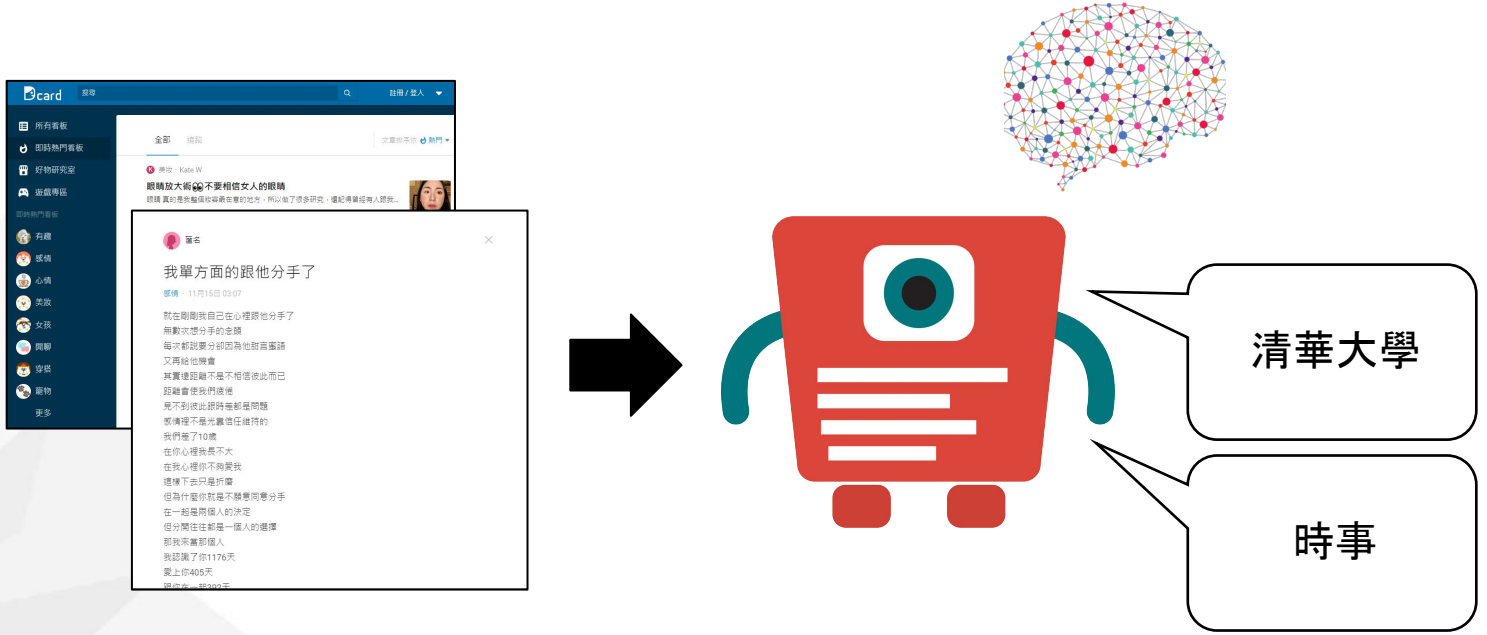
Dcard Article Analysis

Team18 / 何政儒 江岷錡 林奕鑫 蔡政諺 吳宗逸

Introduction

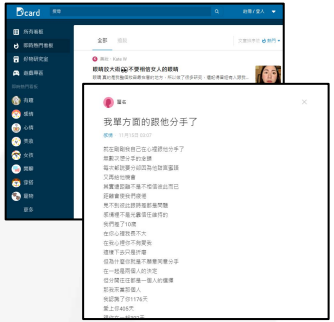
Dcard Article Analysis

- Analyze the articles in Dcard, and predict author information.
- Analyze the tags and comments in Dcard.

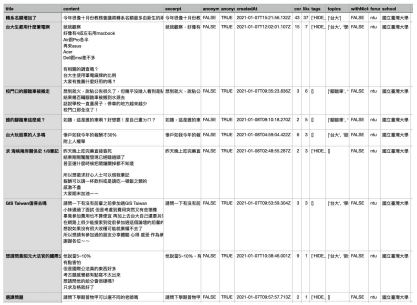


Overall Process

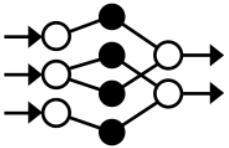
Dcard Article Analysis



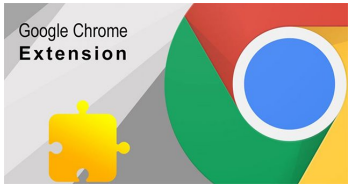
Web Crawler



NLP Model



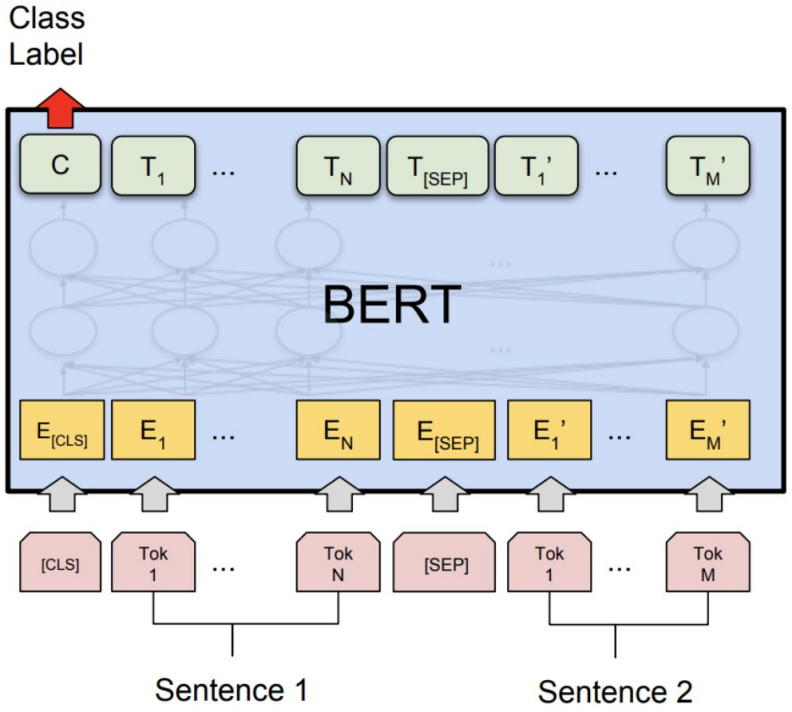
Web Deployment



NLP Model

Dcard Article Analysis

- BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers, 2018)

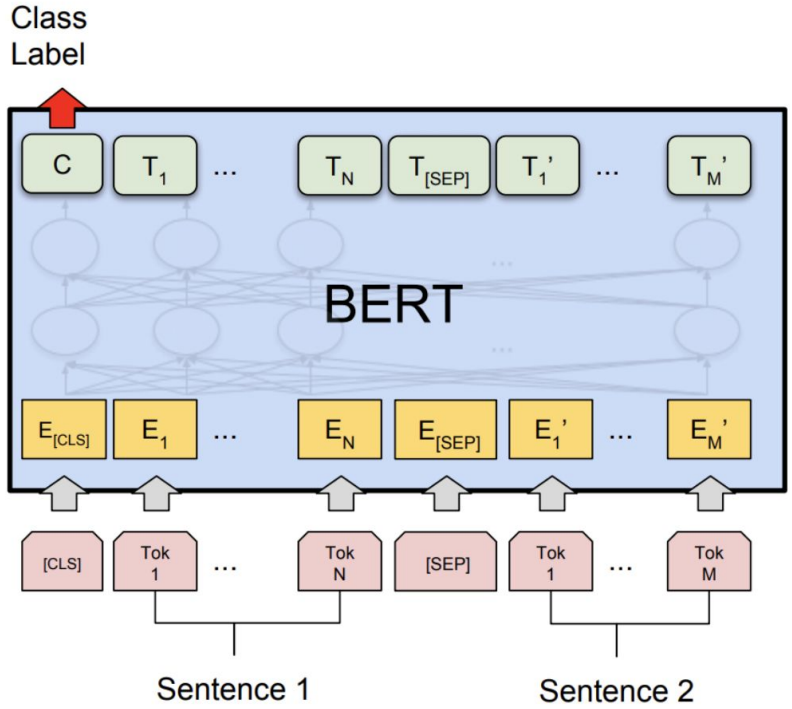


BERT: What and Why?

Dcard Article Analysis

- BERT = Encoder of Transformer
- Self attention
  - Compute in parallel
- Positional encoding
  - **Elven** hit **the ball**
  - **The ball** hit **Elven**
- Contextualized word embedding
  - money **bank**
  - river **bank**

BERT learns different meaning

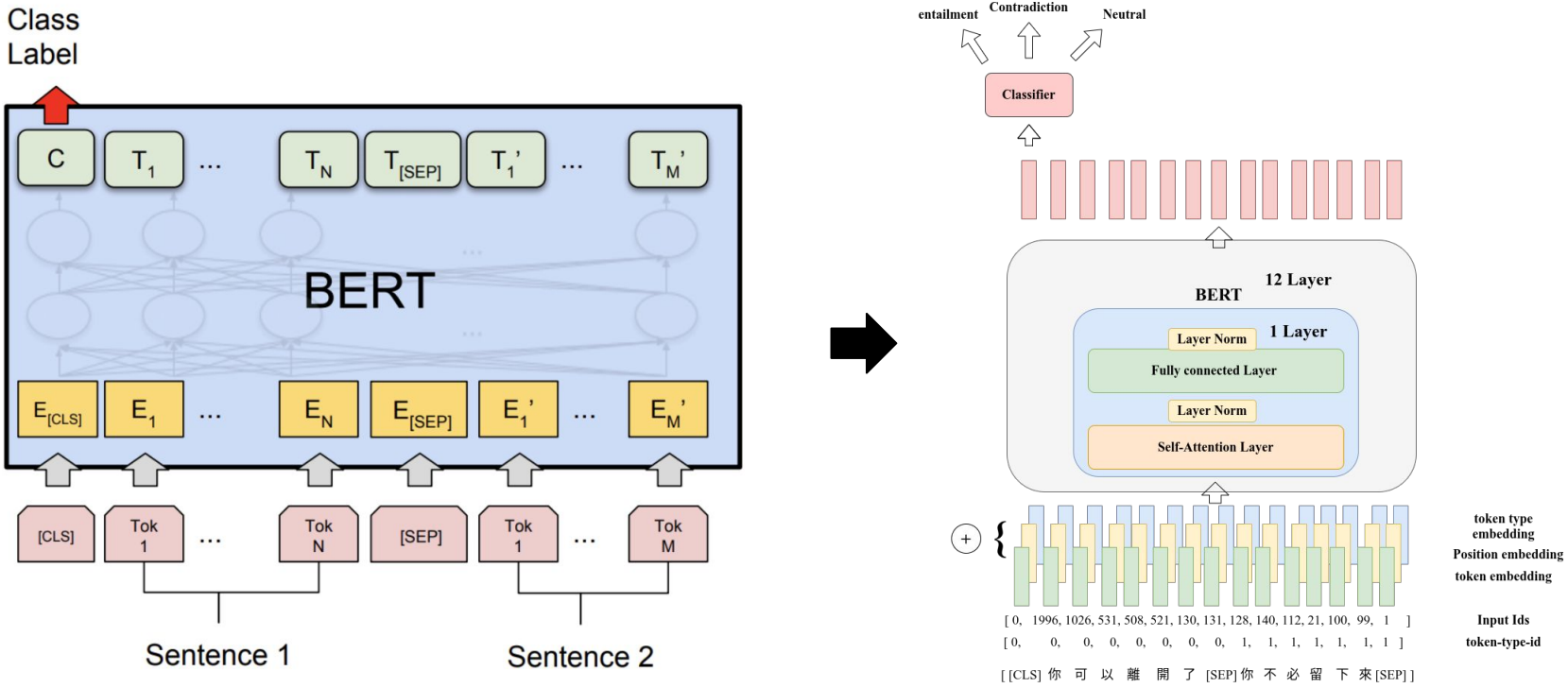


Reference:
https://www.youtube.com/watch?v=ugWDIIOHtPA
https://www.youtube.com/watch?v=UYPa347-DdE
https://www.youtube.com/watch?v=1_gRK9EIQpc

BERT: How to Train?

Dcard Article Analysis

- Fine-tune the **pre-trained Chinese BERT** and train the **classifier**.



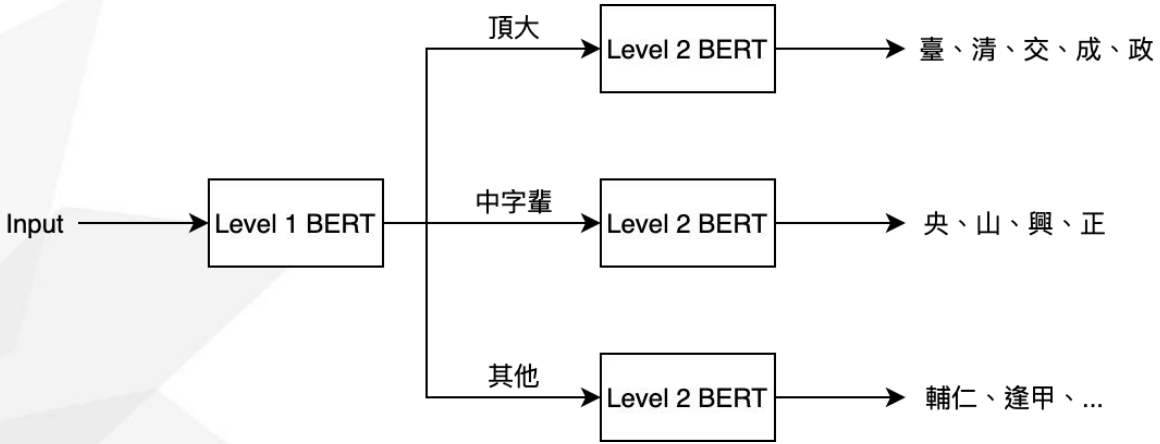
School Classification

Dcard Article Analysis

- Multi-class classification (14 schools + others)

```
labels = {'國立臺灣大學': 0, '國立清華大學': 1, '國立交通大學': 2, '國立成功大學': 3, '國立政治大學': 4,
          '國立中央大學': 5, '國立中山大學': 6, '國立中興大學': 7, '國立中正大學': 8,
          '輔仁大學': 9, '逢甲大學': 10, '東海大學': 11, '義守大學': 12, '中原大學': 13, '其他': 14}
```

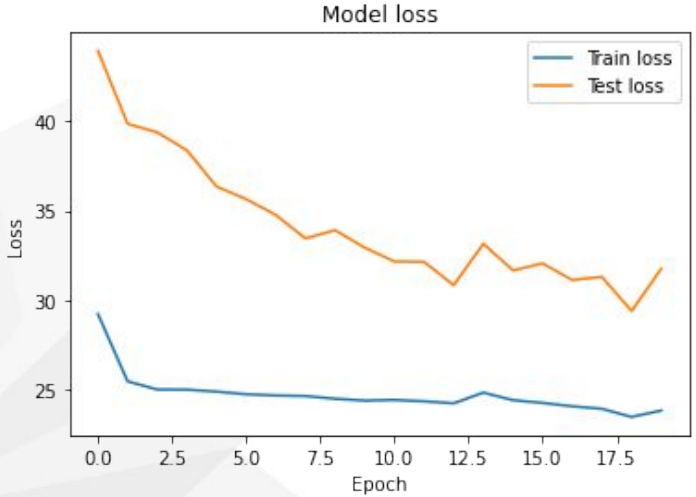
- Hierarchical Structure

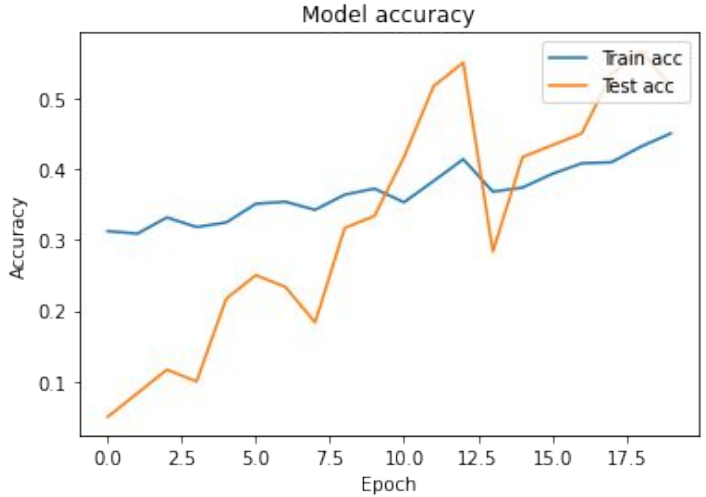


School Classification

Dcard Article Analysis

- Dataset
  - **15,000 labeled articles** from school forums (NTU, NTHU, ...) for training.
  - **1,000 labeled articles** from popular forums (funny, mood, ...) for validation.

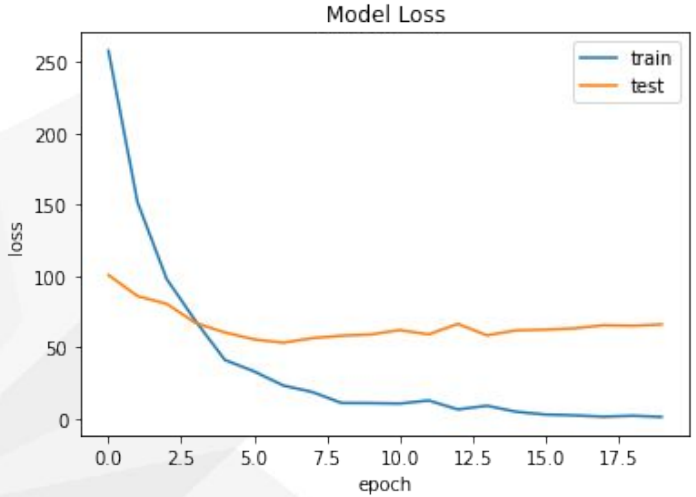


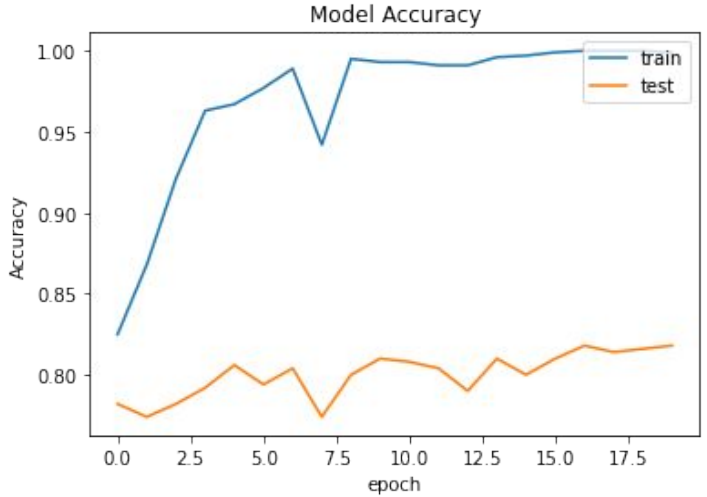


Tag Classification

Dcard Article Analysis

- Dataset
  - Label: Mood, Love, Starsign, CurrentEvents
  - **1,000 labeled title for each label forums** for training. (Total: 4,000)
  - **200 labeled title for each label forums** for validation. (Total: 1,000)

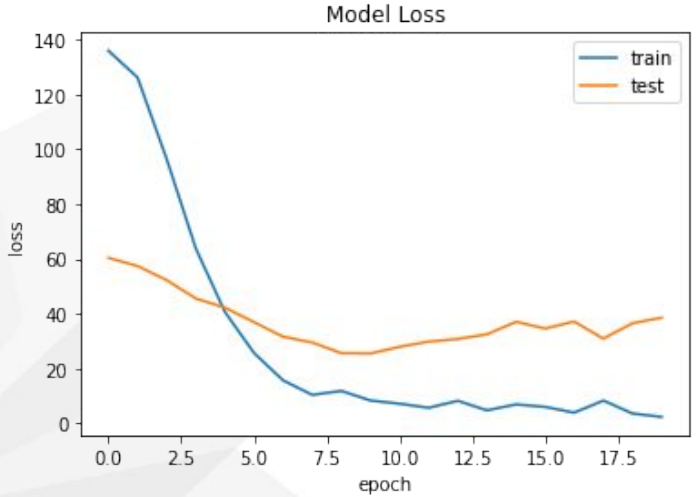


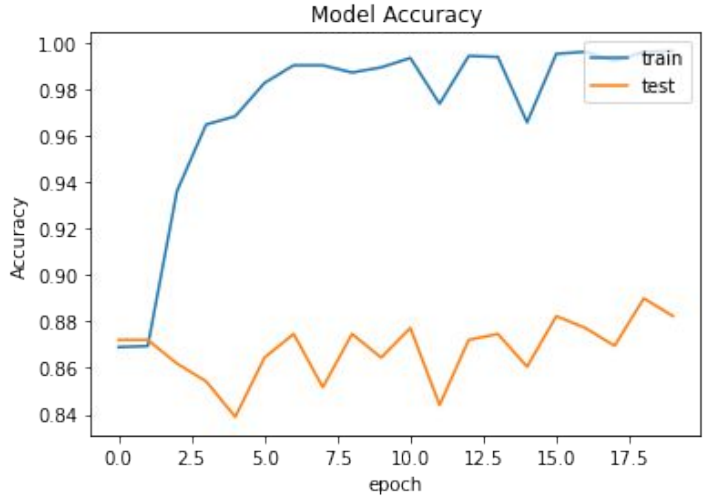


Comment Classification

Dcard Article Analysis

- Dataset
  - Label: Positive, Negative, Non-Correlated
  - **1,000 labeled articles** from popular forums (funny, mood, ...) for training.
  - **200 labeled articles** from popular forums (funny, mood, ...) for validation.





Live Demo

Dcard Article Analysis