A Comparative Study of Compression-Based Text Classification Against Deep Neural Networks

Zijian Zhang, Yiming Wang, Peiling Yu, Yunjie Qu

CIS 5300 Fall 2023 Project Presentation

Penn Engineering

Introduction

Penn Engineering

Models to Explore

Gzip + k-NN

Other Compressors + k-NN

DNN

Dataset 1 - AG News

- Dataset Description:
  AG is a collection of more than 1 million news articles, which are gathered from more than 2000 news sources by ComeToMyHead in more than 1 year of activity. And we are using the AG's news topic classification dataset from Hugging Face.

- Number of Rows (Training): 120,000
- Number of Rows (Testing): 7,600

- Classification Labels: 0 - World, 1 - Sports, 2 - Business, 3 - Sci/Tech

- Dataset Example:
  {'text': "Wall St. Bears Claw Back Into the Black (Reuters) Reuters - Short-sellers, Wall Street's dwindling band of ultra-cynics, are seeing green again.",
  'label': 2}

Dataset 2 - Sogou News

- Dataset Description:

  The Sogou News dataset is a mixture of 2,909,551 news articles from the SogouCA and SogouCS news corpora, in 5 categories. The number of training samples selected for each class is 90,000 and testing 12,000. Note that the Chinese characters have been converted to Pinyin. Classification labels of the news are determined by their domain names in the URL. And we are using the Sogou news topic classification dataset from Hugging Face.

- Number of Rows (Training): 450,000
- Number of Rows (Testing): 60,000

- Classification Labels: 0 - sports, 1 - finance, 2 - entertainment, 3 - automobile, 4 - technology

- Dataset Example:

  {'title': '2008 di4 qi1 jie4 qi1ng da3o guo2 ji4 che1 zha3n me3i nv3 mo2 te4',

  'content': '2008di4 qi1 jie4 qi1ng da3o guo2 ji4 che1 zha3n yu2 15 ri4 za4i qi1ng da3o guo2 ji4 hui4 zha3n zho1ng xi1n she4ng da4 ka1i mu4 . be3n ci4 che1 zha3n jia1ng chi2 xu4 da4o be3n yue4 19 ri4 . ji1n nia2n qi1ng da3o guo2 ji4 che1 zha3n shi4 li4 nia2n da3o che2ng che1 zha3n gui1 mo2 zui4 da4 di2 yi1 ci4 , shi3 yo4ng lia3o qi1ng da3o guo2 ji4 hui4 zha3n zho1ng xi1n di2 qua2n bu4 shi4 ne4i wa4i zha3n gua3n . yi3 xia4 we2i xia4n cha3ng mo2 te4 tu2 pia4n .',

  'label': 3}

Gzip + k-NN Exploration

Penn Engineering

Baseline Method - Gzip + k-NN

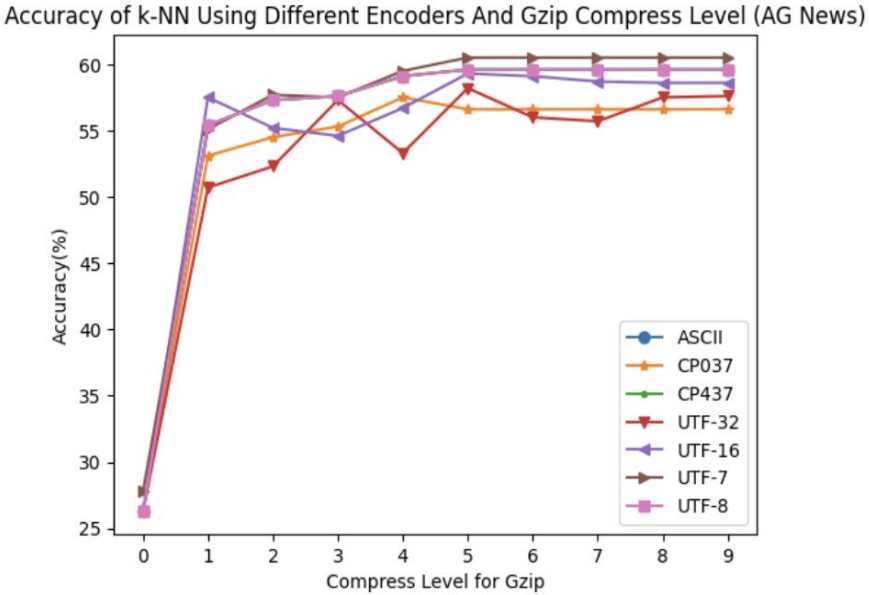
- Step 1: Compress training and test data using Gzip under the same hyper-parameter setting.

- Step 2: Classify each test data datapoint using k-NN with the following definition for the distance of two data points (NCD, Normalized Compression Distance):

$$NCD(x,y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}$$

where x and y are two strings, xy is the concatenation of x and y, and C(x), C(y), C(xy) mean the length of compressed x, y, xy respectively.

Gzip Hyper-Parameter - Encoding & Compress Level



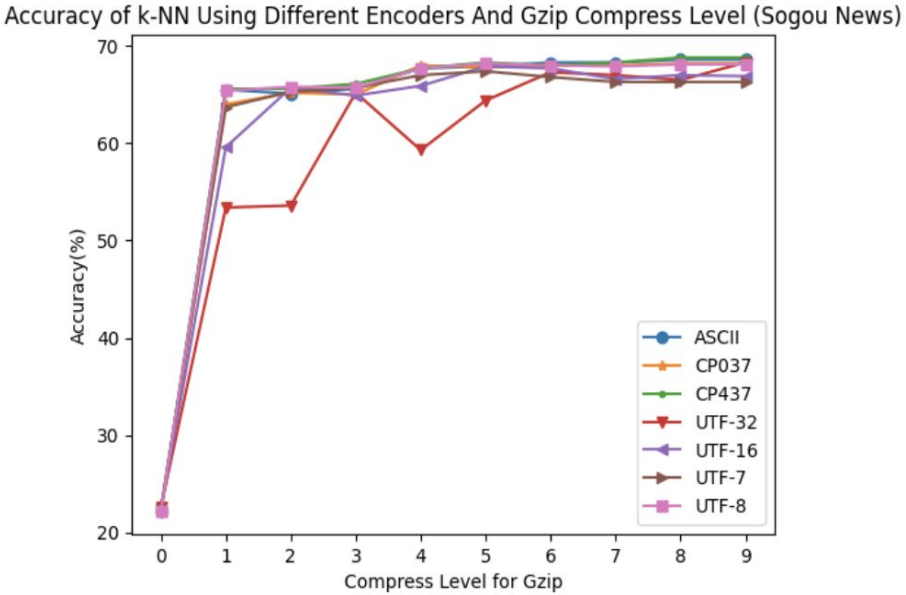
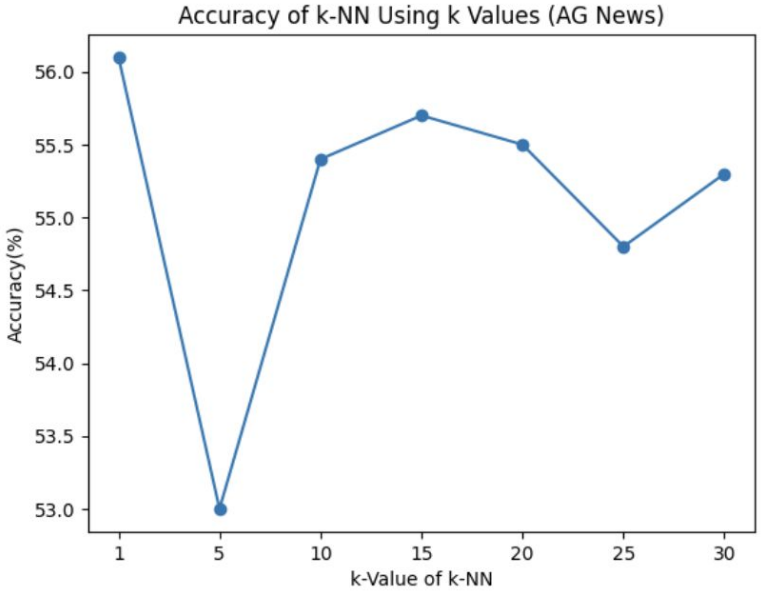


Figure 1: k-NN Performance under Different Encoders and Gzip Compress Level for AG News

k-NN Hyper-parameter - k-Value



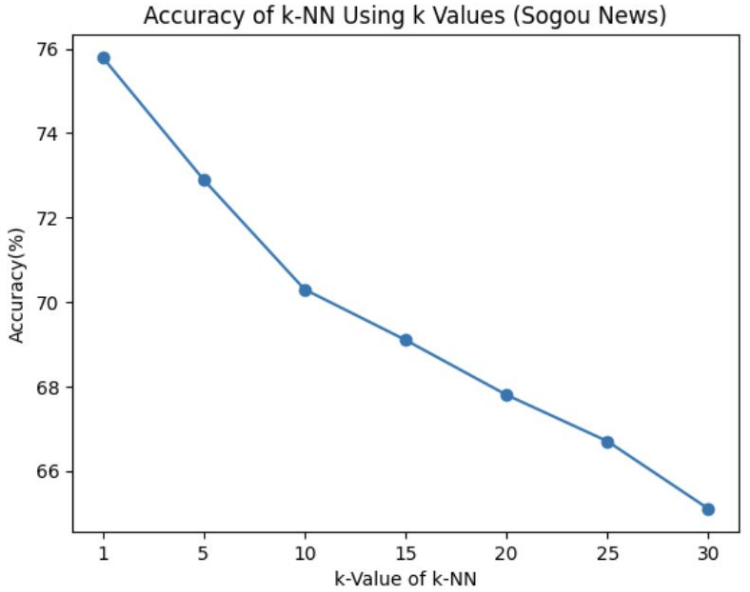


Figure 2: k-NN Performance under Different Encoders and Gzip Compress Level for AG News

Additional Compression Methods

Penn Engineering

Gzip, Lz4 and Hashed n-grams

Gzip

Gzip is based on the DEFLATE algorithm, which is a combination of LZ77 and Huffman coding. The LZ77 Compression finds and eliminates duplicate strings of data. After LZ77 compression, Huffman coding is applied to replace common sequences with shorter representations and less common sequences with longer representations.

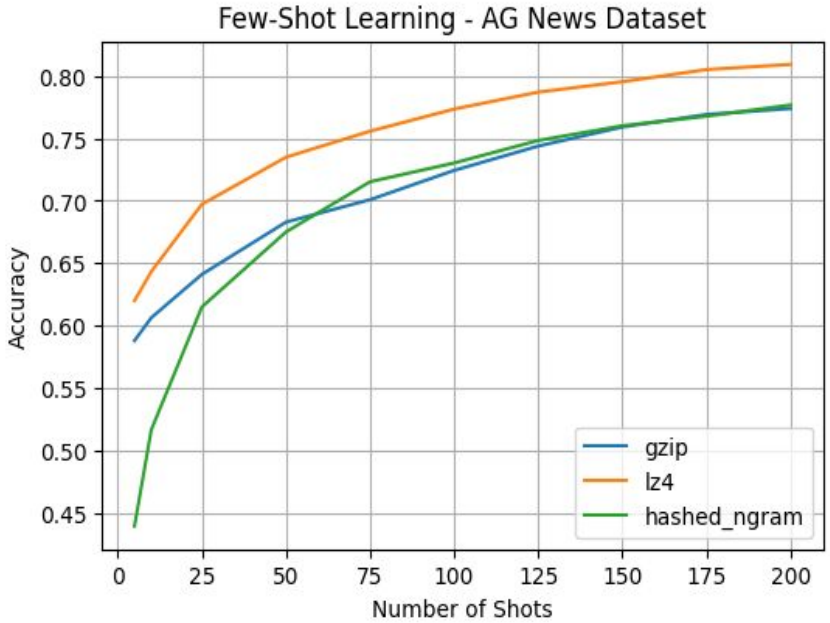
Lz4

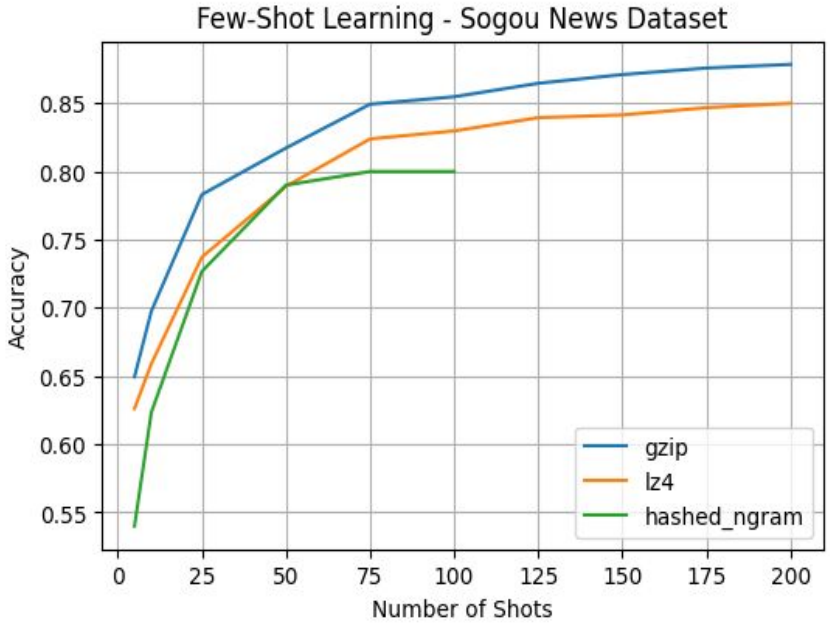
LZ4 only uses a dictionary-matching stage (LZ77), and unlike other common compression algorithms does not combine it with an entropy coding stage (e.g. Huffman coding in DEFLATE).

Hashed n-grams

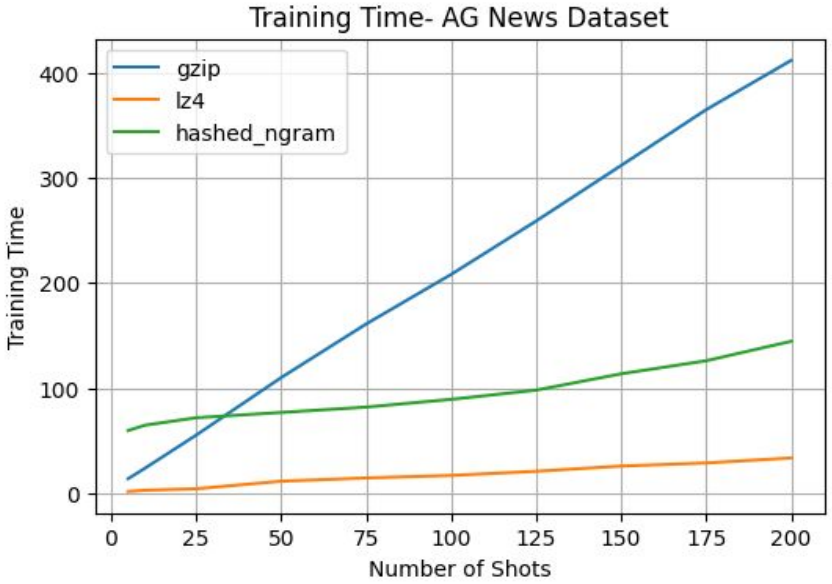
Processes text data by focusing on character-level n-grams, which are sequences of adjacent characters within a text, ranging in length from 5 to 50 characters.

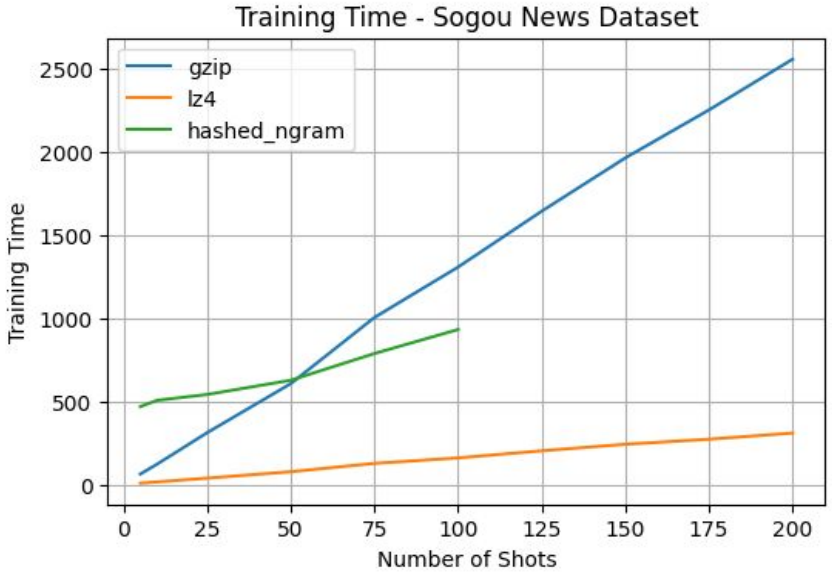
Few-shots Learning Performance Comparison



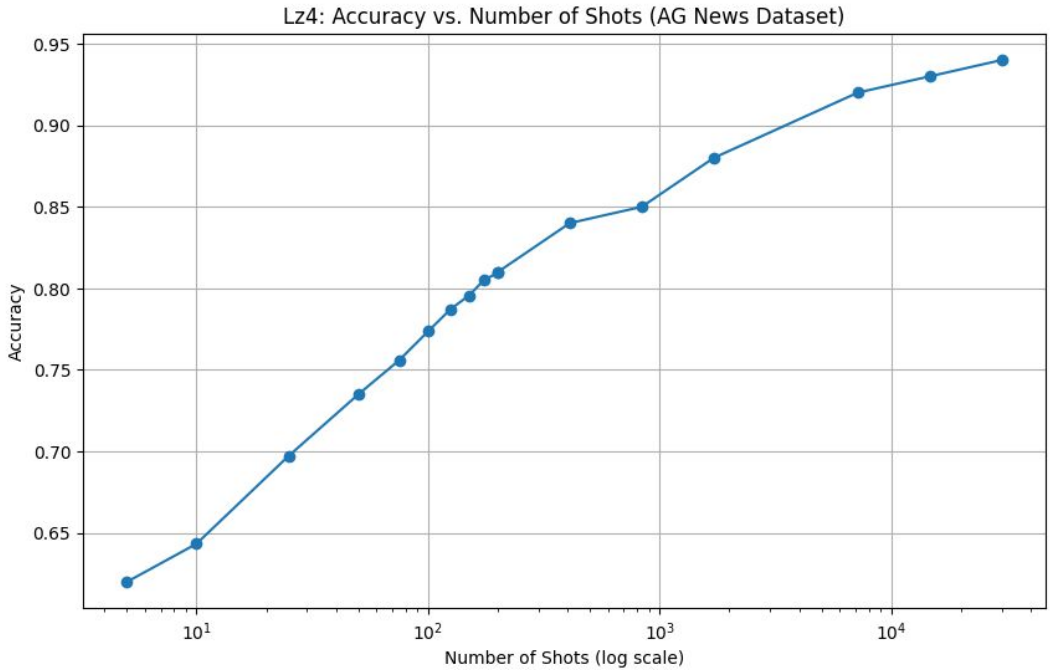


Few-shots Learning Training Time Comparison





Most Efficient Compressor - Lz4



Considering both accuracy and computational efficiency, and given the constraints of time and computational resources, we chose the lz4 compressor to process the entire training dataset. The accuracy reached **93.86%**.

Penn Engineering

DNN Methods

Penn Engineering

LSTM and BERT

LSTM

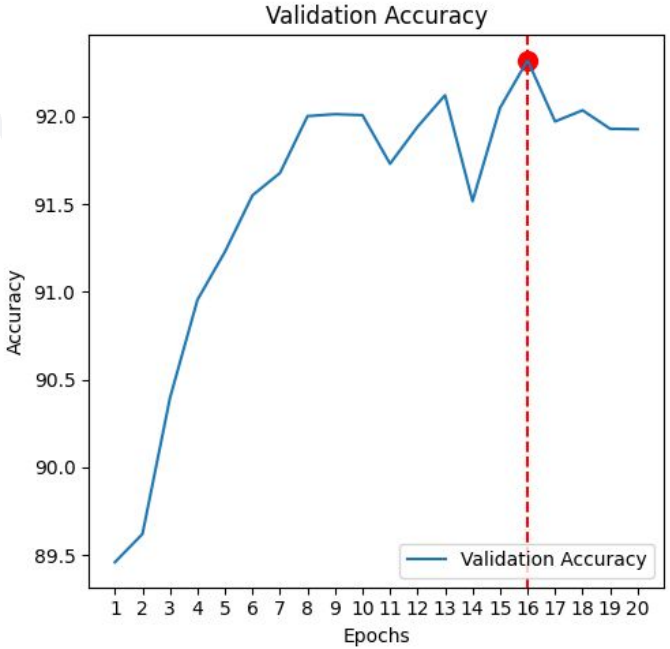
Bidirectional LSTM model, paired with word2vec embeddings, effectively processes and learns from text by interpreting context from both past and future data points, offering a robust solution for diverse NLP tasks.

BERT

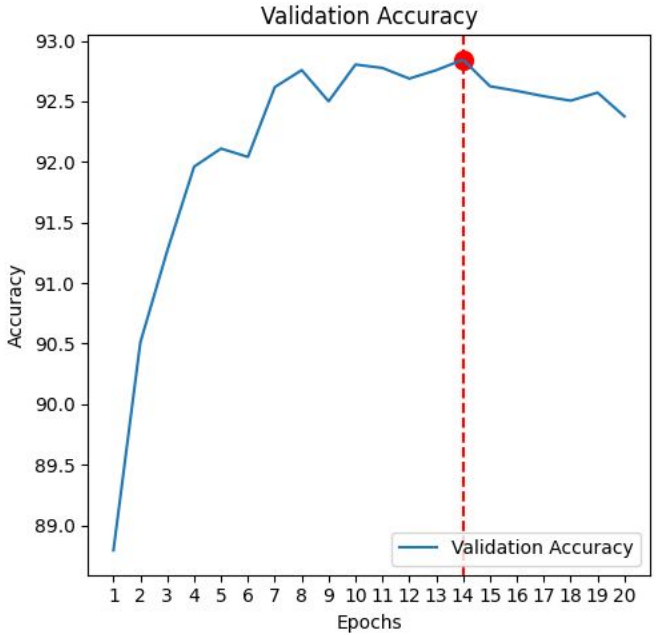
BERT model, tailored for text classification, uses the Transformer architecture for nuanced language interpretation. It employs a specialized tokenizer with attention masks to enhance focus and accuracy by filtering out irrelevant details.

LSTM Results

AG News Acc: 92.3%



Sogou News Acc: 92.8%



Penn Engineering

BERT Results

BERT Results on AG News

Class	Precision	Recall	F1-Score
0	0.94	0.96	0.95
1	0.99	0.98	0.99
2	0.91	0.92	0.91
3	0.92	0.91	0.91
Accuracy		0.9404	
Macro Avg	0.94	0.94	0.94
Weighted Avg	0.94	0.94	0.94

BERT Results on Sogou News

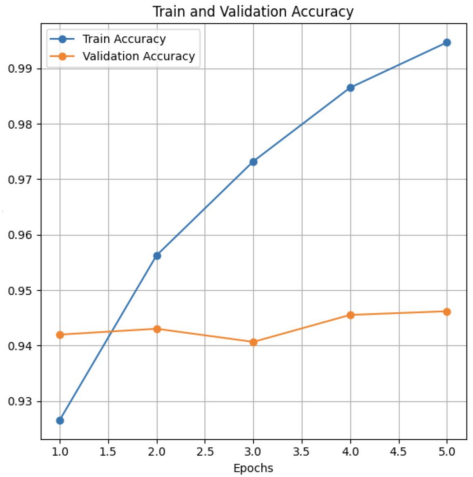
Class	Precision	Recall	F1-Score
0	0.93	0.91	0.92
1	0.77	0.92	0.84
2	0.91	0.90	0.91
3	0.92	0.89	0.91
4	0.97	0.86	0.91
Accuracy		0.8964	
Macro Avg	0.90	0.90	0.90
Weighted Avg	0.90	0.90	0.90

Hyperparameter optimization for BERT

	learning_rate	batch_size	num_epochs	dropout_rate	weight_decay	warmup_steps	scheduler_type	optimizer	final_output
32	0.00005	32	5	0.1	0.0001	0	linear	AdamW	0.90100
41	0.00005	64	5	0.1	0.0001	0	cosine	AdamW	0.90000
37	0.00005	32	5	0.3	0.0001	0	cosine	AdamW	0.89950
36	0.00005	32	5	0.3	0.0001	0	linear	AdamW	0.89925
40	0.00005	64	5	0.1	0.0001	0	linear	AdamW	0.89875

Class	Precision	Recall	F1-Score
0	0.95	0.96	0.96
1	0.99	0.99	0.99
2	0.92	0.91	0.92
3	0.93	0.92	0.92
Accuracy		0.9462	
Macro Avg	0.95	0.95	0.95
Weighted Avg	0.95	0.95	0.95

Table 4: BERT Results on AG News



Gzip+BERT Results

Example: are seeing green again. Gzip compressed length: 136

Dataset with Gzip Length

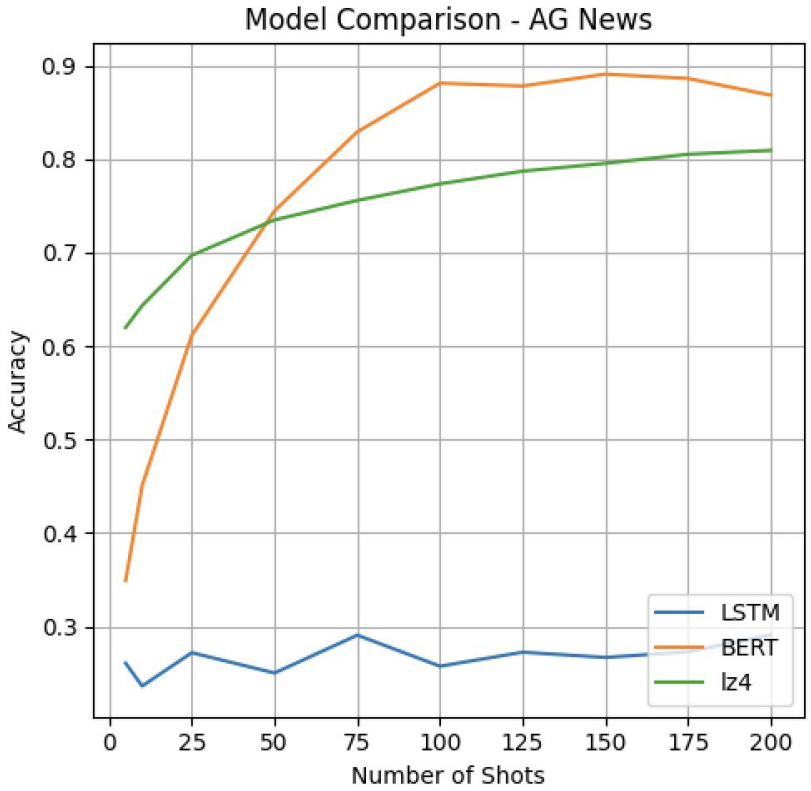
```
Epoch 2/5: 100%|██████████| 3750/3750 [37:29<00:00,  1.67it/s]
Epoch 2/5, Train, Loss: 0.1290, Acc: 0.9553
Evaluating: 100%|██████████| 238/238 [00:47<00:00,  4.96it/s]
Epoch 2/5, Val, Loss: 0.1610, Acc: 0.9451
```

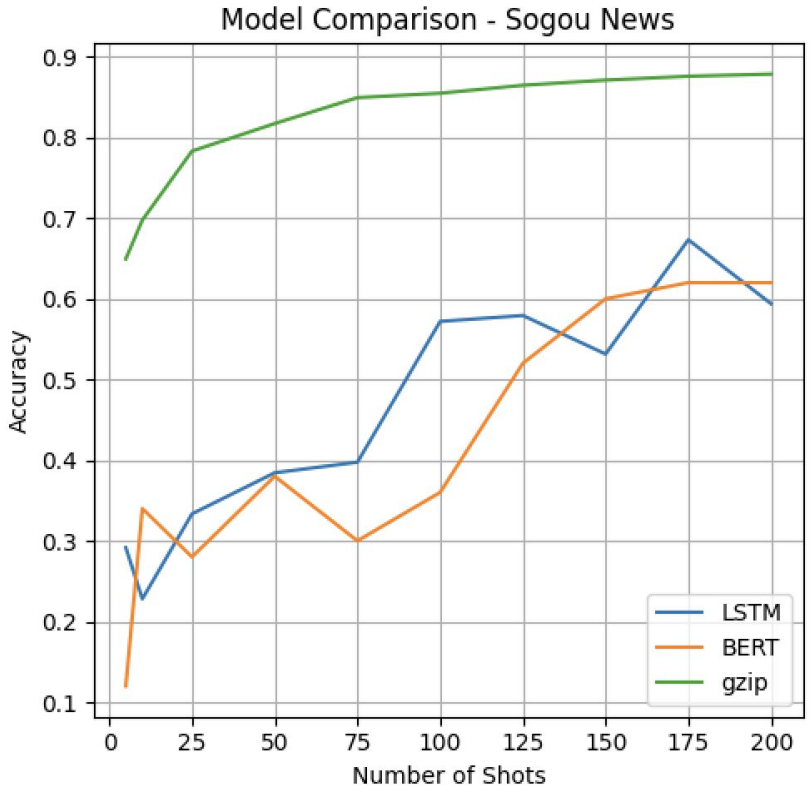
Original

```
Epoch 2/5: 100%|██████████| 3750/3750 [40:44<00:00,  1.53it/s]
Epoch 2/5, Train, Loss: 0.1270, Acc: 0.9563
Evaluating: 100%|██████████| 238/238 [00:54<00:00,  4.34it/s]
Epoch 2/5, Val, Loss: 0.1638, Acc: 0.9430
```

Next: Concatenate the compressed sentence onto the original one and feed into the network. Will require higher computation power.

Few-shots learning application





Summary

Penn Engineering

Results Table

Method	AG News	Sogou
LSTM	92.3%	92.8%
Gzip + kNN	93.7%	-
BERT	94.0%	89.6%
Gzip + BERT(3 epoch)	94.4%	-
BERT(param-tuned)	94.6%	-