Transfer Learning and Textual Analysis of Accounting Disclosures:

Applying Big Data Methods to Small(er) Data Sets

Federico Siano Peter D. Wysocki

BOSTON UNIVERSITY

Department of Accounting
Questrom School of Business

Motivation

- We **introduce** the **concept** of **Machine Transfer Learning** and **demonstrate** its **application** to content modeling of accounting disclosures

Motivation

- We **introduce** the **concept** of **Machine Transfer Learning** and **demonstrate** its **application** to content modeling of accounting disclosures
- Machine Transfer Learning represents a **powerful** yet **accessible** methodological tool **to understand language**, thus a dimension of human behavior, in the business context

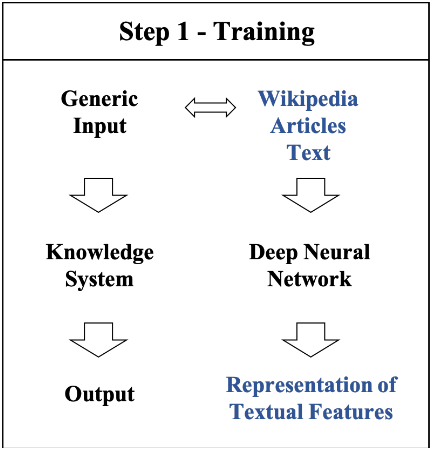
Motivation

- We **introduce** the **concept** of **Machine Transfer Learning** and **demonstrate** its **application** to content modeling of accounting disclosures
- Machine Transfer Learning represents a **powerful** yet **accessible** methodological tool **to understand language**, thus a dimension of human behavior, in the business context
- **Textual analysis** seems to have reached a **plateau**: current approaches largely employ **dictionary-based** and **statistical techniques** that are **unable** to account for **context** within corporate communications

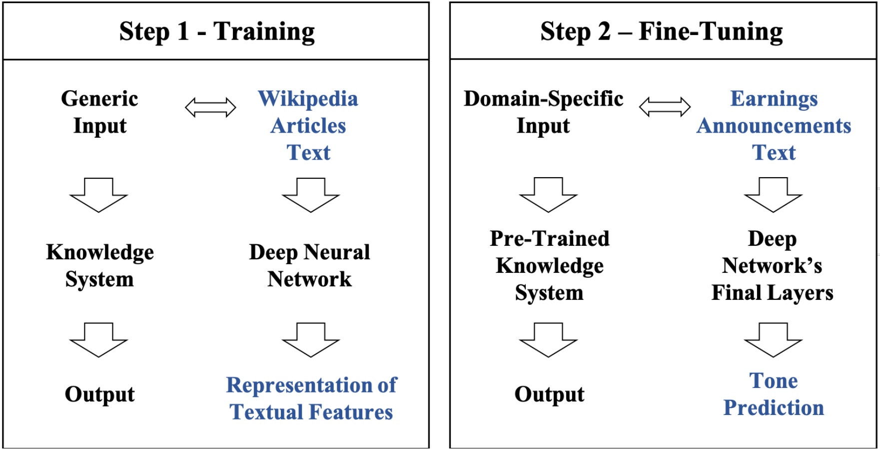
Machine Transfer Learning

- Consists in **transferring** deep neural **networks' knowledge** (i.e., features representation), acquired through generic inputs, **to domain-specific tasks** (i.e., classification, regression, dimensionality reduction)

Machine Transfer Learning



Machine Transfer Learning



Machine Transfer Learning

- It has two main advantages, **high prediction accuracy** and **modest resource intensiveness** (i.e., computation time)

Machine Transfer Learning

- It has two main advantages, **high prediction accuracy** and **modest resource intensiveness** (i.e., computation time)
- High prediction accuracy is achieved through **large-scale inputs** and a **deep learning architecture**

Machine Transfer Learning

- It has two main advantages, **high prediction accuracy** and **modest resource intensiveness** (i.e., computation time)
- High prediction accuracy is achieved through **large-scale inputs** and a **deep learning architecture**
- Low resource intensiveness is achieved through **fine-tuning only the last few layers** of the network using **small-scale inputs**

Machine Transfer Learning

- It has two main advantages, **high prediction accuracy** and **modest resource intensiveness** (i.e., computation time)
- High prediction accuracy is achieved through **large-scale inputs** and a **deep learning architecture**
- Low resource intensiveness is achieved through **fine-tuning only the last few layers** of the network using **small-scale inputs**
- An **applied researcher** can exploit a **powerful** pre-trained model for **the modest cost** of task-specific fine-tuning

Which Knowledge System? *BERT*

- ***BERT* is an open-source pre-trained language model** based upon a deep unsupervised network architecture

Which Knowledge System? *BERT*

- ***BERT* is an open-source pre-trained language model** based upon a deep unsupervised network architecture
  - Base version has **12** layers, **768** hidden unites and **110 Million** parameters
  - **One week** of training time with **4 TPUs**

Which Knowledge System? *BERT*

- ***BERT* is an open-source pre-trained language model** based upon a deep unsupervised network architecture
  - Base version has **12** layers, **768** hidden unites and **110 Million** parameters
  - **One week** of training time with **4 TPUs**
- Introduced **by Google Research** (Devlin et al. 2019), it has shown **superior performance** in a number of tasks

Which Knowledge System? *BERT*

- ***BERT* is an open-source pre-trained language model** based upon a deep unsupervised network architecture
  - Base version has **12** layers, **768** hidden unites and **110 Million** parameters
  - **One week** of training time with **4 TPUs**
- Introduced **by Google Research** (Devlin et al. 2019), it has shown **superior performance** in a number of tasks
- *BERT*'s recipe combines **bi-directionality** and **transformers** to model **non-sequential, long-range semantic associations**

Bi-directionality

$\texttt{We use }\textcolor{red}{\texttt{[MASK]}}\texttt{ }\overbrace{\texttt{to portray business performance}}^{\textit{context}}$

$\underbrace{\texttt{What should I expect when studying}}_{\textit{context}}\texttt{ }\textcolor{red}{\texttt{[MASK]}}\texttt{?}$

Accounting is predicted based on a richer left-to-right and right-to-left characterization

Transformers

Accountants

Serve

A

Crucial

Role

In

Modern

Societies

Because

They

Help

Portraying

The

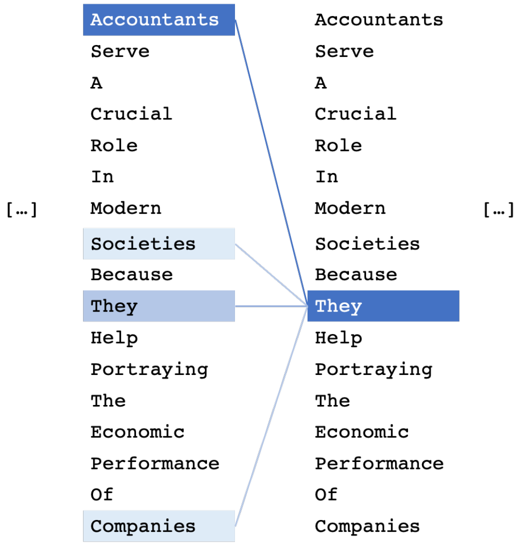
Economic

Performance

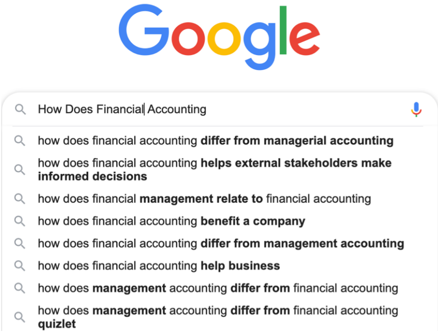
Of

Companies

Transformers



A Real World Application



The Accounting Task

- We **classify the tone** of earnings announcements

The Accounting Task

- We **classify the tone** of earnings announcements
- A long **literature in accounting** attempts to **characterize and quantify** the information contained within the **text** of corporate disclosures (see Li, 2010; Loughran and McDonald, 2016)

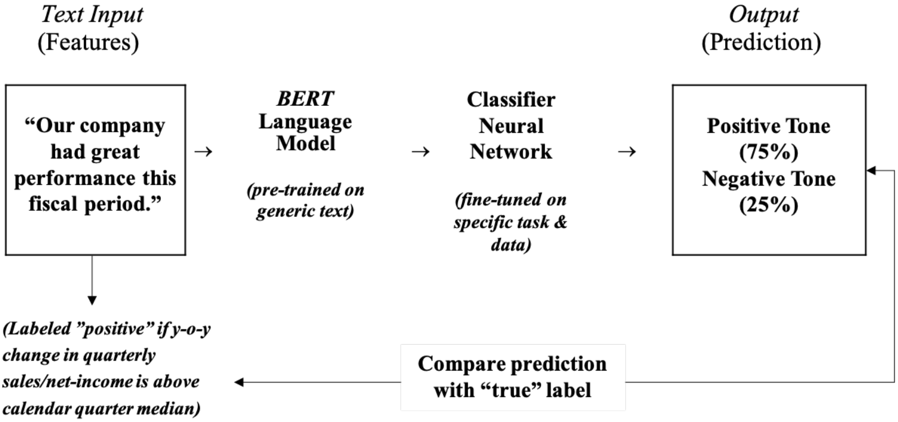
The Accounting Task

- We **classify the tone** of earnings announcements
- A long **literature in accounting** attempts to **characterize and quantify** the information contained within the **text** of corporate disclosures (see Li, 2010; Loughran and McDonald, 2016)
- **Recent works** have focused on **the tone** of earnings announcements or 10-K filings

The Accounting Task

- We **classify the tone** of earnings announcements
- A long **literature in accounting** attempts to **characterize and quantify** the information contained within the **text** of corporate disclosures (see Li, 2010; Loughran and McDonald, 2016)
- **Recent works** have focused on **the tone** of earnings announcements or 10-K filings
- **Tone** is a multi-dimensional, **context-specific** construct that can be hardly characterized through **word-based approaches**

The Accounting Task



The Setting

- We investigate *Business Wire* **Q1 Earnings Announcements** and **exclude** disclosures about **extra-ordinary transactions** (e.g., M&A, Equity Issuances)

The Setting

- We investigate *Business Wire* **Q1 Earnings Announcements** and **exclude** disclosures about **extra-ordinary transactions** (e.g., M&A, Equity Issuances)
- We study **the first 10 and the first 15 sentences**

The Setting

- We investigate *Business Wire* **Q1 Earnings Announcements** and **exclude** disclosures about **extra-ordinary transactions** (e.g., M&A, Equity Issuances)
- We study **the first 10 and the first 15 sentences**
- We make this **choices** for **three reasons**:

The Setting

- We investigate *Business Wire* **Q1 Earnings Announcements** and **exclude** disclosures about **extra-ordinary transactions** (e.g., M&A, Equity Issuances)
- We study **the first 10 and the first 15 sentences**
- We make this **choices** for **three reasons**:
    1. to **maximize** the **signal-to-noise** ratio

The Setting

- We investigate *Business Wire* **Q1 Earnings Announcements** and **exclude** disclosures about **extra-ordinary transactions** (e.g., M&A, Equity Issuances)
- We study **the first 10 and the first 15 sentences**
- We make this **choices** for **three reasons**:
  1. to **maximize** the **signal-to-noise** ratio
  2. to **show** the **power** of Transfer Learning using a **small dataset** of earnings announcements

The Setting

- We investigate *Business Wire* **Q1 Earnings Announcements** and **exclude** disclosures about **extra-ordinary transactions** (e.g., M&A, Equity Issuances)
- We study **the first 10 and the first 15 sentences**
- We make this **choices** for **three reasons**:
    1. to **maximize** the **signal-to-noise** ratio
    2. to **show** the **power** of Transfer Learning using a **small dataset** of earnings announcements
    3. due to **technical reasons** related to transformers and attention

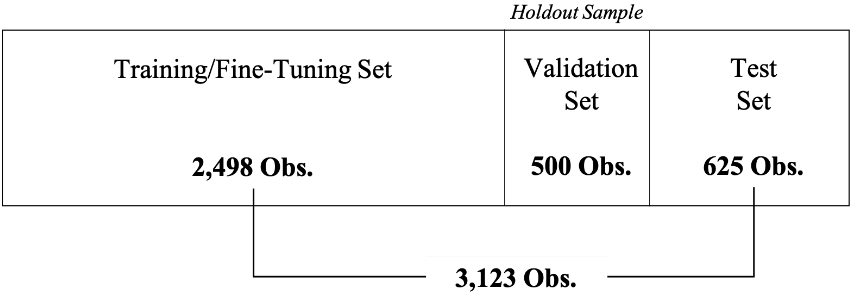
The Empirical Strategy

- We first evaluate *BERT*'s test **accuracy** in **absolute and relative terms** (i.e., benchmarking)

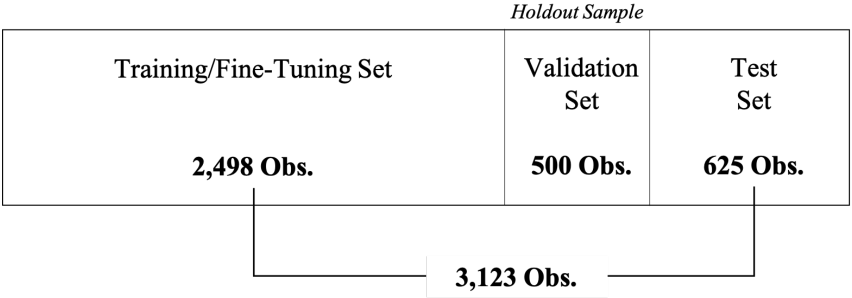
The Empirical Strategy

- We first evaluate *BERT*'s test **accuracy** in **absolute and relative terms** (i.e., benchmarking)
- We then perform **"masking" tests** to gather insights into **BERT's functioning**

The Sample


Holdout Sample
Training/Fine-Tuning Set
2,498 Obs.
Validation Set
500 Obs.
Test Set
625 Obs.
3,123 Obs.

The Sample



Labels (i.e., positive and negative) are almost **balanced** across sample sets

Model's Validation

- We select a **random sample of 500 observations** to identify the proper models' **hyper-parameters**

Model's Validation

- We select a **random sample of 500 observations** to identify the proper models' **hyper-parameters**
- The validated *BERT* model:
  1. *BERT* Base Uncased with 12 layers and 768 hidden states
  2. 2e-5 learning rate
  3. 3 training epochs

Main Results

BERT Model Selection and Cross-Validation Results on the Sample of Selected Business Wire Q1 Earnings Announcements

	Training Obs. (n= 2,498) Positive Labels	Training Obs. (n= 2,498) Negative Labels	Test Obs. (n=625) Positive Labels	Test Obs. (n=625) Negative Labels	AUC	Evaluation Accuracy	F1 Score
Panel A: 10-sentence Documents							
Average Test Results on 10-sentence Documents using Ten-fold Cross-Validation	**1,380**	**1,118**	**343**	**282**	**0.76**	**0.77**	**0.75**
Panel B: 15-sentence Documents							
Average Test Results on 15-sentence Documents using Ten-fold Cross-Validation	**1,380**	**1,118**	**343**	**282**	**0.78**	**0.79**	**0.78**
Panel C: Benchmarking Models							
Loughran-McDonald Dictionary Classification	**NA**	**NA**	**343**	**282**	**NA**	**0.64**	**0.64**
Naïve Bayes Sentiment Classifier (TextBlob)	**NA**	**NA**	**343**	**282**	**NA**	**0.53**	**0.68**
Random (Dummy) Classifier	**1,380**	**1,118**	**343**	**282**	**0.49**	**0.50**	**0.54**

Main Results

BERT Model Selection and Cross-Validation Results on the Sample of Selected Business Wire Q1 Earnings Announcements

	Training Obs. (n= 2,498)		Test Obs. (n=625)		AUC	Evaluation Accuracy	F1 Score
	Positive Labels	Negative Labels	Positive Labels	Negative Labels			
Panel A: 10-sentence Documents							
Average Test Results on 10-sentence Documents using Ten-fold Cross-Validation	**1,380**	**1,118**	**343**	**282**	**0.76**	**0.77**	**0.75**
Panel B: 15-sentence Documents							
Average Test Results on 15-sentence Documents using Ten-fold Cross-Validation	**1,380**	**1,118**	**343**	**282**	**0.78**	**0.79**	**0.78**
Panel C: Benchmarking Models							
Loughran-McDonald Dictionary Classification	**NA**	**NA**	**343**	**282**	**NA**	**0.64**	**0.64**
Naïve Bayes Sentiment Classifier (TextBlob)	**NA**	**NA**	**343**	**282**	**NA**	**0.53**	**0.68**
Random (Dummy) Classifier	**1,380**	**1,118**	**343**	**282**	**0.49**	**0.50**	**0.54**

Main Results

BERT Model Selection and Cross-Validation Results on the Sample of Selected Business Wire Q1 Earnings Announcements

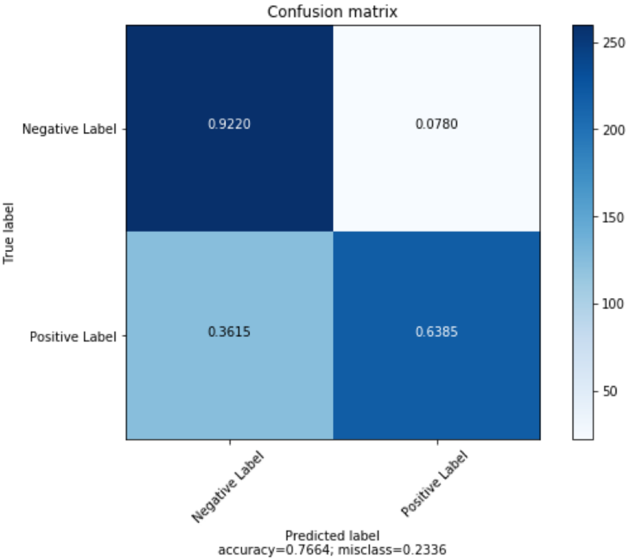
	Training Obs. (n= 2,498) Positive Labels	Training Obs. (n= 2,498) Negative Labels	Test Obs. (n=625) Positive Labels	Test Obs. (n=625) Negative Labels	AUC	Evaluation Accuracy	F1 Score
Panel A: 10-sentence Documents							
Average Test Results on 10-sentence Documents using Ten-fold Cross-Validation	**1,380**	**1,118**	**343**	**282**	**0.76**	**0.77**	**0.75**
Panel B: 15-sentence Documents							
Average Test Results on 15-sentence Documents using Ten-fold Cross-Validation	**1,380**	**1,118**	**343**	**282**	**0.78**	**0.79**	**0.78**
Panel C: Benchmarking Models							
Loughran-McDonald Dictionary Classification	**NA**	**NA**	**343**	**282**	**NA**	**0.64**	**0.64**
Naïve Bayes Sentiment Classifier (TextBlob)	**NA**	**NA**	**343**	**282**	**NA**	**0.53**	**0.68**
Random (Dummy) Classifier	**1,380**	**1,118**	**343**	**282**	**0.49**	**0.50**	**0.54**

Main Results

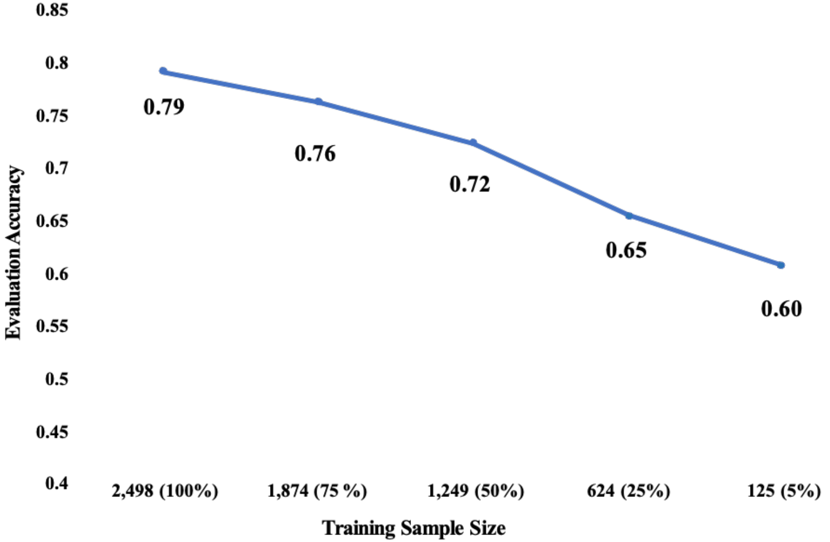
BERT Model Selection and Cross-Validation Results on the Sample of Selected Business Wire Q1 Earnings Announcements

	Training Obs. (n= 2,498) Positive Labels	Training Obs. (n= 2,498) Negative Labels	Test Obs. (n=625) Positive Labels	Test Obs. (n=625) Negative Labels	AUC	Evaluation Accuracy	F1 Score
Panel A: 10-sentence Documents							
Average Test Results on 10-sentence Documents using Ten-fold Cross-Validation	**1,380**	**1,118**	**343**	**282**	**0.76**	**0.77**	**0.75**
Panel B: 15-sentence Documents							
Average Test Results on 15-sentence Documents using Ten-fold Cross-Validation	**1,380**	**1,118**	**343**	**282**	**0.78**	**0.79**	**0.78**
Panel C: Benchmarking Models							
Loughran-McDonald Dictionary Classification	**NA**	**NA**	**343**	**282**	**NA**	**0.64**	**0.64**
Naïve Bayes Sentiment Classifier (TextBlob)	**NA**	**NA**	**343**	**282**	**NA**	**0.53**	**0.68**
Random (Dummy) Classifier	**1,380**	**1,118**	**343**	**282**	**0.49**	**0.50**	**0.54**

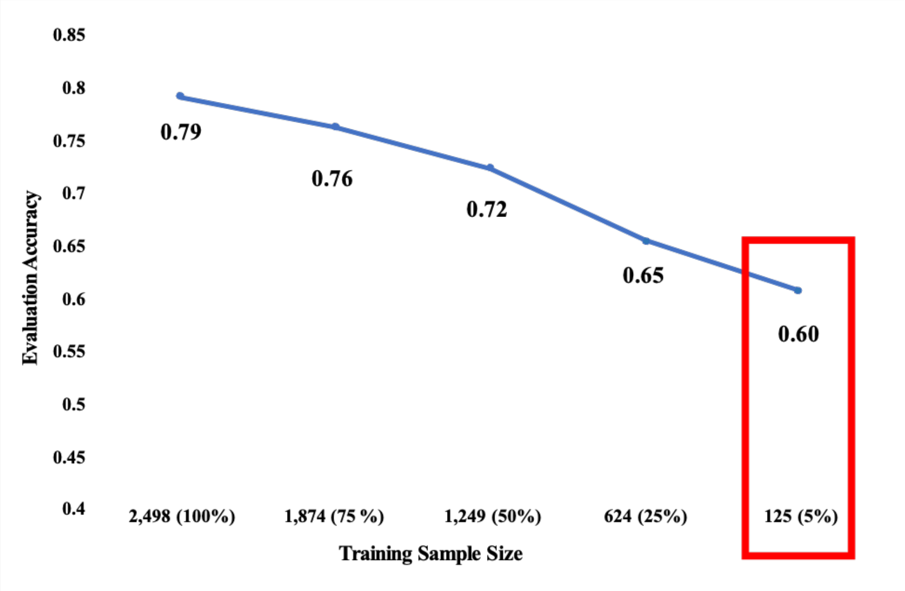
BERT Confusion Matrix



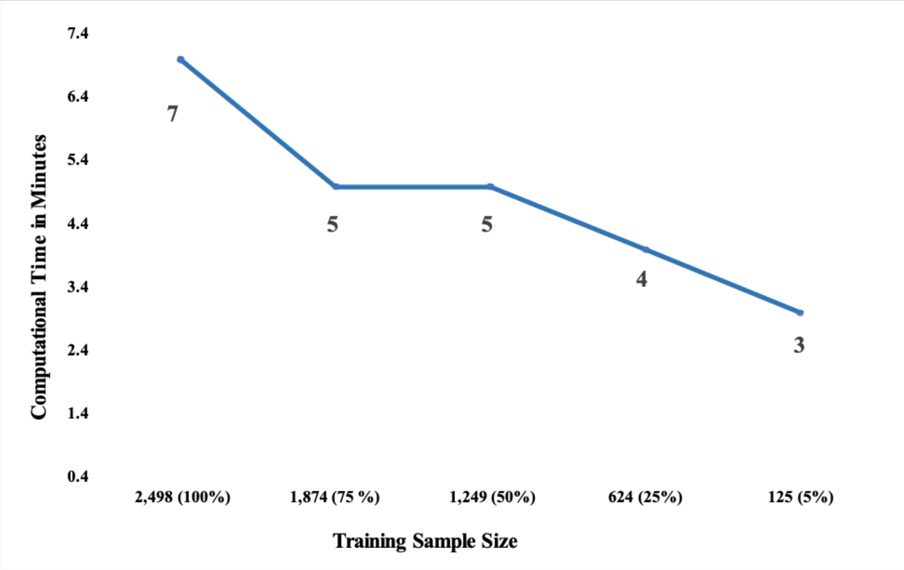
Evaluation Accuracy Across Distinct Sample Sizes



Evaluation Accuracy Across Distinct Sample Sizes



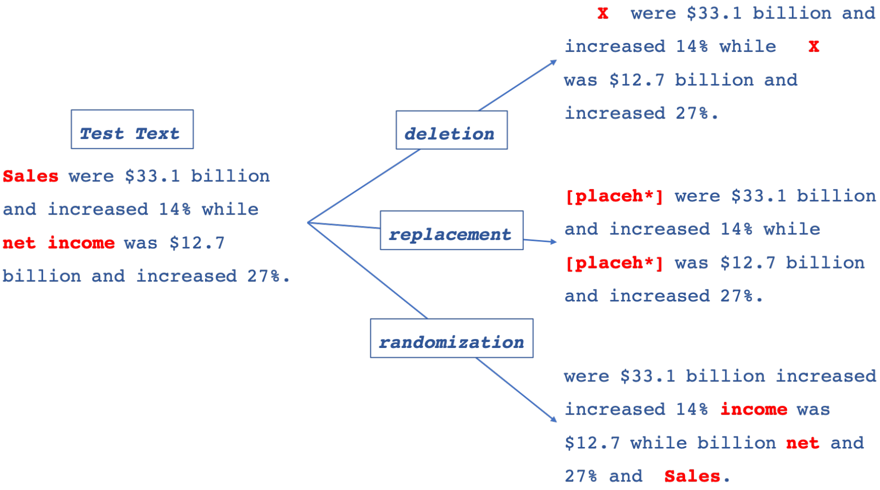
Computational Time Across Distinct Sample Sizes



"Masking" Tests

- Is ***BERT*** achieving **high accuracies** for the **wrong reasons**?

"Masking" Tests



"Masking" Tests

BERT Model Prediction Accuracy on the "Masked" Text of Business Wire Q1 Earnings Announcements

Labeling Benchmark	Masking Test	Average Accuracy	Average AUC	Average F1	Total Observations
Median of y-o-y change in Quarterly Sales (by calendar quarter)	"Unmasked" 15-sentence Documents Benchmark	79%	78%	78%	625
	delete "I/income" and "N/net income"	75%	74%	76%	625
	delete "S/sales", "R/revenues" and "E/earnings"	74%	74%	77%	625
	delete "L/loss" and "L/losses"	68%	67%	75%	625
	delete "I/increase/s/d" and "D/decrease/s/d"	72%	71%	71%	625
	delete all numbers	71%	71%	71%	625
	replace "I/income" and "N/net income"	74%	73%	75%	625
	replace "S/sales", "R/revenues" and "E/earnings"	72%	72%	74%	625
	replace "L/loss" and "L/losses"	65%	64%	73%	625
	replace all numbers	65%	64%	72%	625
	randomize *all words*	66%	60%	48%	625

"Masking" Tests

BERT Model Prediction Accuracy on the "Masked" Text of Business Wire Q1 Earnings Announcements

Labeling Benchmark	Masking Test	Average Accuracy	Average AUC	Average F1	Total Observations
Median of y-o-y change in Quarterly Sales (by calendar quarter)	"Unmasked" 15-sentence Documents Benchmark	79%	78%	78%	625
	delete "I/income" and "N/net income"	75%	74%	76%	625
	delete "S/sales", "R/revenues" and "E/earnings"	74%	74%	77%	625
	delete "L/loss" and "L/losses"	68%	67%	75%	625
	delete "I/increase/s/d" and "D/decrease/s/d"	72%	71%	71%	625
	delete all numbers	71%	71%	71%	625
	replace "I/income" and "N/net income"	74%	73%	75%	625
	replace "S/sales", "R/revenues" and "E/earnings"	72%	72%	74%	625
	replace "L/loss" and "L/losses"	65%	64%	73%	625
	replace all numbers	65%	64%	72%	625
	randomize *all words*	66%	60%	48%	625

"Masking" Tests

BERT Model Prediction Accuracy on the "Masked" Text of Business Wire Q1 Earnings Announcements

Labeling Benchmark	Masking Test	Average Accuracy	Average AUC	Average F1	Total Observations
Median of y-o-y change in Quarterly Sales (by calendar quarter)	"Unmasked" 15-sentence Documents Benchmark	79%	78%	78%	625
	delete "I/income" and "N/net income"	75%	74%	76%	625
	delete "S/sales", "R/revenues" and "E/earnings"	74%	74%	77%	625
	delete "L/loss" and "L/losses"	68%	67%	75%	625
	delete "I/increase/s/d" and "D/decrease/s/d"	72%	71%	71%	625
	delete all numbers	71%	71%	71%	625
	replace "I/income" and "N/net income"	74%	73%	75%	625
	replace "S/sales", "R/revenues" and "E/earnings"	72%	72%	74%	625
	replace "L/loss" and "L/losses"	65%	64%	73%	625
	replace all numbers	65%	64%	72%	625
	randomize *all words*	66%	60%	48%	625

"Masking" Tests

BERT Model Prediction Accuracy on the "Masked" Text of Business Wire Q1 Earnings Announcements

Labeling Benchmark	Masking Test	Average Accuracy	Average AUC	Average F1	Total Observations
Median of y-o-y change in Quarterly Sales (by calendar quarter)	"Unmasked" 15-sentence Documents Benchmark	$\boxed{79\%}$	78%	78%	625
	delete "I/income" and "N/net income"	75%	74%	76%	625
	delete "S/sales", "R/revenues" and "E/earnings"	74%	74%	77%	625
	delete "L/loss" and "L/losses"	68%	67%	75%	625
	delete "I/increase/s/d" and "D/decrease/s/d"	72%	71%	71%	625
	delete all numbers	71%	71%	71%	625
	replace "I/income" and "N/net income"	74%	73%	75%	625
	replace "S/sales", "R/revenues" and "E/earnings"	72%	72%	74%	625
	replace "L/loss" and "L/losses"	65%	64%	73%	625
	replace all numbers	65%	64%	72%	625
	randomize *all words*	$\boxed{66\%}$	60%	48%	625

Conclusions

1. With **small fine-tuning data** and for **modest computational time**, ***BERT*** **achieves** classification **accuracies** that are far **above** those of **benchmarking methods**

Conclusions

1. With **small fine-tuning data** and for **modest computational time**, ***BERT* achieves** classification **accuracies** that are far **above** those of **benchmarking methods**

2. *BERT* seems to **provide** with a **feature representation** of text that captures **fundamental semantic connections in language**

Conclusions

1. With **small fine-tuning data** and for **modest computational time**, ***BERT*** **achieves** classification **accuracies** that are far **above** those of **benchmarking methods**

2. *BERT* seems to **provide** with a **feature representation** of text that captures **fundamental semantic connections in language**

3. Data proliferation, high power, modest resource intensiveness and easiness of use likely make **Transfer Learning a long-lasting phenomenon**

THANK YOU!