|  |
| --- |
| IE4483 |
| Sentiment Analysis |
|  |

|  |
| --- |
| KOH JIAHUI, SHAO SIYING, PYAE SONE KHIN#  10-29-2023 |

## BERT

### Feature Format

For the task of sentiment analysis using a BERT model, the chosen feature format is tokenized and encoded text data. This format is a standard practice in natural language processing (NLP) and ensures that raw textual data can be effectively utilized by the model for sentiment classification.

### Data Preprocessing Procedures

1. Text Cleaning:

* Lowercasing: All text data is converted to lowercase to maintain consistency and prevent case-related variations.
* Removing Punctuation: Special characters, punctuation marks, and symbols are removed from the text to focus on meaningful words.
* Removing Numbers: Numerical digits or symbols are excluded from the text data, as they may not be relevant for sentiment analysis.
* Removing Stopwords: Common, non-discriminative words (stopwords) are eliminated from the text, such as "the," "and," "is," etc.
* Removing Duplicated Rows: Duplicate entries, if present, are removed to ensure data quality.

1. Tokenization:

Breaking down text into words or subword units is an important step in NLP. When using BERT, words are represented as subword pieces through subword tokenization. This process relies on a trained tokenizer which can be found in the Hugging Face Transformers library.

1. Padding and Truncation:

BERT models require input sequences with a fixed length. To achieve this, padding is used to extend shorter sequences and truncation is used to shorten longer sequences. This ensures that all input sequences have the same length.

1. Encoding:

Text data is converted into numerical values that can be understood by the BERT model. The encoding process involves;

* input\_ids: A series of integer IDs that represent the tokens.
* attention\_mask: A binary mask that distinguishes between actual tokens and padding tokens.
* token\_type\_ids: Utilized to differentiate between two distinct sentence inputs. It is not used for BERT but included for compatibility.

1. Dataset Creation:

The structured dataset contains encoded text as well as target labels representing sentiments, where 0 indicates a negative sentiment and 1 represents a positive sentiment.

1. Data Splitting:

The dataset is divided into two sets, one for training and the other for validation. It is commonly done with an 80% ratio for training and a 20% ratio for validation purposes.

1. Batching:

During the training and prediction phase, the data is organized into batches for efficient processing. To manage these batches effectively, data loaders like PyTorch's DataLoader are utilized.

By following these data preprocessing procedures, the unprocessed text data undergoes a transformation to a suitable format that can be seamlessly fed into a BERT model. This format allows the model to comprehend and generate predictions based on the encoded text information.