Illustrated BERT

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

Semi-supervised Learning Step

Model:

Dataset:

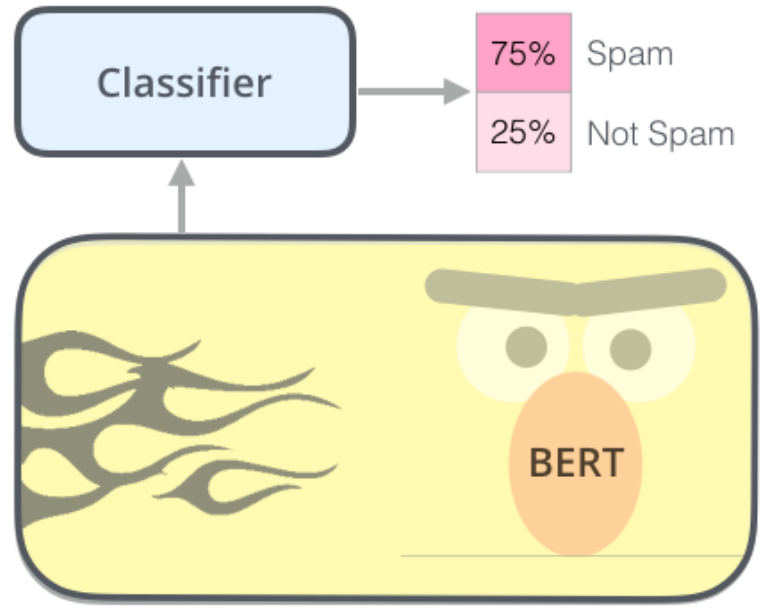
Objective:

Predict the masked word (langauge modeling)

2 - Supervised training on a specific task with a labeled dataset.

Supervised Learning Step

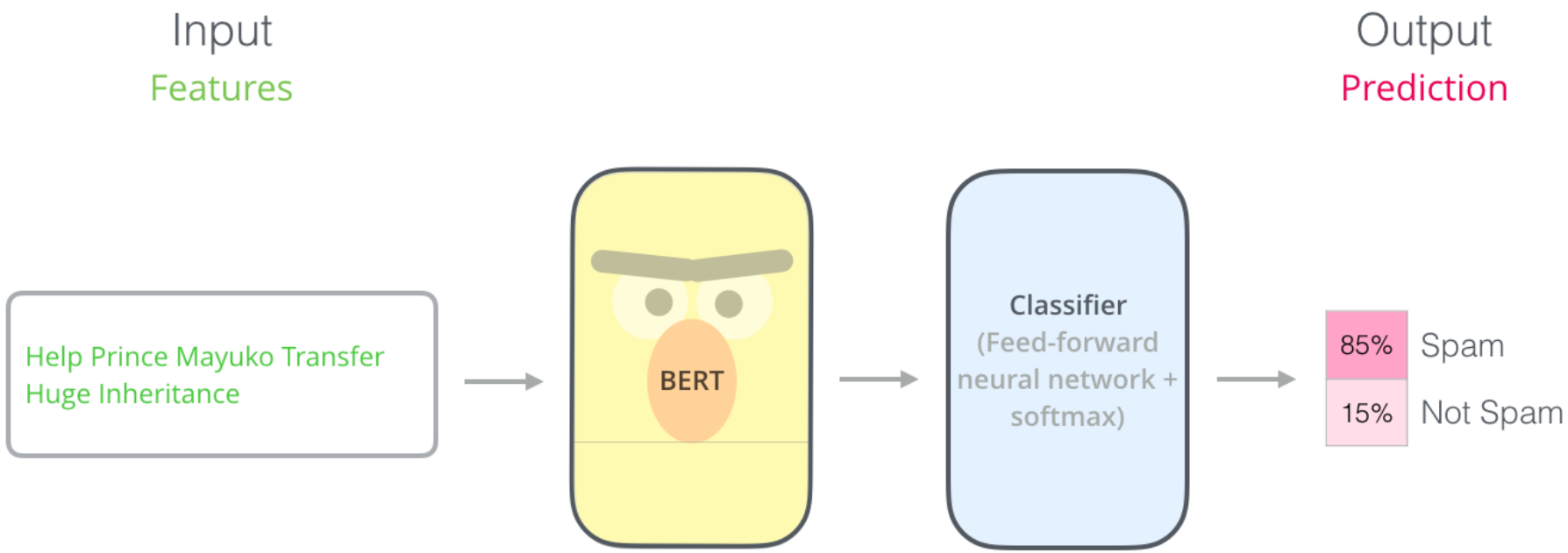
Model: (pre-trained in step #1)



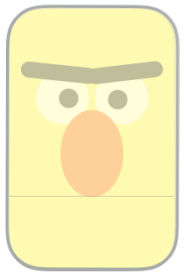
Dataset:

Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached…	Not Spam

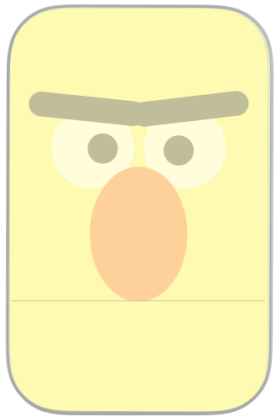
Sentence Classification



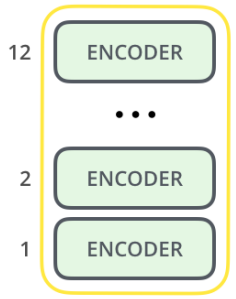
Architecture



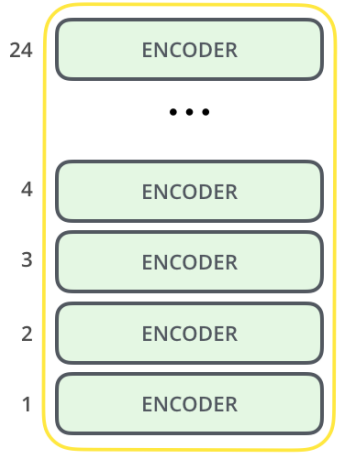
BERT$_{BASE}$



BERT$_{LARGE}$

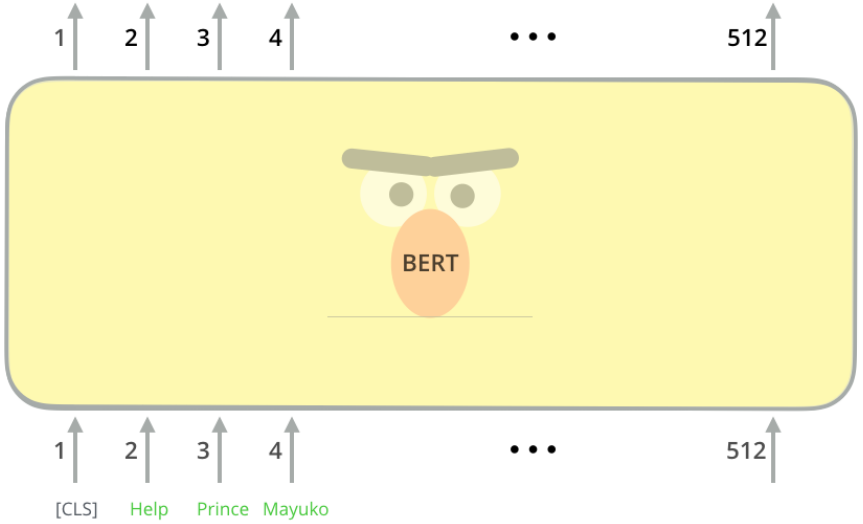


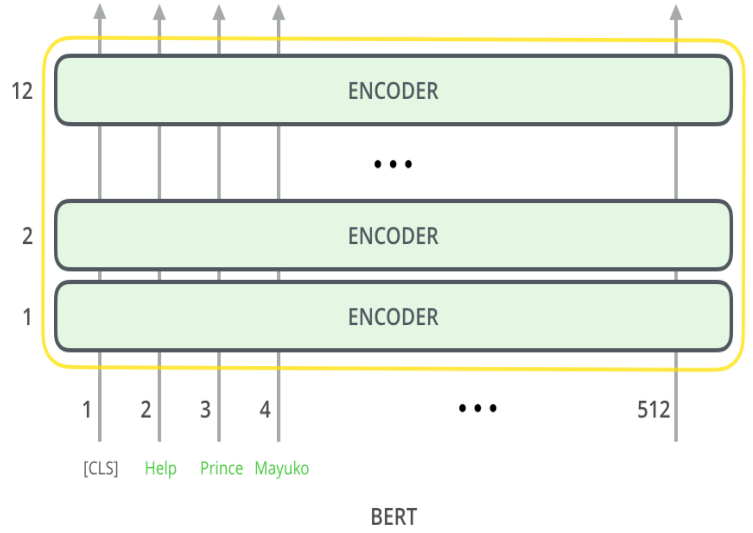
BERT$_{BASE}$



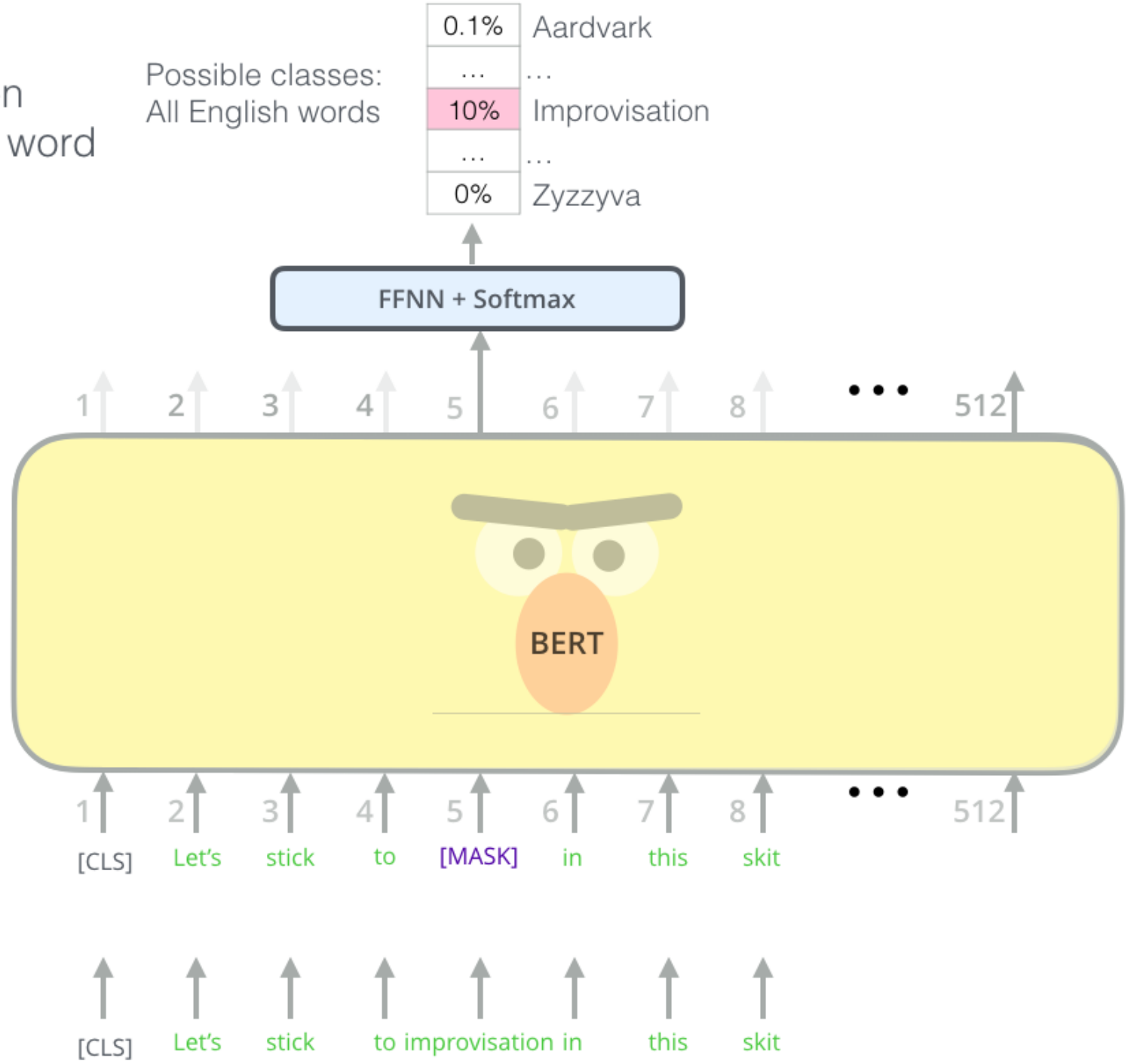
BERT$_{LARGE}$

Model Input

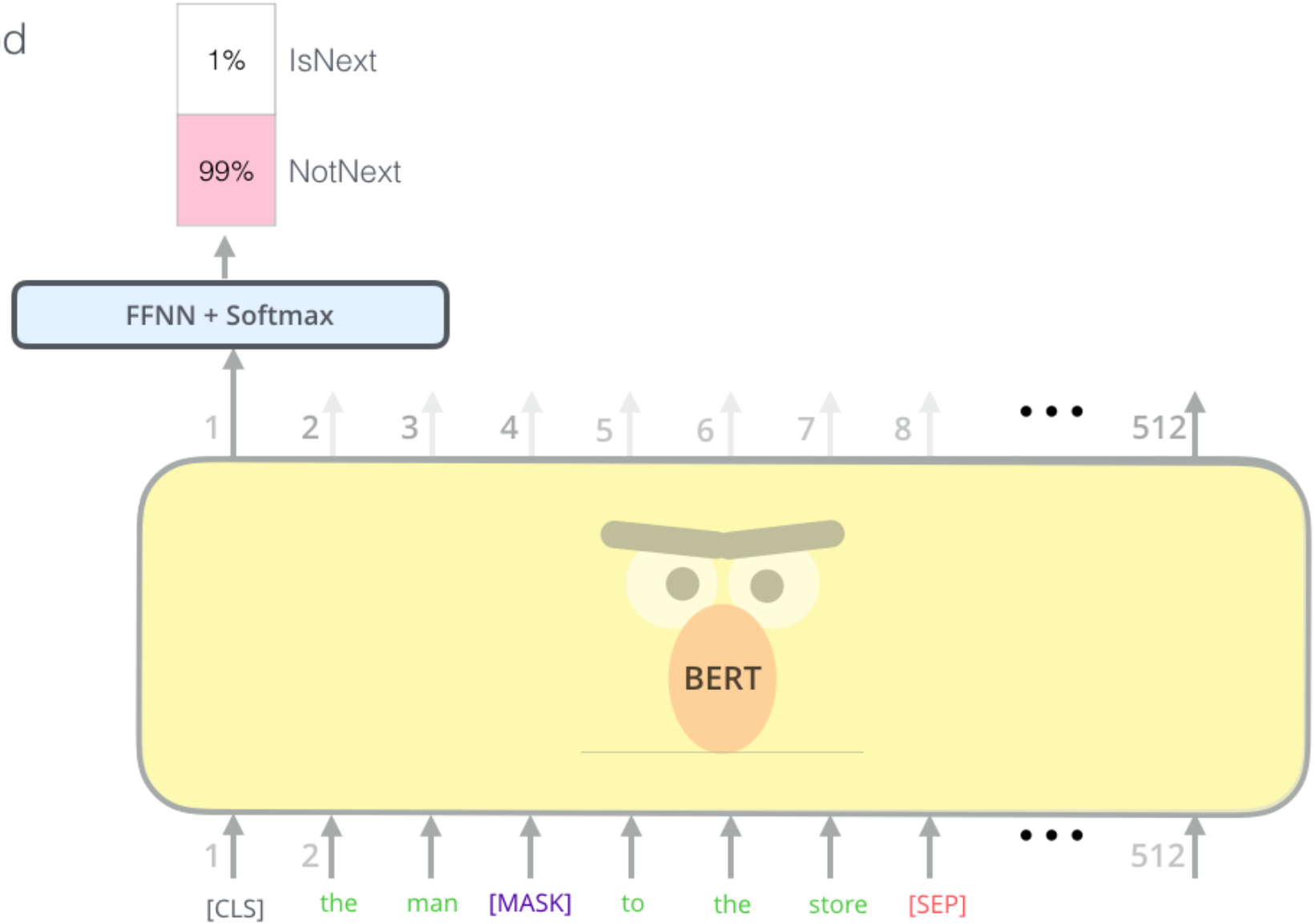

1
2
3
4
• • •
512
BERT
1
2
3
4
• • •
512
[CLS]
Help
Prince
Mayuko


12
ENCODER
• • •
2
ENCODER
1
ENCODER
1
2
3
4
• • •
512
[CLS]
Help
Prince
Mayuko
BERT

Use the output of the masked word's position to predict the masked word



Predict likelihood that sentence B belongs after sentence A



Tokenized Input

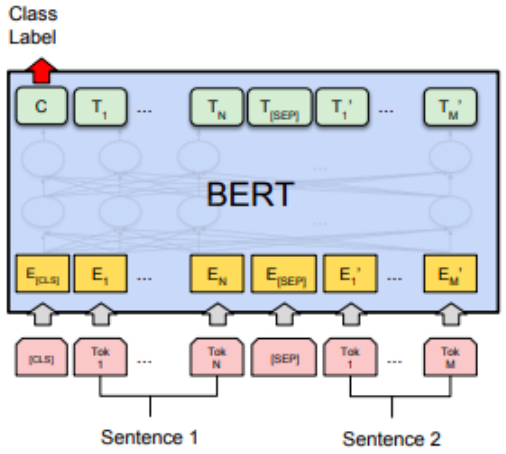
Input

[CLS] the man [MASK] to the store [SEP] penguin [MASK] are flightless birds [SEP]

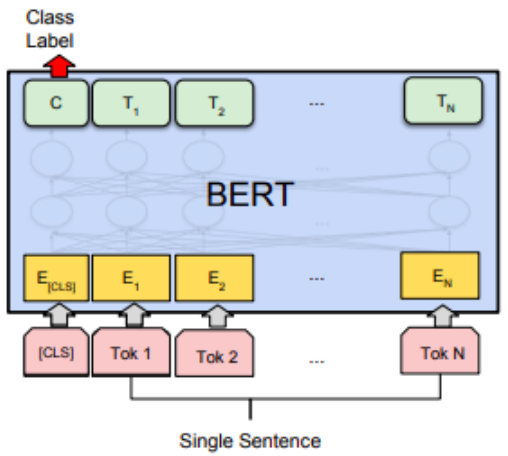
Sentence A

Sentence B

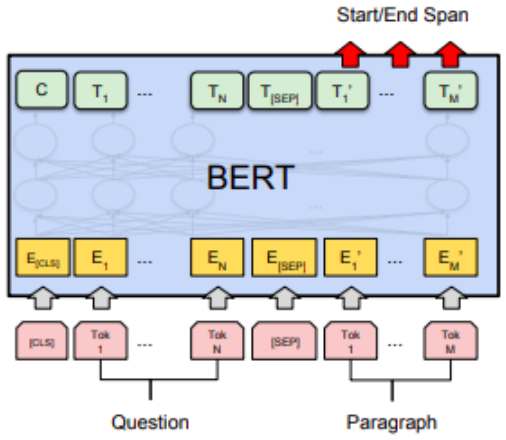
Task specific-Models



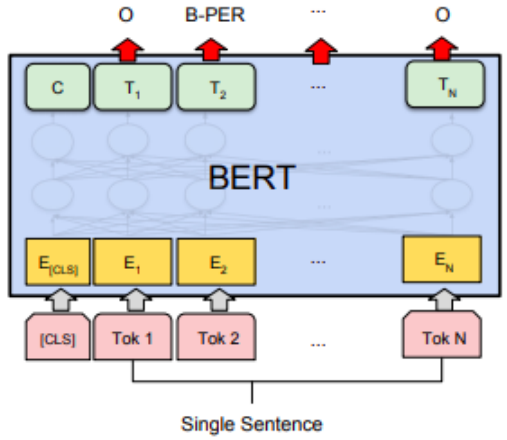
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Sequence Pair Classification Datasets

Dataset	Classes	Sample
MNLI (Multi-Genre Natural Language Inference)	Entailment, Contradiction, and Neutral.	**Premise**: A woman is smiling and talking to a man. **Hypothesis**: A woman is happily chatting with a man. **Label**: Entailment
QQP (Quora Question Pairs)	Duplicate or Not Duplicate	**Question 1:** How can I be a good geologist? **Question 2:** What should I do to be a great geologist? **Label:** Duplicate
QNLI (Question-answering Natural Language Inference)	Entailment, Contradiction, and Neutral.	**Question:** What is the capital of France? **Sentence:** The capital of France is Paris. **Label**: Entailment
STS-B (Semantic Textual Similarity Benchmark)	Similarity Scores from 0 to 5	**Sentence 1**: A man is playing a saxophone. **Sentence 2**: A person is playing the flute. **Similarity Score:** 2.5
MRPC (Microsoft Research Paraphrase Corpus)	Paraphrase or Not Paraphrase	**Sentence 1:** The cat is on the mat. **Sentence 2:** There is a cat on the mat. **Label:** Paraphrase
RTE (Recognizing Textual Entailment)	Entailment, Contradiction or Unknown	**Premise**: The cat is sitting on the windowsill. **Hypothesis**: The cat is outside. **Label**: Contradiction
SWAG (Situations With Adversarial Generations)	Each example in SWAG consists of a context sentence and four possible choices about what could happen next in the given situation.	**Context**: A child is riding a bike in the park. He approaches a puddle. What happens next? **Choices:** **A)** He swerves to avoid it. **B)** He jumps into the puddle. **C)** He speeds up to splash through it. **D)** He stops and turns around. Correct Answer: **C)** He speeds up to splash through it.

Single Sequence Classification Datasets

Dataset	Classes	Sample
SST-2 (Stanford Sentiment Treebank)	Positive or Negative	**Sentence:** "This movie is fantastic!" **Label:** Positive
CoLA (Corpus of Linguistic Acceptability)	Acceptable or Not Acceptable	**Sentence:** The cat sat on the mat. **Label: 1 (**Acceptable)

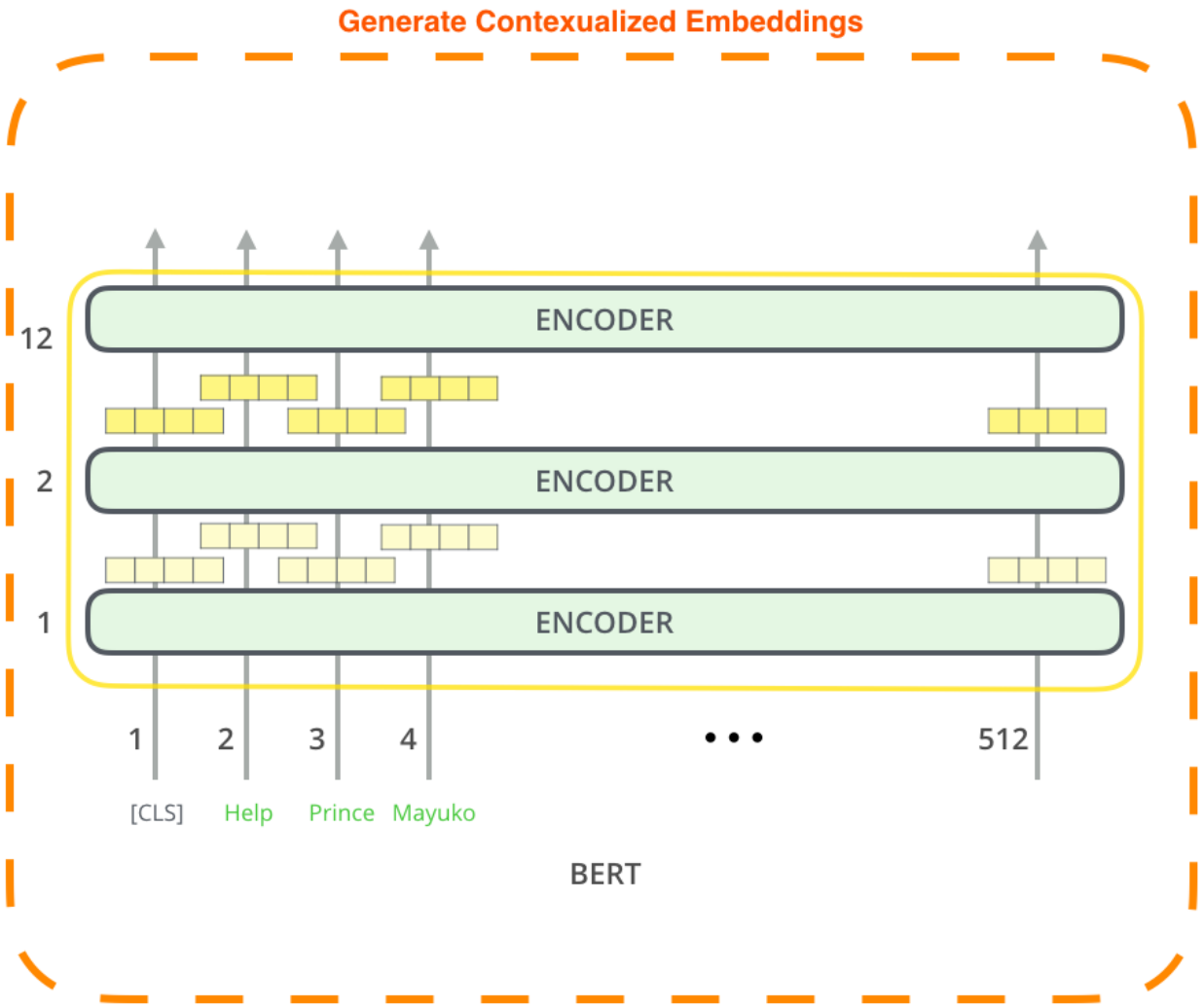
Question Answering Datasets

Dataset	Sample
SQuAD v1.1 (Stanford Question Answering Dataset)	**Context:** "The quick brown fox jumps over the lazy dog." **Question:** "What jumps over the lazy dog?" **Answer:** "The quick brown fox."

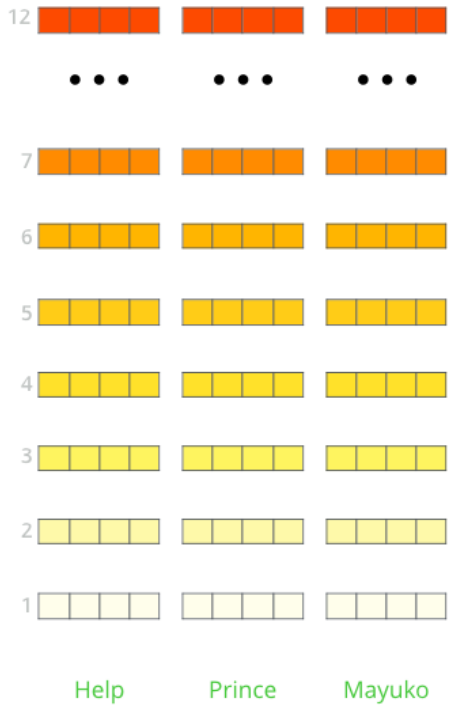
Question Answering Datasets

Dataset	Labels	Sample
CoNLL-2003 NER	Provides annotations for named entity recognition, including four types of named entities: PER (persons), LOC (locations), ORG (organizations), and MISC (miscellaneous entities).	**Sentence**: Thousands of demonstrators have marched through London to protest the war in Iraq. **Annotations:** Thousands - O of - O demonstrators - O have - O marched - O through - O London - B-LOC to - O protest - O the - O war - O in - O Iraq - B-LOC . – O ***In this example, "London" and "Iraq" are tagged as B-LOC (beginning of location), while other words are tagged with "O" indicating they are outside named entities.***

Feature Extraction



The output of each encoder layer along each token's path can be used as a feature representing that token.



But which one should we use?

Using Bert Embeddings for Classification

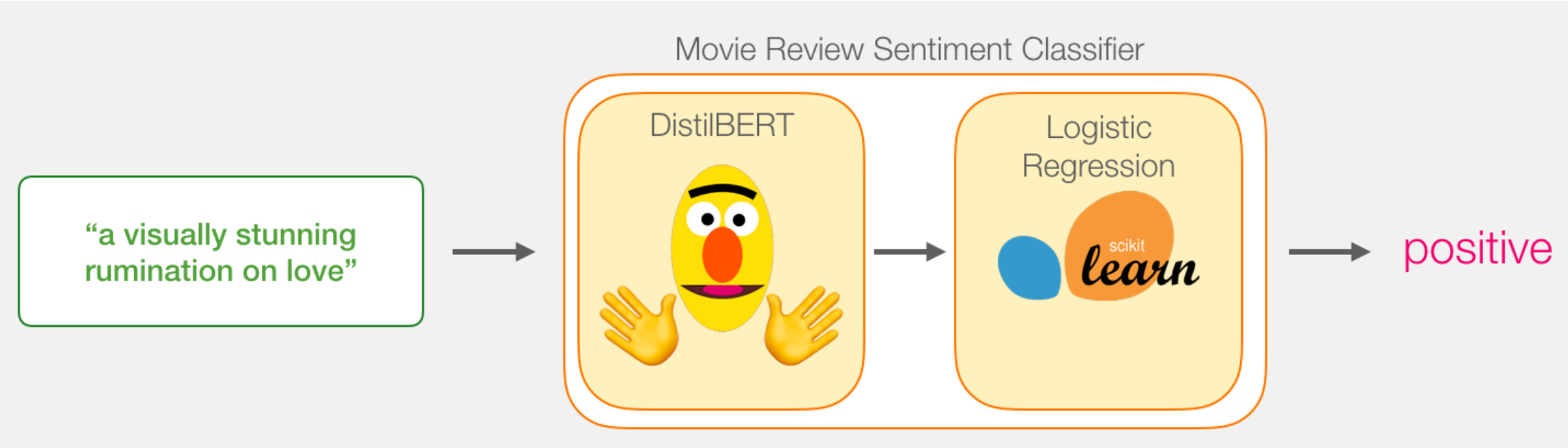
Sentence Sentiment Classification

Dataset

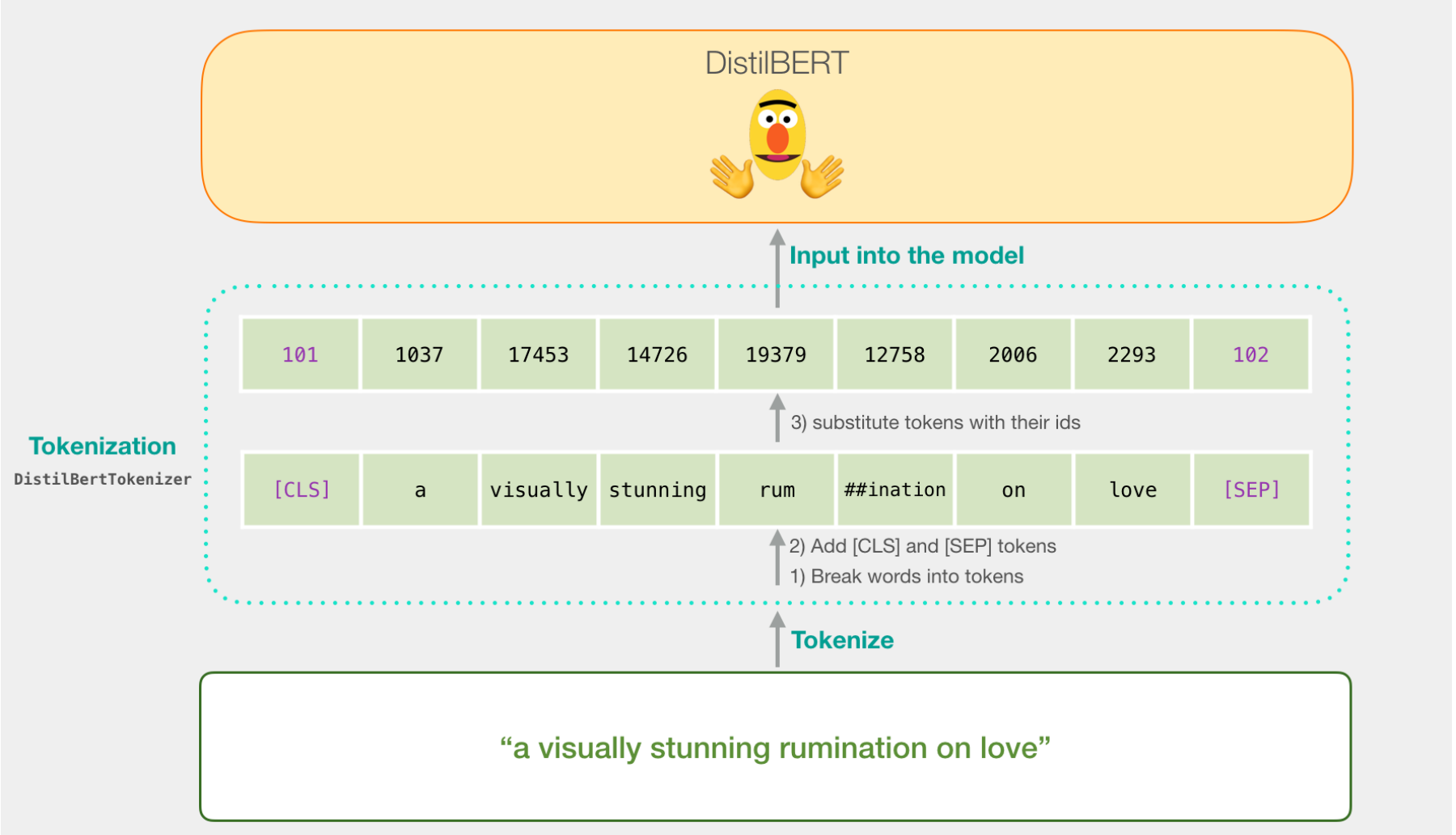
- The dataset we will use in this example is SST2, which contains sentences from movie reviews, each labeled as either positive (has the value 1) or negative (has the value 0):

sentence	label
a stirring , funny and finally transporting re imagining of beauty and the beast and 1930s horror films	1
apparently reassembled from the cutting room floor of any given daytime soap	0
they presume their audience won't sit still for a sociology lesson	0
this is a visually stunning rumination on love , memory , history and the war between art and commerce	1
jonathan parker 's bartleby should have been the be all end all of the modern office anomie films	1

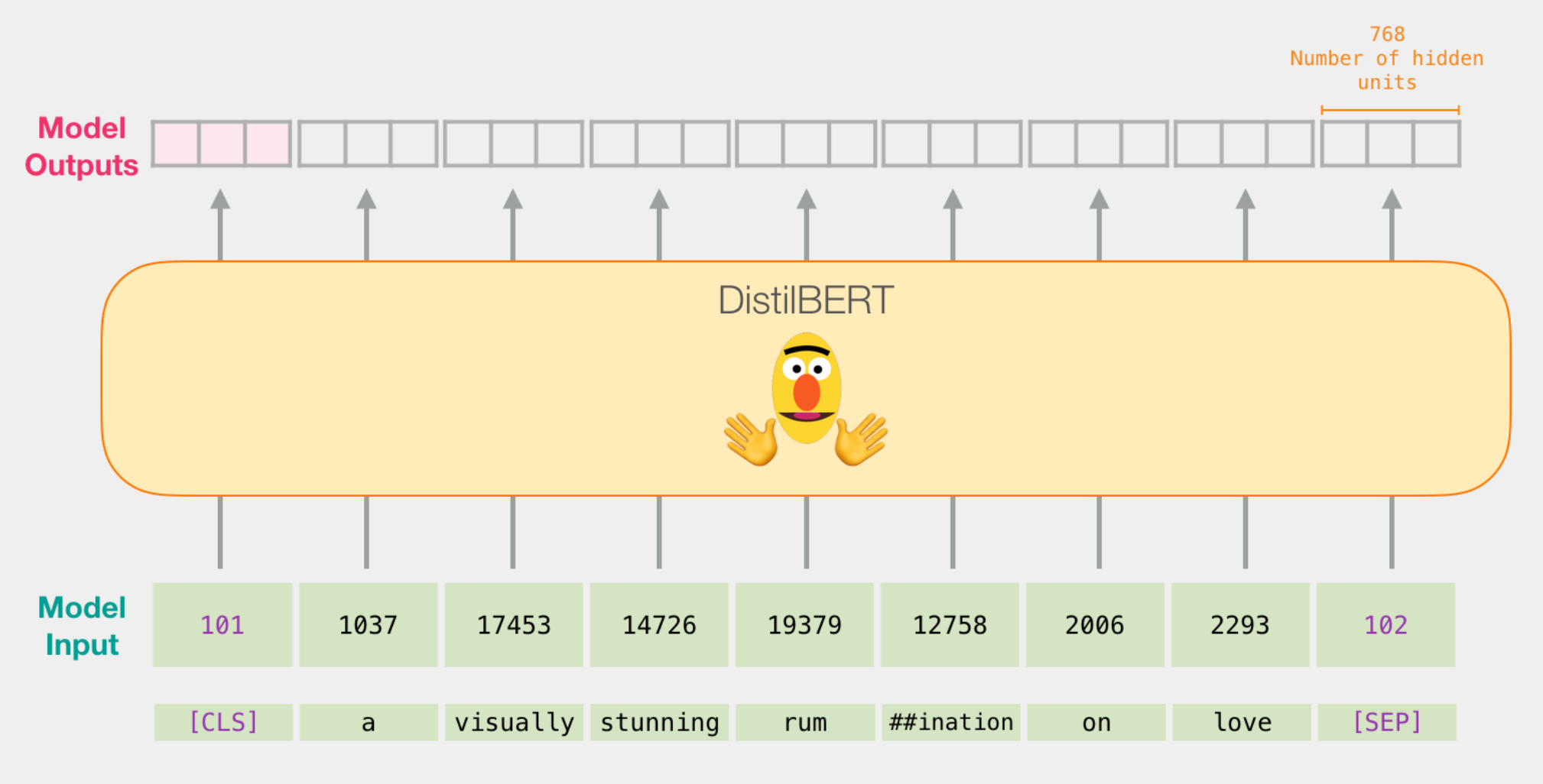
Model


Movie Review Sentiment Classifier
"a visually stunning rumination on love"
DistilBERT
Logistic Regression
scikit
learn
positive

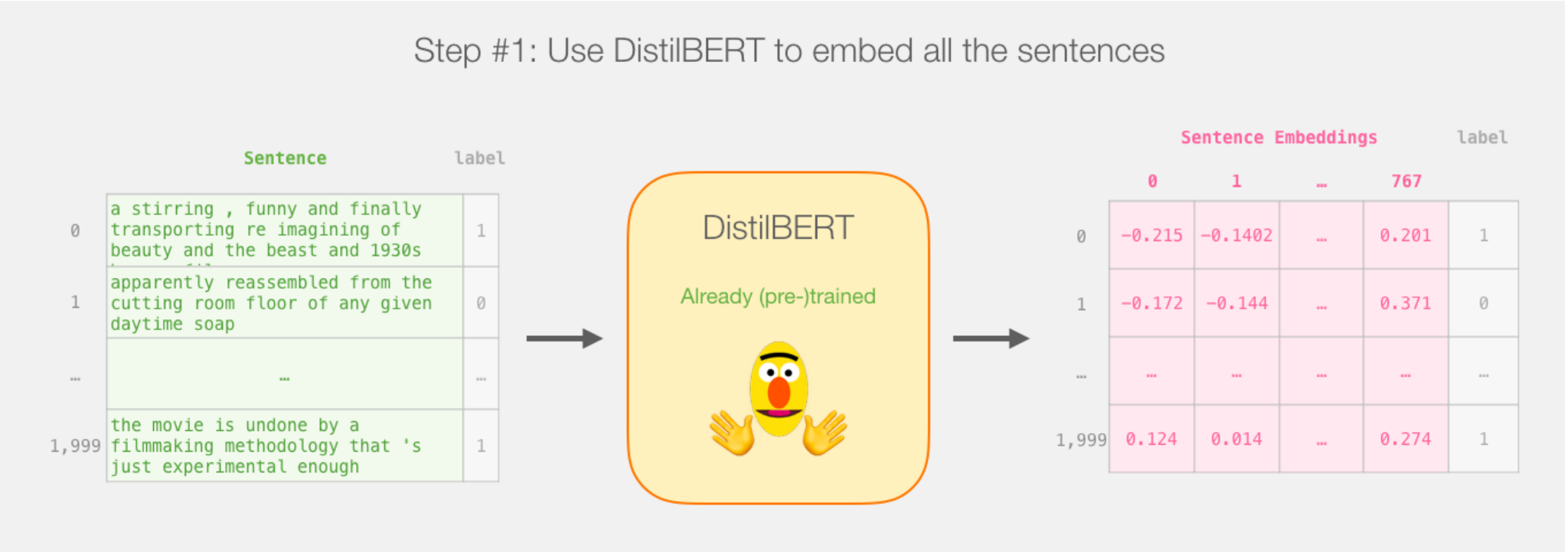
Tokenization



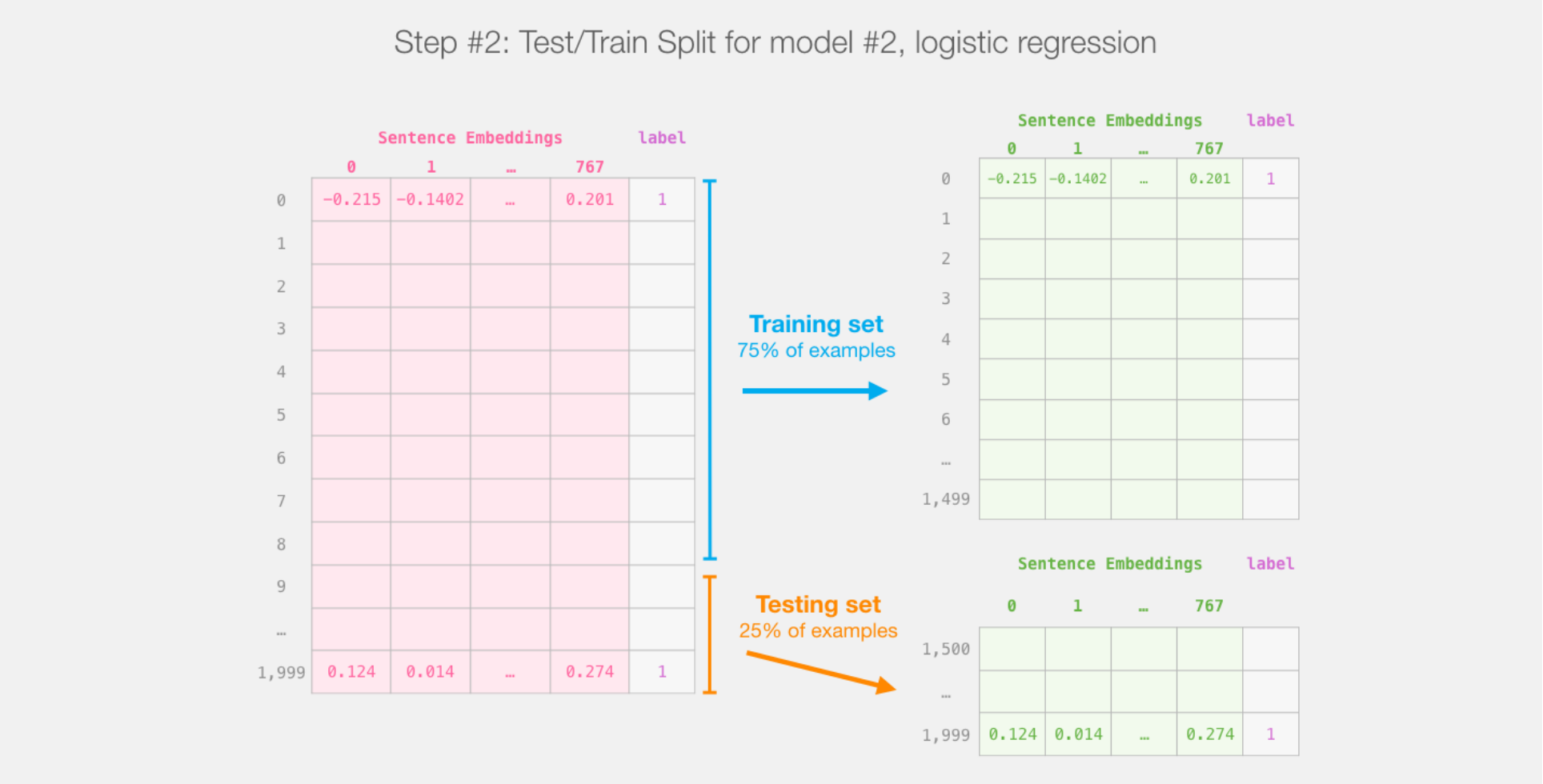
Sentence embedding



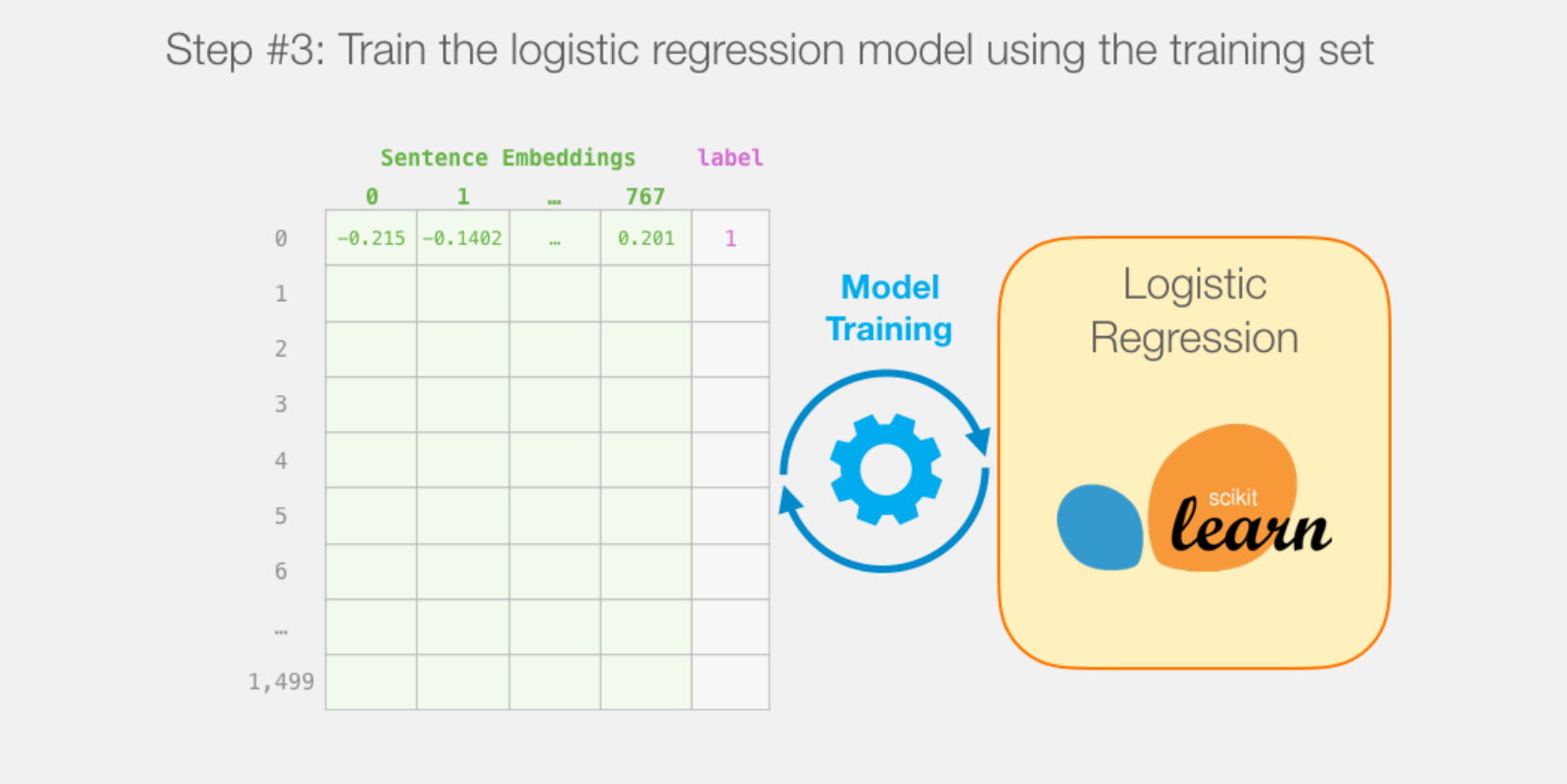
Generating data

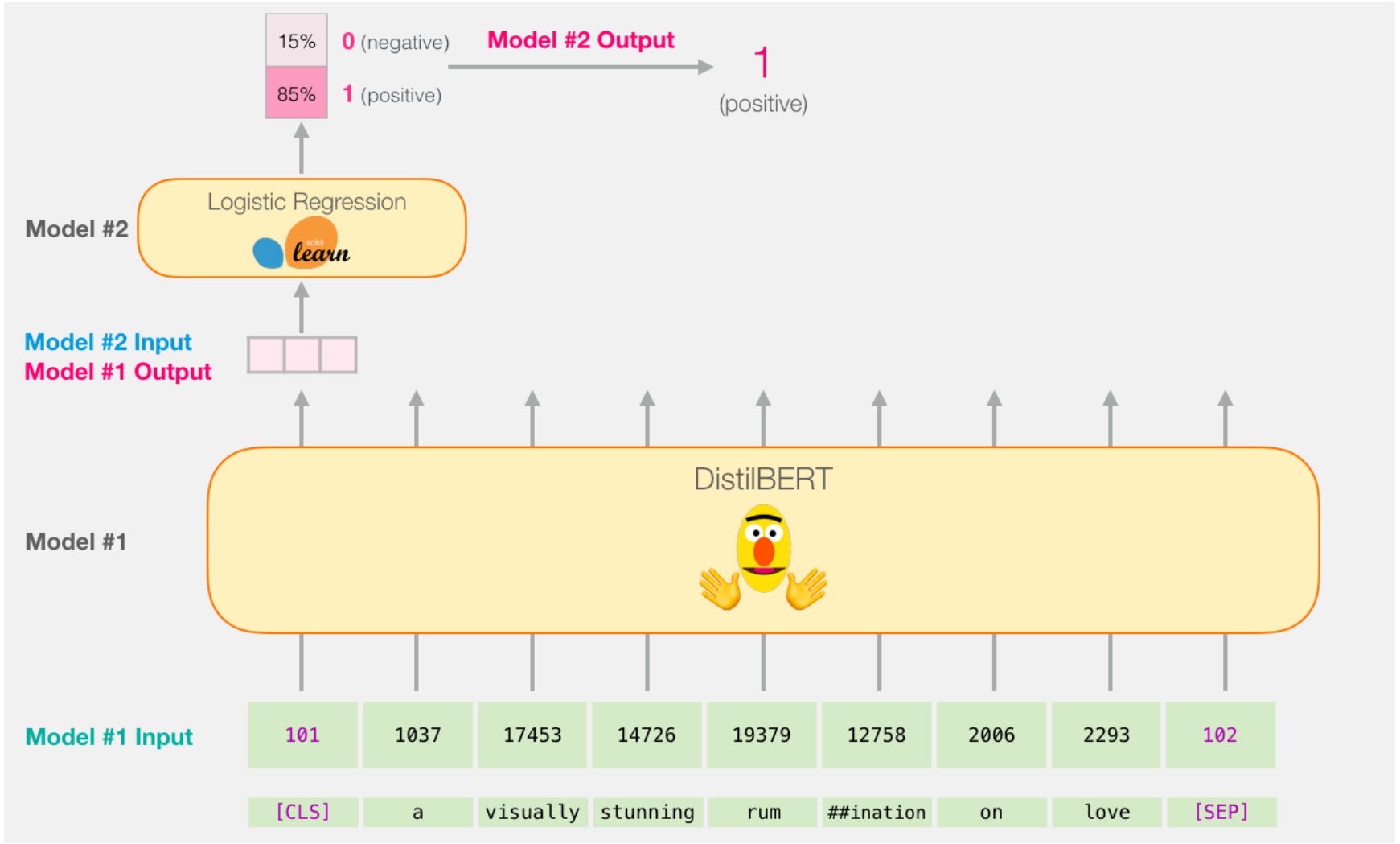


Train-test Split



Training Logistic regression




15%
0 (negative)
Model #2 Output
1
(positive)
85%
1 (positive)
Model #2
Logistic Regression
Model #2 Input
Model #1 Output
Model #1
DistilBERT
Model #1 Input
101
1037
17453
14726
19379
12758
2006
2293
102
[CLS]
a
visually
stunning
rum
##ination
on
love
[SEP]

References

- https://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/
- https://jalammar.github.io/illustrated-bert/