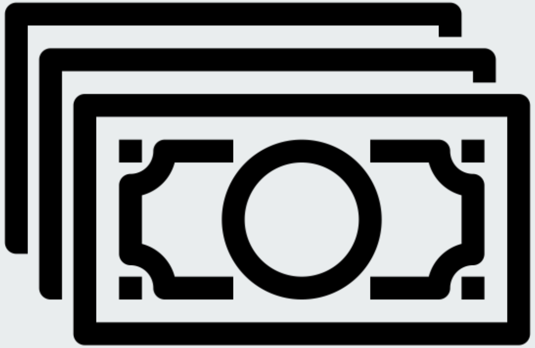
Financial Domain Specific Word Embeddings



Bhumika Kapur, Jun Tao [John] Luo, Nathan Luskey, Amey Patel

Outline

Problem Description

Task

Data Pipeline

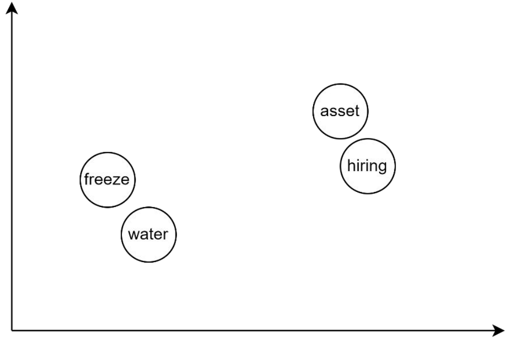
Methodology

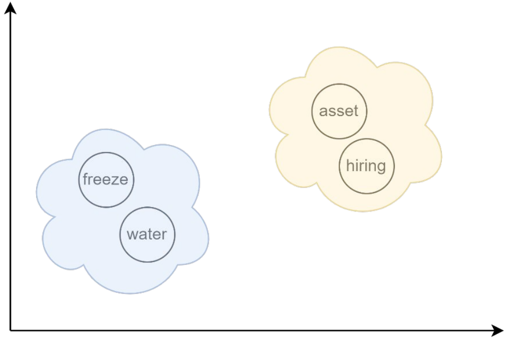
Results and Discussion

Conclusions

Problem Description

Word Embeddings - General





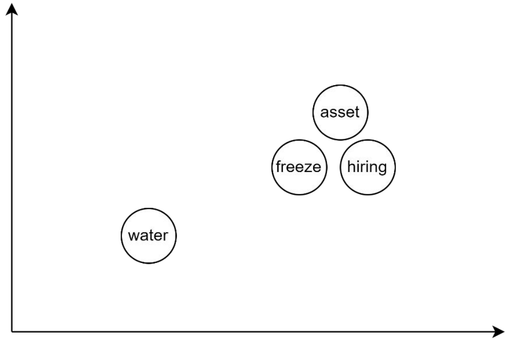
<u>Data Sources:</u>

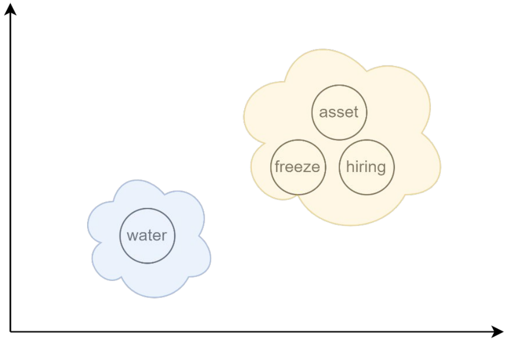
Wikipedia
Books
Google News

<u>Downstream Tasks:</u>

Question Answering
Semantic-Syntactic Word Relationship
Sentence Completion

Word Embeddings - Financial





Data Sources:

SEC Filings
Earnings Call Transcripts
Analyst Reports

Downstream Tasks:

Sentiment Analysis
Stock Volatility Predictions

Task

Embedding Data

SEC Filings

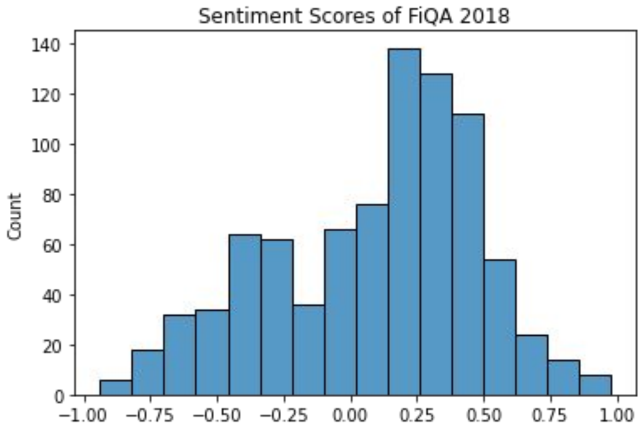
- Standardized regulatory documents
- Filed Quarterly
- Earnings, Assets, Key Management

Downstream Data

FiQA 2018:

1111 Dataset Samples

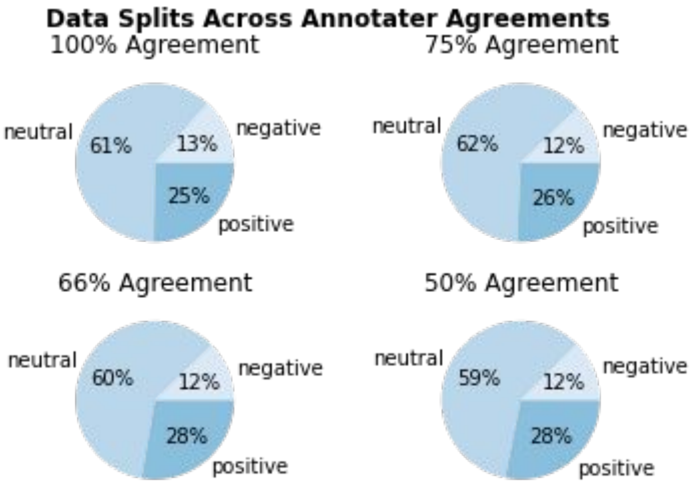
Sentiment Score between 1 and -1



Phrasebank:

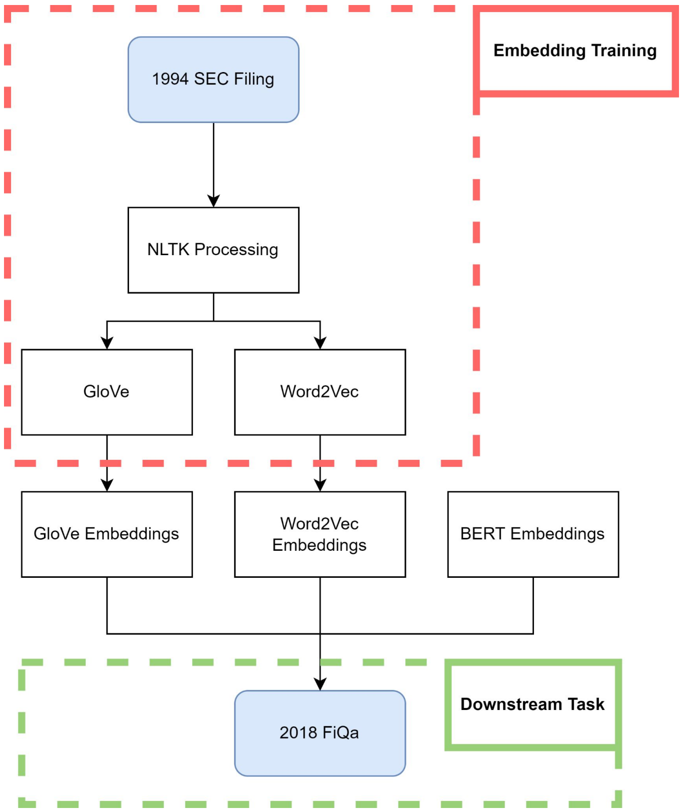
4840 Dataset Samples

Sentiment Scores as -1, 0, 1

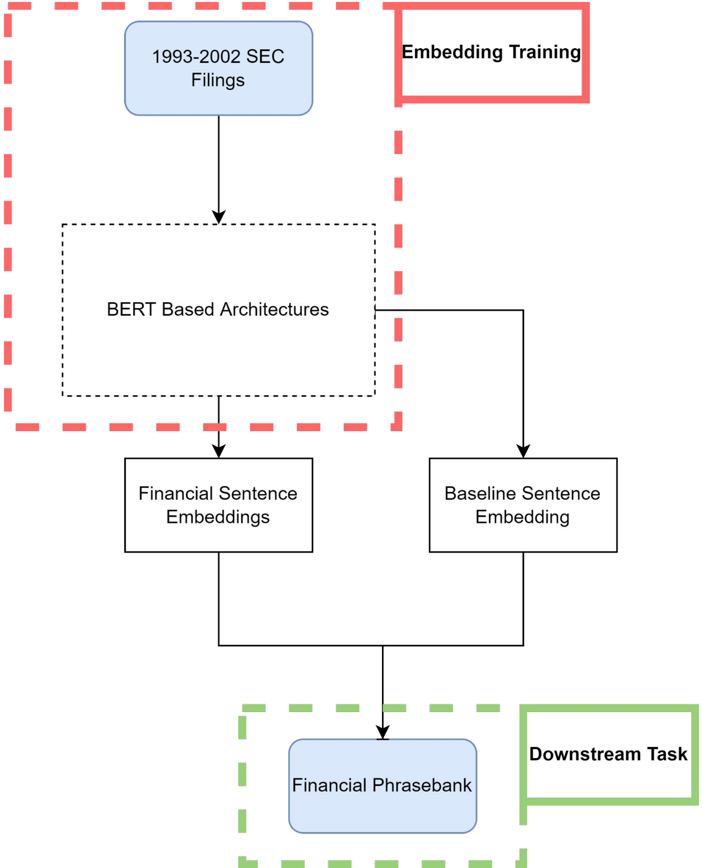


Data Pipeline

Initial Pipeline

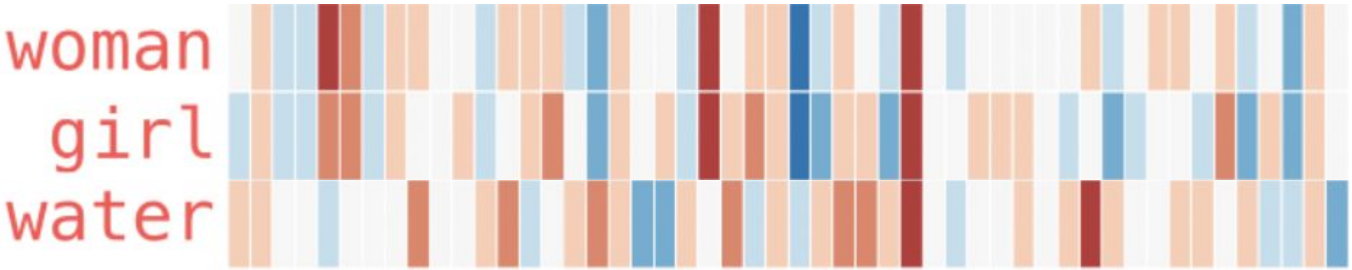


Final Pipeline



Methodology:
Word Embeddings

Word Vectors



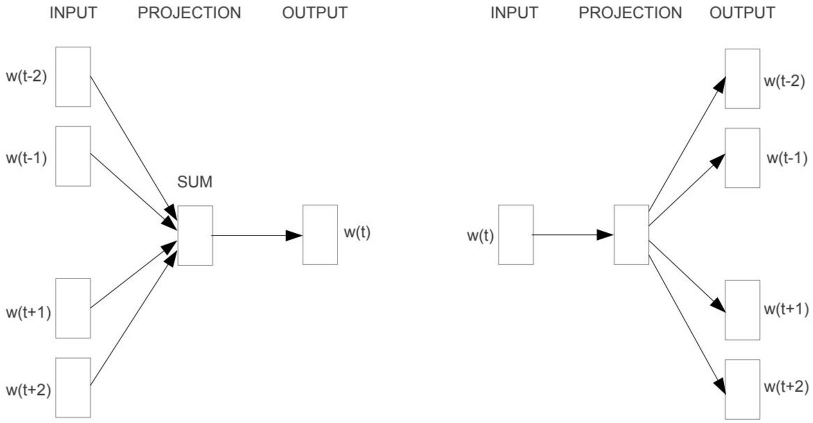
Word2Vec

- Continuous Bag of Words

$$L_{CBOW}(\theta) = \prod_{t=1}^{T} \prod_{\substack{-m \leq j \leq m \\ j \neq 0}} P(w_t | w_{t+j}; \theta)$$

- Skip-Gram

$$L_{Skip-gram}(\theta) = \prod_{t=1}^{T} \prod_{\substack{-m \leq j \leq m \\ j \neq 0}} P(w_{t+j} | w_t; \theta)$$



CBOW

Skip-gram

GloVe: Global Vectors for Word Representations

- Matrix factorization using global word-word co-occurrence counts

- Weighted least squares objective

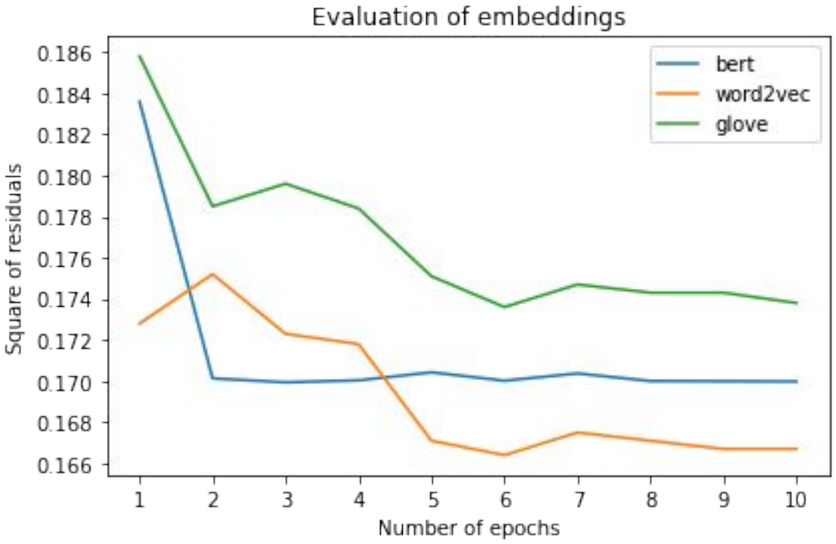
$$J=\sum_{i,j=1}^{V} f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - log X_{ij})^2$$

GloVe: Global Vectors for Word Representations

- Weighted least squares model that trains on global word-word co-occurrence counts
- Uses a probe word to deduce the relationship between two words
- $$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2 , \quad (8)$$
- Complexity ~ O(|C|)
- Implementation: 20 epochs trained on 1994 SEC filings
- Paper results: 75% word analogy

Midterm Results

- Performance on 2018 FiQA sentiment analysis task
- Regression task, lower residual is better
- Training on financial data improved downstream performance significantly



Methodology:
Contextualized Word Embeddings

BERT

Why BERT?

- Does not dynamically capture relationship based on the context of the words
- Doesn't distinguish words with different meanings (river **bank** vs **bank** of America)

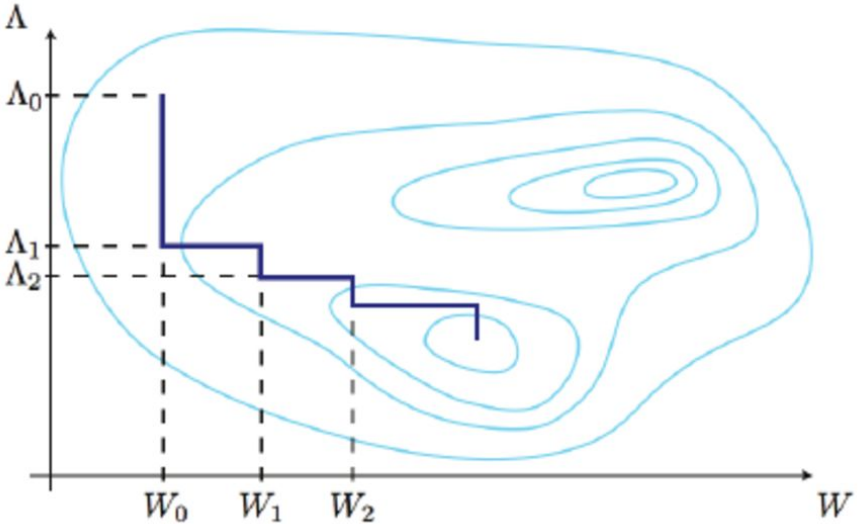
Encoder side of transformer

- BERT uses transformer encoders to get bidirectional conditioning

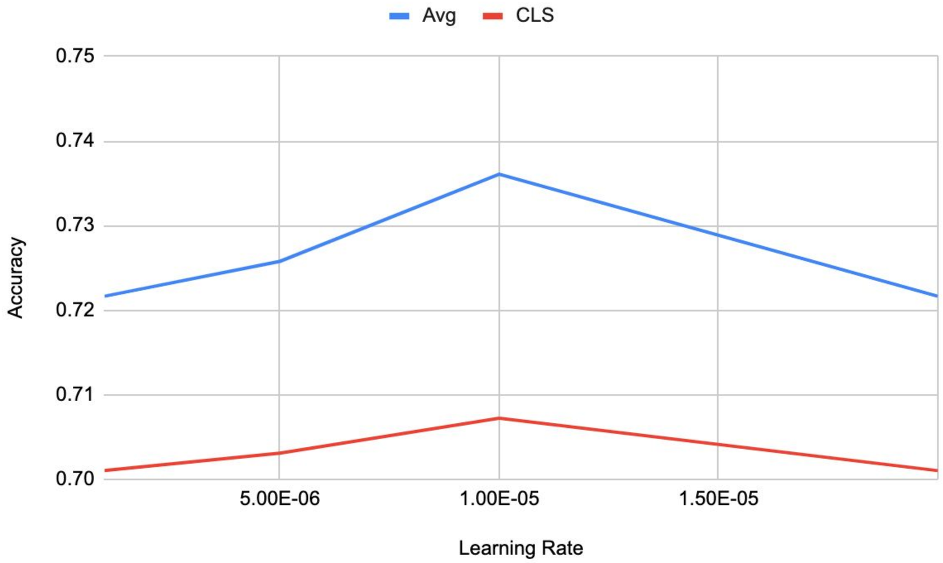
Ablation Studies

Block Coordinate Descent

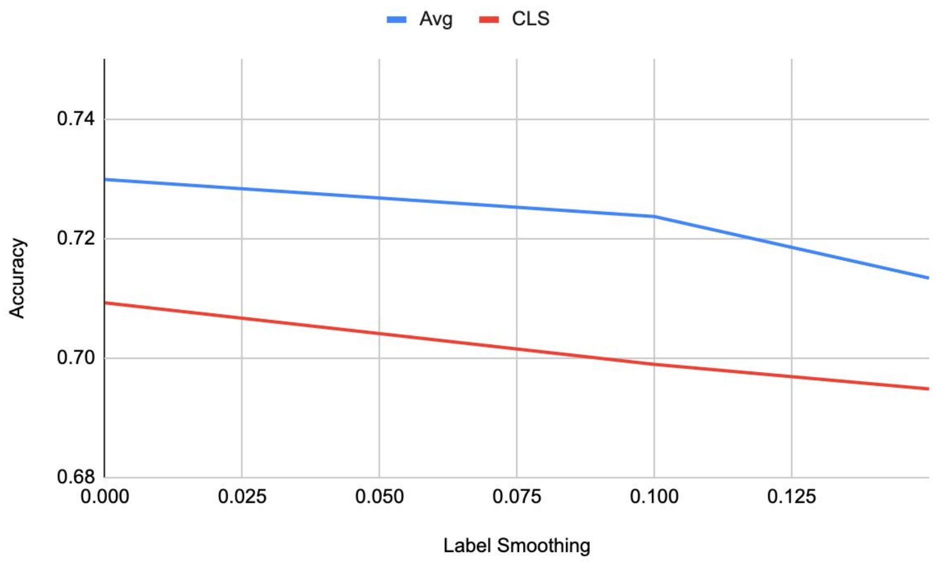
- Iteratively vary one parameter while keeping other constant
- Efficiently search through parameter space given limited computation resource



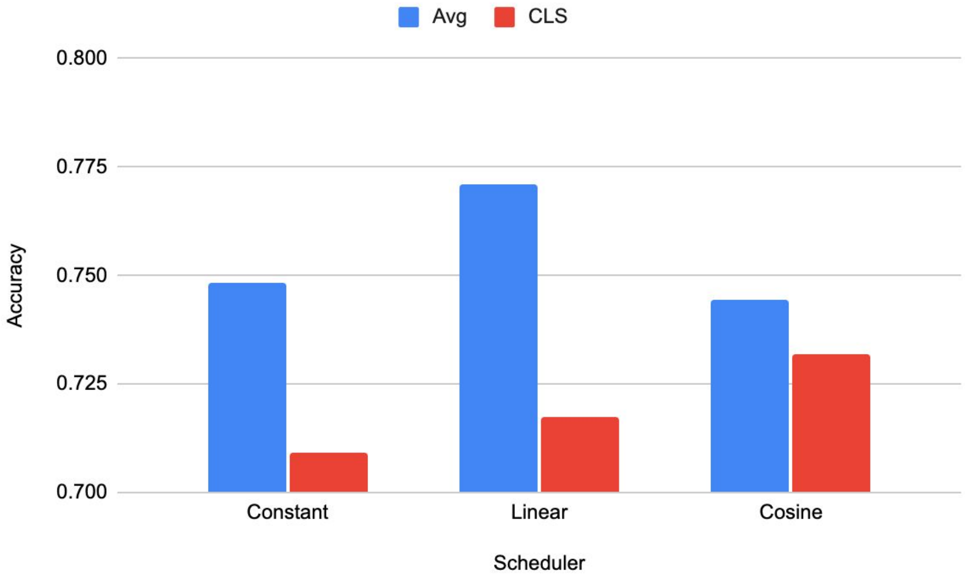
Learning Rate



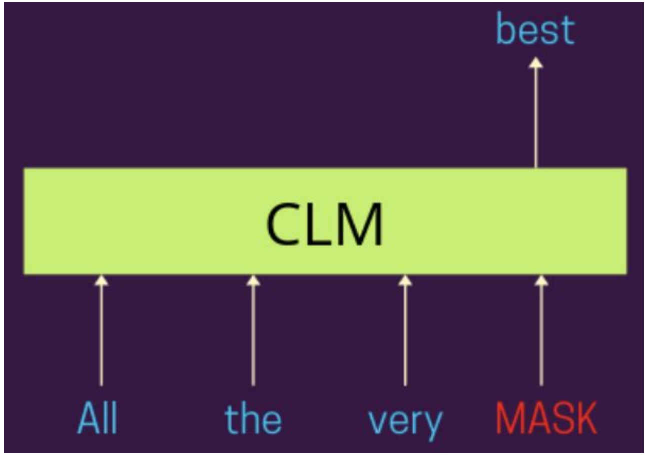
Label Smoothing

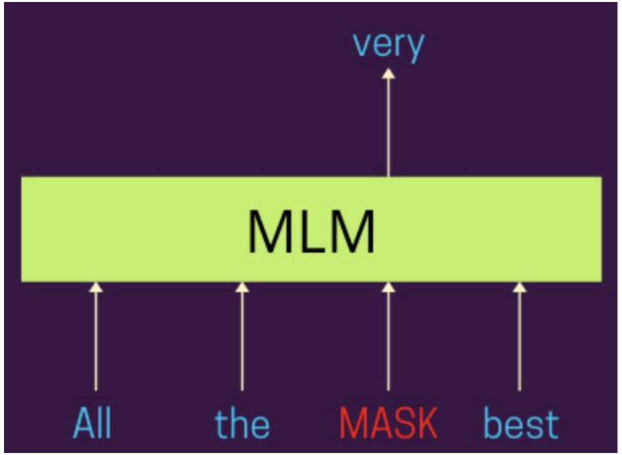


Scheduler


Avg
CLS
0.800
0.775
0.750
0.725
0.700
Accuracy
Constant
Linear
Cosine
Scheduler

Ablation Studies: CLM vs MLM



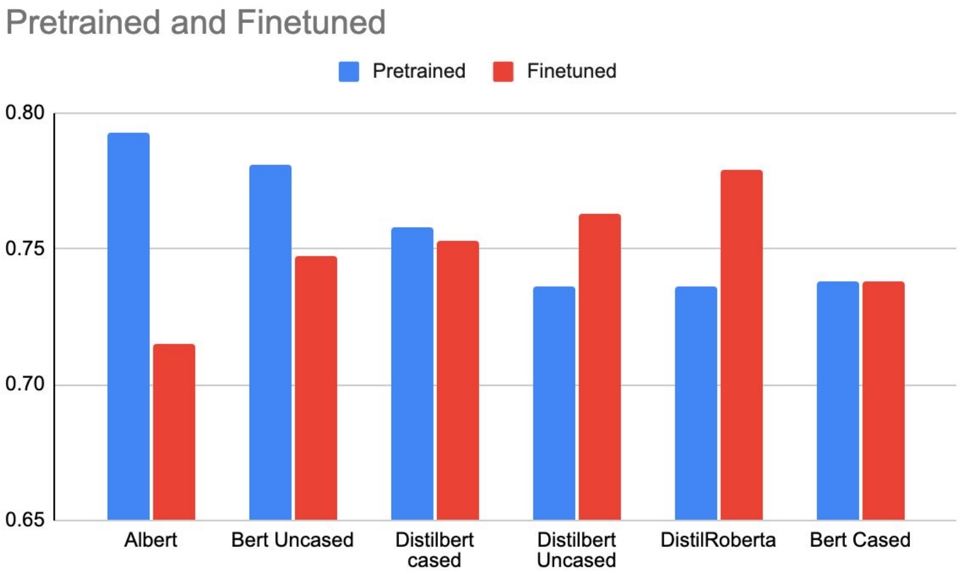


CLM Results

Base	Fine-tuned
62.68%	65.78%

3% improvement!

Ablation Studies: Models

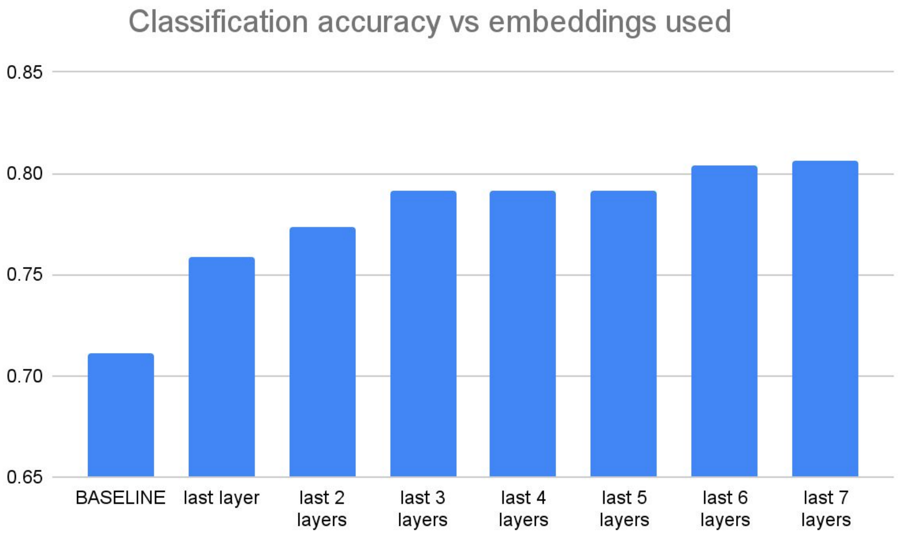


Ablation Studies

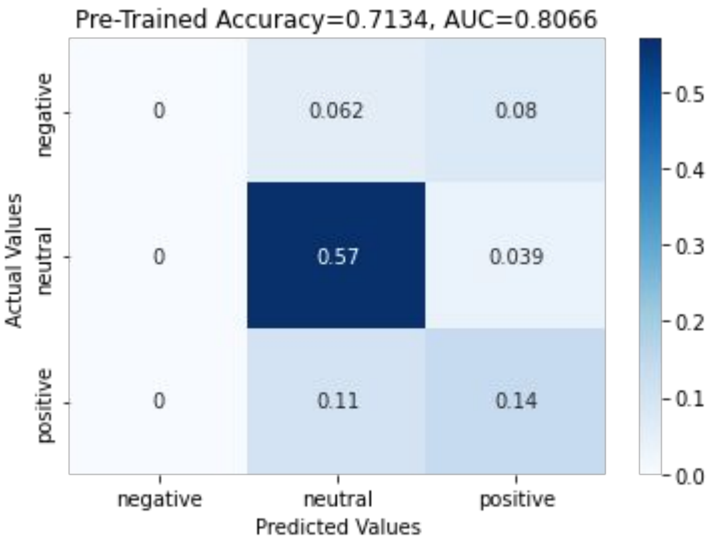
Use information contained in intermediate layers:

- attention layers
- hidden states

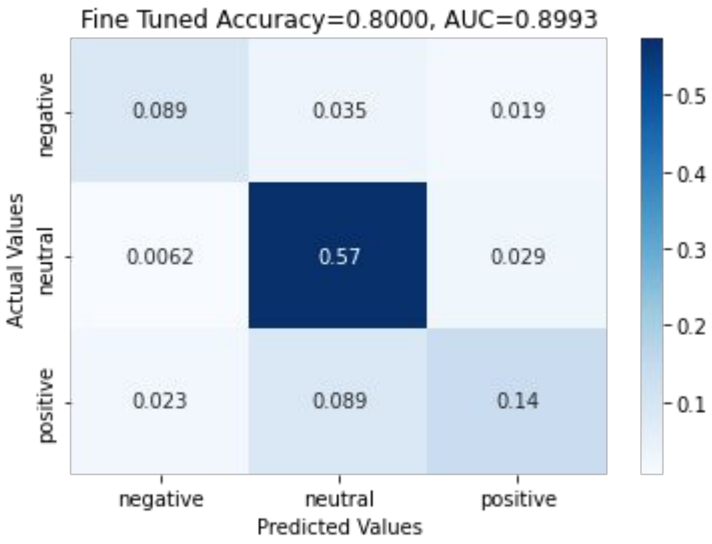
8% improvement!!



Results



Pre-Trained Accuracy=0.7134, AUC=0.8066
Actual Values
negative
neutral
positive
0
0.062
0.08
0
0.57
0.039
0
0.11
0.14
Predicted Values
0.0
0.1
0.2
0.3
0.4
0.5



Fine Tuned Accuracy=0.8000, AUC=0.8993
Actual Values
negative
neutral
positive
0.089
0.035
0.019
0.0062
0.57
0.029
0.023
0.089
0.14
Predicted Values
0.1
0.2
0.3
0.4
0.5

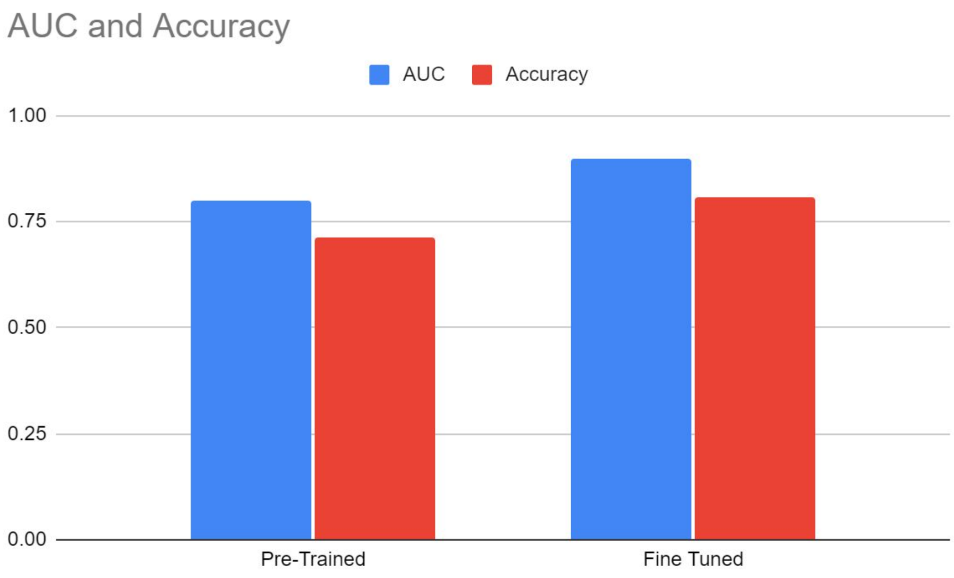
Conclusions

Conclusions

Data Importance

Training Optimization

BERT’s Information Encoding for Embeddings



TODO: (John) t-sne or PCA

Thanks to Akshat Gupta & Amelia Kuang!

Sources

Jay Alammar - The Illustrated Word2vec

Hugging Face - Financial Phrasebank Dataset

FiQA 2018

Investopedia - SEC Filings

"Efficient Estimation of Word Representations in Vector Space"

MLM vs CLM

“BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”

BERT Variants

“GloVe: Global Vectors for Word Representation”

“Block Coordinate Descent”