Dataset Section:

benchmark datasets used in this project:

IMDB movie reviews dataset

Classification using Naïve Bayes model

BERT model for comparison

1. IMDB movie reviews dataset:

Contains 50,000 movie reviews, split equally into 25,000 training and 25,000 testing samples

Reviews are labeled as positive (1) or negative (0) based on the sentiment expressed

Data preprocessing steps:

Lowercased all text

Removed HTML tags

Removed punctuation marks

Tokenized the text using NLTK's word\_tokenize function

1. BERT model for comparison:

Same preprocessing steps as naïve bayes model

Pre-trained model 'bert-base-uncased' used for text classification

Tokenizer from Hugging Face's Transformers library was used to tokenize and encode the text

Maximum length of 256 tokens set for input sequences

Input sequences were padded and truncated to the specified maximum length

Attention masks generated to differentiate between actual tokens and padding

Exploratory analysis:

Class distribution was examined to ensure balance in the dataset

Vocabulary size was calculated after preprocessing steps

Attention weights from the BERT model were analyzed to understand the model's focus on different parts of input sequences

Discussion:

Comparison of Naive Bayes and BERT models:

Naive Bayes:

Simple probabilistic model

Quick training and prediction times

Relies on the bag-of-words representation, losing context and word order information

BERT:

Deep learning model based on transformers

Requires more computational resources and time for training and prediction

Captures context and word order information through self-attention mechanism

Benefits from pre-training on large external corpora

Performance on IMDB Reviews classification task:

Naive Bayes generally achieves lower accuracy compared to BERT

BERT demonstrates superior performance due to its ability to capture context and word order information

Examination of attention matrices:

Correctly predicted documents show focused attention on relevant words or phrases

Incorrectly predicted documents show scattered or misplaced attention

Conclusion:

Pretraining on an external corpus benefits BERT:

Helps the model capture general language patterns and semantics

Transfers knowledge gained from pretraining to the movie review prediction task

Improves the model's ability to understand context and sentiment in movie reviews

Performance differences between deep learning and traditional machine learning methods:

* Deep learning models like BERT generally outperform traditional methods like Naive Bayes on complex NLP tasks
* Deep learning models are more capable of capturing contextual information and complex patterns in the data
* Traditional machine learning methods can be more computationally efficient, but may struggle with complex tasks due to limitations in feature representation and modeling assumptions