111550079 方品仁 Report

[GitHub with data](https://github.com/Fang0105/AI_Capstone/tree/main/HW1/data)

# Topic

The research question is about NLP. The task for supervised learning models is sentiment analysis, and for unsupervised learning model is topic modeling.

# Description Of Dataset

The training dataset is obtained from comments of a video on YouTube. The topic of the video is the election debate between Joe Biden and Donald Trump. These comments are fetched from Google API, filling the id of the video and save into a csv file. Then remove emojis in the comments and get rid of those comments that are not in English. Finally, use Textblob model to label each comment, 1 for Positive, 0 for Neutral, -1 for Negative. This dataset is for training and testing during the training process. Besides the comments from the debate, I downloaded other comments from a video whose topic is the discussion of the debate between NewJeans, a Korean girl group, and its companies, Hybe and Ador in court. This dataset is for testing the model only.

|  |  |  |
| --- | --- | --- |
|  | Training | Testing |
| Election debate | 55877 | 13970 |
| Court debate | X | 123 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Positive | Neutral | Negative |
| Election debate | 27917 | 25364 | 16566 |

# Supervised Learning

## Logistic Regression with Tfidf

Term Frequency-Inverse Document Frequency (Tfidf) is a statistical measure used in information retrieval and text mining to evaluate the importance of a word in a document relative to a collection of documents. Logistic Regression model is a model for classification problem. In this task, it takes the Tfidf value of a comment as input. Both Tfidf and Logistic Regression model are implemented in Scikit-Learn (sklearn), am open-source machine learning library for Python, built on top of Numpy, SciPy, and Matplotlib.

### Result

* Cross-Validation Scores:

[0.82202935, 0.82784538, 0.82487696, 0.82290828, 0.82917226]

* Mean Accuracy: 82.54%
* Standard Deviation: 0.0028
* Final Accuracy: 82.763%

|  |  |
| --- | --- |
|  |  |
|  |  |

## BERT with Word2Vec

Word2Vec is a shallow, two-layer neural network to create word embeddings. It learns relationships between word based on contest. It creates fixed-size word embeddings. But doesn’t handle polysemy. Bidirectional Encoder Representations from Transformers (BERT) is a deep neural network based on the transformer architecture. It revolutionized NLP by understanding context bidirectionally, means it considers both left and right context when processing words. In this work, I don’t train a BERT model from begin. Instead, I used pretrained BERT model to do transfer learning. This approach can save me a lot of time. Word2vec is implemented in sklearn; while BERT can be done with PyTorch transformer.

### Training Parameters

* Batch Size: 16
* Epochs: 5
* Initial Learning Rate: 5e-5
* Weight Decay: 0.01

### Result

* Accuracy: 96.13%

|  |  |
| --- | --- |
|  |  |
|  |  |

# Unsupervised Learning

## LDA

Latent Dirichlet Allocation (LDA) is a generative probabilistic model for topic modeling in NLP. IT helps uncover the hidden topics in a collection of documents. LDA model is implemented in sklearn.

### Result

一張含有 文字, 螢幕擷取畫面, 字型, 軟體 的圖片

AI 產生的內容可能不正確。

### Analysis

The result from LDA is not as my expected. I expected that the topics the model could have found was terms in some specific aspect such as economic, civilian, transport, and war. But it turned out that a lot of key word are “Biden” and “Trump”, which are obviously not I wanted to see.

# Experiment

## The Influence of The Size of The Dataset

The Logistic Regression model and BERT model are trained with different size dataset, 100%, 50%, 25%. The expected result was that the accuracy went down as the size of the dataset shrink. The real result was that the accuracy indeed went down as the size of the dataset shrink. But it didn’t decrease significantly, for both Logistic Regression model and BERT model. Therefore, the conclusion for my experiment is that the amount of training data doesn’t really influent the performance.

|  |  |  |  |
| --- | --- | --- | --- |
| LR model | 100% | 50% | 25% |
| Accuracy | 82.763% | 81.789% | 80.103% |

|  |  |  |
| --- | --- | --- |
| LR model | Classification Report | Learning Curve |
| 100% |  |  |
| 50% |  |  |
| 25% |  |  |

|  |  |  |
| --- | --- | --- |
| LR model | Confusion Matrix | ROC Curve |
| 100% |  |  |
| 50% |  |  |
| 25% |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| BERT | 100% | 50% | 25% |
| Accuracy | 96.13% | 94.69% | 93.44% |

|  |  |  |
| --- | --- | --- |
| BERT | Confusion Matrix | ROC Curve |
| 100% |  |  |
| 50% |  |  |
| 25% |  |  |

## Performance on different topic dataset

The goal of this experiment is to check whether the model trained with politic debate dataset can also perform well on course debate dataset. I expected that the accuracy might decrease due to the different topic.

The result fits what I expected. Both models have different level of decrease in accuracy. The accuracy of Logistic Regression model drops from 82.763% to 76.422%. And the accuracy of BERT model drops from 96.13% to 86.17%.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Politic Debate | Course Debate | Difference |
| LR model | 82.763% | 76.422% | 6.341 |
| BERT model | 96.13% | 86.17% | 9.96 |

# Discussion

* The result of BERT and LR model are much better than I have thought. But after seeing the performance, I thought the performance of LDA model would have also been great. But it turned that the performance was not ideal. I thought the topics the model found would be from more deeper concepts.
* The good performance for both Logistic Regression model and BERT model can be attributed to the topic of the dataset. When it comes to politics, people tend to be fierce then give out comments with emotional words. For example, humiliating words. These kinds of words can be easily learned by models.
* I want to do more experiments on different dataset with different topics, some topics that people can give more objective comments, checking whether the performance can also be good.
* This is the first NLP project for me. I learned a lot about NLP, such as models and tokenization. I also learn how to use tensorboard to visualize the training process and Hugging face trainer to train a model.

# Appendix

* Logistic Regression

一張含有 文字, 螢幕擷取畫面, 軟體 的圖片

AI 產生的內容可能不正確。

* BERT

一張含有 文字, 螢幕擷取畫面, 軟體, 作業系統 的圖片

AI 產生的內容可能不正確。

一張含有 文字, 螢幕擷取畫面, 軟體, 多媒體軟體 的圖片

AI 產生的內容可能不正確。

* LDA

一張含有 文字, 螢幕擷取畫面, 軟體, 作業系統 的圖片

AI 產生的內容可能不正確。