

Enhancing the Dependency Mechanism of RoBERTa

Natural Language Processing:
EDM-RoBERTa

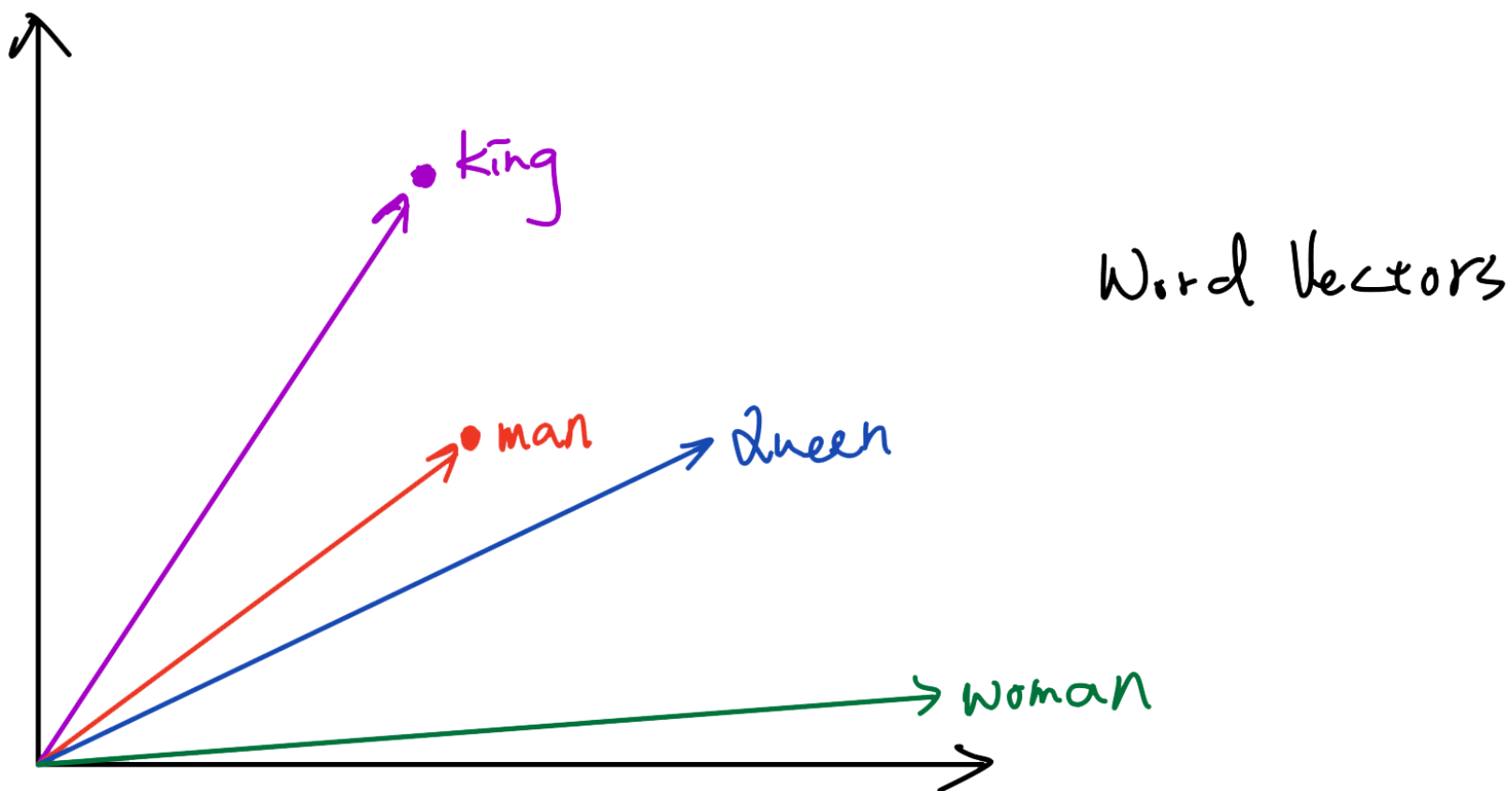
PRJ2020-002 Team Members: 李昱廷 郭為軒 曹仲辰 吳岳霖 林裕峰

Instructor: 張炎清 教授



EDM-RoBERTa: Enhancing the Dependency Mechanism of RoBERTa

- Learning Language Representations: Word Vectors (Word Embeddings)



General Language Model

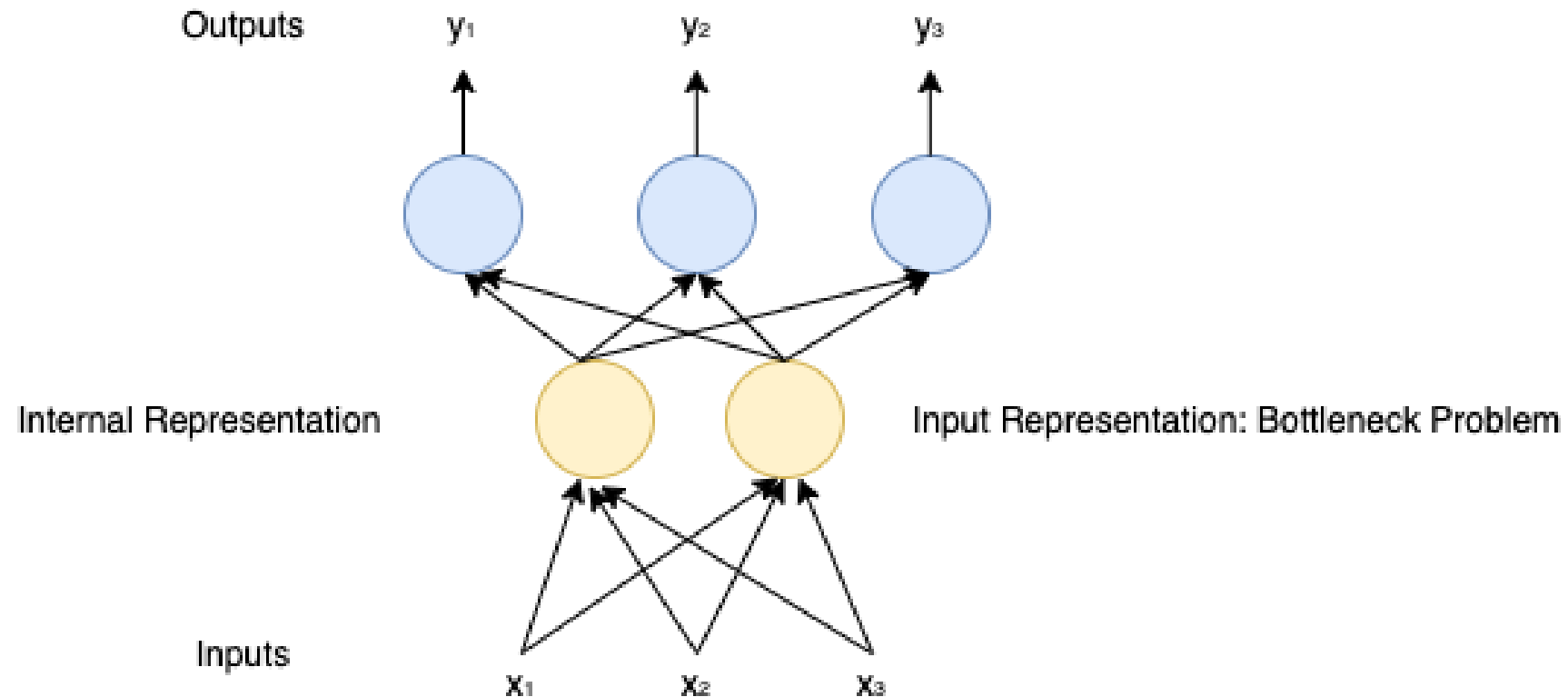
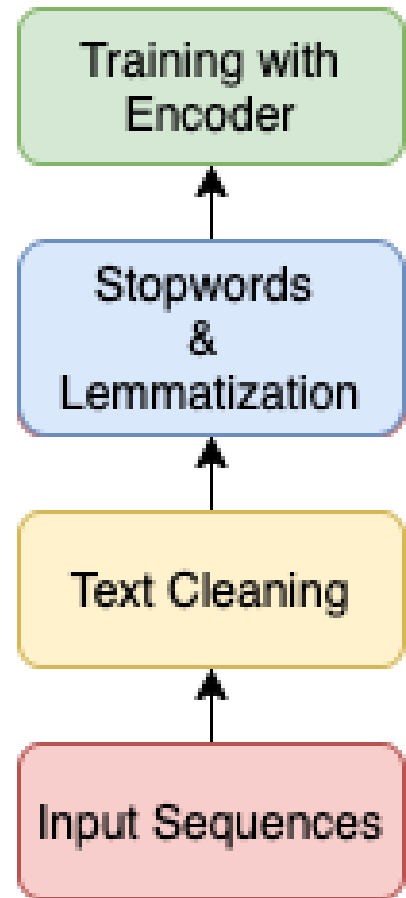
Goal: Build a general language model to process with natural language.

The model can be adapted to different NLP tasks by transfer learning.

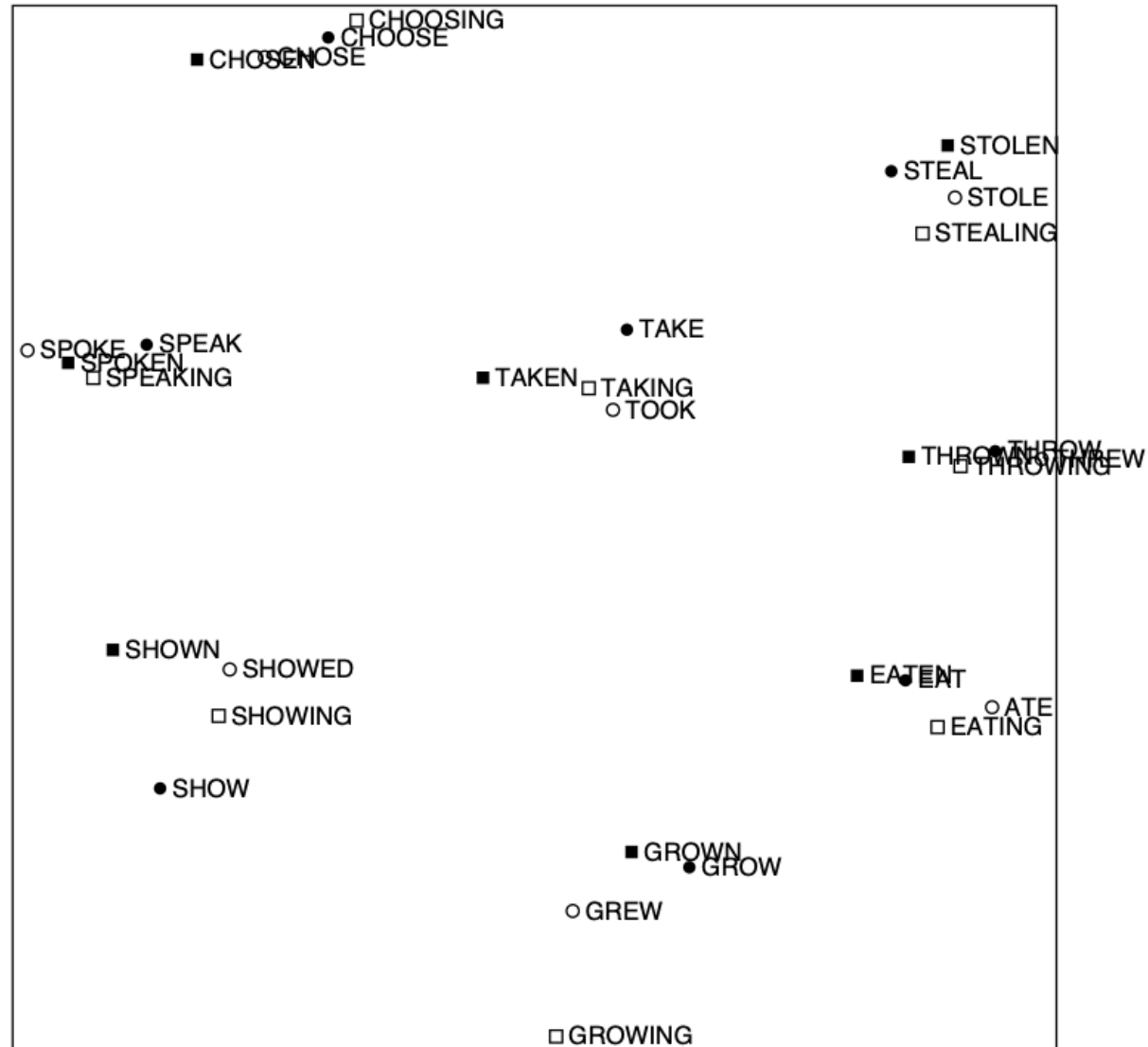
Transfer Learning: Pre-train and fine-tune the language model.

Conventional Language Model: Sequence to Sequence model (Seq2Seq)

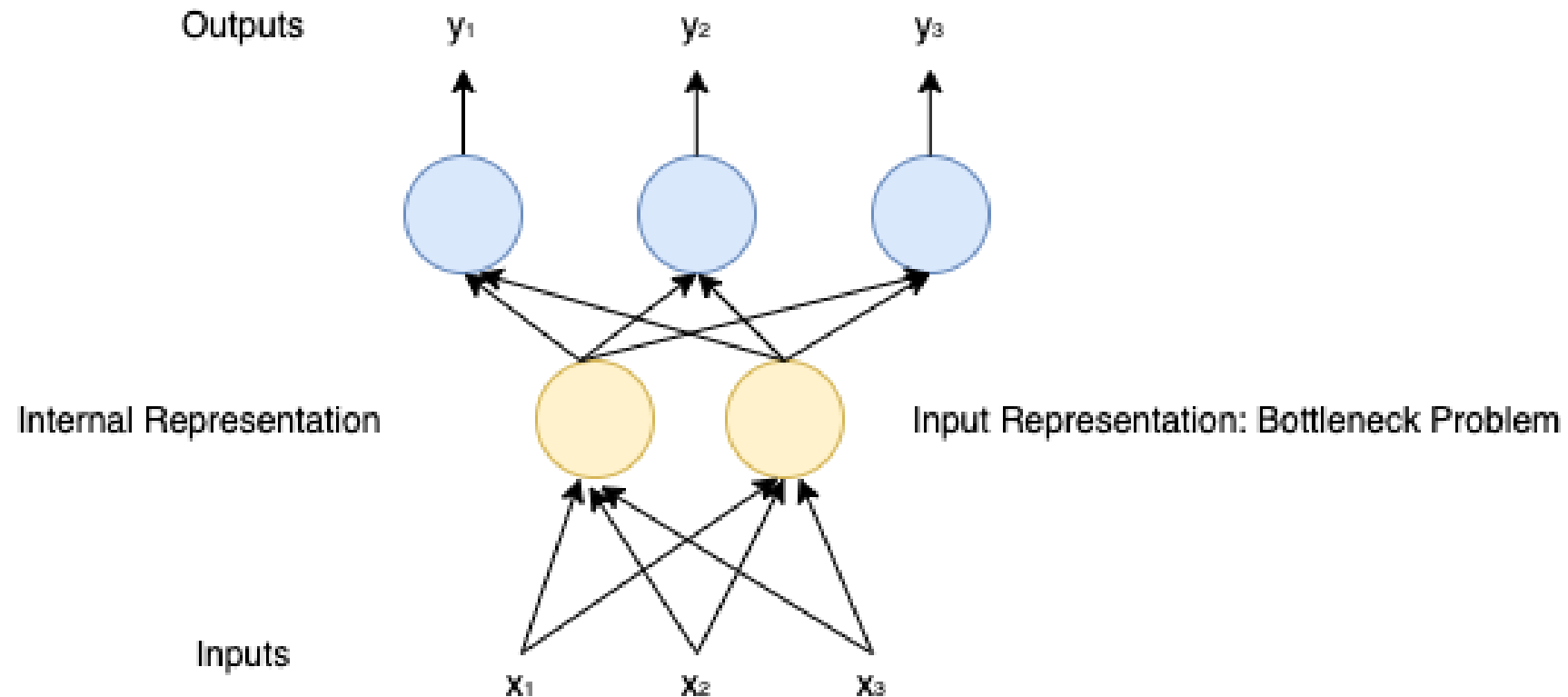
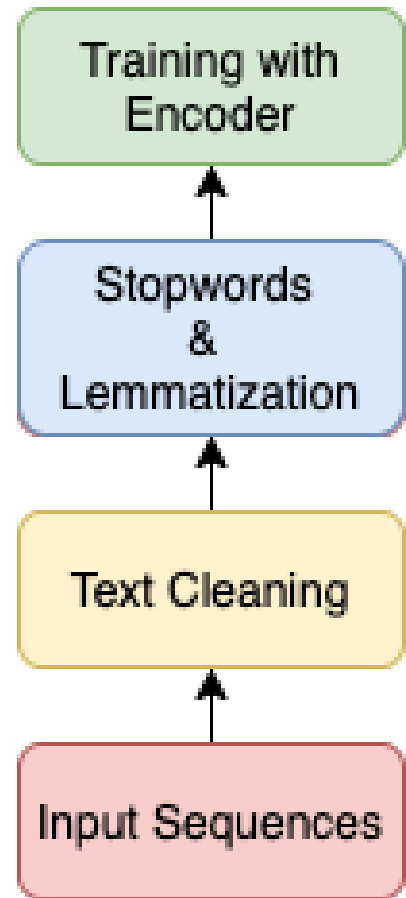
Flowchart of Text Sequences Processing and Bottleneck Problem



Syntactic Patterns Emerge in Word Vectors



Flowchart of Text Sequences Processing and Bottleneck Problem



Encoder-Decoder Structure

- Natural Language Understanding (NLU) with Encoder: Sentiment Analysis, Named Entity Classification, etc.
- Natural Language Generation (NLG) with Decoder: Neural Machine Translation, Question Answering, etc.
- Context Vector: Causing “**Bottleneck Problem**”

Encoder-Decoder Structure: Transformer-based Models

- **Models: BERT, RoBERTa, XLNet, DistilBERT**

Bidirectional Encoder Representation from Transformers (**BERT**)

Robustly optimized BERT approach (**RoBERTa**)

- **Task: Sentiment Analysis**

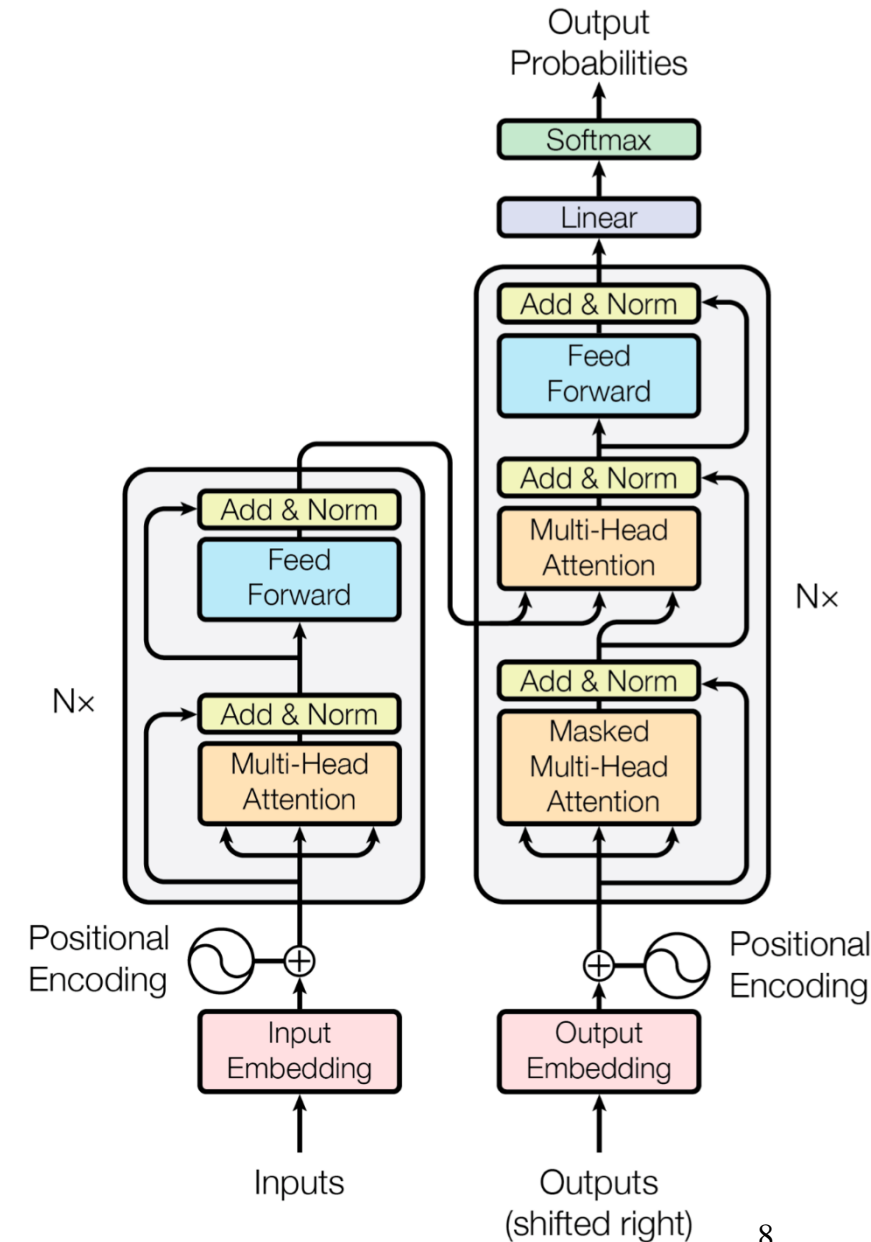
- **Benchmark Datasets:**

- ✓ First GOP Debate Twitter Sentiment
- ✓ Tweets from verified users concerning stocks traded on the NYSE, NASDAQ & SNP
- ✓ SST-2: IMDb Movies Reviews
- ✓ SST-5: Rotten Tomatoes Movies Reviews

Structure: Transformer-based Models

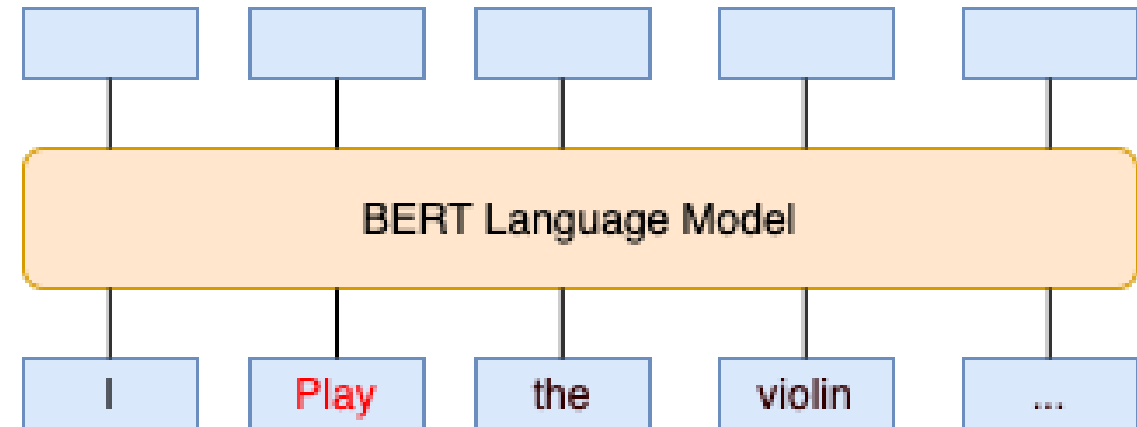
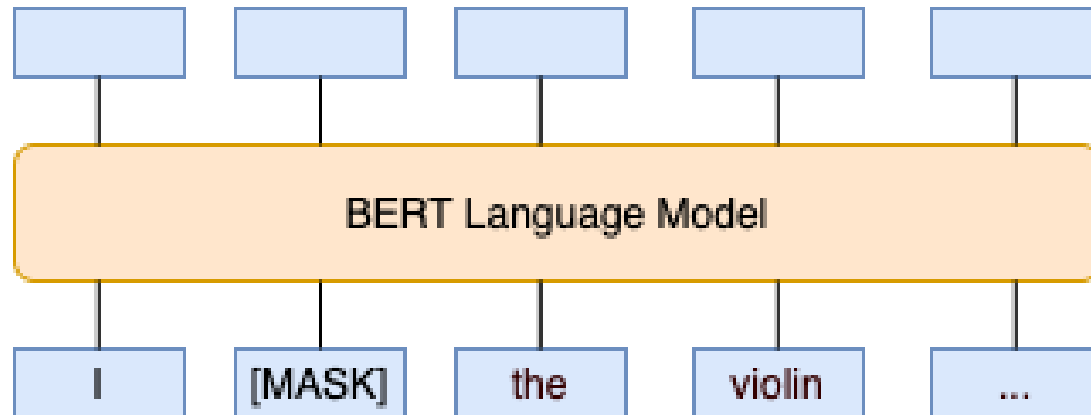
Pre-training approaches

- **BERT:** Maskd LM & Next Sentence Prediction
- **RoBERTa:**
 - Trained on **More Corpus**
(WikiText103, BookCorpus, CCNews)
 - Trained with **Bigger batch sizes**



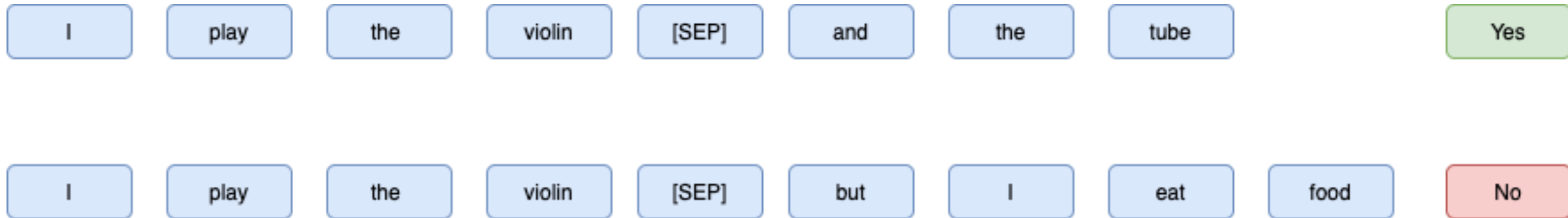
Masked Language (Masked LM)

- Masked LM



Next Sentence Prediction (NSP)

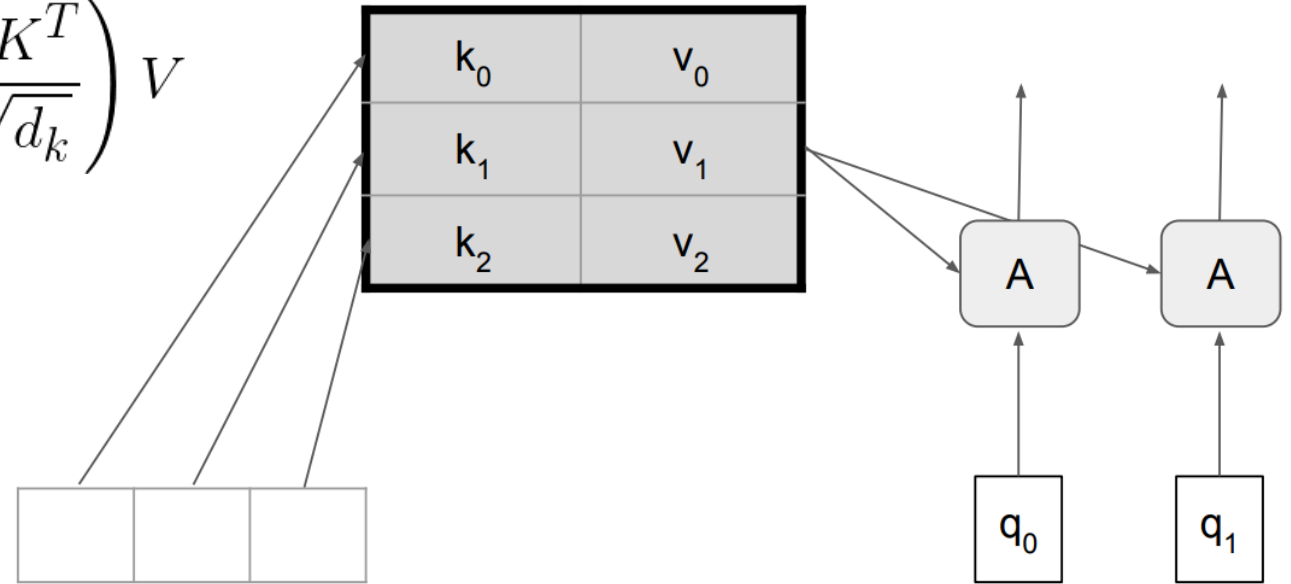
- NSP



Attention Mechanism

- Scaled Dot-Product Self Attention

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

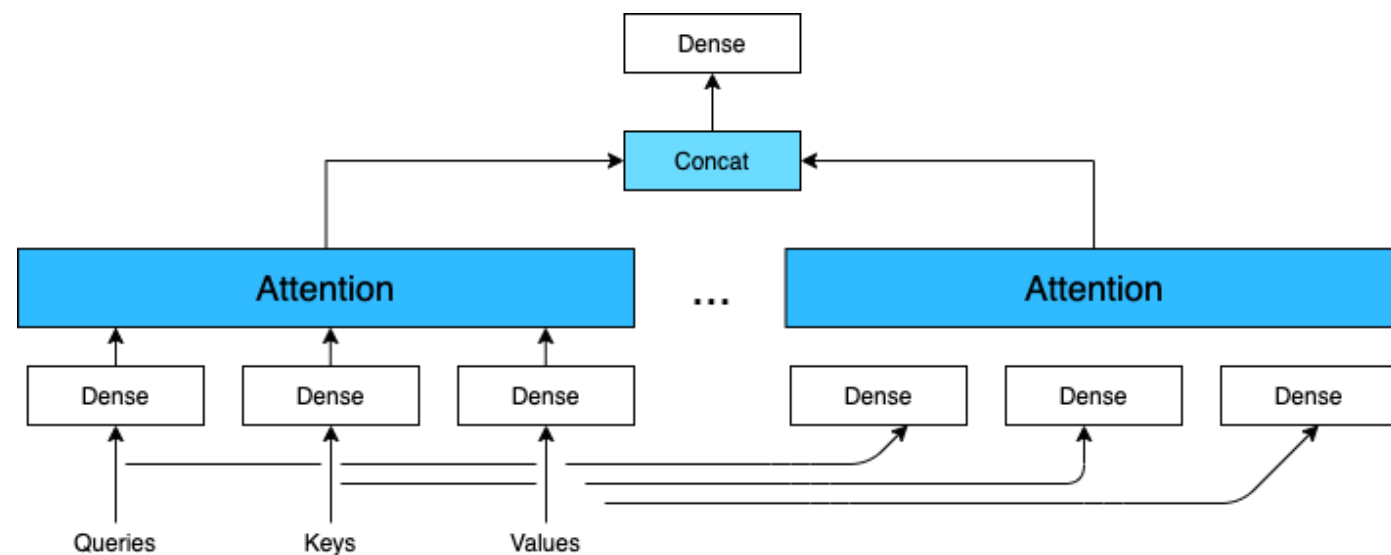
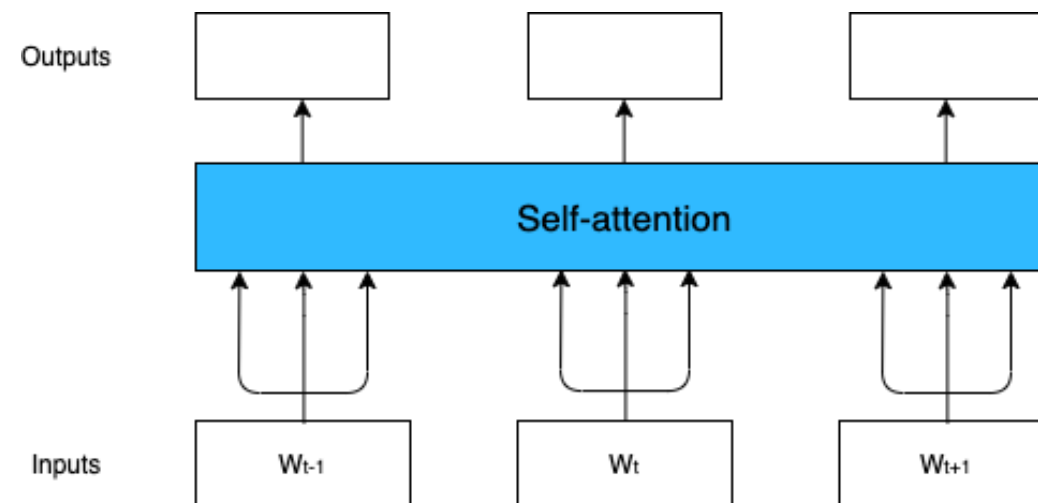


Attention Mechanism

- Scaled Dot-Product Attention

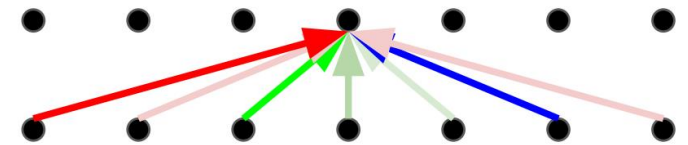
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- Multi-Headed Attention



Attention Mechanism: The Fix – Multi-Headed Attention

- **Multi-Headed Attention**
 - Multiple attention layers (heads) in parallel (shown by different colors)
 - Each layer uses different linear transformations.
 - Different heads can learn different relationships.



Problems with Attention Mechanism

- Too many heads -> Hard to process queries from multiple positions in parallel.
- **Valid heads are unknown**
- Processing with “Valid heads” problem:

We propose EDM-RoBERTa to optimize the attention process

Optimization

- Single-headed Attention RNN (SHA-RNN)

- **Boom Layer**

$$v \in \mathbb{R}^H \rightarrow \mu \in \mathbb{R}^{N \times H} \rightarrow \omega \in \mathbb{R}^H$$

- **Activation Function:** Gaussian Error Linear Units (GELUs)

- Models: BERT, RoBERTa, DistilBERT, XLNet
- **Solving the short-term dependency problem from Transformer-based Models**

Optimization: Sentimental Ambiguity

Sentiment Analysis



My Experience so far has
been fantastic!

Positive



The product is ok I guess

Neutral

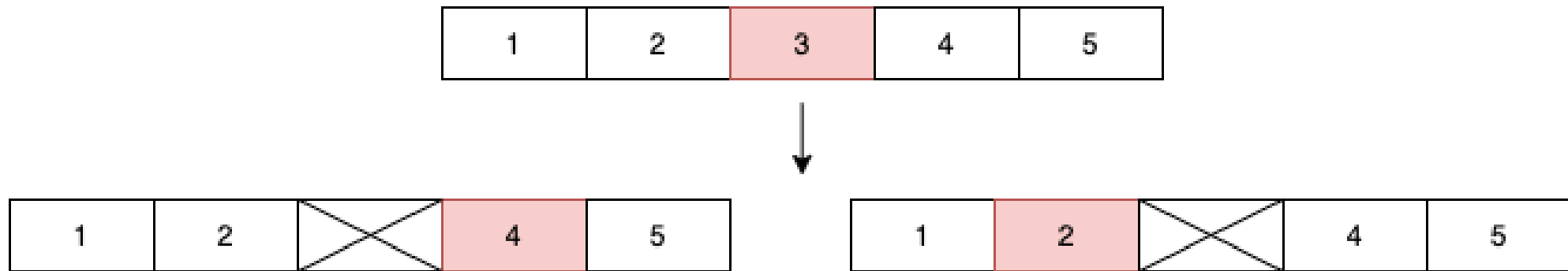


Your support team is
useless

Negative

Optimization: Sentimental Ambiguity

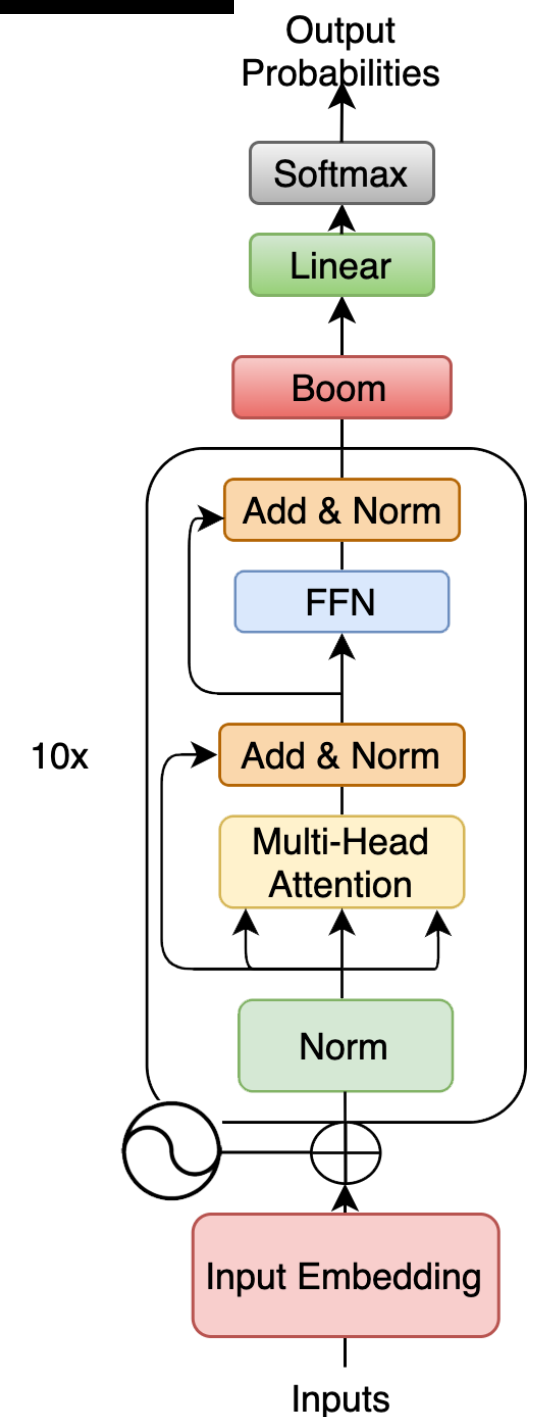
- Why using both two-dimensional and multi-dimensional sentiment analysis datasets to train EDM-RoBERTa ?
 - Deal with the sentences with sentimental ambiguity



Improved Learning Model: EDM-RoBERTa

- Choose the best-performed model: **RoBERTa**
- **Optimize the RoBERTa model with Boom Layer**

Enhancing the Dependency Mechanism of RoBERTa



Optimization: Statistics

Fine-tuning Transformer-based Models with IMDB Dataset

	Epoch	Accuracy	train loss	valid loss	error rates
BERT _{LARGE}	6	92.6	0.35	0.55	0.29
RoBERTa _{LARGE}	6	93.17	0.22	0.53	0.26
XLNet	6	89.53	0.28	0.69	0.37
DistilBERT	6	86.48	0.32	0.74	0.35
EDM-RoBERTa	6	94.76	0.27	0.49	0.2

Fine-tuning Transformer-based Models with Rotten Tomatoes Dataset

	Epoch	Accuracy	train loss	valid loss	error rates
BERT _{LARGE}	5	66.21	0.64	0.68	0.3
RoBERTa _{LARGE}	5	68.91	0.67	0.7	0.29
XLNet	5	62.83	0.73	0.79	0.38
DistilBERT	5	54.65	0.8	0.77	0.44
EDM-RoBERTa	5	76.18	0.64	0.62	0.26

EDM-RoBERTa: Enhancing the Dependency Structure of RoBERTa

- Detailed parameters of EDM-RoBERTa

EDM-RoBERTa (Enhance the Dependency Mechanism of RoBERTa)

bsz	steps	lr	ppl	SST-2	SST-5
256	1M	1.00E-05	3.83	92.6	74.57
2K	125K	2.00E-04	3.61	94.76	76.18
8K	31K	1.00E-03	3.72	92.1	74.31

Runtimes & Environments

Google Colaboratory Pro	GPU: NVIDIA Tesla V100-SXM2-16GB
	CPU: Intel Xeon(R) @2.00GHz OS: Ubuntu 18.04.5 LTS RAM: 32GB
MacBook Pro (16-inch Late 2019)	GPU: AMD Radeon Pro 5600M-8GB-HBM2
	CPU: Intel Core i9-9980HK @2.4GHz Dual Boot OS: Ubuntu 18.04.5 LTS with macOS Catalina 10.15.7 (19H2) RAM: 64GB
MacBook Pro (13-inch, M1, 2020)	GPU & CPU: Apple M1 Chip with 8-core CPU, 8-core GPU NPU: Apple M1 Chip with 16-core Neural Engine Environments: CreateML, Tensorflow-mac RAM: 16GB

Conclusions

We introduced a language representation model called EDM-RoBERTa.

EDM-RoBERTa is designed to improve the dependency mechanism, and fine-tune the whole model with sentiment analysis datasets.

Experiments and statistics show our proposed model successfully enhance the dependency mechanism on local context.

EDM-RoBERTa outperforms conventional pre-trained models, including Seq2Seq, BERT, RoBERTa, XLNet, and DistilBERT.