. This paper described the development, testing, and unique features of the

NLP and CV models used by Safe Kids AI. Due to their efficiency and accuracy, these models provide a significant benefit, allowing for high performance with no perceptible impact to device performance.

References

Ayo, F. E., Folorunso, O., Ibharalu, F. T., & Osinuga, I. A. (2020). Machine learning techniques for hate speech classification of Twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review, 38*. https://doi.org/10.1016/j.cosrev.2020.100311

Basile, V., Bosco, C., Fersini, E., Debora, N., Patti, V., Rosso, P., & Sanguinetti, M. (2019). Multilingual detection of hate speech against immigrants and women in Twitter at SemEval-2019 Task 5: Frequency analysis interpolation for hate in speech detection. Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN.

Bhattacharya, S., Singh, S., Kumar, R., Bansal, A., Bhagat, A., Dawer, Y., Lahiri, B., & Ojha, A. K. (2020, May 11–16). Developing a multilingual annotated corpus of misogyny and aggression. Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, Marseille, France.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May 15-17). Automated hate speech detection and the problem of offensive language. Proceedings of the International AAAI Conference on Web and Social Media, Montreal, Quebec, Canada.

de Gibert, O., Perez, N., García-Pablos, A., & Cuadros, M. (2018). *Hate speech dataset from a White Supremacy forum*.

Kim, J. Y., Ortiz, C., Nam, S., Santiago, S., & Datta, V. (2020). Intersectional bias in hate speech and abusive language datasets. *ArXiv, abs/2005.05921*.

Cornell University Library, arXiv.org. (2022). *Hatexplain: A benchmark dataset for explainable hate speech detection*.

Mollas, I., Chrysopoulou, Z., Karlos, S., & Tsoumakas, G. (2022). ETHOS: A multi-label hate speech detection dataset. *Complex & Intelligent Systems*. https://doi.org/10.1007/s40747-021-00608-2

Saha, P., Mathew, B., Goyal, P., & Mukherjee, A. (2018). Hateminers: Detecting hate speech against women. https://doi.org/10.13140/RG.2.2.33967.18081

Toraman, C., Şahinuç, F., & Yilmaz, E. H. (2022, June 20-25). Large-scale hate speech detection with cross-domain transfer. Proceedings of the 13th Language Resources and Evaluation Conference, Marseille, France.

Waseem, Z. (2016, November 5). Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. Proceedings of 2016 EMNLP Workshop on Natural Language Processing and Computational Social Science, Austin, TX.

Waseem, Z., & Hovy, D. (2016, June 12-17). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. Proceedings of NAACL-HLT 2016, San Diego, CA.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019, June 2-7). Predicting the type and target of offensive posts in social media. Proceedings of NAACL-HLT 2019, Minneapolis, MN.