Using BERT to investigate cognitive biases in thinking of everyday unexpected events

Molly S. Quinn, University College Dublin

1 | Example Language Model Use Case

2 | Resources for Getting a Jump-Start on Text Classification with BERT

Contents

Predicting Unexpected Events

In a foreign city and worried about missing a flight?

Q: What unexpected things could occur?

- Traffic is really bad.
- Buses are unsafe.
- Taxis are unreliable.
- Police outriders escort me to the airport.

What is the Unexpected?

- Subjective probability?
- What objects do people mutate in the unexpected?
- What are the features of the event?

  Valence, controllability, goals...?

Example Material

Sentence Type	Scenario
Goal	Louise wants to shop at an expensive clothes store.
Intermediate Event	She is wearing her favourite dress and matching shoes.
Plan Step	Louise draws money from the ATM.

Instructions

Unexpected: Something unexpected occurred. What do you think happened?

Expected: Assuming nothing unexpected occurs, what do you expect to typically happen next?

Example Responses

Unexpected

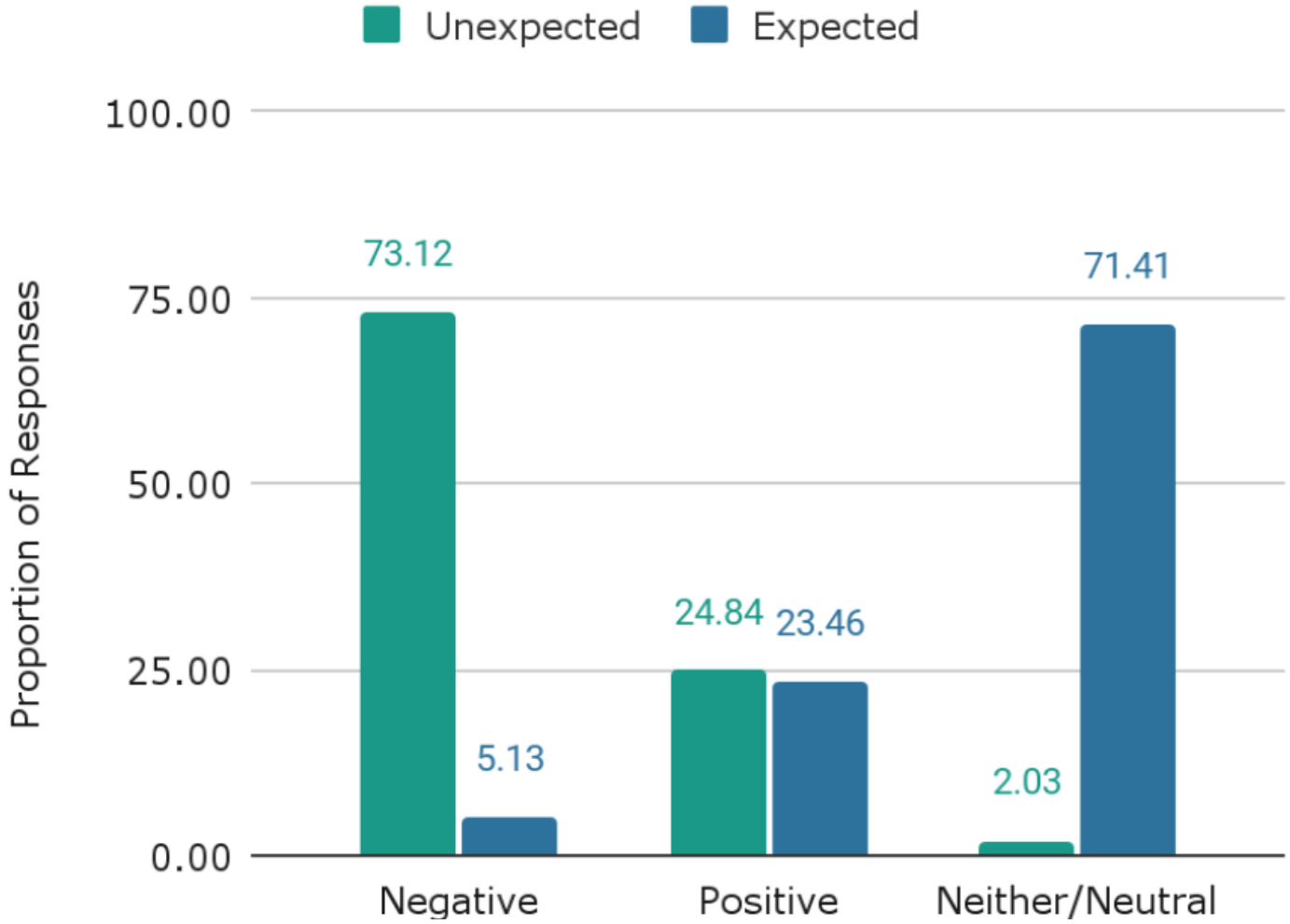
- “She saw the account balance as she withdrew money and realized she didnt have enough to justify the shopping trip”
- “She is robbed at the ATM”
- “She had more money than she realised so could buy something expensive.”

Expected

- “She will go into the expensive clothes store and look around. If there is something she likes she may try it on and she may buy some clothes”
- “I would expect Louise to take her money to the clothes store and spend her money.”
- “she buys something to match her outfit”

Valence

Prediction of unexpected events is valenced and skewed towards the **negative**.



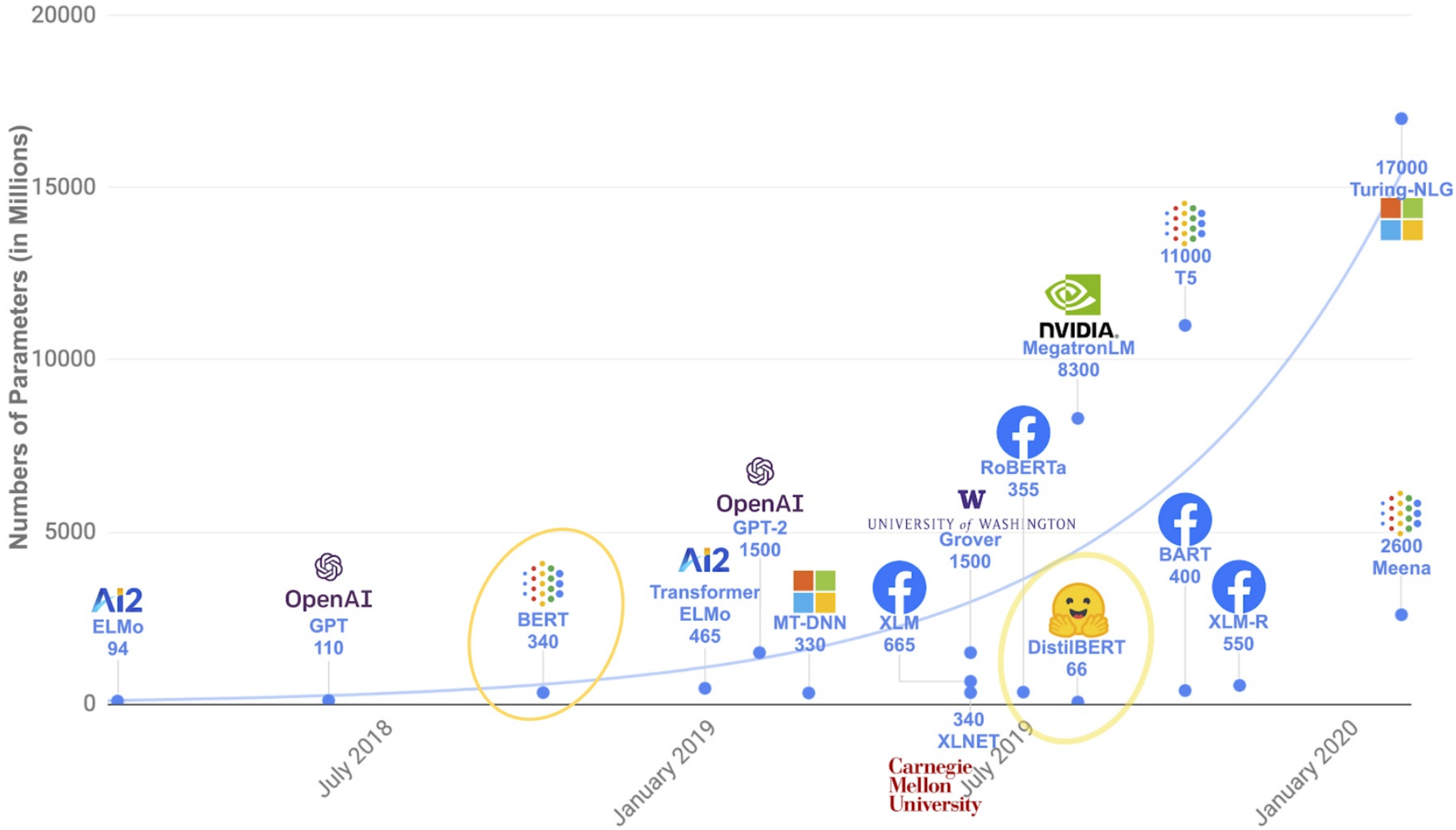
Contents

What is the Goal?

- Some say that language models can serve as models of “psychological semantics”
  - Meaning computational word representations are similar to human mental word representations
- But others say this is the wrong way to look at it.

“. . . our critique is that the current [Language Models] are too strongly linked to complex text-based patterns, and too weakly linked to world knowledge”

LAKE, B. M. & MURPHY, G. L. (2021). WORD MEANING IN MINDS AND MACHINES. PSYCHOLOGICAL REVIEW. ADVANCE ONLINE PUBLICATION. HTTPS://DOI.ORG/10.1037/REV0000297


20000
15000
10000
5000
0
Numbers of Parameters (in Millions)
ELMo
94
OpenAI
GPT
110
BERT
340
Transformer
ELMo
465
OpenAI
GPT-2
1500
MT-DNN
330
XLM
665
UNIVERSITY of WASHINGTON
Grover
1500
340
XLNET
Carnegie
Mellon
University
RoBERTa
355
NVIDIA
MegatronLM
8300
DistilBERT
66
11000
T5
BART
400
XLM-R
550
17000
Turing-NLG
2600
Meena
July 2018
January 2019
July 2019
January 2020

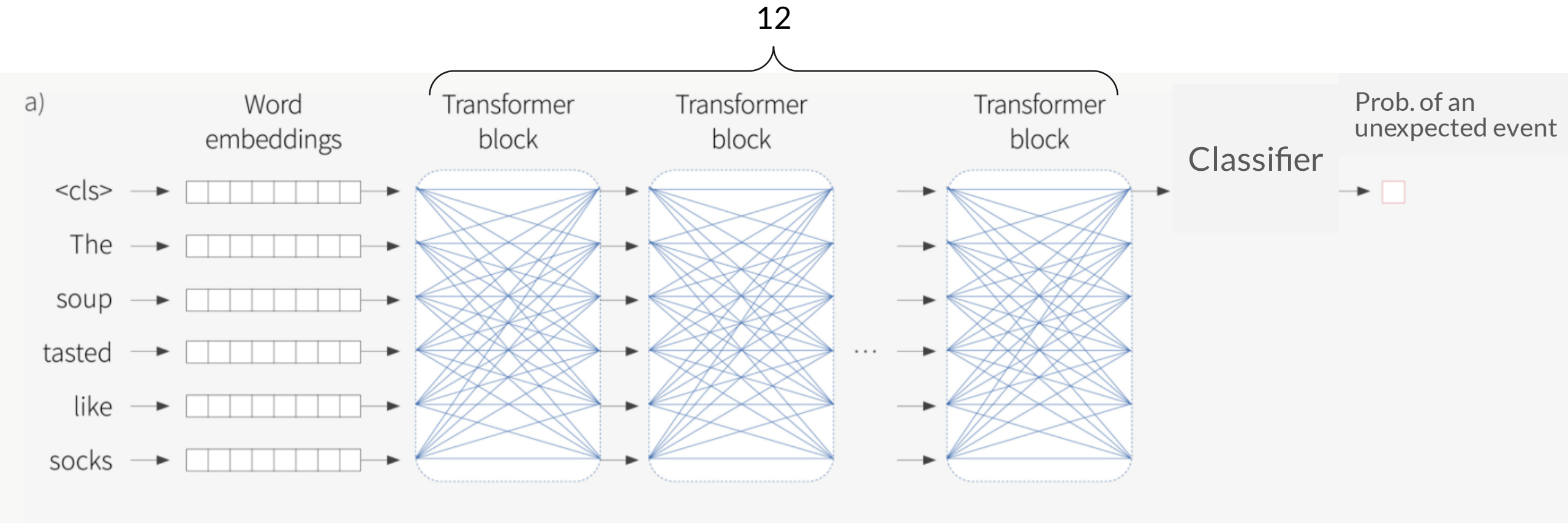
Why BERT, Specifically?

- Short texts (max 512 tokens)
- Convenient! (open-source)

How Does BERT Work?

a cursory runthrough

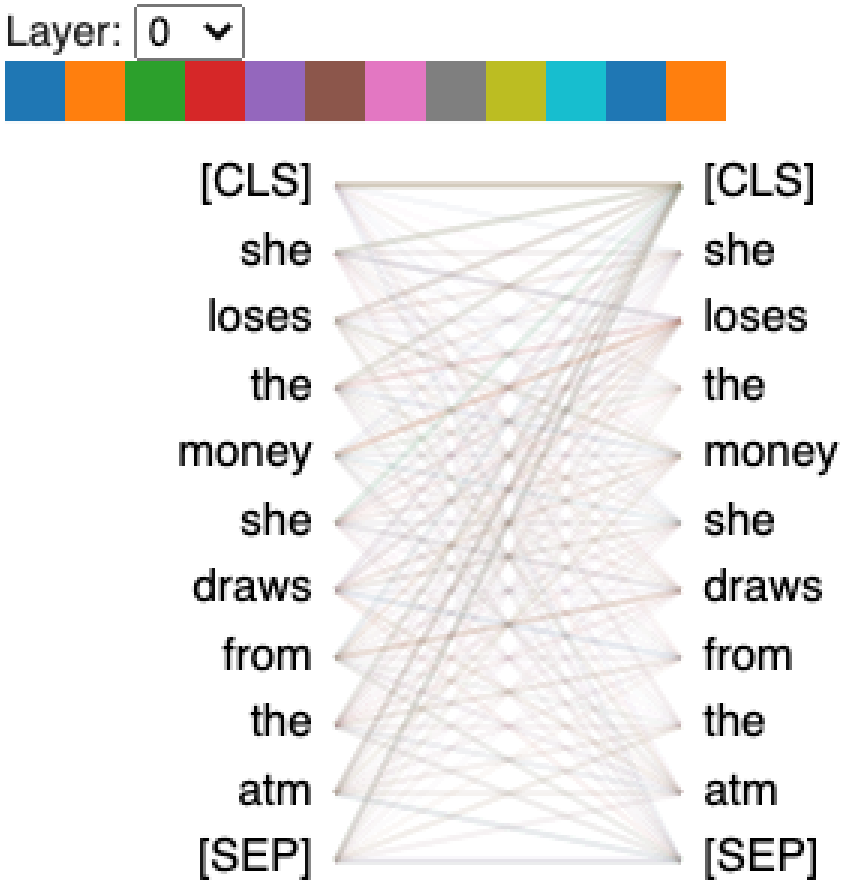
BERT Layers



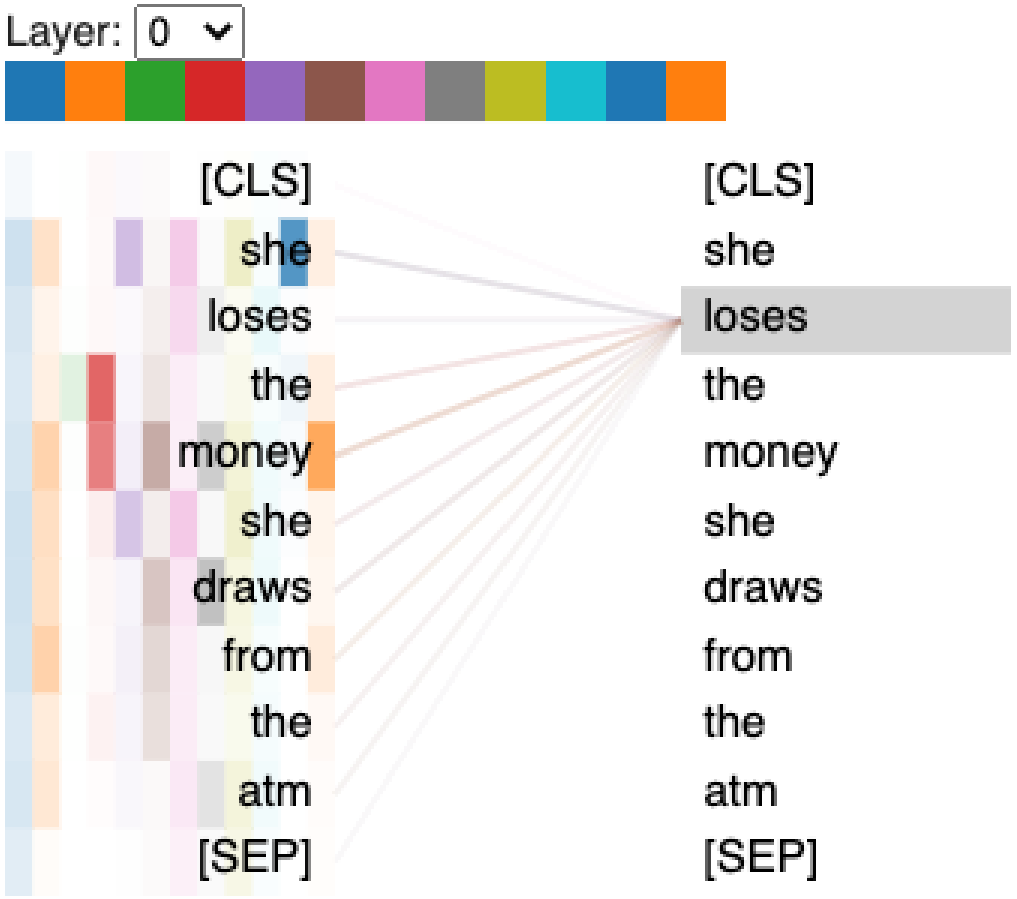
https://www.borealisai.com/en/blog/tutorial-14-transformers-i-introduction/

Each word is informed, to some degree, by every word

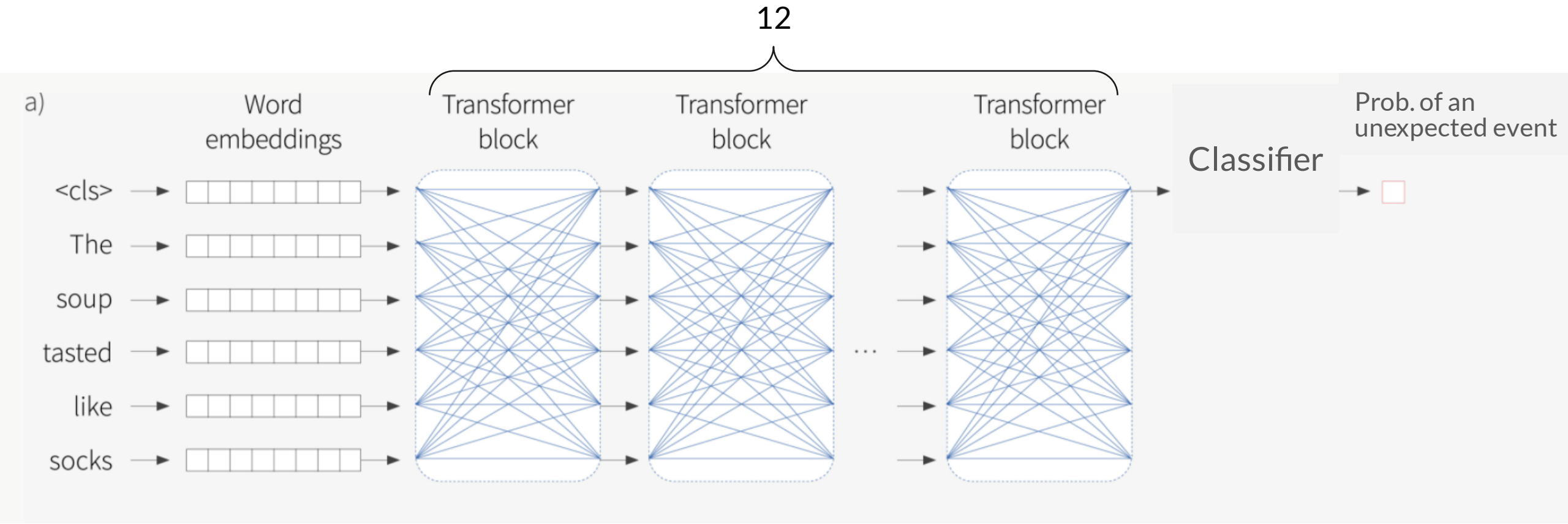
All Attention Weights



Attention Weights for “loses”



BERT Layers



https://www.borealisai.com/en/blog/tutorial-14-transformers-i-introduction/

Resources

huggingface Transformers

https://huggingface.co/docs/transformers/

BERT for Humanists

http://www.bertforhumanists.org/

https://www.youtube.com/watch?v=UmyOhI9AciI

https://colab.research.google.com/drive/19jDqa5D5XfxPU6NQef17BC07xQdRnaKU?usp=sharing

Current Experiments

1 | Logistic Regression (baseline)

2 | Fine-Tuned BERT

3 | Fine-Tuned DistilBERT

4 | LIME Explanations

Train/Test Split

Test Set	Training Subset 1	Training Subset 2	Training Subset 3
steve_gardening	rebecca_swimming	katie_kitten	anna_interview
louise_shopping	sally_wine	lucy_loan	michael_tea
alan_plane	karen_bus	belinda_meeting	robert_essay
edith_exam	bob_job	peter_college	sam_driving
mary_food	bill_holiday	john_party	sean_call

General Method

1 | Logistic Regression

- Using BERT (WordPiece) Encodings
- TF-IDF Vectors
- Binary logistic regression model
- Trained on all 3 training sets
- Tested on test set

2 | BERT Models

- Hyperparameter search by crossvalidation:
- Fine-Tune on 2 training sets, and validate on the 3rd training set
- Choose best hyperparemeters for final model training
- Final Models:
- Fine-Tuned on all 3 training sets
- Tested on test set

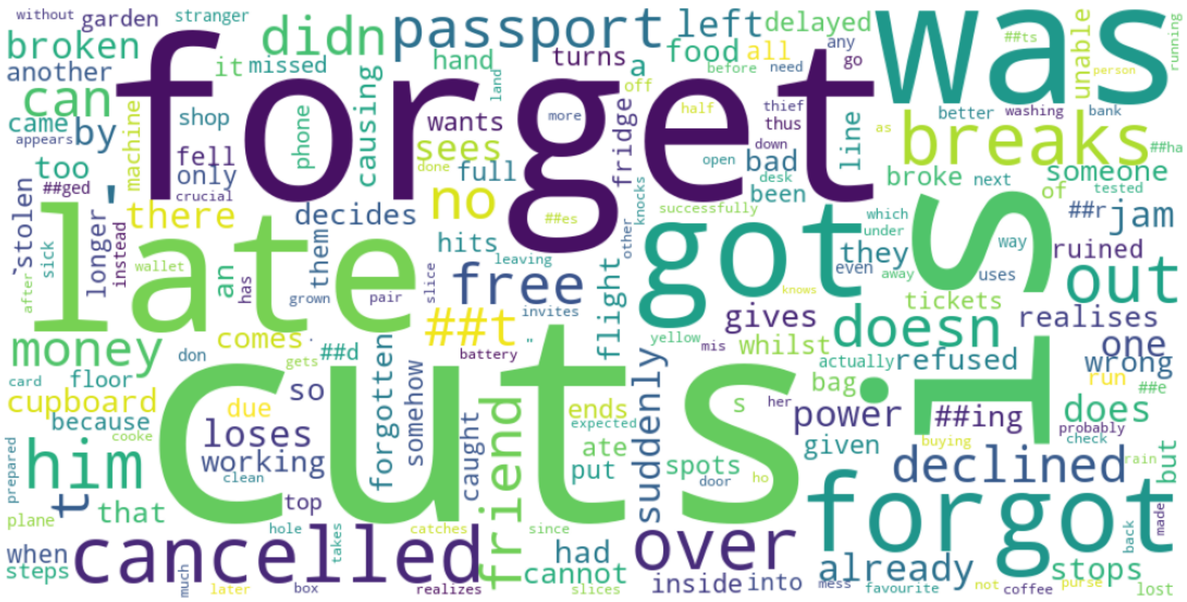
Results

	Logistic Regression			BERT Not Fine-Tuned			
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**	**Support**
expected	0.89	0.68	0.77	0	0	0	195
unexpected	0.70	0.89	0.78	0.45	**1.00**	0.62	160
accuracy			0.78			0.45	355
macro avg	0.79	0.79	0.78	0.23	0.50	0.31	355
weighted avg	0.80	0.78	0.78	0.20	0.45	0.28	355
	DistilBERT Fine-Tuned			**BERT Fine-Tuned**			
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**	**Support**
expected	**0.95**	0.85	**0.90**	0.92	**0.87**	0.89	195
unexpected	0.84	0.95	**0.89**	**0.85**	0.91	0.88	160
accuracy			**0.90**			0.89	355
macro avg	0.90	0.90	**0.90**	0.89	0.89	0.89	355
weighted avg	0.90	0.90	**0.90**	0.89	0.89	0.89	355

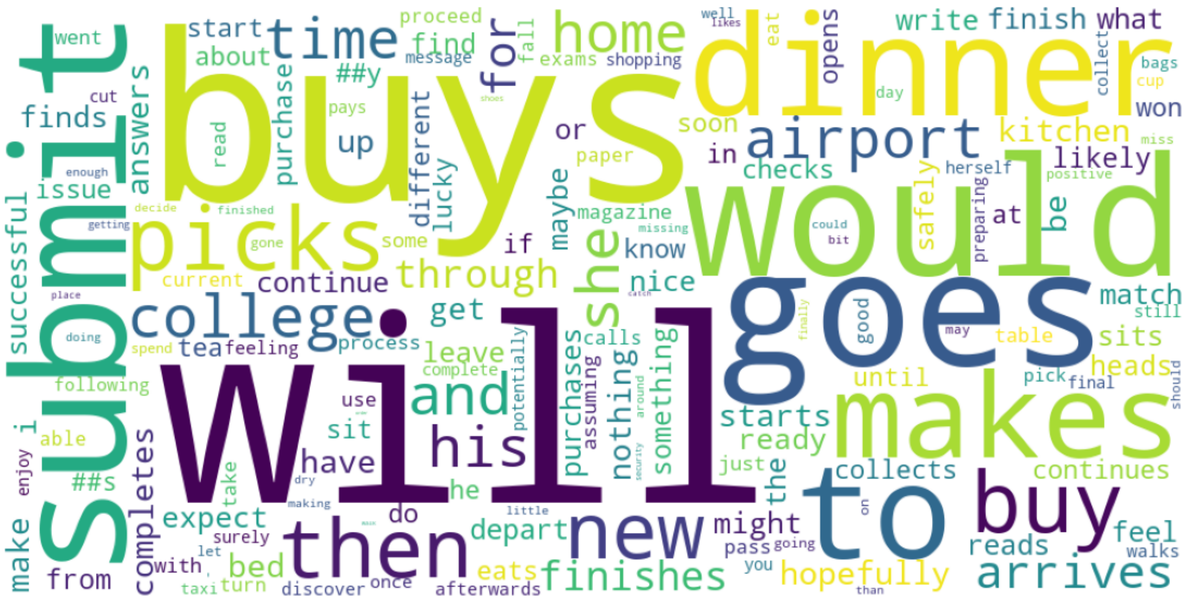
Contents

Logistic Regression: TF-IDF Weighted Coefficients

Unexpected



Expected



LIME (Locally Interpretable Model-Agnostic Explanations)

id	text	
1 (input)	Someone robs her	
2	[MASK] robs her	
3	Someone [MASK] her	
...	...	
n	Someone robs [MASK]	

LIME (Locally Interpretable Model-Agnostic Explanations)

id	text	prediction
1 (input)	Someone robs her	unexpected
2	[MASK] robs her	unexpected
3	Someone [MASK] her	expected
...	...	...
n	Someone robs [MASK]	unexpected

LIME (Locally Interpretable Model-Agnostic Explanations)

id	text	prediction
1 (input)	Someone robs her	unexpected
2	[MASK] robs her	unexpected
3	Someone [MASK] her	expected
...	...	...
n	Someone robs [MASK]	unexpected

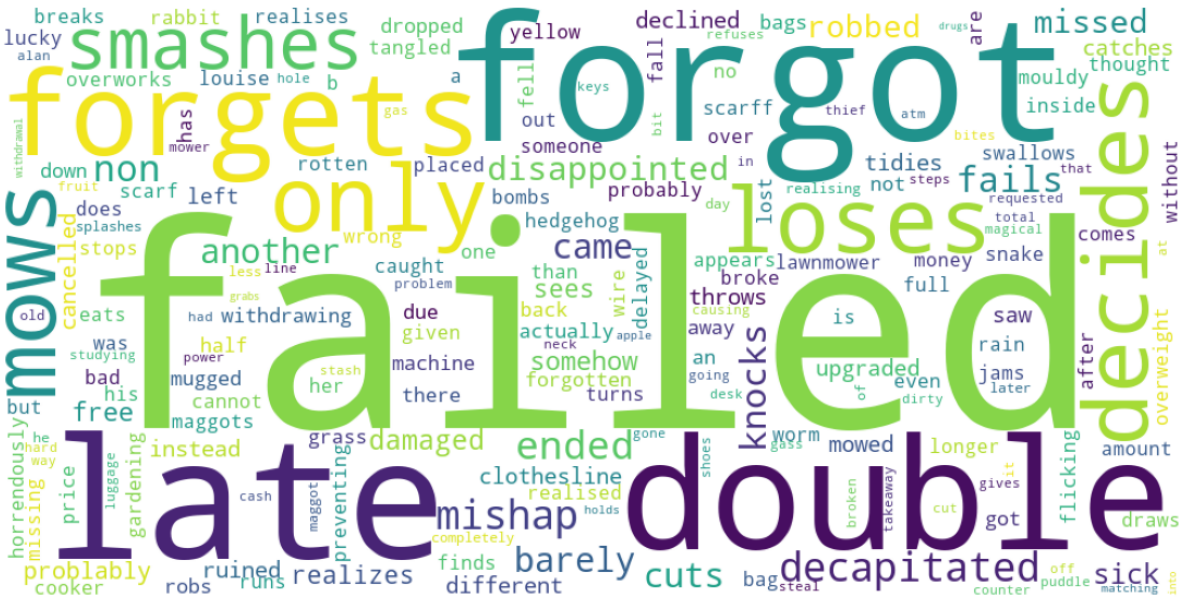
BERT Average LIME Weights

Unexpected

Expected

DistilBERT Average LIME Weights

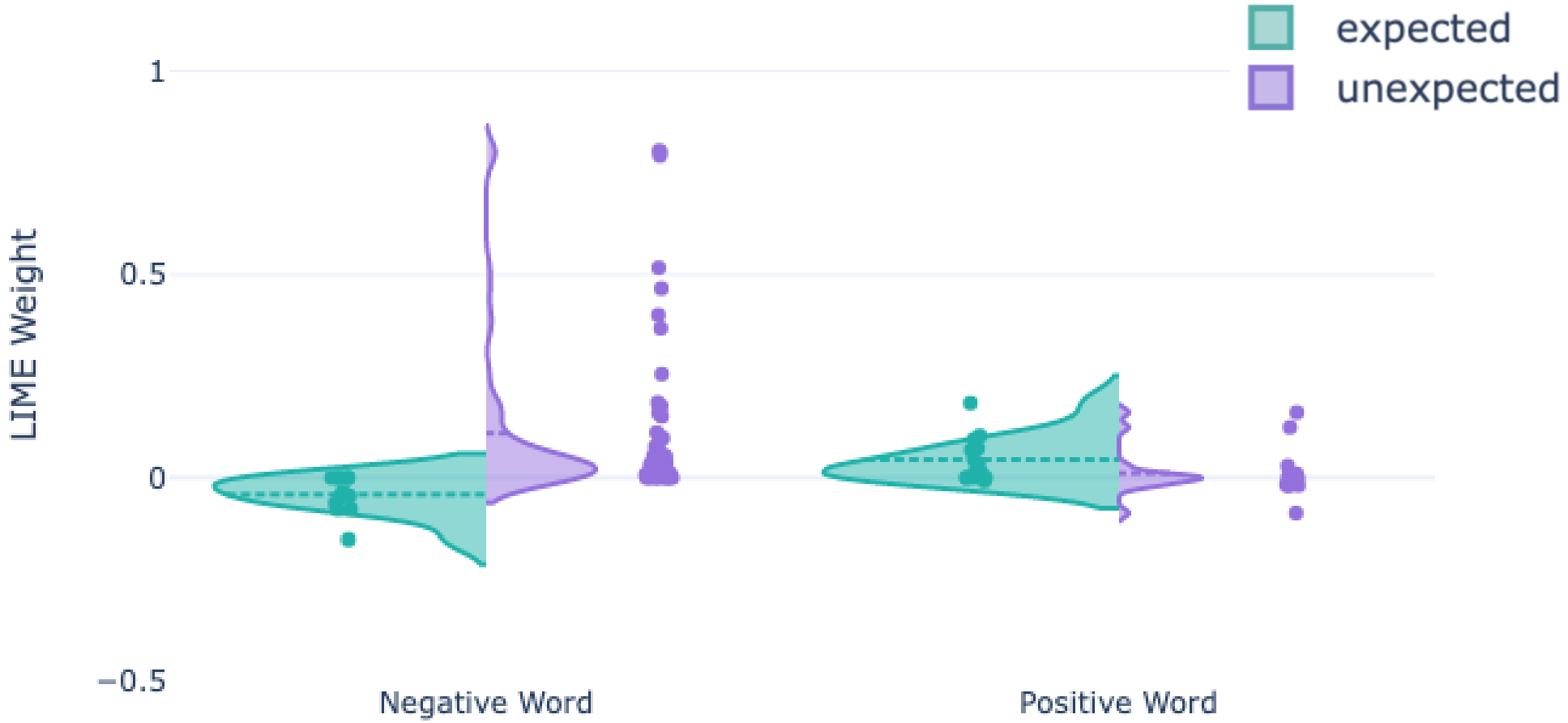
Unexpected



Expected

BERT Classifications

DistilBERT Classifications



True Class: unexpected
BERT Predicted Class: unexpected
Log Predicted Class: unexpected
DistilBERT Predicted Class: unexpected

	[CLS]_0	she_1	loses_2	the_3	money_4	she_5	draws_6	from_7	the_8	atm_9
ic Regression	0.0	-0.55	0.27	-0.23	0.36	-0.55	0.0	-0.19	-0.23	0.0
IME on BERT	0.0	0.0	0.89	-0.01	-0.03	0.0	0.0	0.01	-0.01	0.0
on DistlBERT	0.0	-0.15	0.47	-0.09	0.07	-0.15	0.07	0.0	-0.09	0.0

True Class: expected
BERT Predicted Class: expected
Log Predicted Class: expected
DistilBERT Predicted Class: expected

	[CLS]_0	she_1	draws_2	the_3	money_4	from_5	the_6	atm_7	and_8	goes_9	shopping_10
ic Regression	0.0	-0.26	0.0	-0.22	0.34	-0.18	-0.22	0.0	-0.38	-0.71	-0.06
IME on BERT	0.0	0.0	0.0	0.01	0.01	0.0	0.01	0.0	0.02	0.02	0.02
on DistlBERT	0.0	0.0	0.0	0.16	0.0	0.0	0.16	0.0	0.12	0.16	0.19

Contents

Resources

huggingface Transformers

https://huggingface.co/docs/transformers/

BERT for Humanists

http://www.bertforhumanists.org/

https://www.youtube.com/watch?v=UmyOhl9Acil

https://colab.research.google.com/drive/19jDqa5D5XfxPU6NQef17BC07xQdRnaKU?usp=sharing

Interpretable Machine Learning

https://christophm.github.io/interpretable-ml-book/

Thank you!

1 | People believe the unexpected is negative and uncontrollable.

2 | BERT models this negativity bias during fine-tuning.

Contact me:


molly.quinn@ucdconnect.ie


@Molly_S_Quinn


scholar.google.com/citations?user=M677etQAAAAJ

Watch Dr. Maria Antoniak's Tutorial:

BERT for Computational Social Scientists

https://www.youtube.com/watch?v=UmyOhl9AcIl

Limitations

- May be overfitting to training data
- Explanation method does not explain the model itself
- Questionable generalizability
  - to other datasets
  - other languages, and
  - behavioral/cultural factors

Future Directions

- Other explanation methods…
- Generating unexpected events
- Explicitly incorporating sentiment/etc. into classification/generation of unexpected events.