News Article Classification Using NLP

19K41A0543 K. SRIDHAR SAI

19K41A0544 K. SRIKAR SAI

19K41A0451 P. SAI GANESH

🕮 INTRODUCTION

- Reading news is a regular hobby for almost all the people in this world. Different people likes different categories of news like sports, entertainment, political etc.
- It is recommended to news websites to organize them into categories so that users may quickly get access to the categories of news that they are interested.
- In Natural Language Processing (NLP), text classification is a classical problem which aims assigning labels or tags to vectors has various applications like text summarization, question answering, sentiment analysis, news categorization and so on.

PROJECT OBJECTIVE

The main objective of the project is to classify a news article into different categories like politics, entertainment etc.

PROJECT BENEFICIARIES

- It simplifies the task of classifying the news based on different categories.
- It also helps to get access to the specific article quickly rather than searching manually.

DATASET INSIGHTS

Each record in the dataset consists of the following attributes:

category: category in which the article was published.

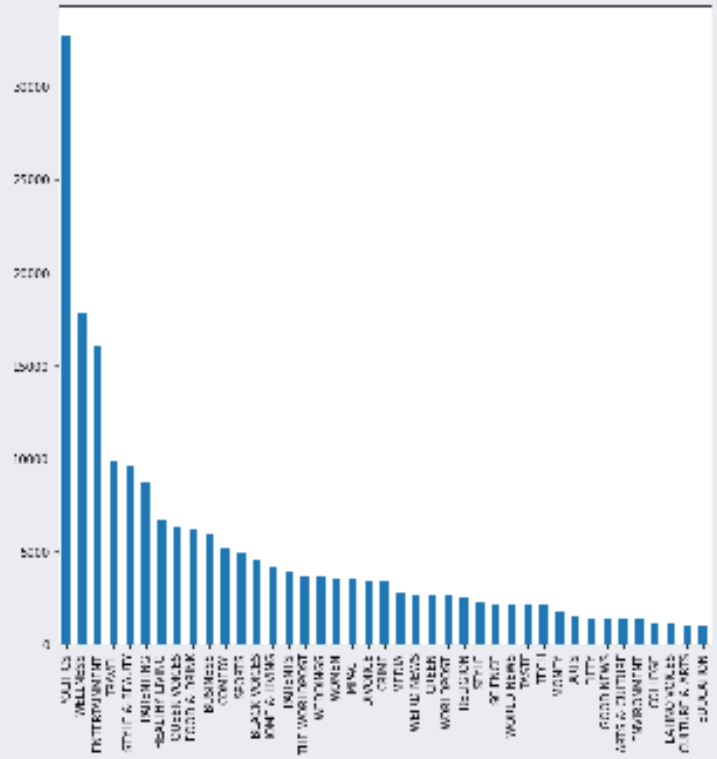
headline: the headline of the news article.

authors: list of authors who contributed to the article.

link: link to the original news article.

short_description: Abstract of the news article.

date: publication date of the article.



ALGORITHM USED

- ✓ Bi-Directional Encoder Representation From Transformers
- ✓ BERT Large — 24 layers (transformer blocks), 16 attention heads, 340 million parameters, and has an output size of 1024-dimensions.

- Loss : categorical_crossentropy
- Activation function : softmax
- Optimizer : Adam
- Total params: 335,182,888
- Trainable params: 335,182,888
- Non-trainable params: 0

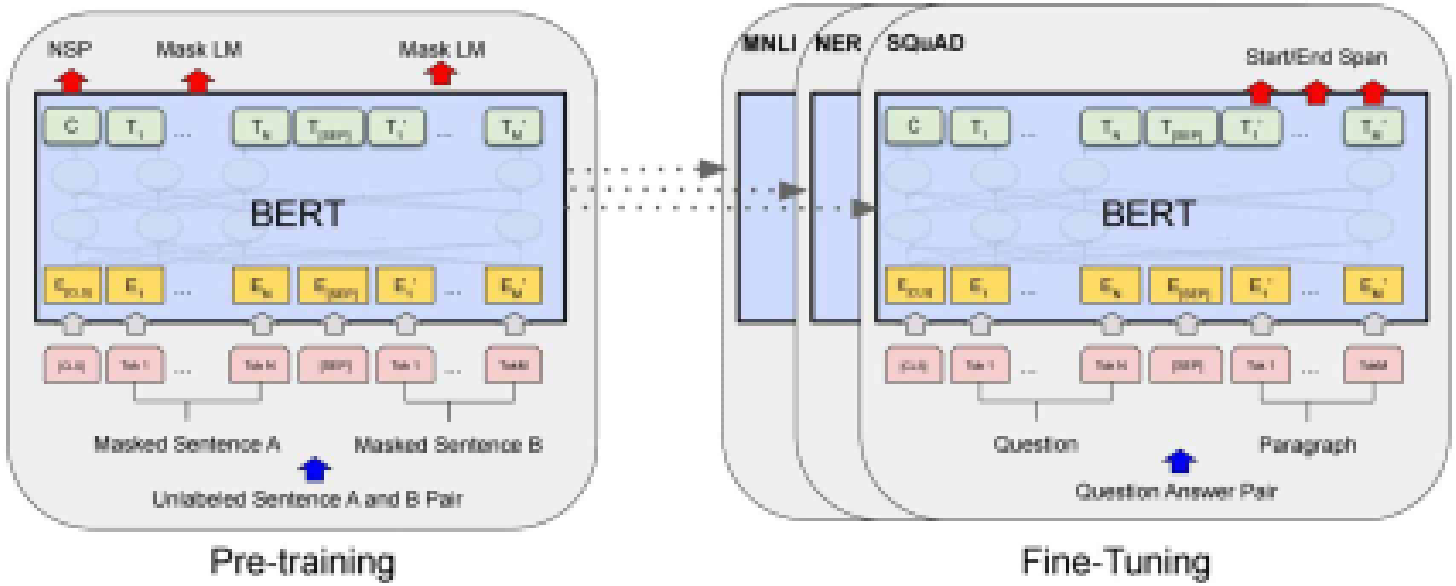


Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. `[CLS]` is a special symbol added in front of every input example, and `[SEP]` is a special separator token (e.g. separating questions/answers).

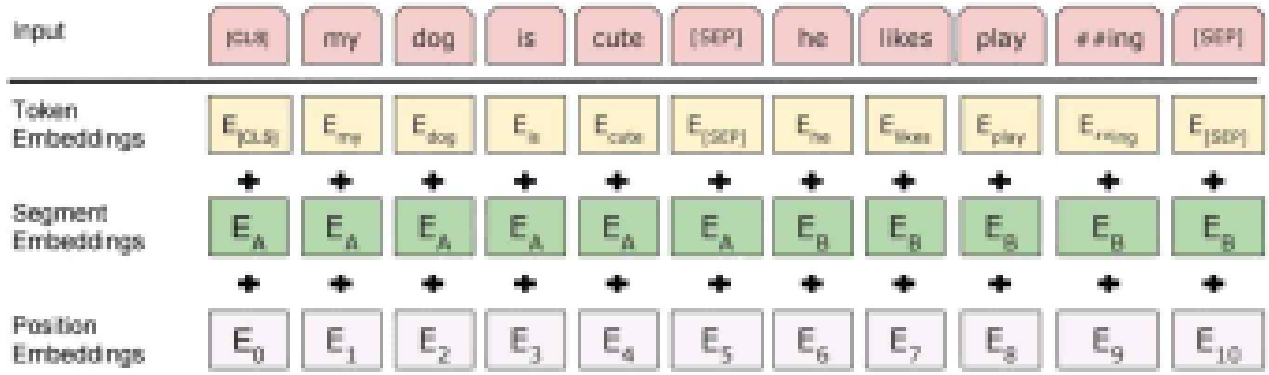
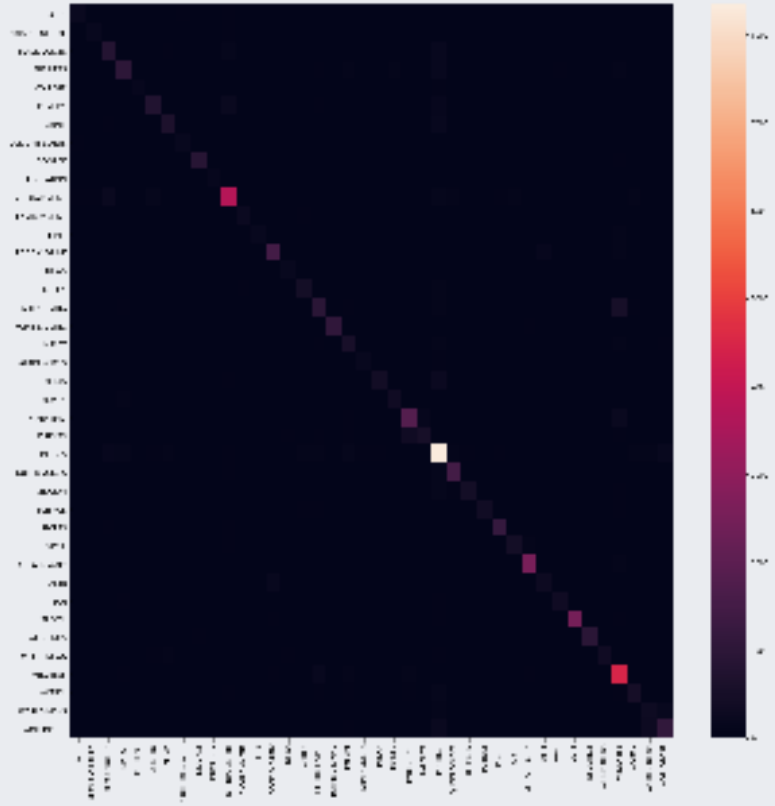
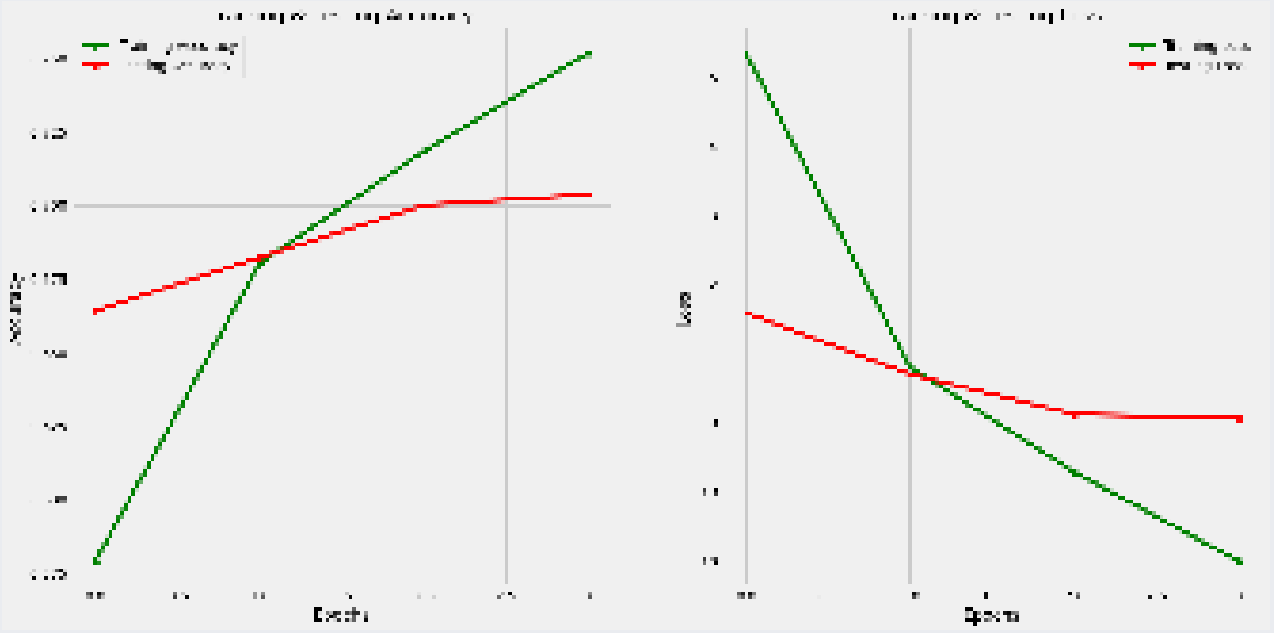


Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

CONFUSION MATRIX



RESULTS



THANK YOU