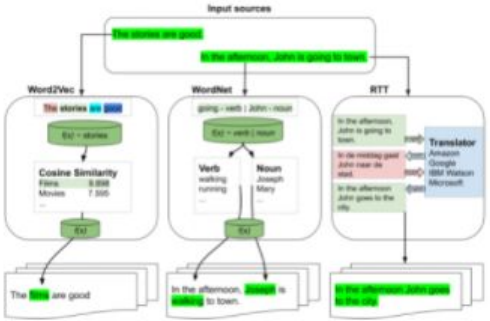
Group Meeting

December (2021)

Remember

Remember



EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks

Jason Wei[1,2] **Kai Zou**[3]
[1]Protago Labs Research, Tysons Corner, Virginia, USA
[2]Department of Computer Science, Dartmouth College
[3]Department of Mathematics and Statistics, Georgetown University
jason.20@dartmouth.edu kz56@georgetown.edu

Remember

- Create valid texts;
- Create texts maintaining the label;
- Create texts with varying sizes;
- Multilingual / language agnostic;
- Execution in devices with low computing power.

Artigo

Artigo

Autores: Lucas Z. Ladeira, Frances Santos, Lucas Cléopas, Pieter Buteneers, and Leandro Villas

Artigo

NEO-NDA: Neo Natural Language Data Augmentation

Artigo - NEO-NDA

NEO-NDA: Neo Natural Language Data Augmentation

Data Augmentation + Downsampling

Artigo - NEO-NDA

NEO-NDA: Neo Natural Language Data Augmentation

Data Augmentation

- Random Insertion: selects a random word and inserts its synonym in a random position of the sentence;
- Random Switch: selects randomly two words and switch their places;
- Random Deletion: delete a random word of the sentence;
- Synonyms Switch: selects randomly one word and switch it for a synonym;
- Translate Back: translate the sentence to another language and back.

Artigo - NEO-NDA

NEO-NDA: Neo Natural Language Data Augmentation

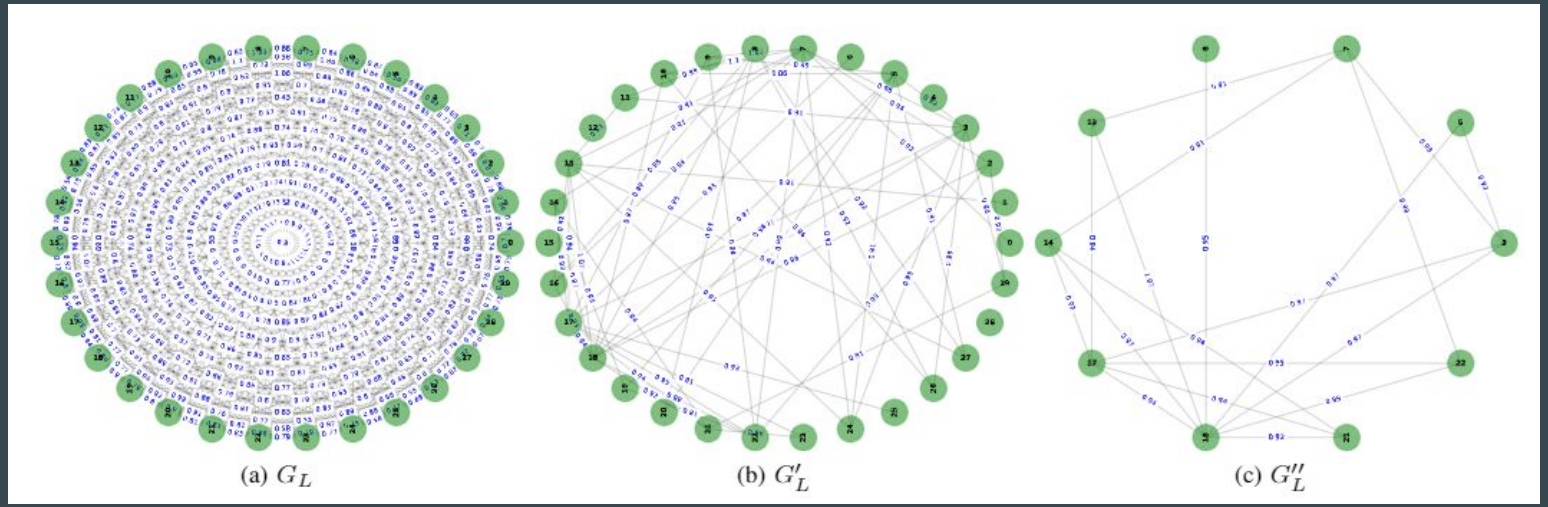
Downsampling

- Uses LaBSE (Language-Agnostic BERT Sentence Embedding) to encode sentences to high dimensional vectors;
- Calculates the similarity using cosine similarity;
- Creates a graph with the similarity;
- Remove edges between two nodes if they are not too similar;
- Nodes are selected according to their degree.

Artigo - NEO-NDA

NEO-NDA: Neo Natural Language Data Augmentation

Downsampling



(a) G_L (b) G'_L (c) G''_L

Artigo - Related Work

TABLE I
RELATED WORK SUMMARY.

Work	Rule-based	Model-based	Multilingual	Distinct Sentence Sizes	Multiple Transformations
Wei et al. [1]	✓	✗	✗	✓	✓
Wei et al. [9]	✓	✗	✗	✓	✓
Li et al. [2]	✗	✓	✓	✓	✗
Kobayashi et al. [4]	✗	✓	✗	✗	✗
Yang et al. [10]	✗	✓	✗	✗	✗
Ciolino et al. [12]	✗	✓	✓	✓	✗
Giridhara et al. [11]	✓	✓	✗	✗	✓
NEO-NDA	✓	✓	✓	✓	✓

Artigo - Experiments

Embeddings

- Distilled Bert trained for multi-language;
- LaBSE;
- XLM.

Models

- Logistic Regression;
- Random Forest;
- Ada Boost.

TABLE IV
DATASETS USED IN OUR EVALUATION.

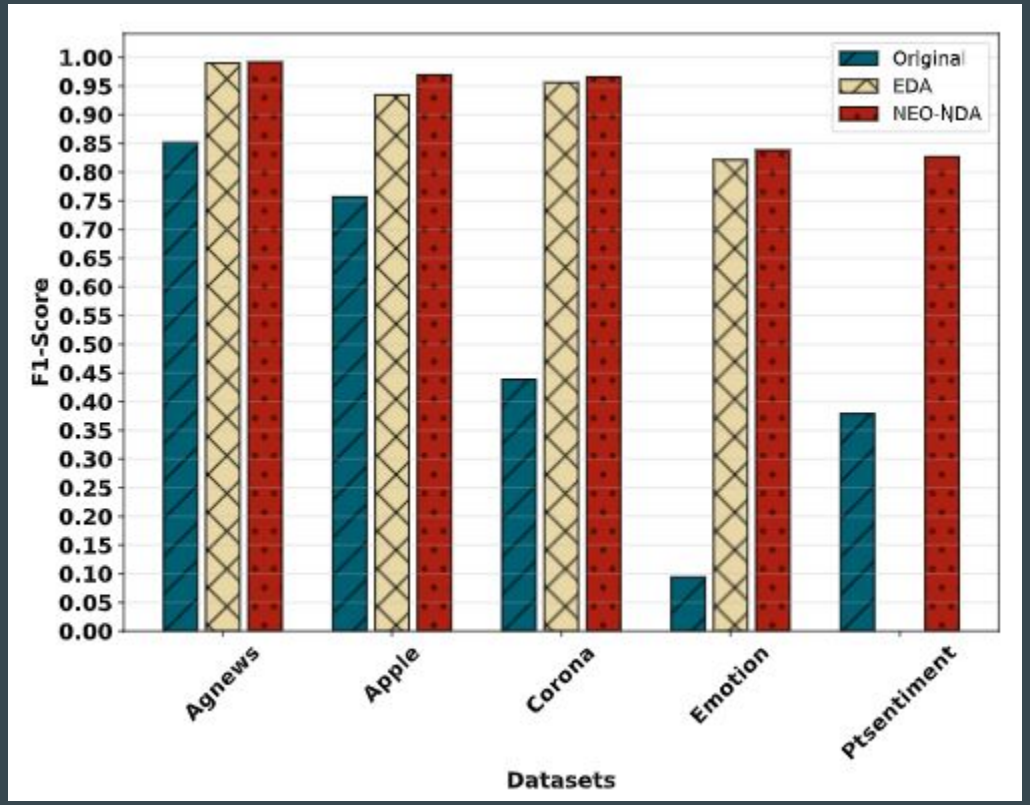
Name	Classes	Distribution	Language
Agnews [18]	4	1286 1270 1204 1240	English
AppleSentiment [19]	3	801 143 686	English
CoronaNLPData [20]	5	889 1263 1340 738 770	English
EmotionData [15]	13	201 272 33 93 22 67 106 981 178 93 1558 280 1116	English
PtSentimentData [21]	5	1041 296 586 1220 1857	Portuguese

Artigo - Results

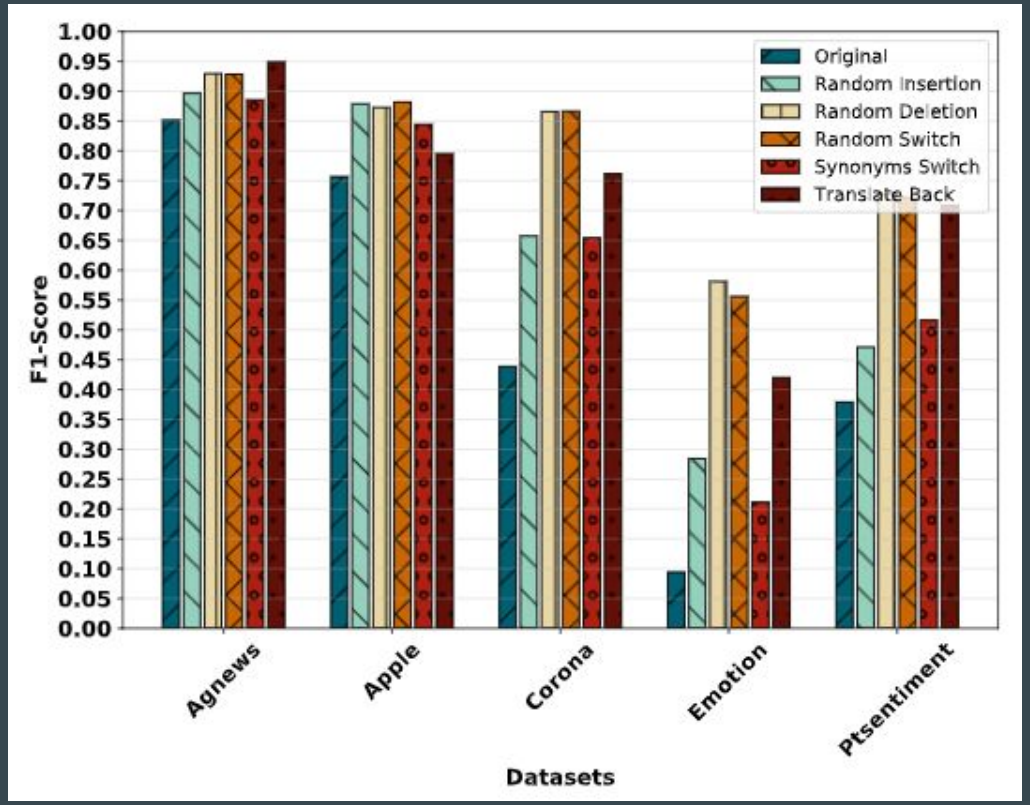
TABLE V
EMBEDDING AND ML ALGORITHMS COMPARISON

Embedding	ML Algorithm	F1-Score
Distil Bert	AdaBoost	0.475
	Logistic Regression	0.581
	Random Forest	0.688
Labse	AdaBoost	0.463
	Logistic Regression	0.586
	Random Forest	0.685
XLM	AdaBoost	0.488
	Logistic Regression	0.648
	Random Forest	**0.718**

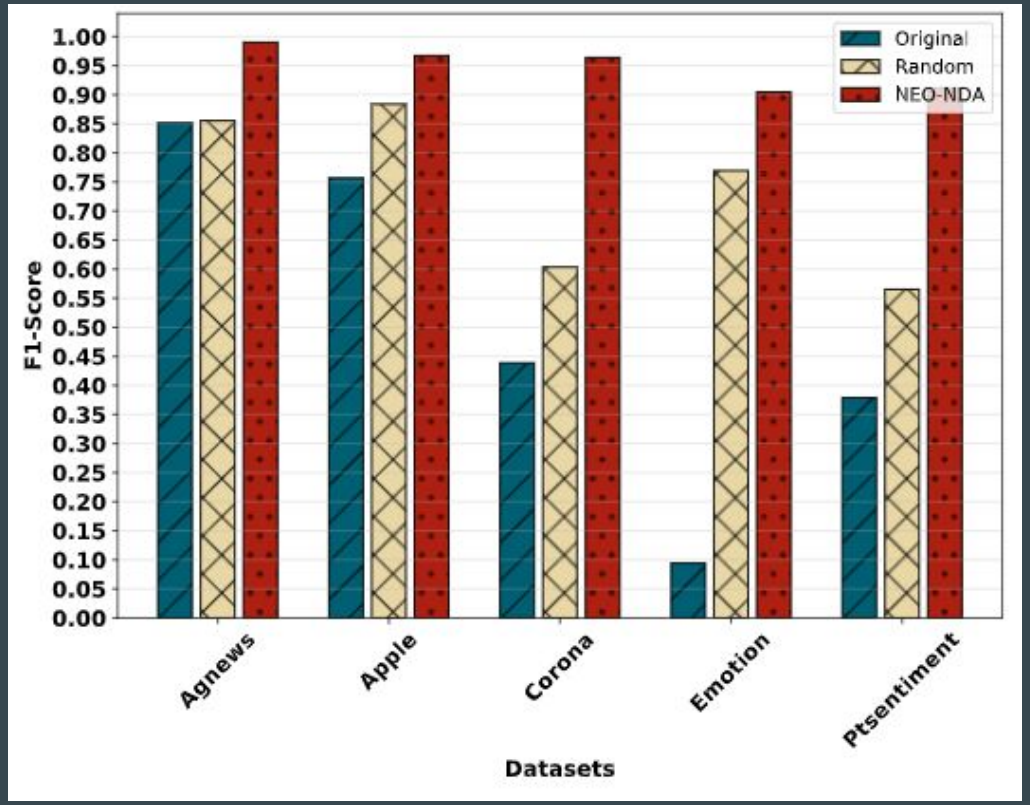
Artigo - Results



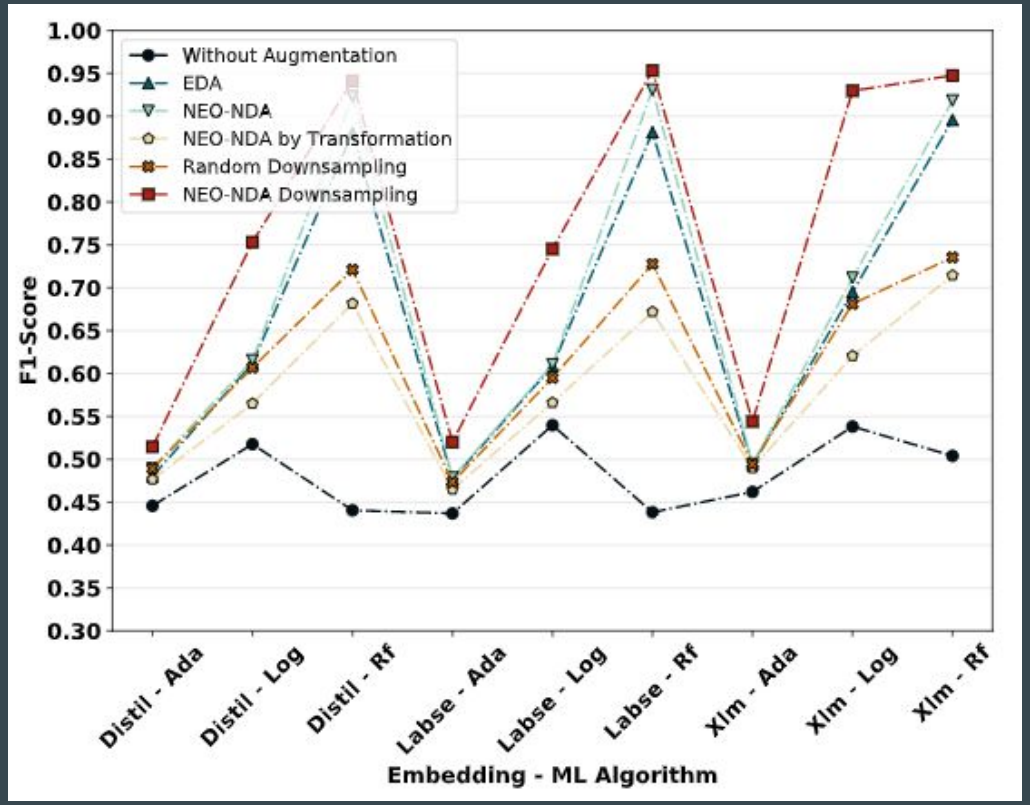
Artigo - Results



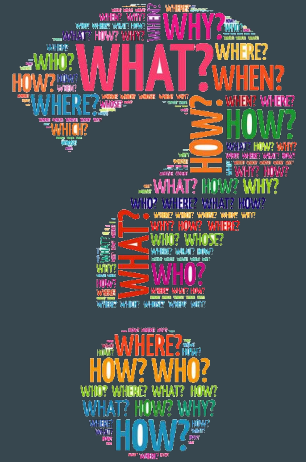
Artigo - Results



Artigo - Results



Thank you!


WHY?
WHERE?
WHO?
WHAT?
WHEN?
HOW?
WHERE?
WHICH?
HOW?
HOW?
WHAT? HOW? WHY?
WHAT?
WHO?
WHERE?
HOW? WHO?
WHO? WHERE? WHAT? HOW?
WHAT? HOW? WHY?
HOW?