You’re not sexist...

...but what you just wrote might be.

Arnaud Blanchard, Carina Prunkl, Marion Gagnadre & Elizabeth den Dulk

Sexism + NLP, why?

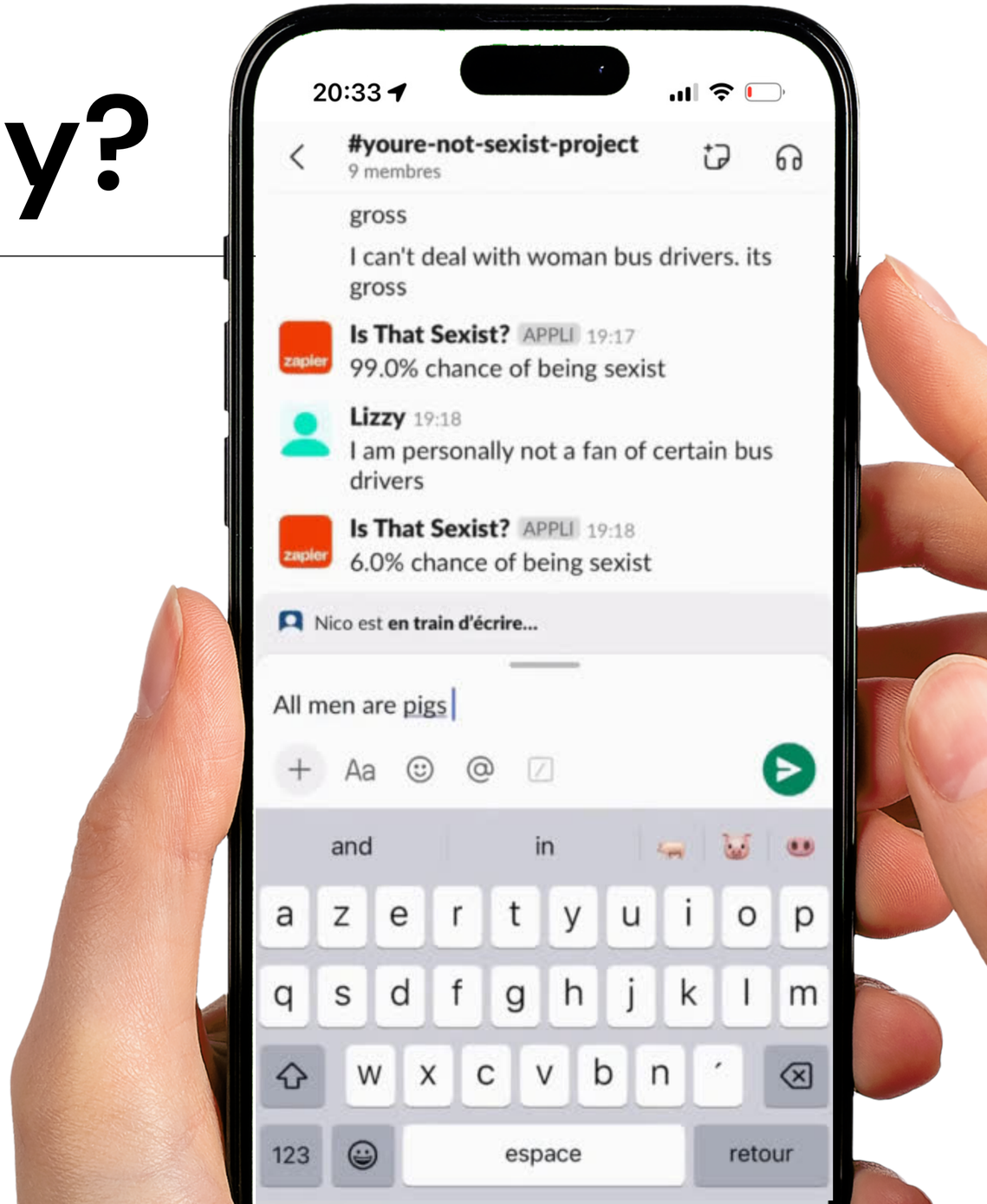
SEX IST

"The proper cure for feminism is hard dick."

"A man who doesn't pay the bill comes across as weak"

"F*%^ing women drivers!!!"

NOT SEX IST



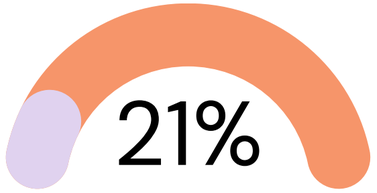
How NLP Models can Help

CONSISTENCY

CONSISTENT RESULTS WITH NO AGENDA

EDUCATION

FEEDBACK LOOP = POSITIVE REINFORCEMENT



labelled Sexist

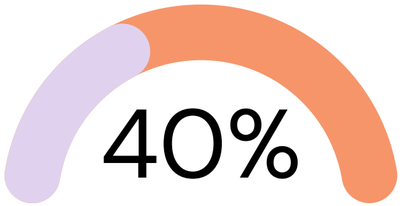
52,245

Rows of unique text

6 Datasets

Merged from Existing Models/Academic Papers

Grammar is SO important! Don't you just hate it when people mix up "Their" and "They're"? 😄 There so dumb!



labelled Sexist

⚖️ **1. Data Balancing**

😜@# **2. Data Cleaning**

grammar is so important dont you just hate it when people mix up their and theyre :grinning: there so dumb

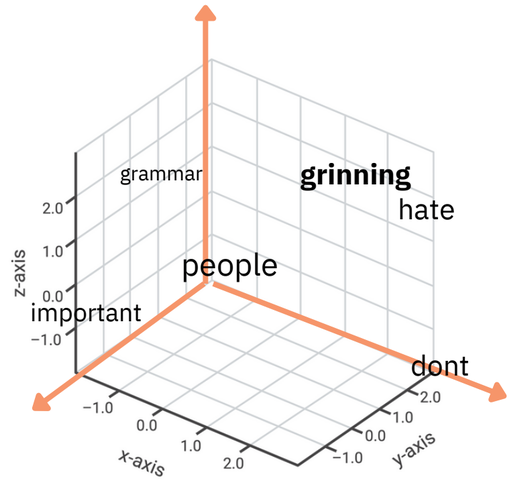
27,920

Rows of unique text

grammar is so important dont you just hate it when people mix up their and theyre :grinning: there so dumb

Preprocessing

- Tokenizing → ['grammar', 'is', 'so', 'important', 'dont', 'you', 'just', 'hate', 'it', 'when', 'people', 'mix', 'up', 'their', 'and', 'theyre', ':', 'grinning', ':', 'there', 'so', 'dumb']
- Remove Stop Words/Lemmatizing → ['grammar', 'important', 'dont', 'hate', 'people', 'mix', ':', 'grinning', ':', 'dumb']
- Vectorizing → [[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]]
- Padding/Truncating → [[1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 0 0 0 0 0 0 0 0]]
- Embedding
  - Word2Vec
  - GloVe



First Models

❌ % of sexist phrases detected by the model that were actually sexist

	CATEGORY	ACCURACY	RECALL	PRECISION
MULTINOMIAL	BASELINE	0.74	0.74	0.67
 GRU	RNN	0.82	0.48	0.62
 BI-LSTM	RNN LSTM WITH BI DIRECTIONAL LAYER	0.80	0.46	0.75

✅ % of sexist phrases correctly detected

BERT-IST

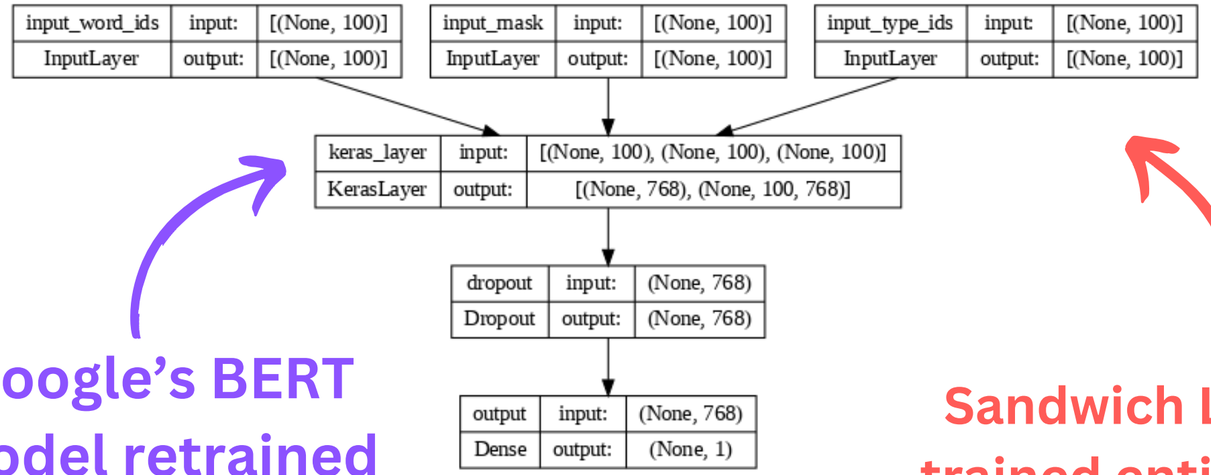
	ACCURACY	RECALL	PRECISION
FINE TUNED BERT → TRANSFORMERS + TRANSFER LEARNING	0.82	0.88	0.83

- ❌ Removed 1 dataset
  - Badly annotated + Accuracy & Recall at 0.28 😱
- ✍️ Added context to workplace dataset
  - He said to her at work “I'll explain to you with simple words so that you understand”
- 🚮 Kept punctuation
  - Adds context, like when someone’s YELLING!!!, sarcasm... etc... 🙄

(for the nerds)

5GB, 1 hour with a V100 GPU and extra RAM... 👹

✅ **88% OF SEXIST PHRASES CORRECTLY DETECTED**



Google’s BERT model retrained on our data

Sandwich Layers trained entirely on our data

Where to go next?

	RECOUNTING	*“People being sexist is wrong - and only adults are wrong”*
	SLANG	*“I'm biting my *** so mf hard this bitch might fall off!!!”*
	AMBIGUITY	*“So at what level of tolerance should selfish women be tolerated?”*
	CONTEXT	*“Nah - women live life without a care!”*

IMPROVED MODEL

FURTHER RESEARCH

MORE TRAINING DATA

Let’s put it to the test.

Sexism Generator + Detector

A live action interface, connected to our model, able to detect and highlight sexism

Slack Bot

- Avoid a #MeToo moment!
- Protect your future political or business career!
- Avoid warding off potential dates by offending their gender!
- Make sure people don’t confuse you with Gerard Depardieu!