SENTIMENT ANALYSIS WITH BERT

By Leck Boon Kong

OVERVIEW

01
WHAT IS BERT?

02
HOW IS BERT TRAINED?

03
WHY USE BERT?

04
SENTIMENT ANALYSIS DEMONSTRATION

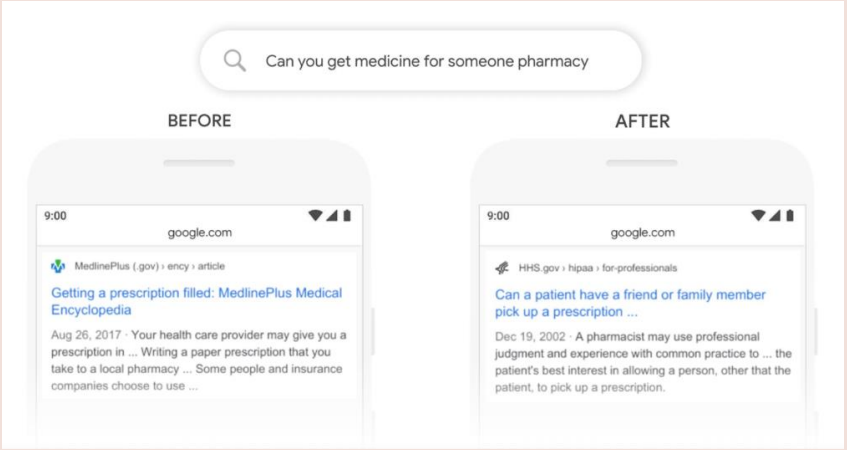
WHAT IS BERT?

History

- Developed by Google AI Language in 2018
- Past models **could not understand natural language** very well and were **single-use**

- Revolutionized the NLP space as it could be used for **11+ language understanding tasks**:
  - Sentiment Analysis
  - Question Answering
  - Text Prediction
  - Summarization and more...

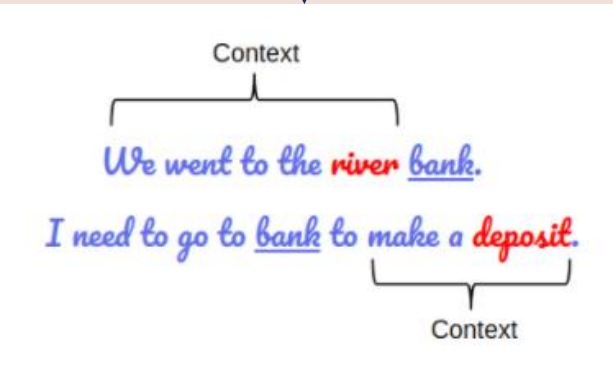
- BERT was trained on **3.3B words**
  - Wikipedia (2.5B words)
  - Google's BookCorpus (800M words)
  - 64 TPUs over **4 days**



WHAT IS BERT?

BERT

Bidirectional Encoder Representations from Transformers



Bi-directional

- Reads text sequence in **both directions**
- Better understanding of **language context** based on surrounding words

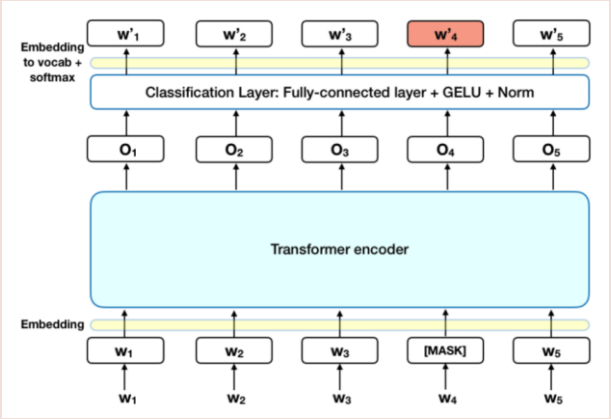
Attention mechanism

- Enable the model to have longer memory
- **Retains context** from previous words
- Greater weights on more **important words**

HOW IS BERT TRAINED?

Masking

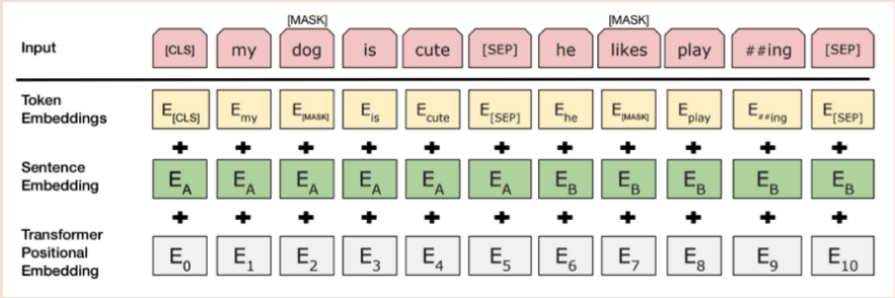
- **15%** of words are replaced with a **[MASK] token**



- Classification layer **predicts probability** of each token
- Multiply vector by **embedding matrix** to get English word
- Use softmax to get word with **highest probability**

Next Sentence Prediction

- Receives **pairs of sentences** as input
- 50% is correct
  - second sentence is **original subsequent sentence** in document
- 50% is wrong
  - second sentence is **random sentence** in corpus



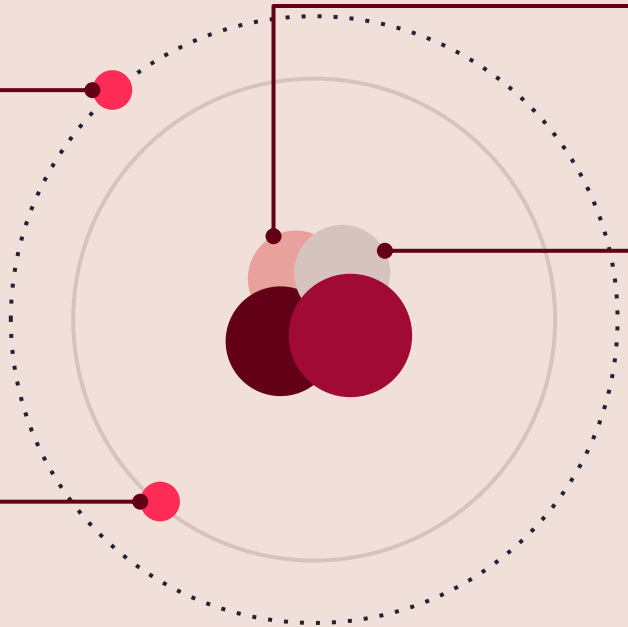
WHY BERT?

Context Understanding

- Past models map every word to a vector (GloVe, word2vec)
- BERT processes a word by looking at **all surrounding words**

Transfer Learning

- Pre-trained on unlabelled text in an **unsupervised manner**
- Serves as a knowledge base to be **fine-tuned** to suit specific tasks



Open Source

- Easy access to **information and code** for implementation
- Creation of **variants of BERT** for different purposes

Ease of Use

- Very **few data cleaning steps** required for BERT model
- **Short time and effort** required to fine-tune model for task

SENTIMENT ANALYSIS DEMONSTRATION

Determine if text is positive, negative or neutral.

DATASET

Seattle Airbnb Open Dataset

https://www.kaggle.com/datasets/airbnb/seattle?select=reviews.csv

CONVENTIONAL NLP DATA CLEANING

STEP 01

Words
"The cat sat on the mat."

Tokens
"The", "cat", "sat", "on", "the", "mat"

Removing punctuations and converting the words to tokens.

STEP 02

Tokens
~~"The"~~, "cat", "sat", ~~"on"~~, ~~"the"~~, "mat"

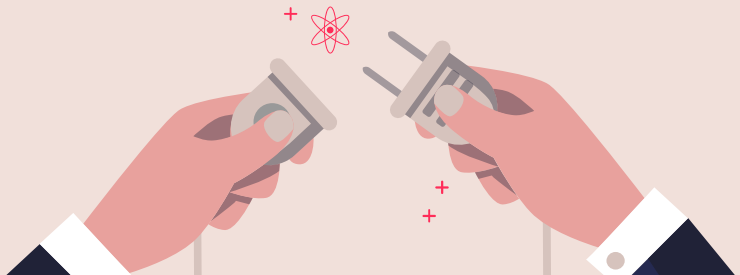
Removing stopwords

STEP 03

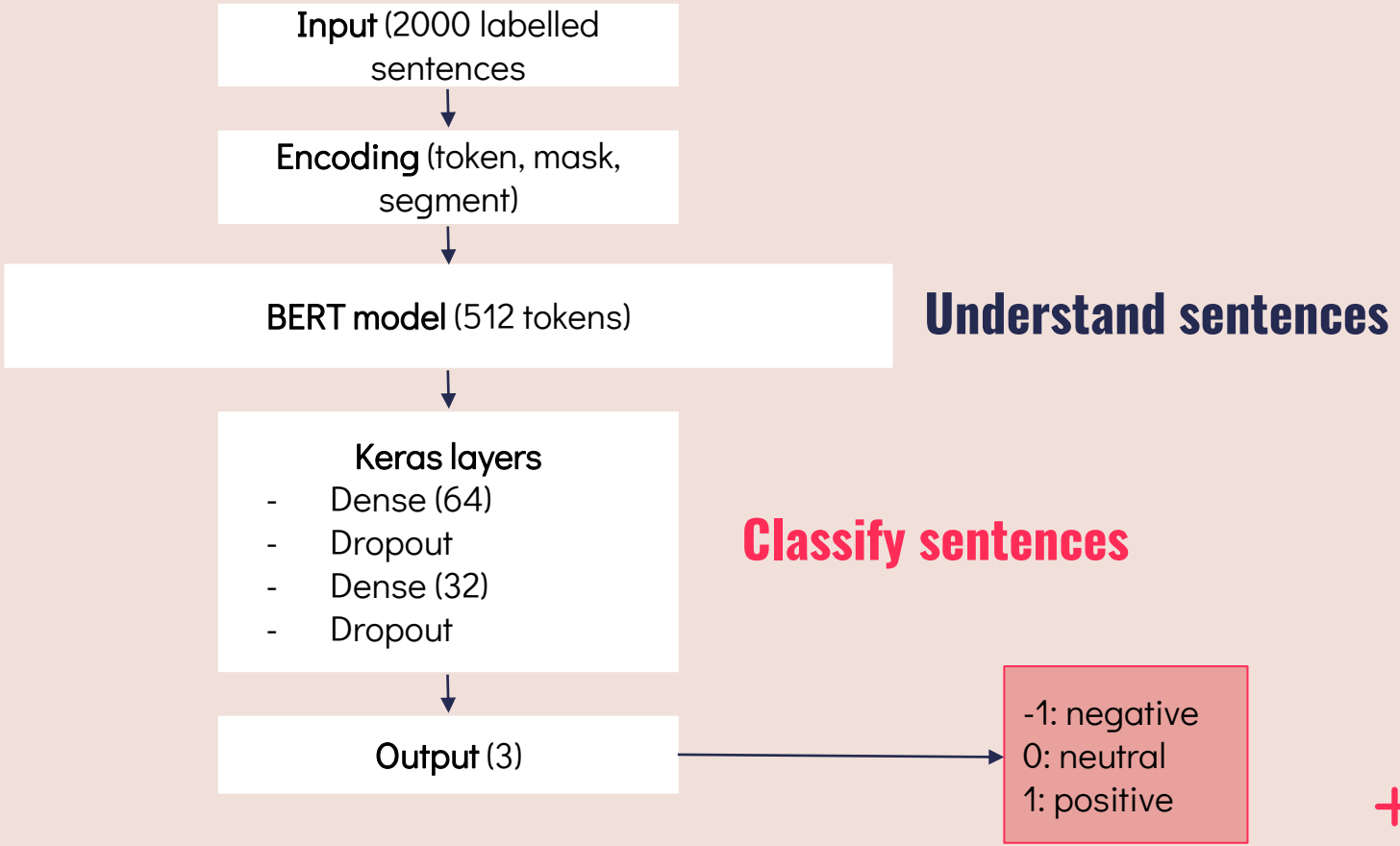
	original_word	lemmatized_word
0	trouble	trouble
1	troubling	trouble
2	troubled	trouble
3	troubles	trouble

	original_word	lemmatized_word
0	goose	goose
1	geese	goose

Lemmatize Words



SUPERVISED SENTIMENT ANALYSIS



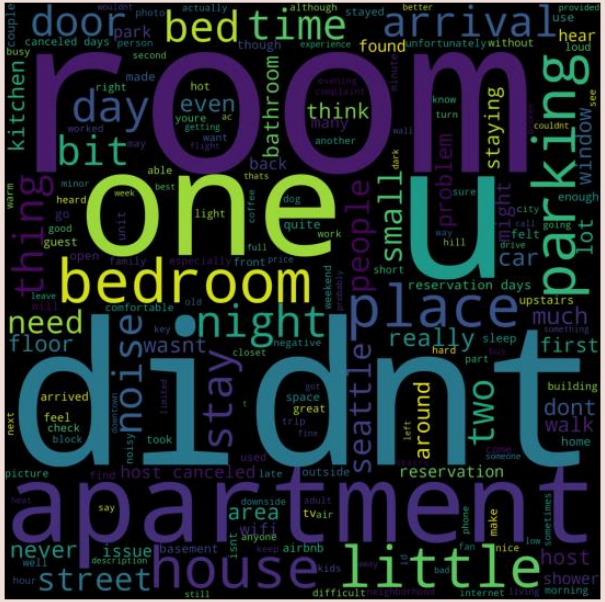
SUPERVISED SENTIMENT ANALYSIS

Accuracy on train set: **90%**

Accuracy on test set: **88%**

Positive word cloud generated

Negative word cloud generated



BERT VARIANTS

DistilBERT

- Smaller, faster and cheaper
- 40% less parameters, run 60% faster and preserves 95% of BERT's performance

DeBERTa

- Decoding enhanced BERT with disentangled attention
- Outperforms other BERT variants in most tasks

FinBERT

- Fine-tuned on Financial PhraseBank dataset
- Used for financial sentiment classification

Multilingual BERT

- Trained on the top 100 languages with the largest Wikipedias
- Perform same functions but with more language types

Thank You!

Any Questions?

Resources

- Google Drive Link
- Kaggle Link
- BERT Research Paper
- BERT Variants