



USING CONTEXTUAL EMBEDDINGS

Sandeep Soni

10/30/2023

PS2

- overall_df: The data frame that you create by merging the data frame of metadata and the data frame of topic probabilities
- per_decade_plot: It's a data frame that contains the average probability of each topic for every decade.

REVIEW OF LANGUAGE MODELS

- Model the probability of sequence directly

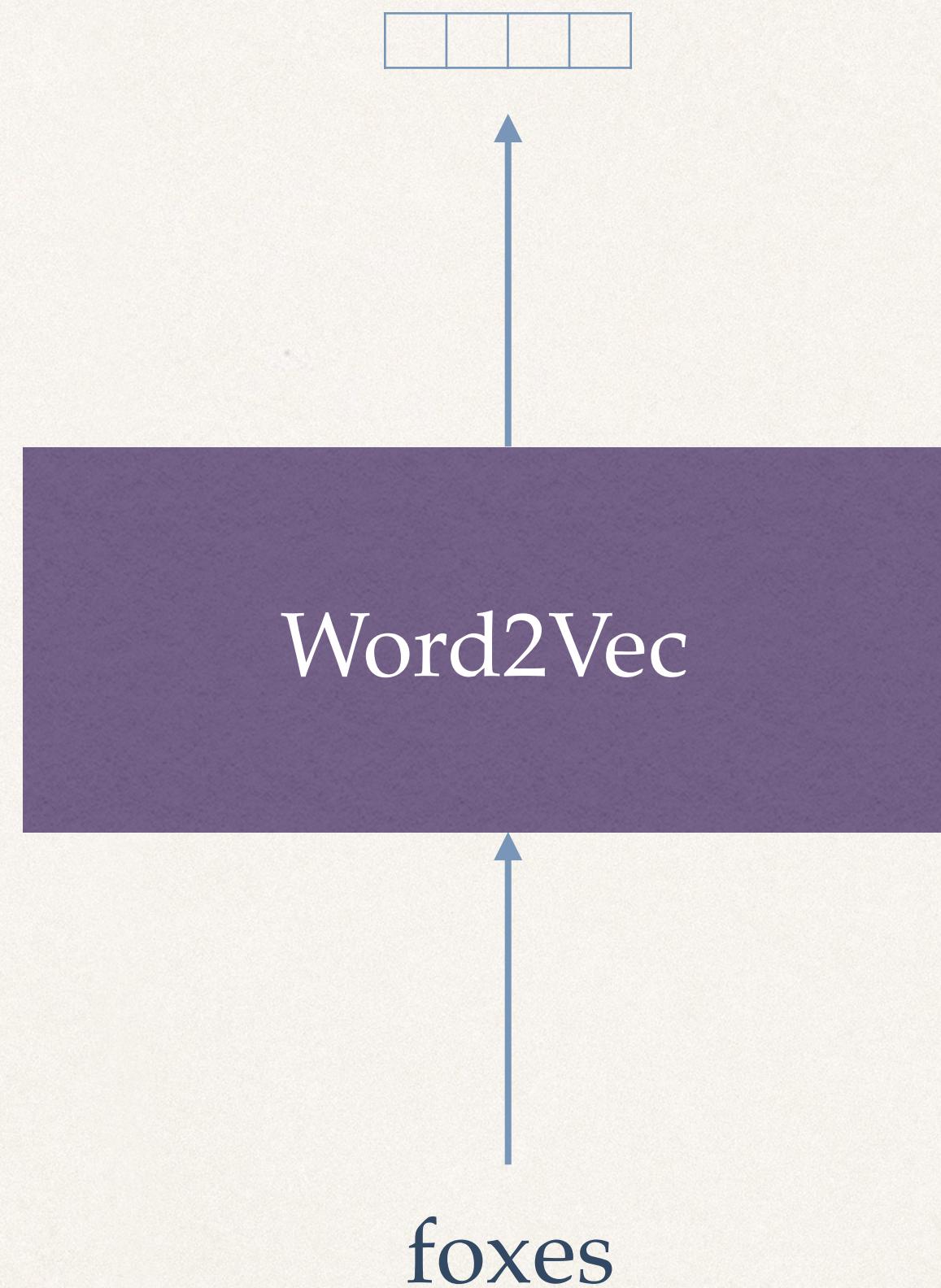
$$P(x) = \prod_i P(x_i | x_{i-1}, \dots, x_{i-n+1})$$

N-gram

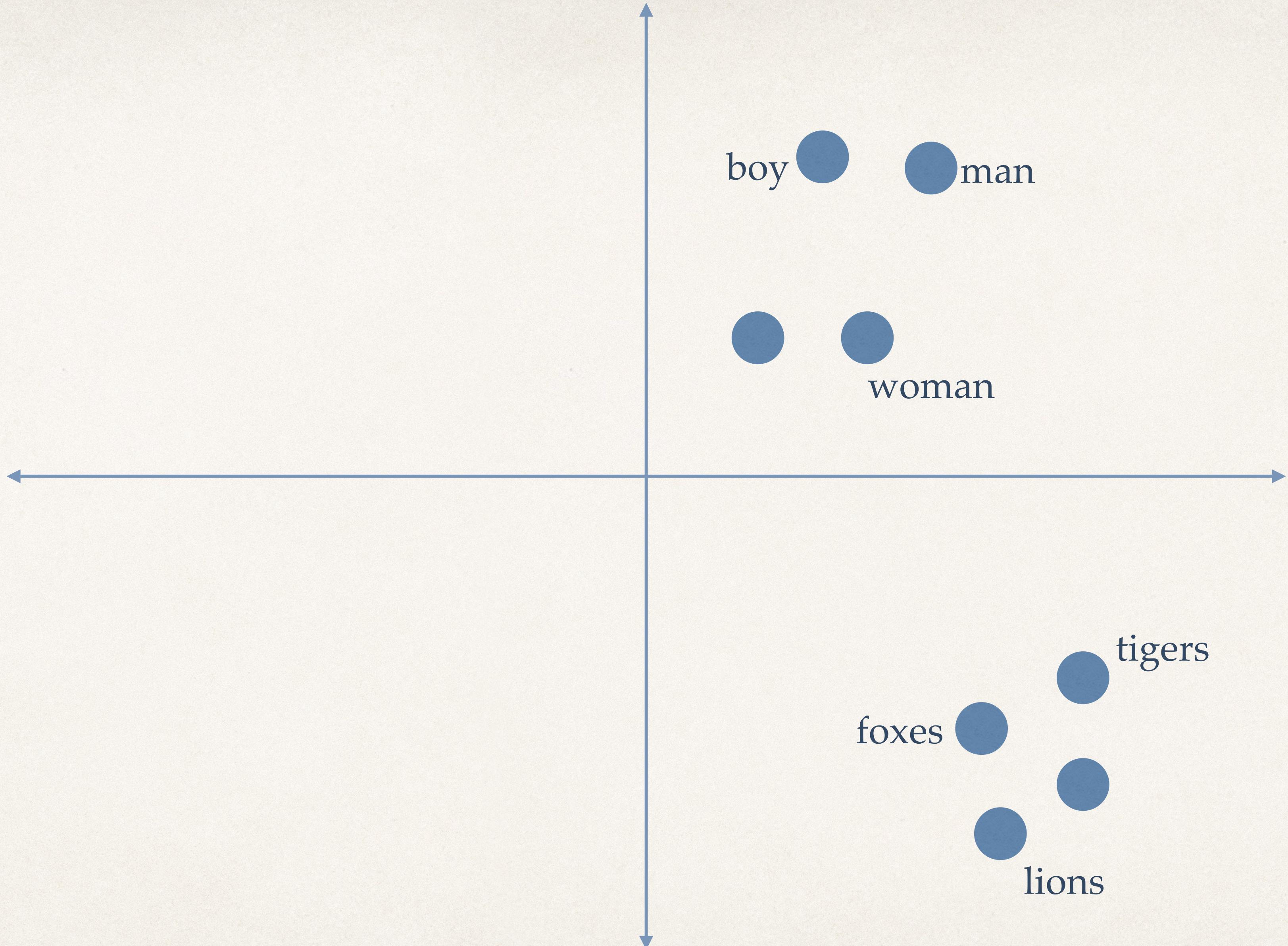
- Model the probability of word given some context

$$P(w | c) = \frac{\exp(\beta_w v_c)}{\sum_{w'} \exp(\beta_{w'} v_c)}$$

CBoW, RNNLM,
Transformers



The abstraction is that once we have a trained word2vec model, then we input a word and we get one embedding for that word



Words that are closer in meaning are geometrically closer

- The foxes ran at us
- The foxes won the soccer game



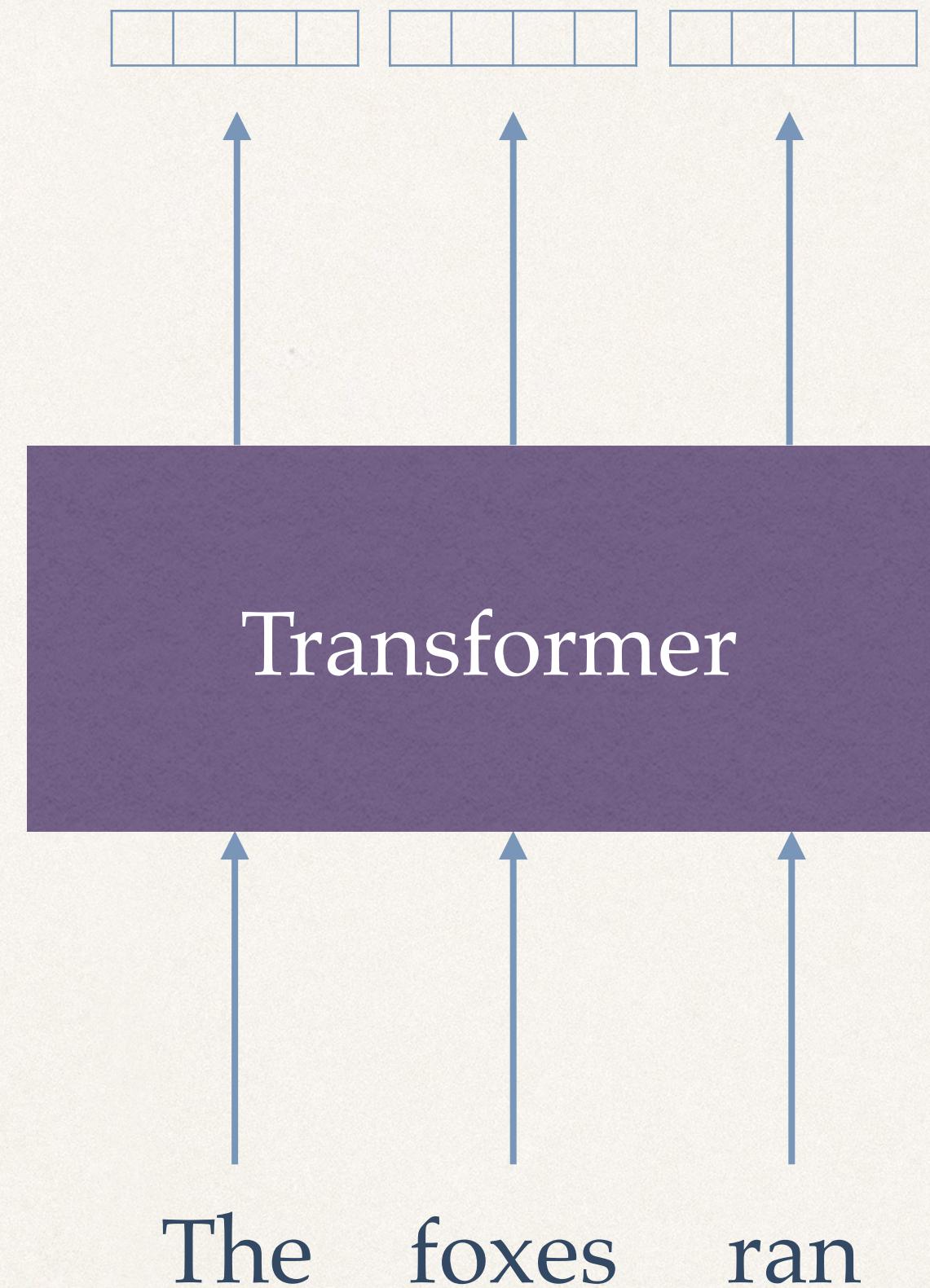
Ideally, we want to give a different vector representation to every instance of a word

TYPES & TOKENS

- Type: foxes
- Tokens:
 - The foxes ran at us
 - The foxes won the soccer game

1.2	-0.1	0.7	-0.5
-----	------	-----	------

-0.4	0.6	0.8	0.4
------	-----	-----	-----



The abstraction is that once we have a trained transformer model, then we input a sequence and we get one embedding for every token in the sequence



Words which
are
contextually
closer in
meaning are
geometrically
closer

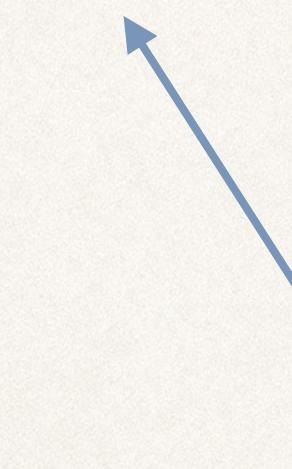
CONTEXTUALIZED WORD VECTORS

Transform the static embedding of a token to an embedding that is sensitive to the local context in which the token appears

We further want these embeddings to be learned from training data and which can be updated for any task

MASKED LANGUAGE MODELING

The _____ won the game



Fill in the blank by
using the surrounding
context

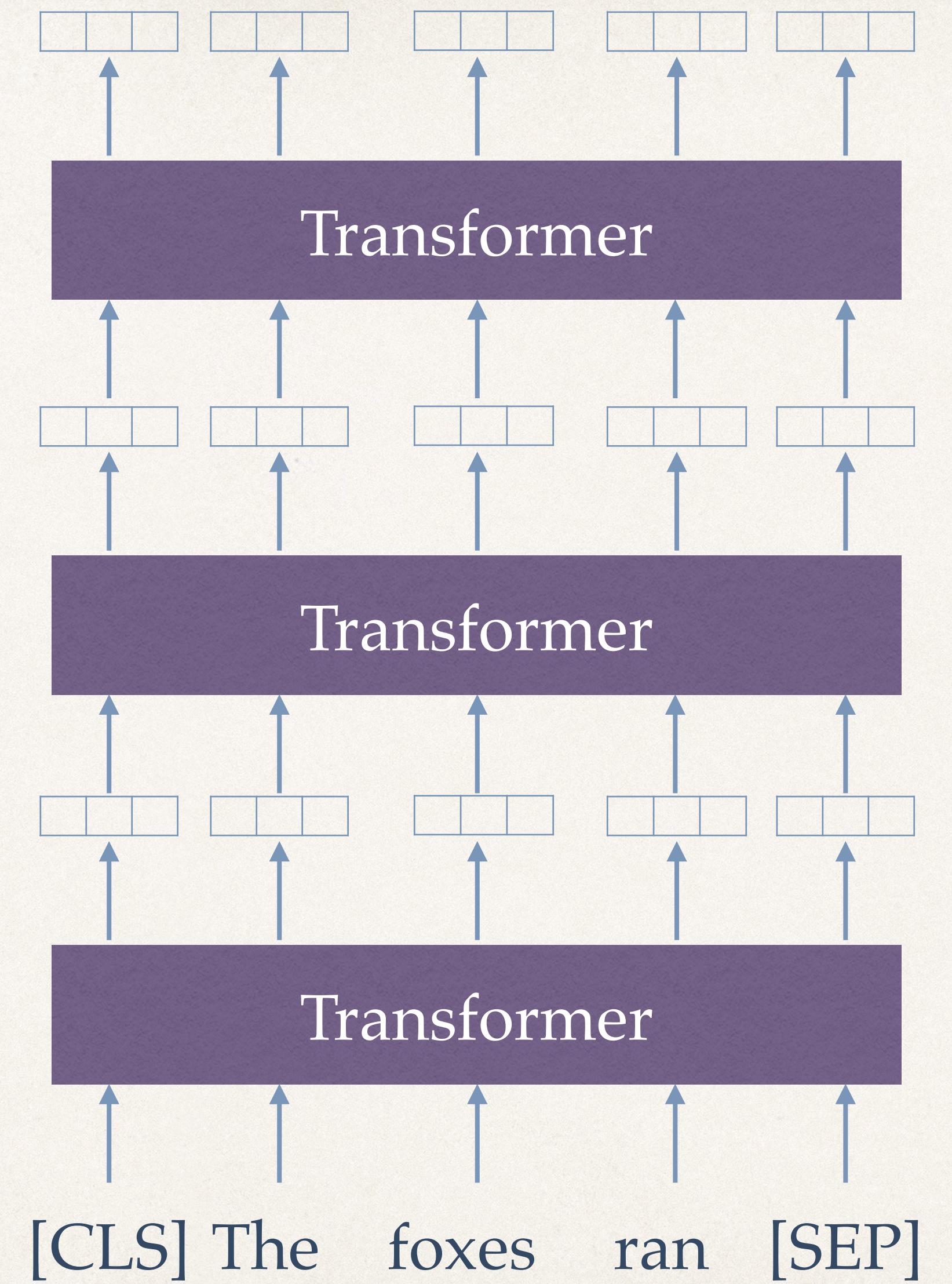
$$P(w_t | \neg w_t)$$

BERT

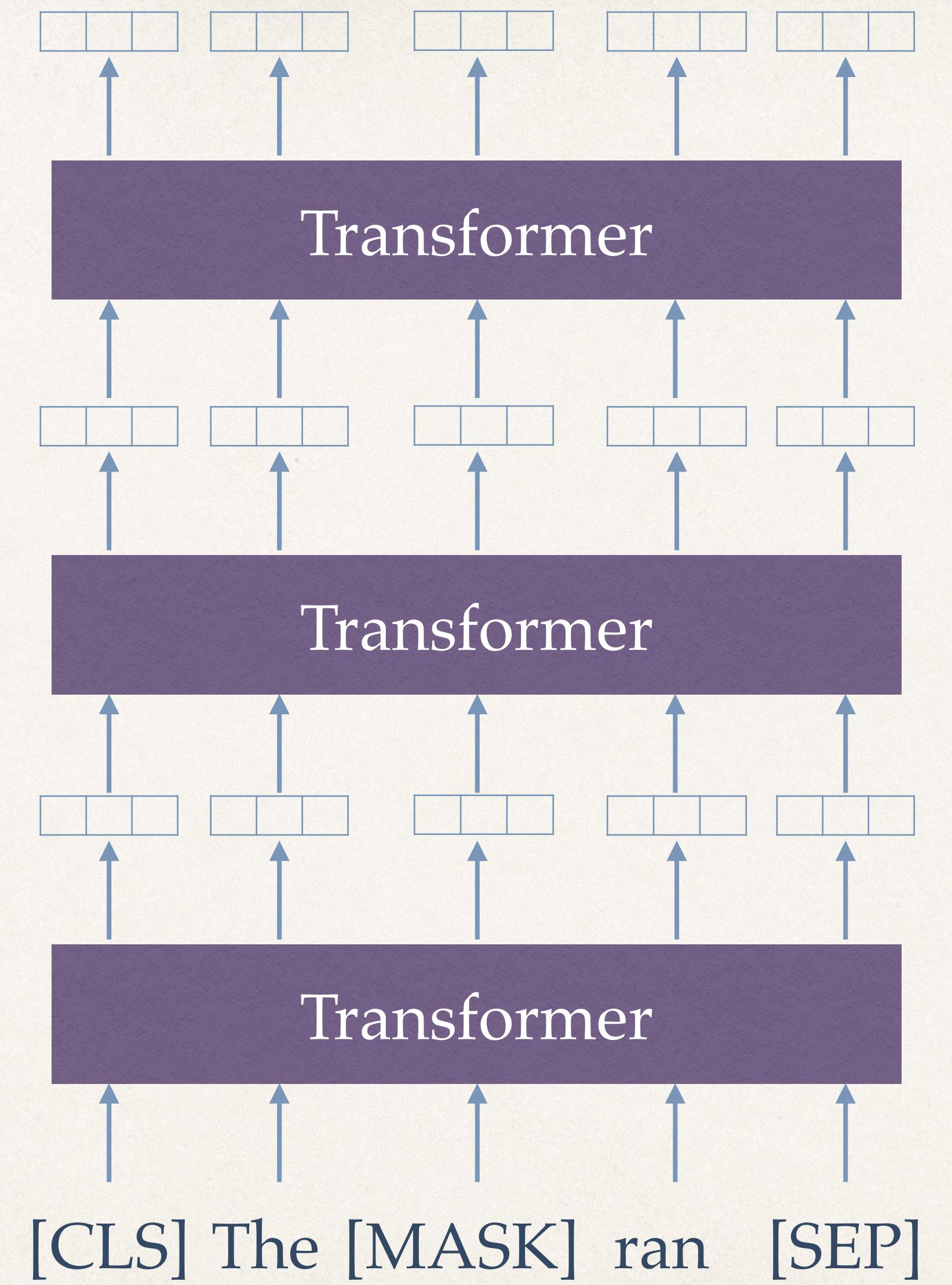
- Transformer based model that predicts masked word based on bidirectional context and next sentence prediction
- Multiple transformer blocks that take sequence of input vectors and give sequence of output vectors

SPECIAL TOKENS

- Every sentence is appended with a special token [CLS] in the beginning and [SEP] at the end
- Embeddings for [CLS] and [SEP] tokens are also learned and can be used as vector representations of the entire sequence



Many such transformer blocks stacked together make the BERT model



At the time of training, we try to predict the original token in place of MASK token

BERT

- Deep networks (12 layers for BERT base, 24 for BERT large)
- Token representations are high dimensional (768 dims for BERT base, 1024 for BERT large)
- Pretrained on large amounts of English text such as Wikipedia (2.5B words) and BooksCorpus (800M words)

WORDPIECE

- Tokens are called wordpieces which allows to limit the vocabulary size and share subword information

this	this
grow	grow
growing	grow + #ing

“What can we do with contextual embeddings?”

TASK PERFORMANCE

Plugging in
the contextual
embeddings
can improve
performance
on linguistic
tasks

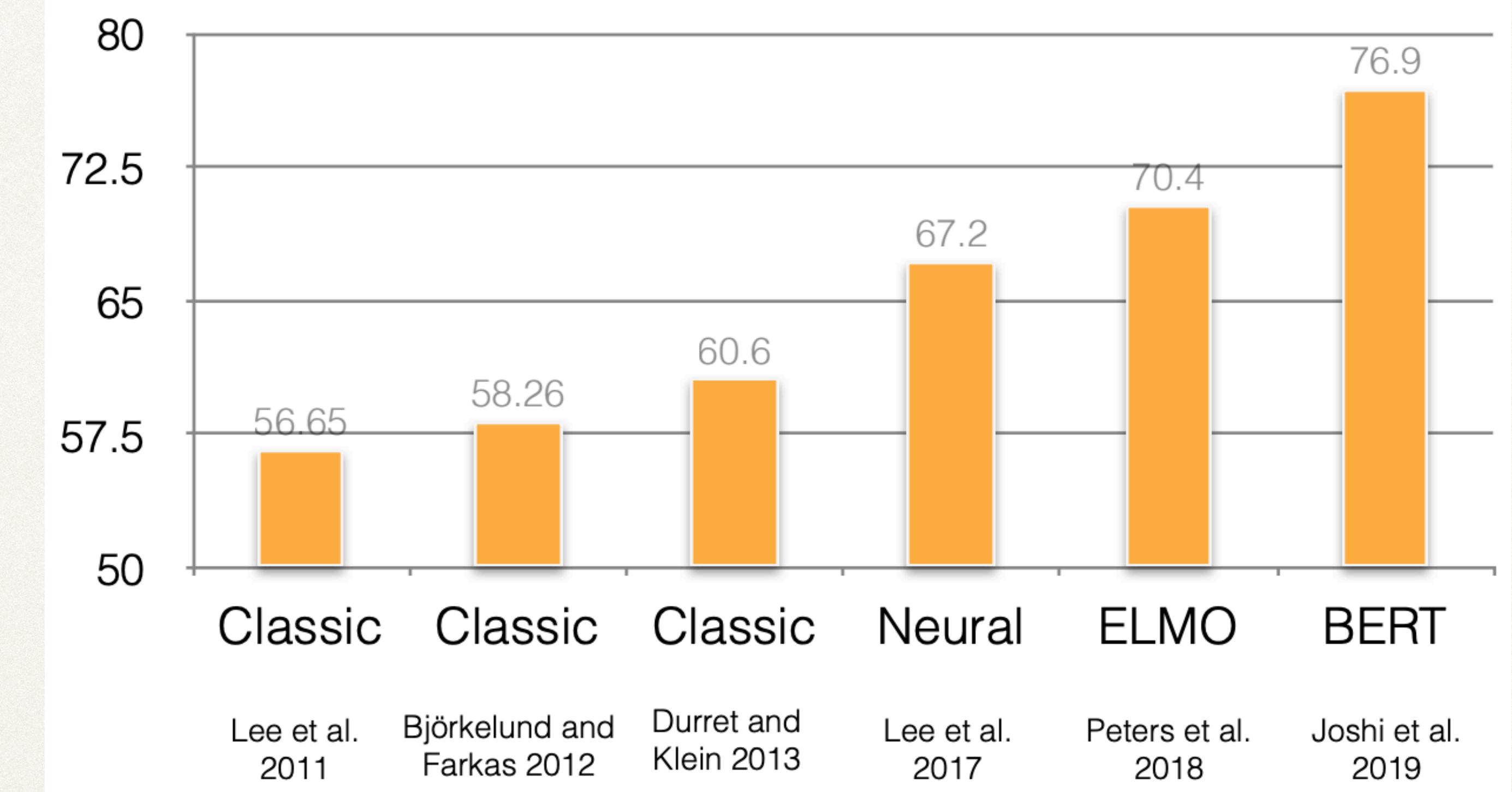
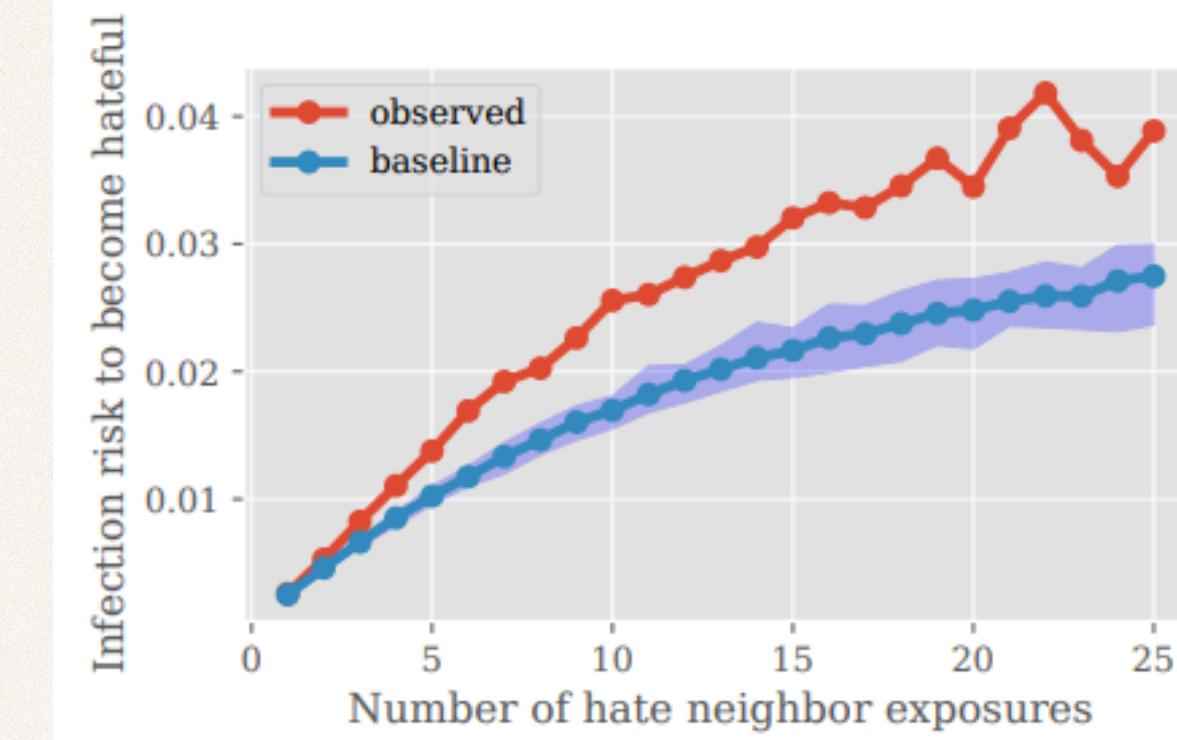


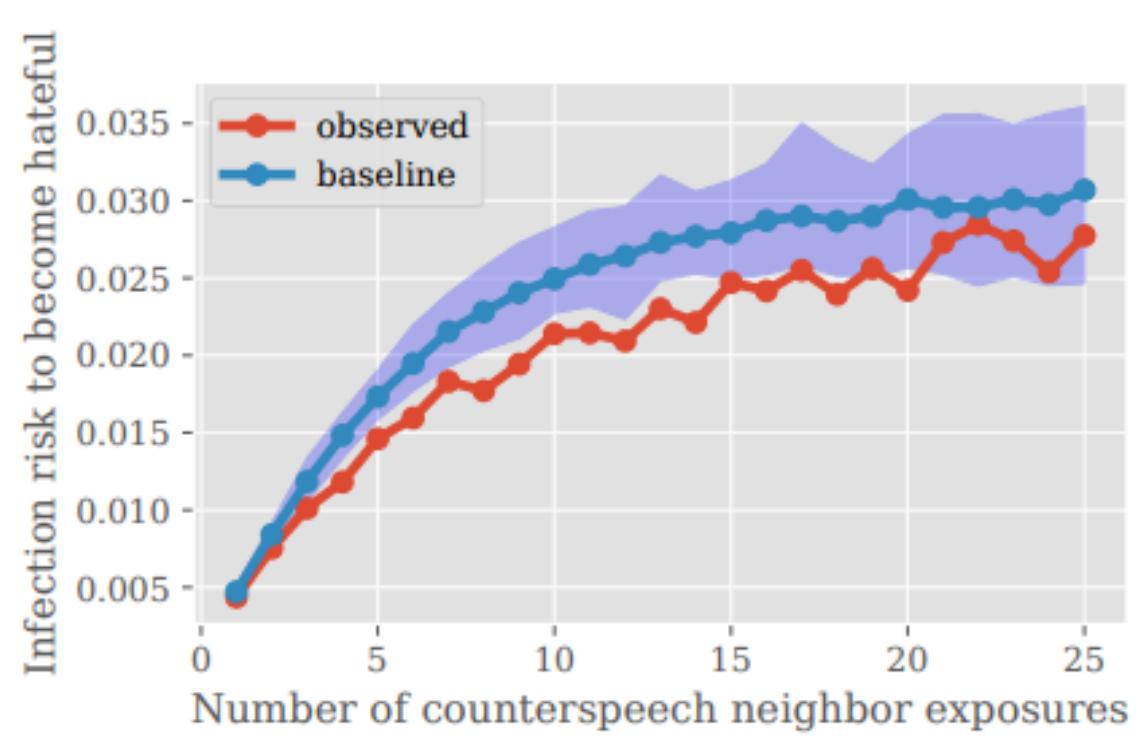
Figure taken from David Bamman's slide

Feature set	Precision	Recall	F1 score
Anti-Asian hate tweet detection			
Linguistic	0.541	0.233	0.323
Hashtag	0.100	0.002	0.005
BERT	0.765	0.760	0.762
Counterspeech tweet detection			
Linguistic	0.483	0.189	0.267
Hashtag	0.800	0.029	0.056
BERT	0.839	0.868	0.853
Neutral tweet detection			
Linguistic	0.632	0.891	0.739
Hashtag	0.591	0.999	0.743
BERT	0.886	0.874	0.880

TABLE III: Tweet classification performance of different feature sets with a neural network classifier. The BERT model has the best classification performance in all three tasks.

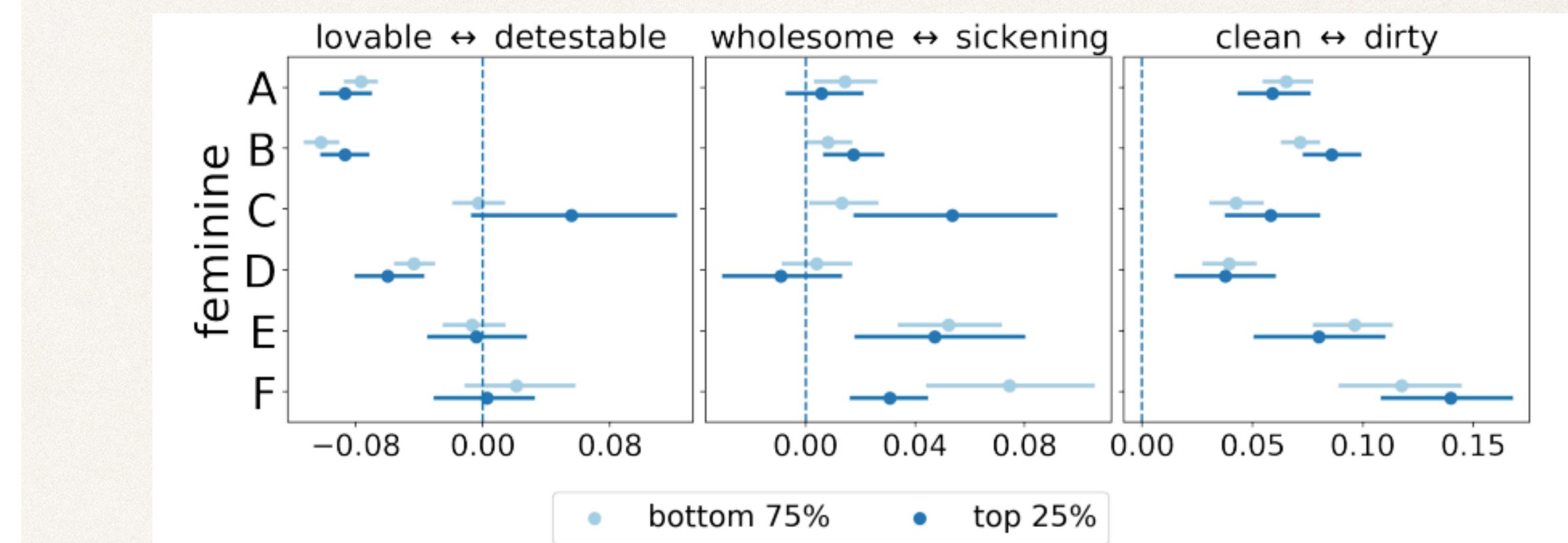
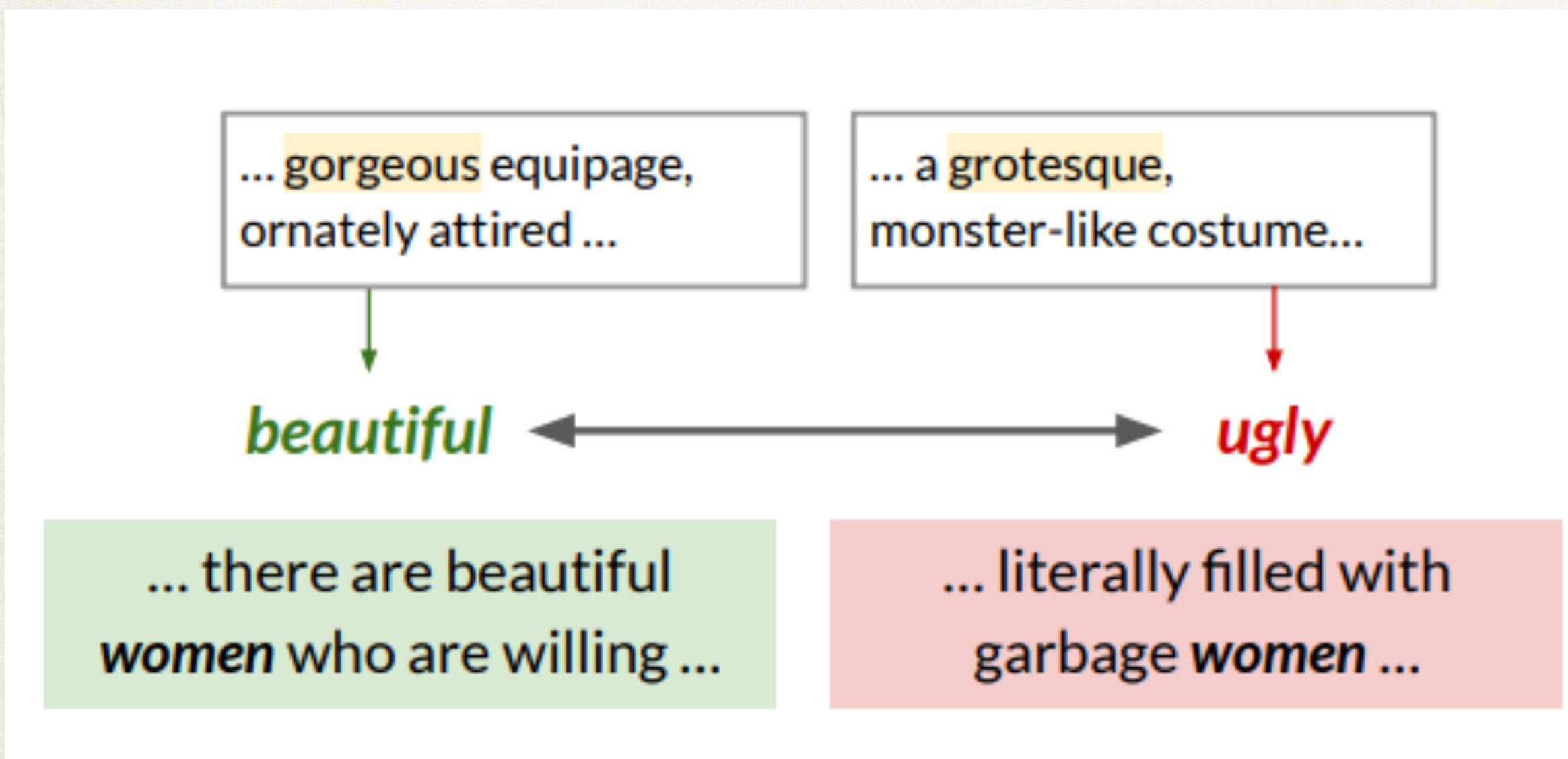


(a) Hate → hate



(b) Counterspeech → hate

CONTEXTUALIZED SEMAXES



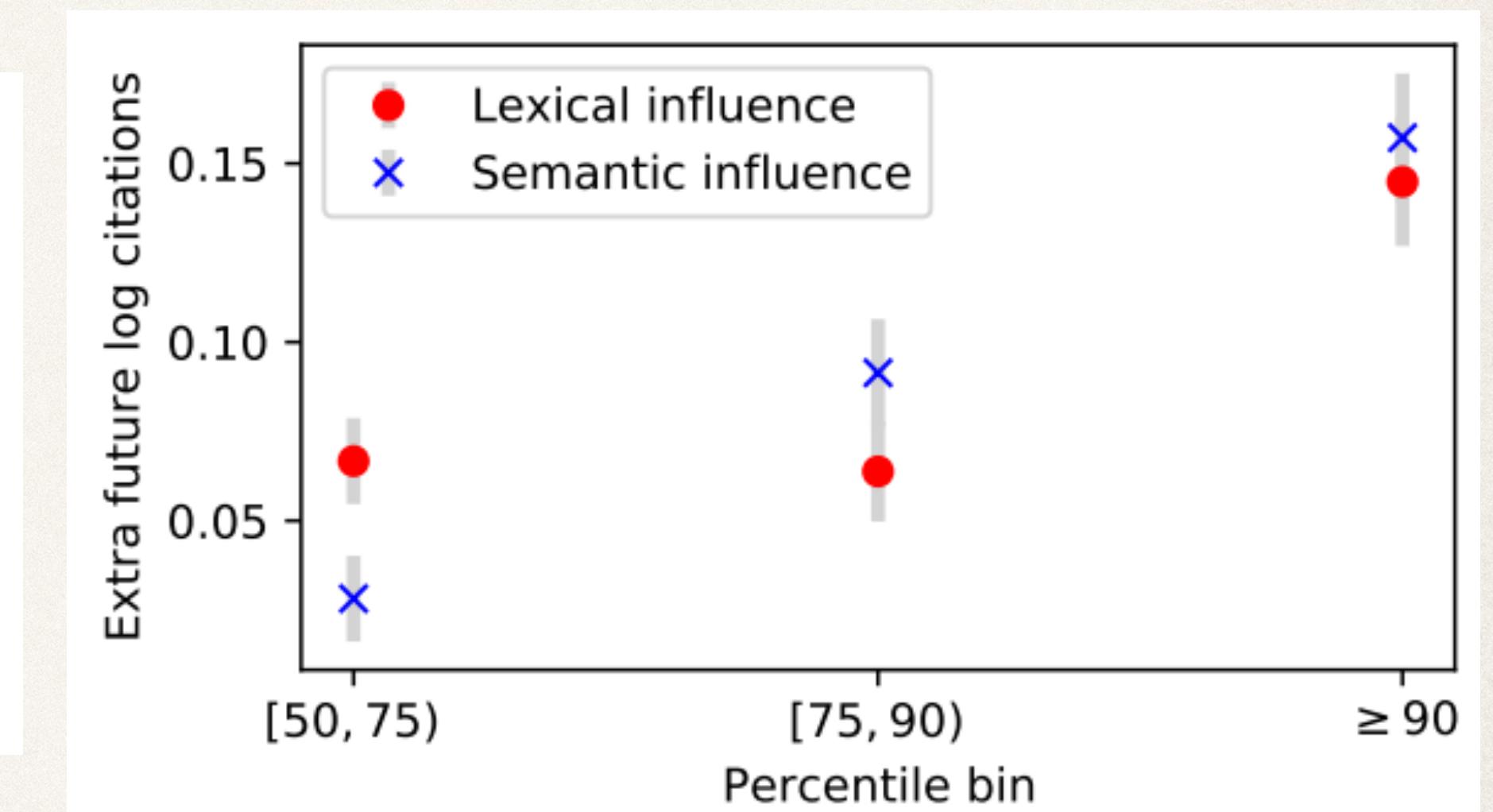
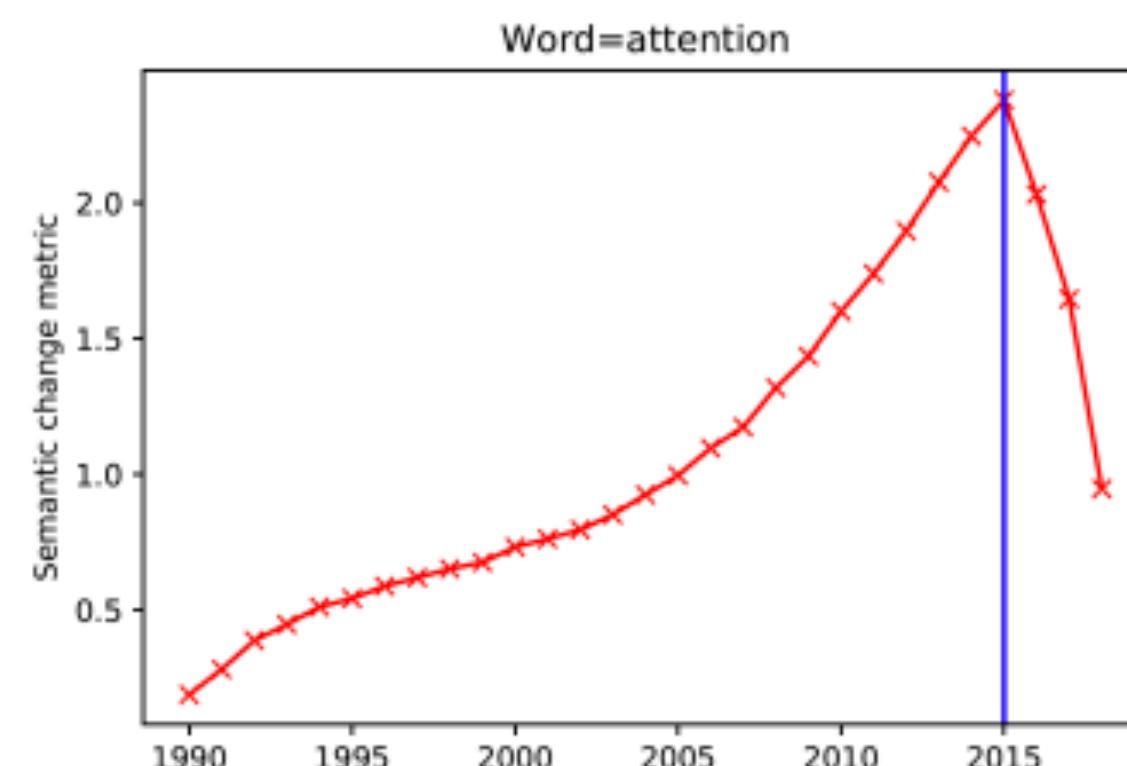
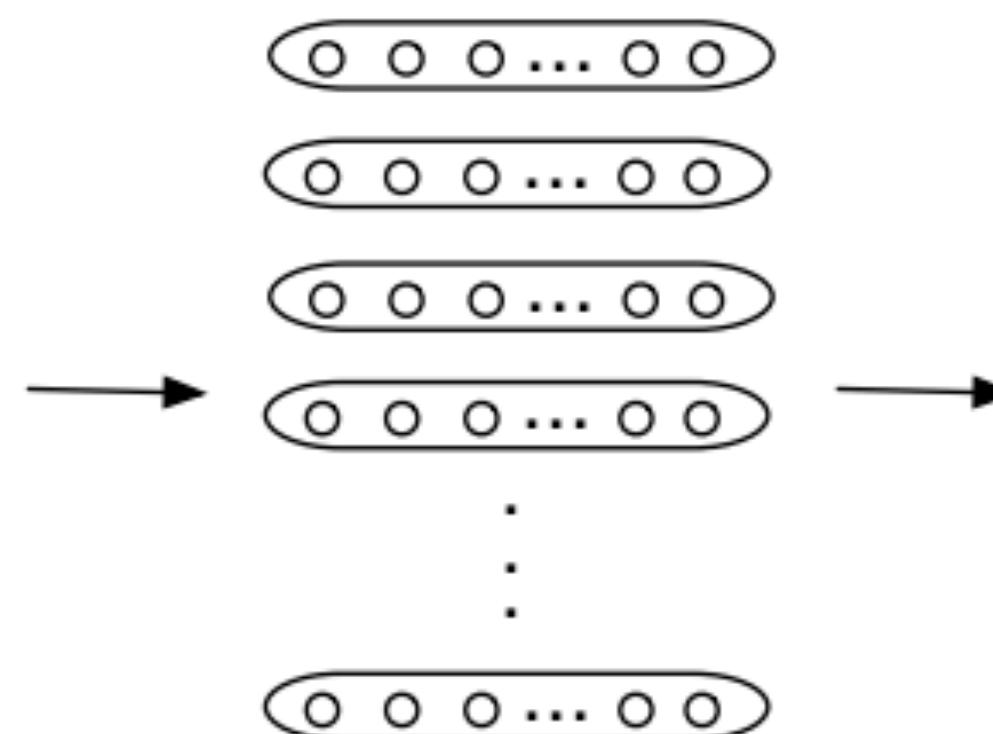
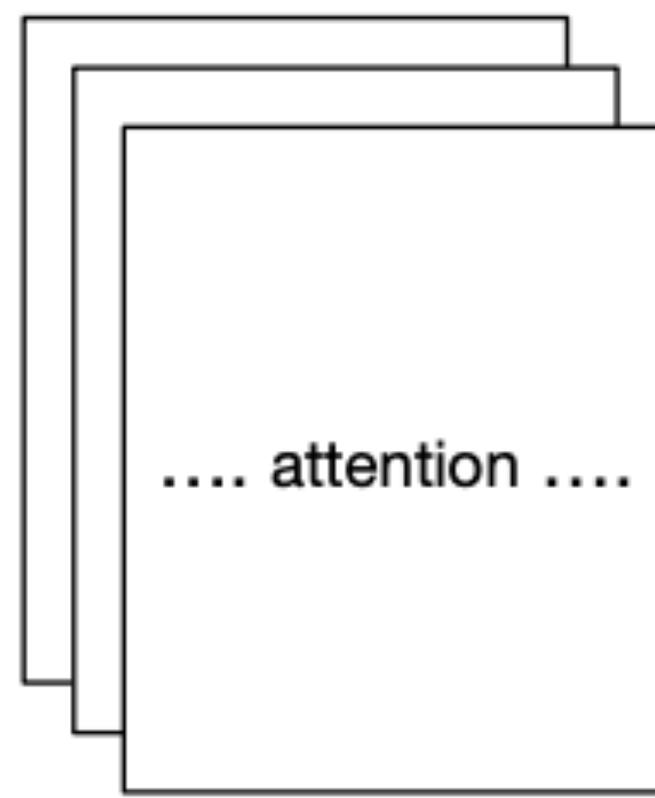
Li Lucy, Divya Tadimeti, and David Bamman. 2022. Discovering Differences in the Representation of People using Contextualized Semantic Axes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3477–3494, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

GEOEMTRY OF CONTEXTUALIZED REPRESENTATIONS

Region	North America	Europe	Middle East	Asia	South America	Oceania	Central America	Africa
BERT-Base	100%	92%	92%	91%	89%	87%	85%	85%
BERT-Base (Artificial Dataset)	100%	89%	92%	89%	88%	88%	87%	87%
BERT-Multilingual	100%	89%	88%	88%	91%	81%	83%	83%

Table 1: % of the average radius of bounding balls relative to the average of radius of bounding balls of North American countries names. Central America also includes countries in the Caribbean.

CHANGE OVER TIME



Soni, Sandeep, David Bamman, and Jacob Eisenstein. "Predicting Long-Term Citations from Short-Term Linguistic Influence." *Findings of the Association for Computational Linguistics: EMNLP 2022*. 2022.