⚠️ **Warning: This presentation contains samples of texts that are offensive and/or hateful in nature.** ⚠️

Probing the impact from fine-tuning on BERT's ability in understanding toxic speech detection

---

Lawrence Lai

Fall 2023

Motivation

Observations during literature review	Reactions while reading these papers
Tons of competitions but researchers tend to **focus solely on outperforming SotA**	"Yay for your SotA performance (for now). **Why did your neural network perform so well**?"
Error analyses to understand what went wrong are either **nonexistent or inadequate**.	"OK… but **what went wrong** and what were your hypotheses?"
Limited understanding of what information is captured by the neural network as long as we get SotA performance.	"**What is actually learned** by the neural network that allows it to perform toxic speech detection?"

Key question

How does the knowledge encoded by the embeddings during pre-training and fine-tuning impact the performance of toxic speech detection?

HateXplain dataset

HateXplain dataset

HateXplain (Mathew et al., 2021) is a benchmark dataset for explainable hate speech detection.

- **19,229** tweets from Twitter and posts from Gab
- **8:1:1** splits for train, validation and test sets.
- Each post has **annotations by multiple annotators** with the following information:
  - **Three-class label**: Hate speech, offensive, normal
  - **Targeted community**
  - **Rationales** behind labeling decision

Data preprocessing

- Convert the **three-class labels into two-class labels** by re-mapping hate speech and offensive as toxic and normal as non-toxic
- Derive **single ground truth label by majority voting** across annotators


```
{
  "post_id":"1179099136349626368_twitter",
  "annotators":[
    {"label":"offensive", "annotator_id": 152, "target":["Women"]},
    {"label":"offensive", "annotator_id": 4, "target": ["Women"]},
    {"label":"offensive","annotator_id": 49, "target":["Women"]}
  ],
  "rationales": [[0, 0, 1, 1], [0, 0, 1, 1], [0, 0, 0, 1]],
  "post_tokens": ["dorothy", "a", "dirty", "slut”]
}
```


Figure: Sample data for one of the posts from HateXplain dataset

Probing tasks 101

Probing tasks

Probing tasks are supervised classification tasks designed to analyze the linguistic knowledge captured and encoded in contextual embeddings.

Related works

- **SentEval by Conneau et al. (2018)**, a toolkit used to evaluate universal sentence representations via binary and multi-class classification, natural language inference and sentence similarity
- **GLUE by Wang et al. (2018)**, a suite of nine tasks designed for probing models for understanding specific linguistic phenomena
- **Edge probing framework by Tenney et al. (2019)** along with a suite of sub-sentence tasks to investigate a range of syntactic, semantic, local and long-range phenomena.

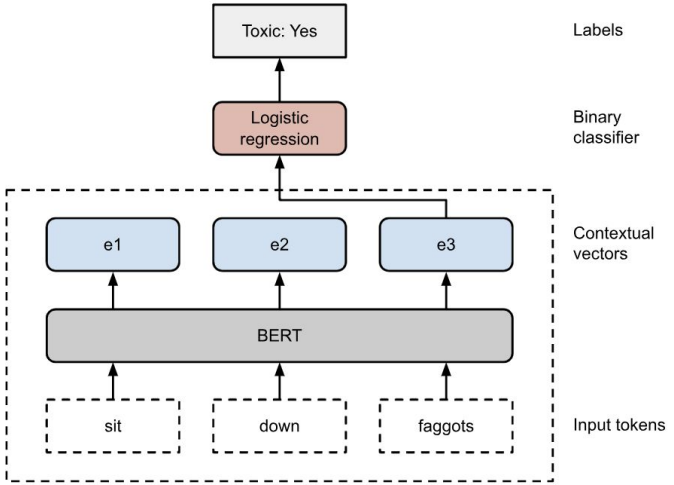

Figure: The architecture for probing tasks. All parameters inside the dashed line are fixed while we train the logistic regression classifier to extract information from the contextual vectors.

Approach

Finetune BERT-based models

- All models are based on **bert-base-uncased with a classification head**.
- **Model 1**: Freeze BERT layers and fine-tune with post-level toxicity
- **Model 2**: Unfreeze all layers and fine-tune with post-level toxicity
- **Model 3**: Unfreeze all layers and fine-tune with post-level toxicity and word-level rationales (provided by authors of hateXplain)

Analysis to inform design of probes

- Conduct **error analysis** to understand what went wrong across the models
- Leverage **LIME (Ribeiro et al., 2016)** to understand the importance of words across the models
- Gather insights to **design probing experiments** using a simple logistic regression probing classifier

Extract contextual embeddings

- Extract **all the hidden states across all layers** (i.e. 13 layers = embedding layer + 12 hidden layers) for each BERT-based subword tokens
- Create **token mapping** from words to BERT-based subword tokens
- Generate **one embedding per word** based on corresponding BERT-based subword tokens

Run probing experiments

- Run **probing experiments** based on **6 topics and 5 tasks**
- **Topics**: Profanity, women, LGBTQ, African, Jewish, Islam
- **Tasks**
  - A: On- vs. off-topic
  - B: On- vs. off-topic in toxic posts
  - C: On- vs. off-topic in non-toxic posts
  - D: Toxic on-topic tokens in toxic posts vs. non-toxic on-topic tokens in non-toxic posts
  - E: Toxic on-topic tokens in toxic posts vs. non-toxic on-topic tokens in toxic posts

Results and discussions

Questions

- How well do the contextual embedding capture information necessary to identify topics that may be associated with toxic speech?
- How does the presence and/or absence of toxicity in a post affect the ability of contextual embeddings in identifying such topics?

Results

- Contextual embeddings from the original BERT model **encode sufficient information** needed to distinguish between on- and off-topic mentions.
- The presence and absence of toxicity have **minimal impact** on the such ability.
- In general, the embeddings **have sufficient information in identifying topics that could potentially be mentioned in a toxic manner**.

Topic	Task	Model 1	Model 2	Model 3
Profanity	A	0.99	0.99	0.99
	B	0.99	0.99	0.99
	C	0.98	0.99	0.99
Women	A	0.97	0.97	0.98
	B	0.96	0.96	0.97
	C	0.97	0.97	0.97
LGBTQ	A	0.98	0.98	0.98
	B	0.98	0.97	0.98
	C	0.98	0.97	0.99
African	A	0.99	0.98	0.98
	B	0.98	0.98	0.98
	C	0.98	0.97	0.98
Jewish	A	0.98	0.97	0.98
	B	0.97	0.97	0.97
	C	0.96	0.95	0.97
Islam	A	0.98	0.98	0.99
	B	0.98	0.97	0.98
	C	0.97	0.97	0.98

Table: F1 scores as averaged across all layers for all topics and Tasks A (on- vs. off-topic), B (on- vs. off-topic in toxic posts) and C (on- vs. off-topic in non-toxic posts)

Results and discussions

Questions

- How well do the contextual embeddings capture information necessary to identify word-level toxicity within a post before and after fine-tuning?

Results

- Contextual embeddings **do not have sufficient capacity** to identify the individual toxic components of a post **even with fine-tuning**.
- Although all three models perform well in identifying toxic posts, they **do not necessarily understand why these toxic posts are toxic**.
- In other words, the models are relying on something else to help them make this determination.

Topic	Task	Model 1	Model 2	Model 3
Profanity	D	0.64	0.71	0.72
	E	0.59	0.59	0.60
Women	D	0.65	0.66	0.68
	E	0.70	0.72	0.72
LGBTQ	D	0.69	0.72	0.70
	E	0.68	0.70	0.69
African	D	0.71	0.73	0.73
	E	0.73	0.76	0.75
Jewish	D	0.72	0.72	0.72
	E	0.71	0.72	0.72
Islam	D	0.67	0.69	0.71
	E	0.63	0.65	0.66

Table: F1 scores as averaged across all layers for all topics and Tasks D (toxic on-topic tokens in toxic posts vs. non-toxic on-topic tokens in non-toxic posts) and E (toxic on-topic tokens in toxic posts vs. non-toxic on-topic tokens in toxic posts.)

Results and discussions

Questions

- How do different fine-tune approaches impact the ability of contextual embeddings in capturing information necessary to identify toxic speech?

Results

- The performance impact from fine-tuning **depends on tasks, topics and data** used in fine-tuning.
- Fine-tuning **may allow new information to be encoded in the earlier hidden layers** of the Transformer stack.

		Profanity		Women		LGBTQ		African		Jewish		Islam	
Task	Layer	Δ2	Δ3	Δ2	Δ3	Δ2	Δ3	Δ2	Δ3	Δ2	Δ3	Δ2	Δ3
D	12	0.13	0.09	0.02	0.04	**0.11**	**0.05**	0.01	0.04	**0.02**	**0.04**	**0.09**	0.08
D	11	**0.19**	**0.16**	0.02	0.02	0.07	0.03	**0.07**	0.04	**0.02**	0.00	**0.09**	**0.10**
D	10	0.16	0.15	0.03	0.04	0.08	**0.05**	0.06	**0.07**	-0.02	0.01	0.05	0.04
D	9	0.12	0.13	**0.04**	0.06	0.08	**0.05**	0.03	0.03	0.01	-0.01	0.01	0.05
D	8	0.09	0.11	**0.04**	**0.09**	0.02	0.00	0.04	0.03	-0.02	-0.01	0.03	0.06
D	7	0.06	0.09	0.00	0.06	0.01	0.00	0.01	0.01	**0.02**	0.01	0.01	0.04
D	6	0.07	0.11	0.02	0.03	0.04	0.03	0.00	0.02	**0.02**	0.03	-0.04	0.03
D	5	0.04	0.10	-0.01	0.01	0.03	0.03	-0.01	0.02	0.01	0.03	0.04	0.05
D	4	0.02	0.02	0.02	-0.01	0.01	0.01	0.04	0.03	0.01	-0.01	-0.01	0.03
D	3	0.04	0.06	0.01	0.04	0.01	0.00	0.00	-0.02	**0.02**	0.03	0.02	0.04
D	2	0.03	0.03	-0.02	0.02	0.02	-0.01	-0.01	0.01	-0.03	0.00	0.00	0.02
D	1	0.00	0.00	-0.01	0.03	-0.01	0.00	0.03	0.00	0.01	-0.01	-0.01	0.03
D	0	-0.01	-0.03	-0.03	-0.01	-0.03	-0.01	-0.01	0.01	0.01	-0.01	-0.03	-0.01
E	12	0.00	0.01	**0.02**	0.04	0.00	0.02	0.02	0.02	**0.04**	**0.04**	**0.06**	0.05
E	11	-0.01	-0.01	0.01	0.02	0.05	**0.05**	0.05	**0.06**	0.01	0.00	0.03	**0.06**
E	10	**0.02**	0.00	0.01	0.04	0.03	0.01	0.05	0.05	0.02	**0.04**	0.03	0.05
E	9	**0.02**	0.02	**0.02**	0.02	-0.01	-0.05	**0.06**	**0.06**	0.02	0.03	0.03	0.03
E	8	0.01	0.02	**0.02**	**0.06**	0.03	-0.02	**0.06**	0.03	-0.02	0.02	0.03	0.05
E	7	0.01	**0.05**	0.01	0.00	0.04	-0.01	0.02	0.00	0.02	0.00	0.00	**0.06**
E	6	**0.02**	0.02	0.00	0.01	-0.01	-0.03	**0.06**	0.01	0.01	-0.05	0.01	-0.01
E	5	-0.03	0.02	**0.02**	0.01	**0.07**	**0.05**	0.01	0.03	0.02	0.02	0.01	0.01
E	4	-0.02	0.00	-0.01	0.01	0.00	0.02	0.02	0.00	0.03	0.00	0.00	-0.04
E	3	0.01	0.02	**0.02**	0.00	0.00	0.03	-0.02	-0.02	0.02	0.00	0.00	0.03
E	2	-0.02	0.01	-0.01	-0.04	0.02	0.03	-0.02	-0.03	0.00	0.02	0.01	0.00
E	1	0.00	0.03	0.01	0.01	-0.02	-0.03	0.01	0.00	0.01	0.01	-0.02	0.00
E	0	-0.02	-0.02	**0.02**	0.01	0.02	**0.05**	0.02	0.04	0.01	**0.04**	**0.06**	0.03

Table: Change in F1 scores from fine-tuning across all layers for all topics and Tasks D (toxic on-topic tokens in toxic posts vs. non-toxic on-topic tokens in non-toxic posts) and E (toxic on-topic tokens in toxic posts vs. non-toxic on-topic tokens in toxic posts.) Δ2 and Δ3 denote performance differences between Model 2 and Model 1 and between Model 3 and Model 1, respectively. Figures in bold with green background denote highest layer-wise performance improvements.

The End

Toxic speech 101

Profanity

Speech that contains awords, phrases, or acronyms that are **impolite, vulgar, or offensive**.

Hate speech

Speech that **criticizes, insults, denounces, or dehumanizes a person or group** on the basis of an identity (such as race, ethnicity, gender, religion, sexual orientation, ability, and national origin).

Sexual

Speech that indicates **sexual interest, activity, or arousal** using direct or indirect references to body parts, physical traits, or sex.

Insults

Speech that includes **demeaning, humiliating, mocking, insulting, or belittling** language. This type of language is also labeled as **bullying**.

Violence

Speech that includes **threats seeking to inflict pain, injury, or hostility** toward a person or group.

Graphic

Speech that uses **visually descriptive** and **unpleasantly vivid imagery**. This type of language is often intentionally verbose to amplify a recipient's discomfort.

Abusive

Speech intended to **affect the psychological well-being** of the recipient, including demeaning and objectifying terms. This type of language is also labeled as **harassment**.

Source: https://docs.aws.amazon.com/transcribe/latest/dg/toxicity.html

Definitions and probing tasks

Definitions

The following is an illustrative example for the topic of women:

Toxic post: “that **girl** is a dirty **slut**”
Non-toxic post: “meryl streep in little **women**”

Red and green highlights denote toxic and non-toxic words, respectively. **Bold** and unbold words denote on-topic and off-topic words, respectively.

Probing tasks

Positive samples
Negative samples

Task A: On-topic words vs. off-topic words

	On-topic word		Off-topic word	
Toxic post	Toxic word	Non-toxic word	Toxic word	Non-toxic word
Non-toxic post	Toxic word	Non-toxic word	Toxic word	Non-toxic word

Task B: Within toxic posts, on-topic words vs. off-topic words

	On-topic word		Off-topic word	
Toxic post	Toxic word	Non-toxic word	Toxic word	Non-toxic word
Non-toxic post	Toxic word	Non-toxic word	Toxic word	Non-toxic word

Task C: Within non-toxic posts, on-topic words vs. off-topic words

	On-topic word		Off-topic word	
Toxic post	Toxic word	Non-toxic word	Toxic word	Non-toxic word
Non-toxic post	Toxic word	Non-toxic word	Toxic word	Non-toxic word

Task D: For on-topic words, toxic words in toxic posts vs. non-toxic words in non-toxic posts

	On-topic word		Off-topic word	
Toxic post	Toxic word	Non-toxic word	Toxic word	Non-toxic word
Non-toxic post	Toxic word	Non-toxic word	Toxic word	Non-toxic word

Task E: For on-topic words in toxic posts, toxic words vs. non-toxic words

	On-topic word		Off-topic word	
Toxic post	Toxic word	Non-toxic word	Toxic word	Non-toxic word
Non-toxic post	Toxic word	Non-toxic word	Toxic word	Non-toxic word

Probing tasks: Illustrative example

The following are illustrative examples for the topic of women. Red and green highlights denote toxic and non-toxic words, respectively:

Task description	Positive samples	Negative samples
A: On-topic words vs. off-topic words	Post: “meryl streep in little women” On-topic tokens: “women”	Post: “sister is a dirty slut” Off-topic tokens: “is”, “a”, “dirty”
B: On-topic words in toxic posts vs. off-topic words in toxic posts.	Toxic post: “sister is a dirty slut” On-topic tokens: “sister”, “slut”	Toxic post: “sister is a dirty slut” Off-topic tokens: “is”, “a”, “dirty”
C: On-topic words in non-toxic posts vs. off-topic words in non-toxic posts.	Non-toxic post: “meryl streep in little women” On-topic tokens: “women”	Non-toxic post: “meryl streep in little women” Off-topic tokens: “meryl”, “streep”, “in”, “little”
D: Toxic on-topic words in toxic posts vs. non-toxic on-topic words in non-toxic posts.	Toxic post: “sister is a dirty slut” Toxic on-topic tokens: “slut”	Non-toxic post: “meryl streep in little women” Non-toxic on-topic tokens: “women”
E: Toxic on-topic words in toxic posts vs. non-toxic on-topic words in toxic posts.	Toxic post: “sister is a dirty slut” Toxic on-topic tokens: “slut”	Toxic post: “sister is a dirty slut” Non-toxic on-topic tokens: “sister”