# Anticipating Hate Speech from Partial Input

Federico Rocca, Timothé Dard, Nevò Mirzai Hamadani
Group 30

## Problem definition

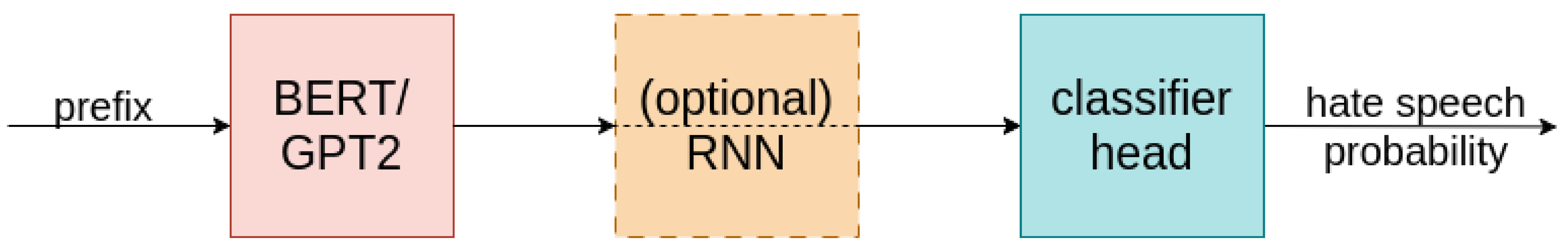
- **What?** Predict the probability a user's draft will become hate speech as they type.
- **Why?** Current detectors only flag finished post; sometimes it is too late to prevent harm, especially if applied to live events.
- **Challenge**: No off-the-shelf dataset of partial sentences labeled by evolving hate-speech risk.

## Key Related Works

- Full-sentence hate speech classifiers:
  - HateBERT: BERT retrained on abusive Reddit posts [1]
- Proactive detection:
  - Predictive analytics on drafts [2]
- Early text classification:
  - Weighted loss for prediction [3]

## Method

- Transformer backbone, **BERT** or **GPT2**
  - Impact of bidirectional (BERT) vs causal (GPT2) self attention on the prediction task
- Optional 3 layer **biLSTM** (h=256)
  - Recurrency bias to put emphasis on the last token

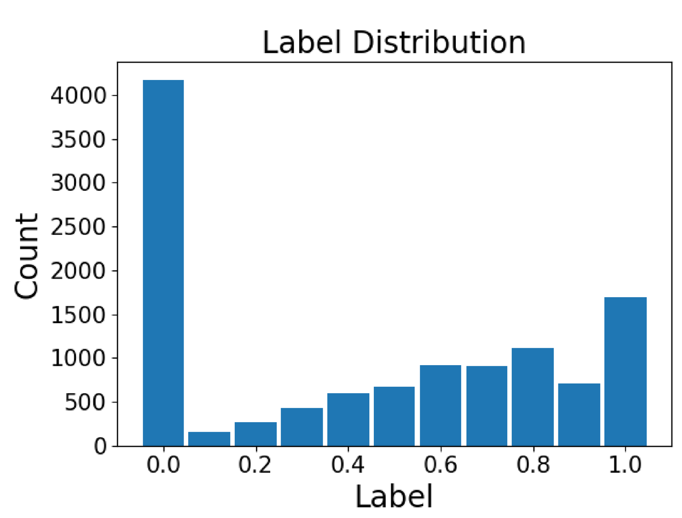



- **Loss**: prefix-length weighting
  - Do not punish wrong prediction on short prefixes
- Finetuning with **probability-labeled dataset**

## Datasets

- Initial training: hate_speech18 (Hugging face) and OLID, large hate speech/offensive speech datasets for general hate speech classification. (≈25000 rows)
- Refinement for implicit hate detection : Implicit-Hate Corpus(MIT), 6347 tweets labeled for explicit/implicit hate
- Task-specific adaptation: **cut and hand-labeled** the Implicit-Hate dataset
- **Final evaluation** : Hand constructed dataset, sentence fragments hand-labeled with our subjectively assessed probability of becoming hate speech

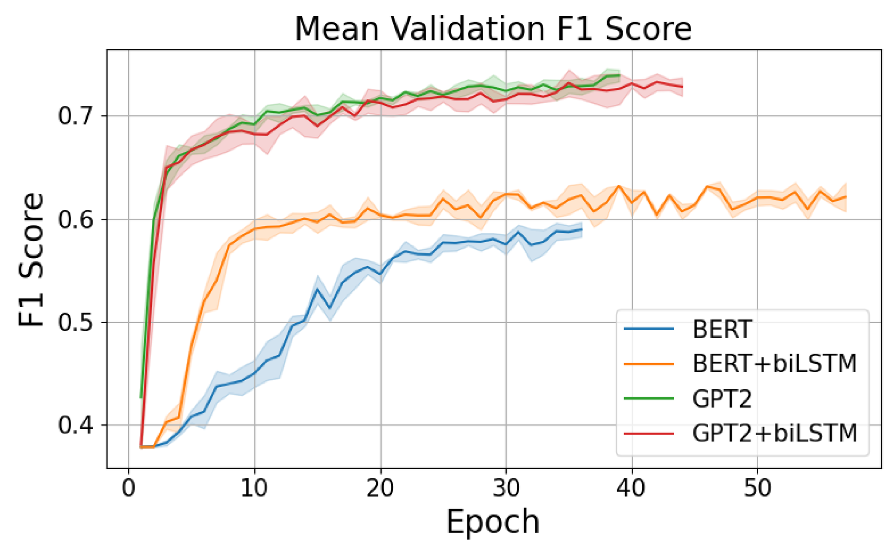
| Sentence | Label |
|---|---|
| ... | - |
| After you strip | 0.0 |
| After you strip off | 0.0 |
| After you strip off his | 0.1 |
| After you strip off his makeup | 0.3 |
| After you strip off his makeup, biologically | 0.6 |
| ... | - |

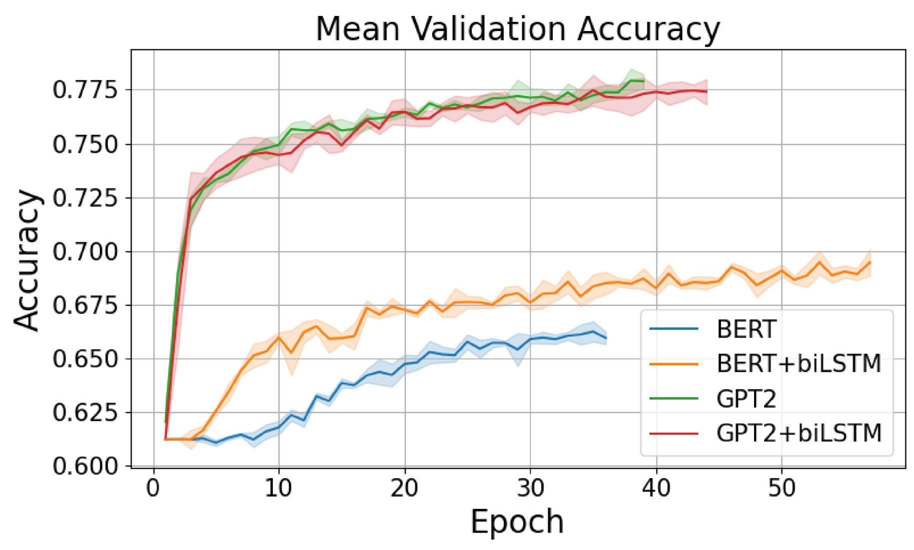



## Validation

- Data split: 85% training, 15% validation
- **Models comparison**, training on binary-labeled datasets:

| Model | Train Accuracy | Validation Accuracy | Train F1 score | Validation F1 score |
|---|---|---|---|---|
| BERT | 0.6552 ± 0.0039 | 0.6653 ± 0.0047 | 0.5726 ± 0.0059 | 0.5999 ± 0.0059 |
| BERT + biLSTM | 0.6877 ± 0.0050 | 0.6902 ± 0.0025 | 0.6156 ± 0.0072 | 0.6206 ± 0.0153 |
| GPT2 | 0.7825 ± 0.0060 | 0.7812 ± 0.0024 | 0.7467 ± 0.0074 | 0.7419 ± 0.0040 |
| GPT2 + biLSTM | 0.7971 ± 0.0077 | 0.7776 ± 0.0060 | 0.7656 ± 0.0099 | 0.7350 ± 0.0092 |

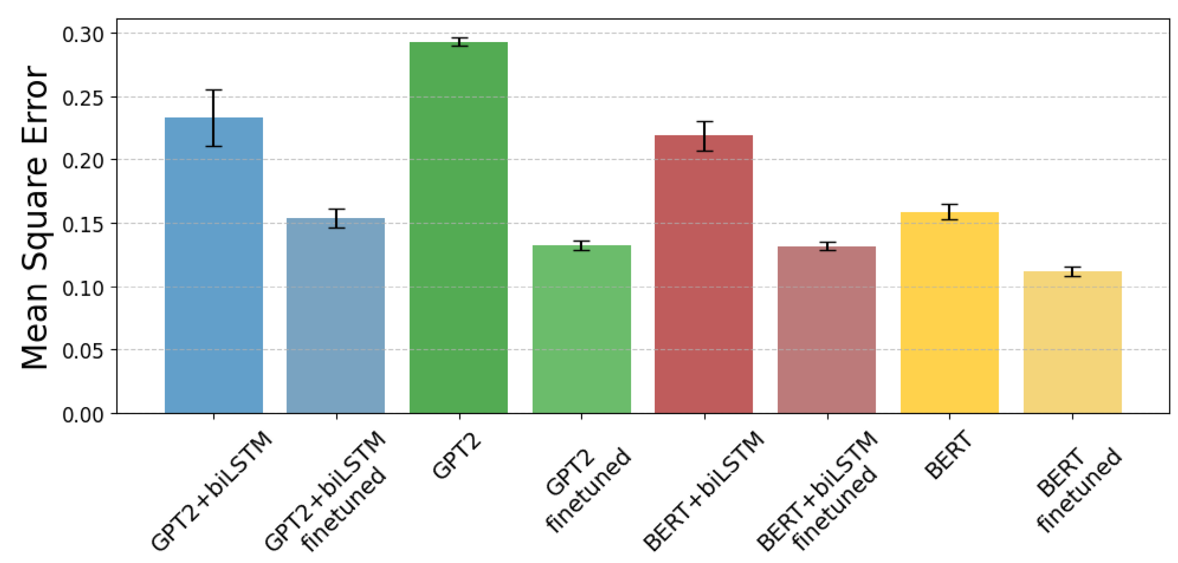



**Key observations:**

- GPT2 models converge faster with higher accuracy and F1
- biLSTM on top of the text encoder seems to improve accuracy and F1 on the validation set
- Training vs validation metric show small generalization gap

**Results on test set:**




## Limitations

- Label mismatch (classification labels and prediction task)
- Dataset size
- Subjectivity in the annotations
- Skewed distribution

## Conclusion

- Incorporating an RNN between transformer and classification head doesn't yield clear benefits at test time.
- Fine-tuning with probability-labeled data improves test-time performance.
- Data improvement is needed.
- The model's current test MSE indicates that further refinement is necessary before practical application.

### References

[1] T. Caselli, V. Basile, J. Mitrovi ´c, and M. Granitzer, "Hatebert: Retraining bert for abusive language detection in english," 2021.
[2] S. Bandara and H. Abeysundara, "A predictive model for anticipated hate and speech violence in social media: Large language model approach," International Journal of Research and Scientific Innovation, vol. XII, pp. 325–332, 03 2025.
[3] A. Cao, J. Utke, and D. Klabjan, "A policy for early sequence classification," 2023.