Lexical Data Augmentation for Text Classification in Deep Learning

Rong XIANG[1], Emmanuele CHERSONI[1], Yunfei LONG[2],
Qin LU[1], Chu-Ren HUANG[1]
The Hong Kong Polytechnic University[1]
University of Essex[2]

Outline

1. Introduction
2. Proposed Method
3. Performance Evaluation
4. Conclusion and Future Work

1. Introduction

- **Data augmentation**

  Improve the performance of machine learning algorithms

  Lexical data augmentation

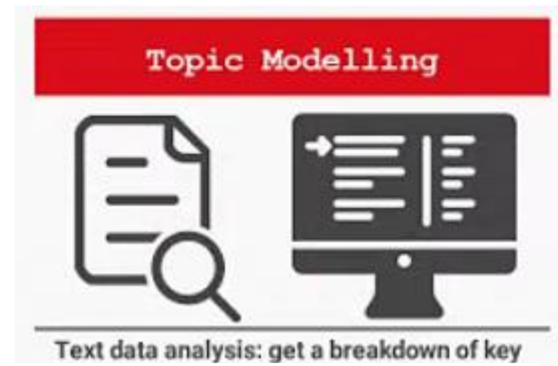
*Without Shakespeare's eloquent **language**, the update is **dreary** and **sluggish**.*

- Without Shakespeare's eloquent **speech**, the update is dreary and sluggish.
- Without Shakespeare's eloquent **speech**, the update is dreary and **dull**.
- Without Shakespeare's eloquent **terminology**, the update is **drab** and sluggish.
- Without Shakespeare's eloquent **speech**, the update is dreary and sluggish.
- Without Shakespeare's eloquent language , the update is dreary and **dull**.

- **Text Classification**

  Widely used

  High performance



Text Classification on IMDb

RANK	METHOD	ACCURACY	PAPER TITLE
1	XLNet	96.21	XLNet: Generalized Autoregressive Pretraining for Language Understanding
2	BERT Finetune + UDA	95.80	Unsupervised Data Augmentation for Consistency Training

Motivation

- **Transformer-based methods**

  Benefited from tremendous training data

  (BERT: 13GB, XLNET: 113GB, etc.)

- **Data augmentation**

  Data scarcity problem

  Effective in traditional machine learning

- **Is data augmentation still useful?**

  Deep learning methods

  Specific domain

2. Proposed Method

- **POS-focused lexical substitution data augmentation**

Two steps augmentation:

- **Substitution Candidate Selection**

  syntactic consistency principle

  uses POS constraints

  select candidate words for substitution.

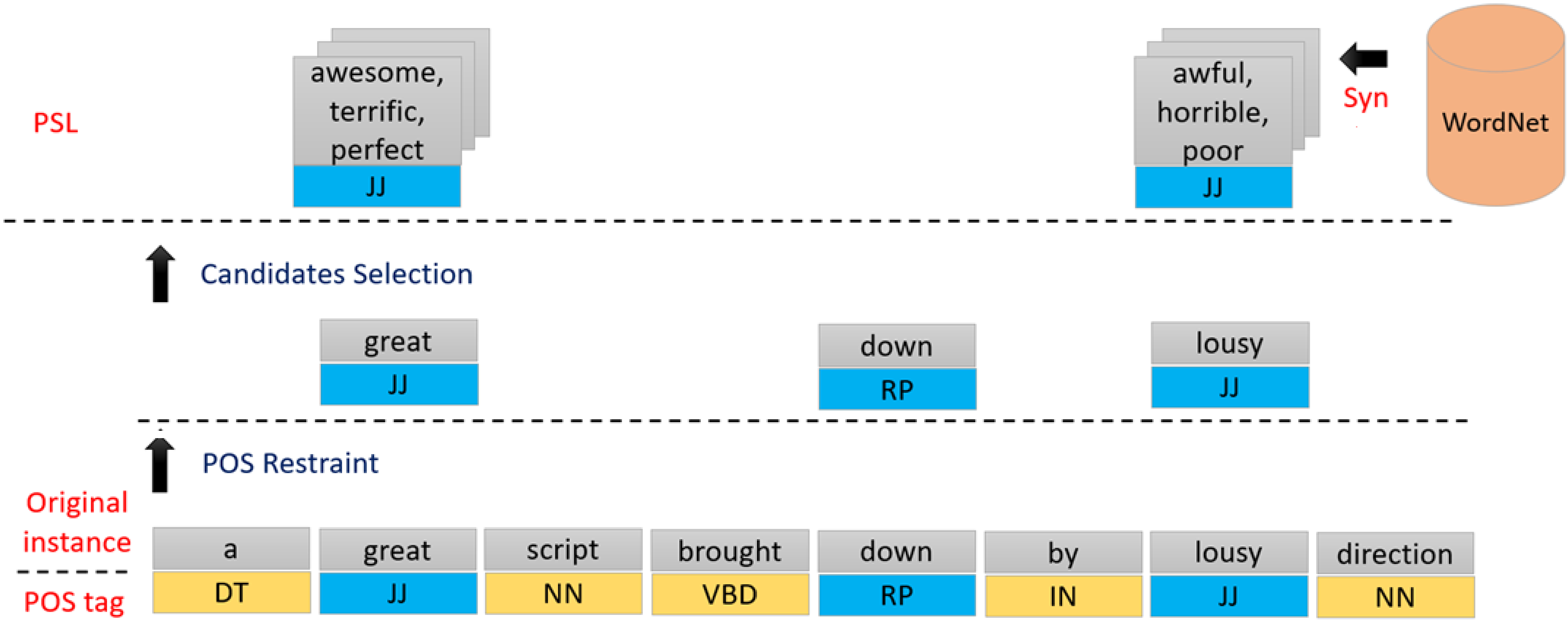
- **Instance Generation**

  semantic consistency principle

  samples from candidate words

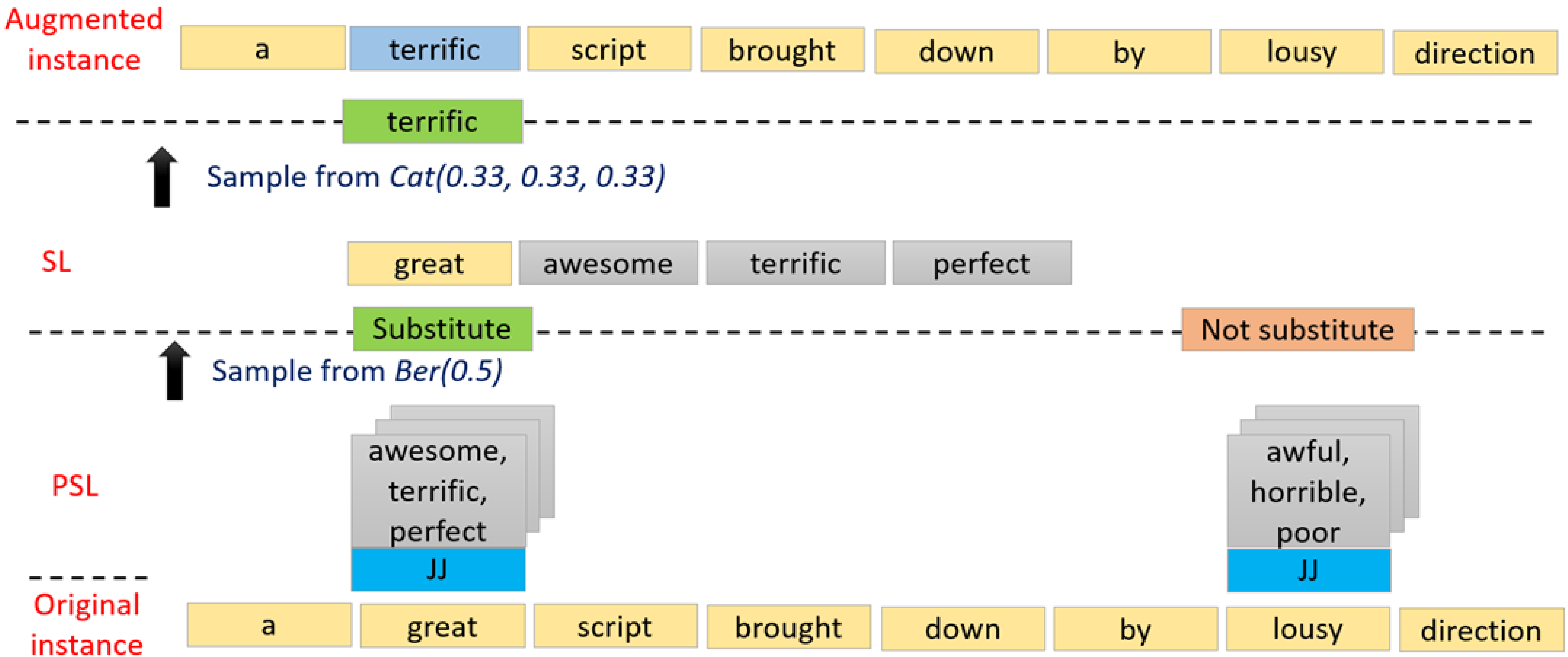
  generate augmented instances

Substitution Candidate Selection



Substitution Candidate Selection

Instance Generation



Instance Generation

3. Performance Evaluation

- ## Classifiers

LSTM a deep learning model uses pre-trained GloVe for word embedding initialization.

BiLSTM-AT uses bidirectional LSTM for document representation. Extended with attention mechanism.

BERT a transformer-based deep learning model. Task-specific fine-tuning is used to achieve best performance.

RoBERTa a refined model based on BERT.

- ## Metric

Accuracy

- ## Baseline method

EDA: Easy Data Augmentation Techniques for Boosting Performance onText Classification Tasks

Dataset

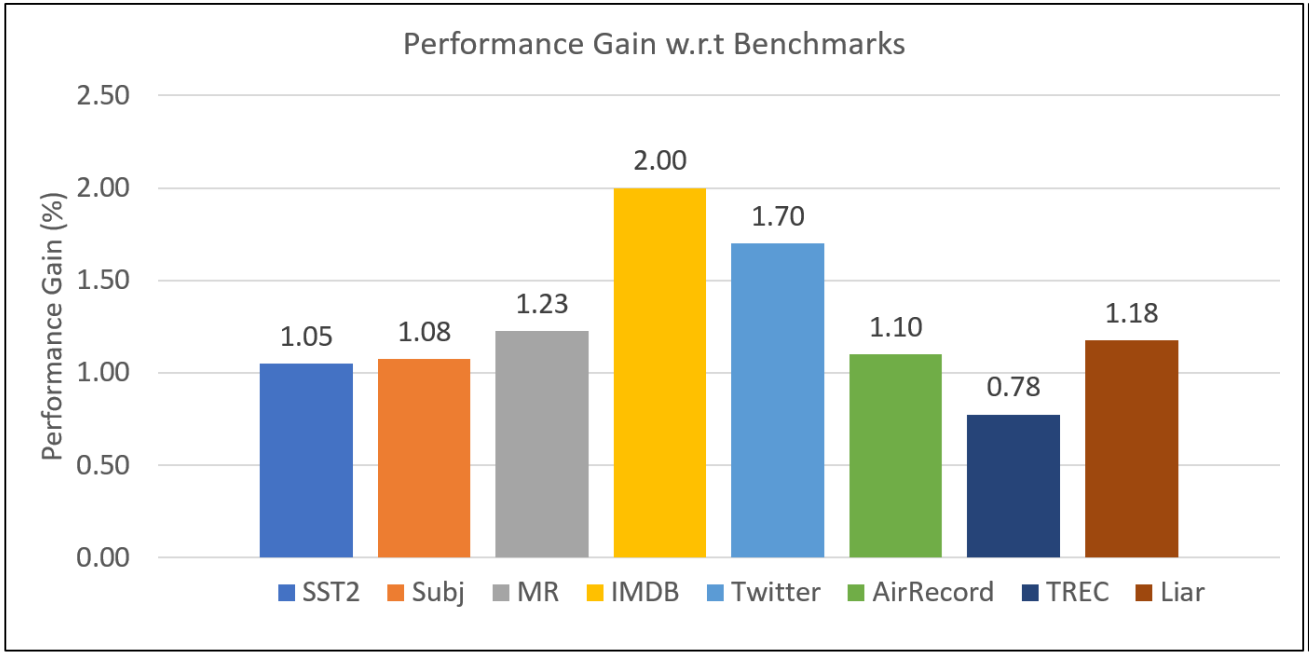
Dataset	N_{train}	N_{test}	L_{avg}	N_{voc}	C
SST-2	6,920	1,821	19	16,185	2
Subj	9,000	1,000	23	21,323	2
MR	9,595	1,067	20	18,765	2
IMDB	22,500	2,500	260	184,885	2
Twitter	89,989	9,999	14	183,645	2
AirRecord	13,172	1.464	18	30,166	3
TREC	5,452	500	10	9,592	6
Liar	10,269	1,283	18	22,765	6

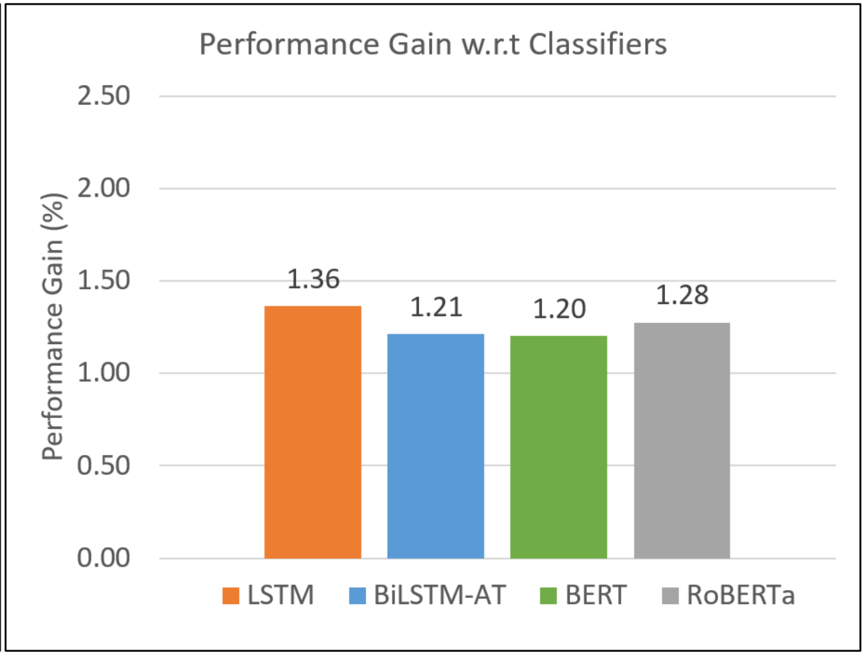
General Performance

	SST-2	Subj	MR	IMDB	Twittter	AirRecord	TREC	Liar
LSTM	80.2	90.8	77.0	80.3	74.7	80.5	88.8	25.3
+EDA	80.9	91.3	77.6	81.2	75.7	81.2	89.3	26.0
+PLSDA	81.0	91.9	78.1	82.6	77.2	81.4	89.3	27.0
BiLSTM-AT	78.2	91.0	75.9	80.5	75.9	81.3	88.3	25.7
+EDA	78.9	91.5	76.6	81.8	76.9	81.9	88.9	26.3
+PLSDA	79.7	92.1	76.8	83.0	77.6	82.0	88.8	26.5
BERT	91.3	97.2	87.1	88.1	82.0	83.2	96.8	27.9
+EDA	92.0	97.4	88.0	88.9	82.7	83.9	97.5	28.2
+PLSDA	92.3	**98.4**	88.7	89.6	83.2	84.4	<u>97.6</u>	**29.0**
RoBERTa	93.0	97.3	90.3	89.1	83.3	84.3	96.5	27.2
+EDA	<u>93.7</u>	97.4	<u>90.7</u>	<u>90.0</u>	<u>84.1</u>	<u>85.5</u>	97.5	27.7
+PLSDA	**93.9**	<u>98.2</u>	**91.6**	**90.8**	**84.7**	**85.9**	**97.8**	<u>28.3</u>

Accuracy of the models: the best is in bold and the second-best is underlined.

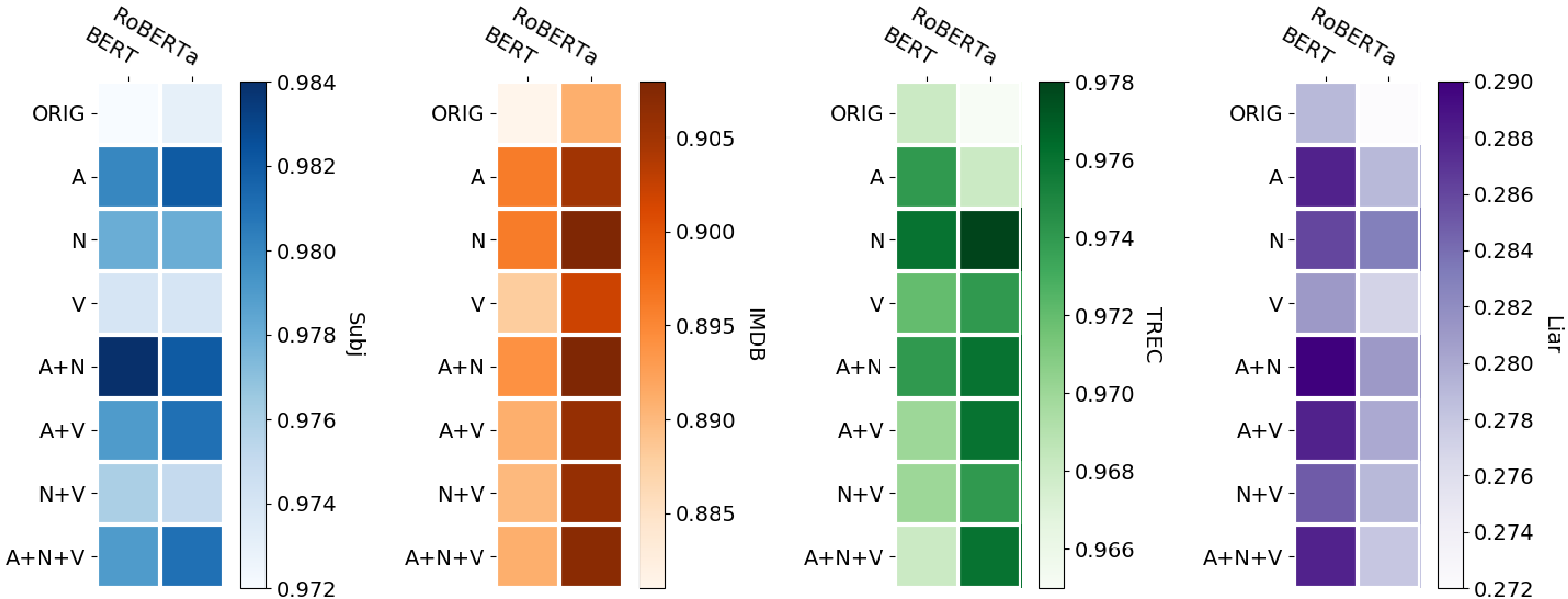
Performance Gain





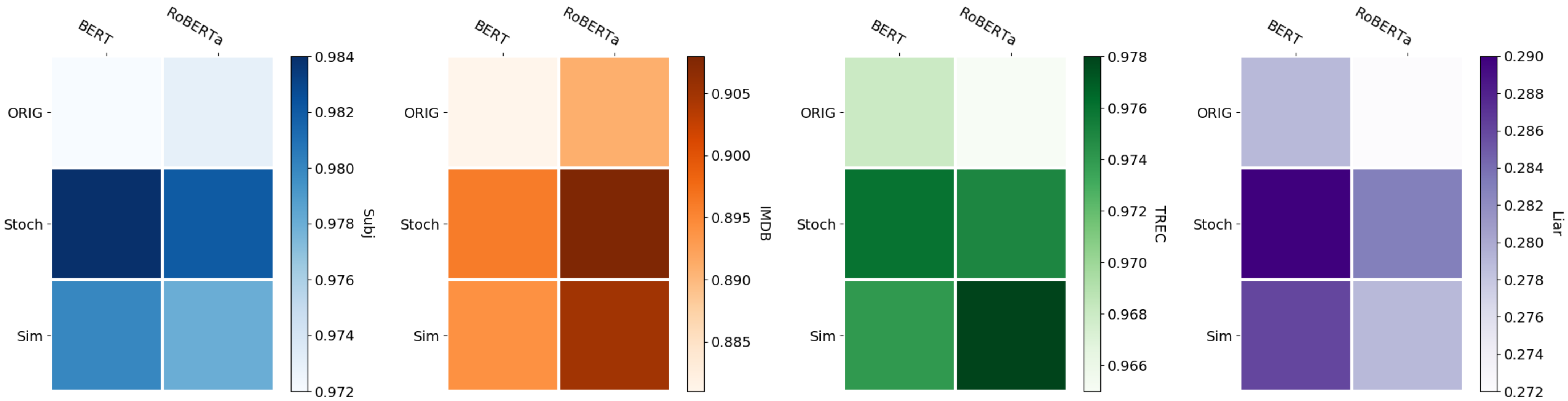
Absolute Performance Gains(%) on
Average Accuracy by PLSDA

Lexicon POS Selection



Heatmaps of Lexicon POS

Sampling Strategy Study



Heatmaps of Sampling Strategy

4. Conclusion & Future Work

- **Conclusions**
  - ✓ Data augmentation further improves the performance of deep learning models.
  - ✓ Nouns and adjectives/adverbs work better as replacement types.
  - ✓ Stochastic sampling outperform similarity-first strategy in finding lexical replacement.

- **Future Work**
  - ✓ Investigate the performance of PLSDA on more publicly accessible datasets.
  - ✓ Explore the feasibility of PLSDA in other NLP tasks.

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Thank you !