Pre-trained Data Augmentation for Text Classification

Hugo Queiroz Abonizio

Prof. Dr. Sylvio Barbon Jr.

Introduction

- Increasing in data generation in recent years
  - Social networks
  - Information systems
  - News agencies
- Automatizing the data processing becomes a challenge in unstructured data

Introduction

- Text mining
  - Merge between data mining and Natural Language Processing
  - Pattern recognition and knowledge extraction
  - Implicit knowledge present in unstructured documents

Introduction

- Text classification problems
    - Automatic categorization of text document into predefined classes
- Maps feature vectors into classes

$$f : \mathscr{D} \rightarrow \boldsymbol{\mathscr{L}}$$

$$f(x_i) = y_i$$

Introduction

- Supervised learning approach
  - Pairs of labeled data
- Large amounts of labeled data
  - Labeling process is costly
  - Prone to overfitting
  - Generalization capacity requires variation on the training set

Introduction

- Usual pipeline for text classification
  a) Data preparation and preprocessing
  b) Feature extraction
  c) Model fitting
  d) Inference/prediction

Introduction

- Data augmentation
    - Generate similar but not identical samples
    - Helps in data scarcity problem
- Widely adopted in Computer Vision field
    - Rotation
    - Equalization
    - Random cropping

Introduction

- Text data augmentation
  - Underexplored techniques when comparing with Computer Vision
  - Recent popularization
  - Label-preserving is difficult

Introduction

- Most explored approaches:
  - Word-level transformations
  - Back-translation (BT)
  - Language model

Introduction

- Word-level transformations
  - Synonym substitution
  - Dictionaries
  - Random insertion/deletion
  - Easy Data Augmentation (EDA) (Wei et al. 2019)

Introduction

- Back-translation

I have no time $\xrightarrow{\text{French}}$ je n'ai pas le temps $\xrightarrow{\text{English}}$ I do not have time

Introduction

- Language models
  - Trained to predict the next word in a sentence

Paris is the capital of

France	0.864
france	0.056
the	0.028
...	

Introduction

- Language models
  - Substitutions based on the context
  - Leverage text generation capacity of pre-trained models

Objective and contributions

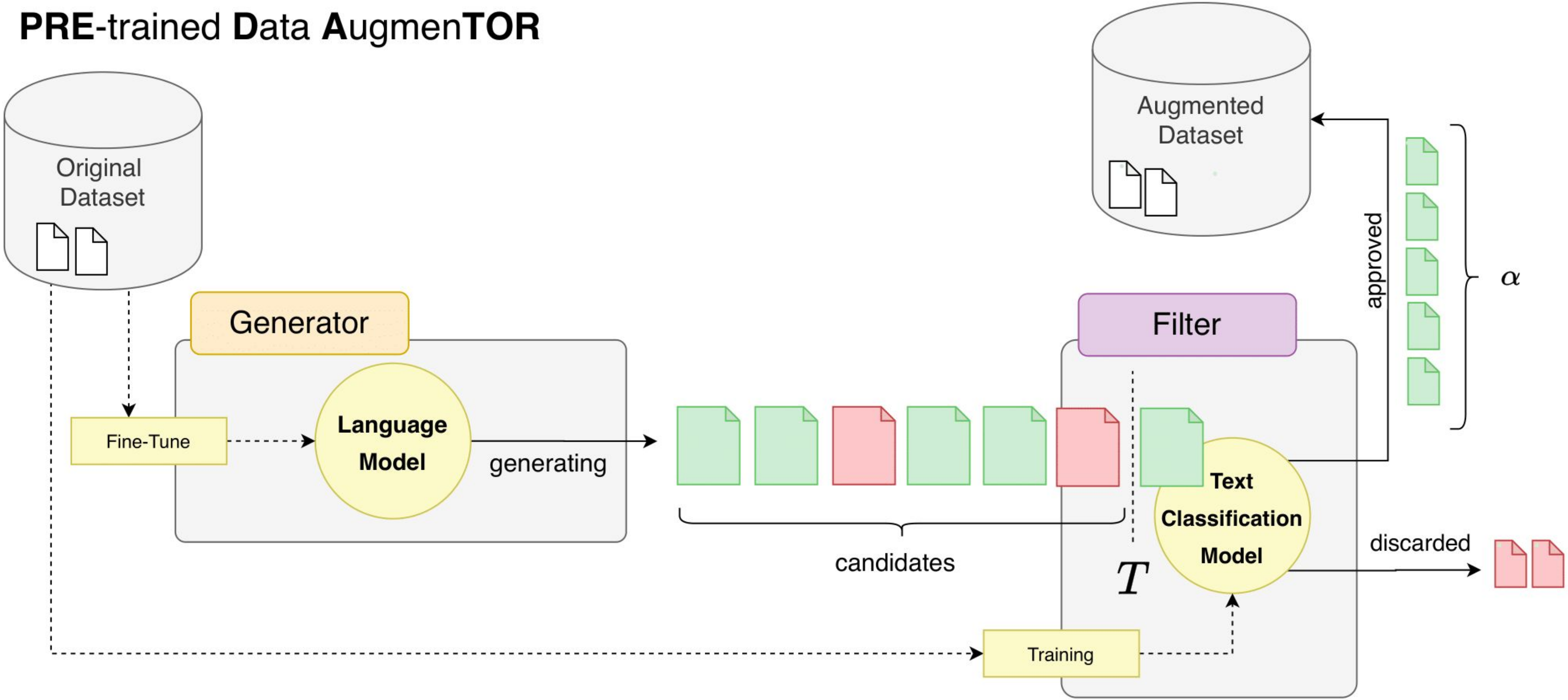
- Propose a method of augmentation for text classification problems that is **robust** and **lightweight**
- Comparison of the proposed method with literature methods
- Investigation of the impact in different domains

PRE-trained **D**ata **A**ugmen**TOR**

Proposed approach

- Data augmentation through generation of new samples leveraged by transfer learning
- Label-preserving through a semi-supervised learning approach
- Low computational cost

Proposed approach



Experiments
Datasets

- AG-NEWS
  - Topic classification
  - News
- CyberTrolls
  - Cyber-aggressive behavior detection
  - Social media posts
- SST-2
  - Movie reviews
  - Sentiment analysis

Experiments

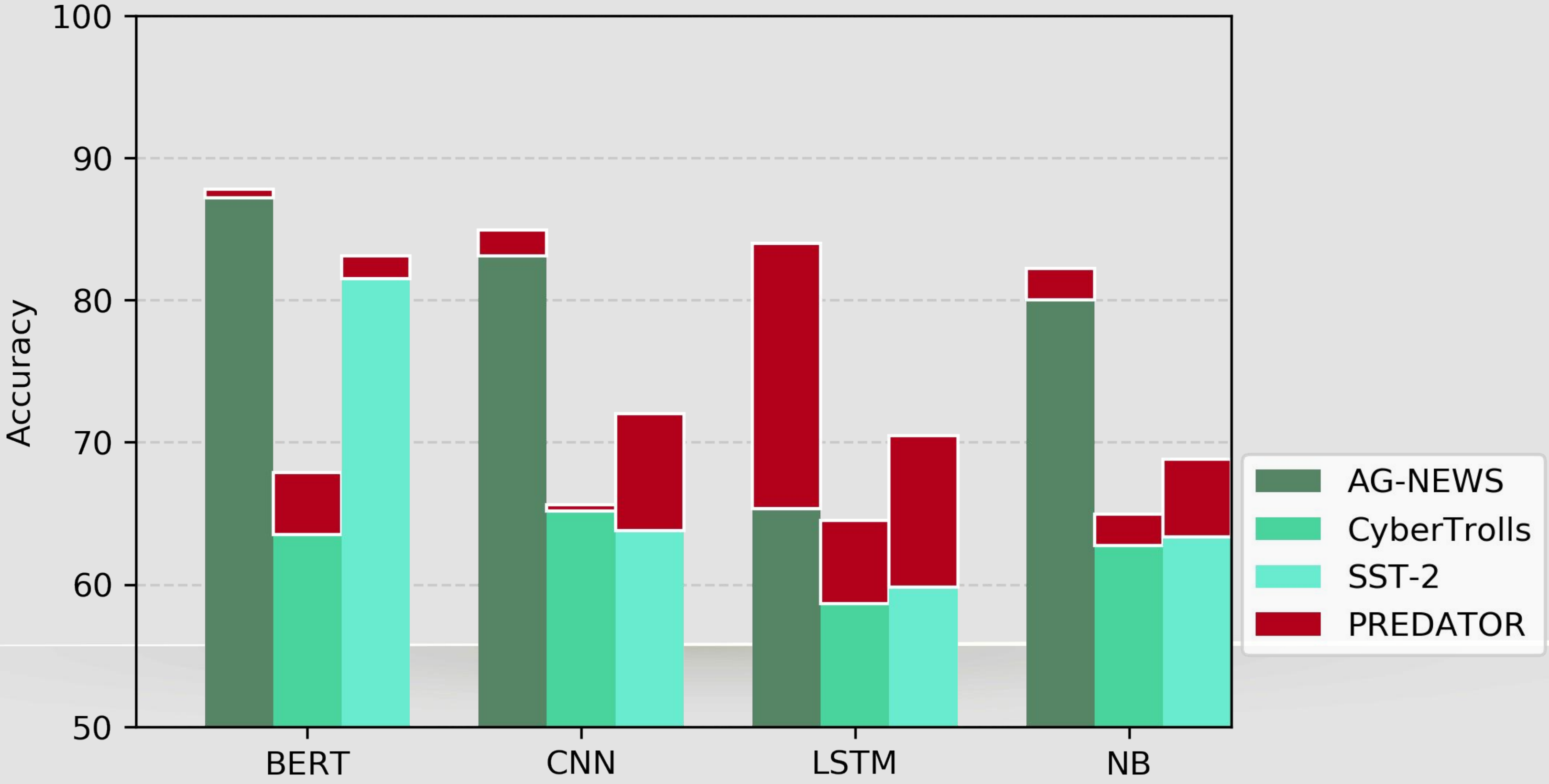
Classifiers

- Traditional
  - Naïve Bayes
- Deep learning
  - CNN
  - LSTM
- State-of-the-art
  - BERT

Experiments
Comparison

- **EDA**
  - Increases the size in 9x
- **Back-translation (BT)**
  - Doubles the dataset size
- **Original labeled data**
  - Samples from original dataset

Results



100
90
80
70
60
50
Accuracy
BERT
CNN
LSTM
NB
AG-NEWS
CyberTrolls
SST-2
PREDATOR

Results

- Comparison with literature methods
  - Accuracy different (average)

Dataset	Method		
	EDA	BT	PREDATOR
AG-NEWS	+4.9% (±3%)	+3.6% (±5%)	**+7.4% (±2%)**
CyberTrolls	-0.9% (±4%)	+1.9% (±2%)	**+5.1% (±2%)**
SST-2	+3.4% (±8%)	+3.7% (±9%)	**+9.7% (±6%)**

Results

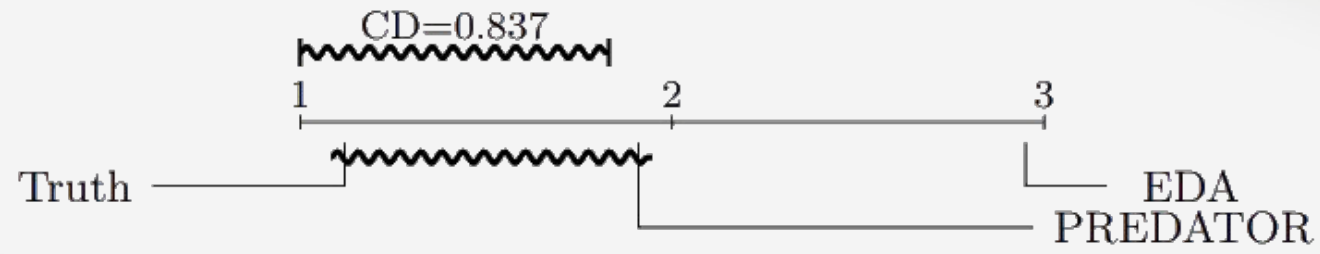
- Comparison with original labeled data
  - Sampling from original dataset with compatible sizes (2x EDA and 10x BT)

Dataset	Método		
	Original 2x	Original 10x	PREDATOR 10x
AG-NEWS	+4.7% (±6%)	**+10.3% (±3%)**	+7.4% (±2%)
CyberTrolls	+3.6% (±2%)	**+14.5% (±3%)**	+5.1% (±2%)
SST-2	+7.6% (±8%)	**+19.5% (±5%)**	+9.7% (±6%)

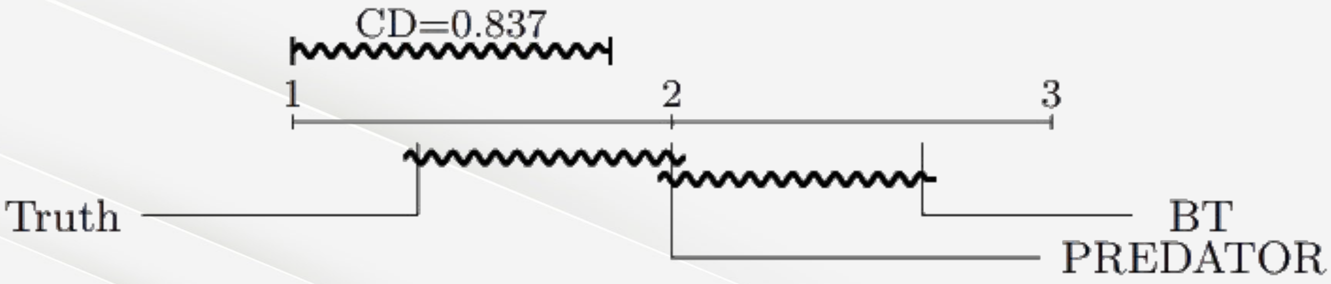
Results

- Statistic comparison of the results

10x



2x



Conclusions

- The proposed method proved effective in improving the performance of models on evaluated scenarios
- Statistically superior to EDA
- Statistically similar to including real labeled data
- Efficient regardless of domain
- Robust against linguistic variations

References

- AGGARWAL, C. C.; ZHAI, C. A survey of text classification algorithms. In: Mining text data. [S.l.]: Springer, 2012. p. 163–222.
- WEI, J.; ZOU, K. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019. p. 6382–6388.
- MINAEE, S. et al. Deep learning based text classification: A comprehensive review. arXiv preprint arXiv:2004.03705, 2020.
- RADFORD, A. et al. Language models are unsupervised multitask learners. 2019.
- KOWSARI, K. et al. Text classification algorithms: A survey. Information, v. 10, n. 4, 2019. ISSN 2078-2489. Disponível em: <https://www.mdpi.com/2078-2489/10/4/150>.