A Benchmark of Unsupervised Off-The-Shelf Anomaly Detection Methods on ADFA-LD

Julia Rolland	UTBM
Quentin Fournier	Polytechnique Montréal
Daniel Aloise	Polytechnique Montréal

Progress Report Meeting
January 21st, 2022

Table of Contents

- Introduction
- Representation Methods
- Outlier Detection Methods
- Introduction to the ADFA-LD Dataset
- Results

Purpose of this internship

- Benchmark unsupervised off-the-shelf methods for anomaly detection
- Experiment with the ADFA-LD dataset
- Introduction to machine learning and deep learning

Representations of the data

Bag of Words

Counts the occurrences of each element in a document with the help of a vocabulary.

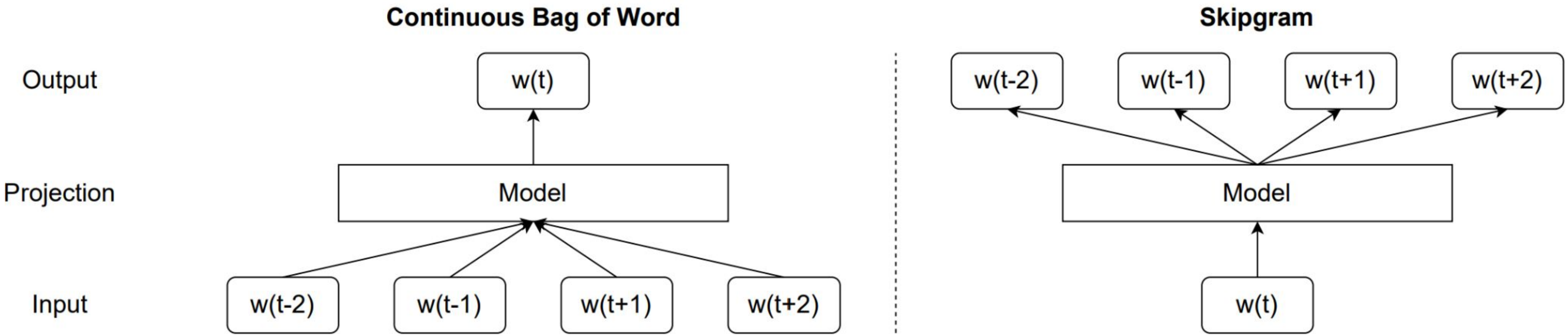
TF-IDF

Determines the importance/rarity of each element within a set of documents.

LDA

Searches for the most popular element in each document and represents the documents with their subjects.

Word2Vec - Skipgram



Word2vec - 2 possible implementations

Neural Network used to represent distributed representations of an element in a set of documents. Predict the context word of a target one.

Outliers Detection Methods

- Cosine Similarity → Geometry Based Method
- k-NN → Distance Based Method
- DBSCAN → Density Based Method
- Isolation Forest → Tree Based Method
- One Class SVM → Pattern Based Method

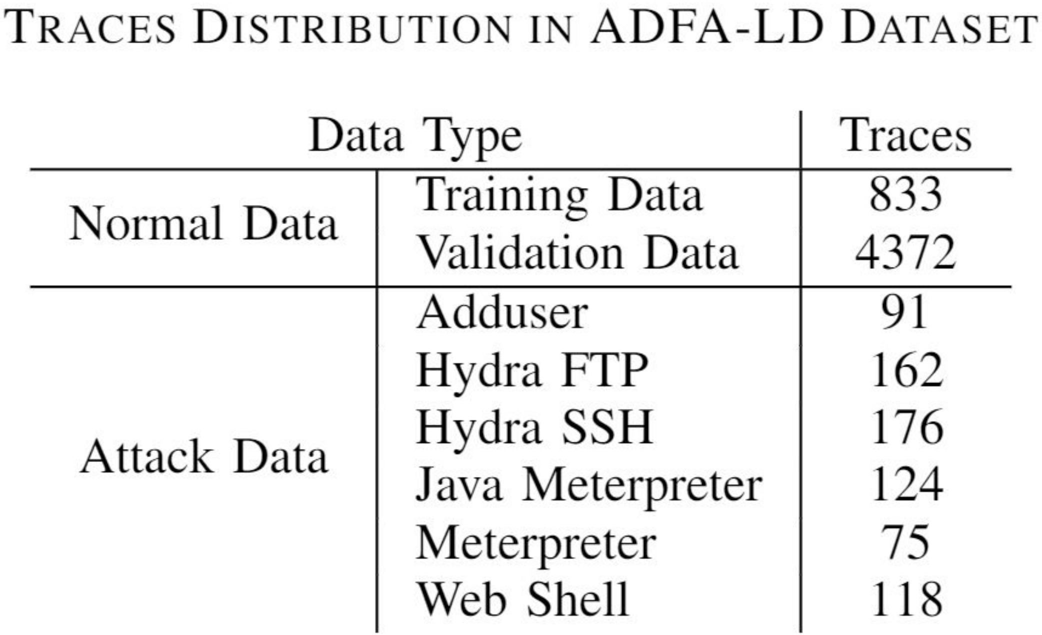
ADFA-LD

Developed on Ubuntu Linux v11.04 in 2014

Publicly available and labelled

Different Trace Categories

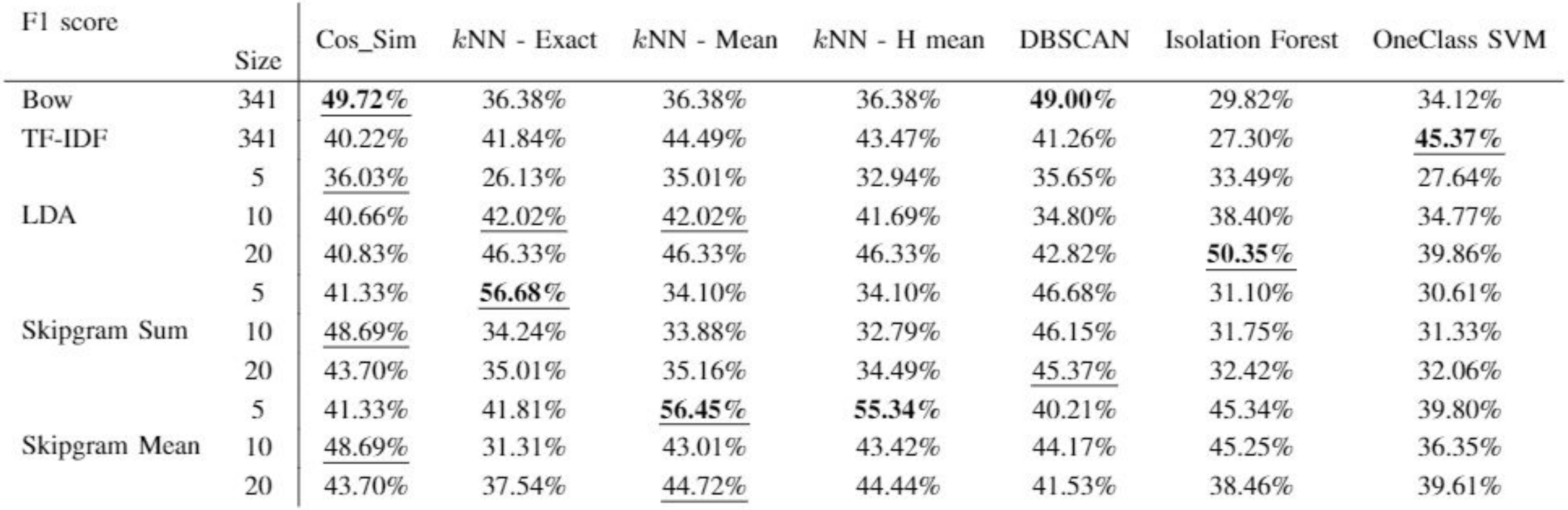
Small Dataset



TRACES DISTRIBUTION IN ADFA-LD DATASET

Data Type		Traces
Normal Data	Training Data	833
	Validation Data	4372
Attack Data	Adduser	91
	Hydra FTP	162
	Hydra SSH	176
	Java Meterpreter	124
	Meterpreter	75
	Web Shell	118

Results



F1 score	Size	Cos_Sim	*k*NN - Exact	*k*NN - Mean	*k*NN - H mean	DBSCAN	Isolation Forest	OneClass SVM
Bow	341	**49.72%**	36.38%	36.38%	36.38%	**49.00%**	29.82%	34.12%
TF-IDF	341	40.22%	41.84%	44.49%	43.47%	41.26%	27.30%	**45.37%**
LDA	5	36.03%	26.13%	35.01%	32.94%	35.65%	33.49%	27.64%
	10	40.66%	42.02%	42.02%	41.69%	34.80%	38.40%	34.77%
	20	40.83%	46.33%	46.33%	46.33%	42.82%	**50.35%**	39.86%
Skipgram Sum	5	41.33%	**56.68%**	34.10%	34.10%	46.68%	31.10%	30.61%
	10	48.69%	34.24%	33.88%	32.79%	46.15%	31.75%	31.33%
	20	43.70%	35.01%	35.16%	34.49%	45.37%	32.42%	32.06%
Skipgram Mean	5	41.33%	41.81%	**56.45%**	**55.34%**	40.21%	45.34%	39.80%
	10	48.69%	31.31%	43.01%	43.42%	44.17%	45.25%	36.35%
	20	43.70%	37.54%	44.72%	44.44%	41.53%	38.46%	39.61%

F1 score on the test set of each combination of representations and outlier detection methods. Bold results denote the best representation score and underlined results denote the best outlier detection method.

Deliverables

- Jupyter Notebook
- Internship report
- This presentation

https://github.com/Julia185/DORSAL_ADFA-LD

Thank you

Do you have any question ?