Using explainers to refine data pre-processing

SAI MUTTAVARAPU

Université Côte d'Azur
930 Rte des Colles, 06410 Biot

Academic Supervisor: **Michel RIVEILL**
Professor of Computer Science UCA (Polytech Nice department)
Computer scientist in the MAASAI team shared by INRIA and the I3S Laboratory

Abstract. Text data classification is becoming high priority in most of the organizations, finding the possibilities to improve the model performance is the main interest of this paper. "Model explainability" provides some methods to explain the model like why the model routed to a particular decision on a given data and how they are working. Such methods are called *Explainers*. Explainers provides the data(feature probabilities) on how the model taking the decisions and feature weights which helps to choose a good performing model from different alternative models. So, this paper provides the work in the direction of comparing the top 3 *explainers* to improve model performance(in text data analysis category) with different text datasets and explore the insights of the results. The main aim follows, 1) Validate the model performance by using the explainer's output. 2) Validate how the main features are estimated based on the explainer. 3) Extract feature's importance, which are impacting the model performance. 4) Therefore the work is carried on data pre-processing by removing explainer defined stop words (in medical data some words have low priority and others has high priority, but the model prioritizes them based on the occurrences). 5) Re-train the model based on the new pre-processed data. In pre-processing step, removing explainer defined stop words from confusion matrix(True positive, True negative / False positive, False negative) estimated by explainers leads to increase the performance of the model for better classification. This paper is refers to Explainable Artificial Intelligence[1].

Keywords: Explainers · Text embeddings · Feature importance · Confusion matrix · XAI (explainable Artificial intelligence).

1 Introduction

In order to explain the black-box model, we need an *explainer* to understand how the model is predicting the outcome based on the input. The explainers will try to understand the importance of the features, how much they impact the output. thus we can plot the different graphs and charts in user understanding format.

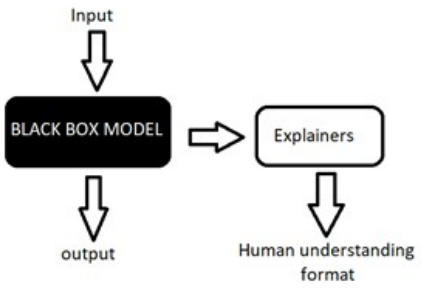

Fig. 1. The model explainability

Usually the model works based on training and assigned weights of the features. Without exploring the assigned feature weights, only relying on the model is not acceptable, for instance considering the medical data analysis. In the figure 1, explainers works with the model while it is categorising the data, providing the information to explain why model made the decision based on the given input[13].

1.1 Types of explainability:

Global explainability: The model gets explained whole, including the training data used to train the model. The algorithm is verified whether used properly or not. The warnings and weakness of the model also will be verified.

Local explainability: The model will be judged on a particular decision for one of the samples in the data. It explains why the model took the decision for that particular input.

1.2 Types of explainers:

Model – agnostic: The tools which can be used for any type of the models for the model explainability.

Model – specific: The tools which help in the explainability of some specific models or some group of models.

2 State of the Art

Explainable AI (XAI): In artificial intelligence, XAI provides the information which results of the solution can be understand by humans. Where the black-box results can not be explained by domain experts, there XAI contributes a lot of

information to understand. XAI algorithms has three principles transparency, interpretability and explainability. Transparency (if the processes that extract model parameters from training data and generate labels from testing data can be described and motivated by the approach designer)[1]. Interpretability (the possibility to comprehend the ML model and to present the underlying basis for decision-making in a way that is understandable to humans)[1]. Explainability (the collection of features of the interpretable domain, that have contributed for a given example to produce a decision)[1].

2.1 Proposed explainers:

In this paper we consider the top known three explainers and they are SHAP, LIME and ELI5. Lets take one full medical text to explain the multiple explainers.
Text: *name -year-old evaluated and examined pediatric oncology clinic today accompanied his mother name was last seen this clinic date for chemotherapy name does not report vomiting after his chemotherapy one day ago and new problems name has been adherent with oral chemotherapy home with prednisone tablets times day* [7]. Lets apply three different explainers on this text.

SHAP This is stands for SHapley Additive exPlanations. SHAP explains the model decisions, by connecting optimal credit allocation with local explanations using the classic Shapley values from game theory[2]. A **game theory** is a decision-making strategy among the competing players(features) for payoffs(importance) as per the contribution in the game (model decision). **Shapley values:** In a game, the shapley value is an average of the marginal contribution across all possible permutations.

Considering that a model gives a prediction to a particular input, by holding some feature importance, the shapley values will be calculated for all features at a particular input. Each feature importance will be calculated by presenting and discarding the feature in the model decisions. In the end, each feature will get it's shapely values, how much exactly is its contribution during the model decision for the particular input.

For example, In the figure 2 Shap calculates the shapley values for all the feature which are represent in the given observation. The strongness of the colour represents, how strongly the feature supports towards the EI class(red colour) or non-EI class(blue).

As SHAP calculates the shapley values for all the significant features so these can be used to present in the local prediction (local interpretability) for a particular input and also representing in the global level impact (global interpretability). Shap value plot helps to represent the global interpretability, where as shap text plot helps in local interpretability. Shap is an agnostic explainer, which supports all kinds of models to examine. There are multiple ways to visualize the model outcome to explain the model decision to the user in human understandable format.

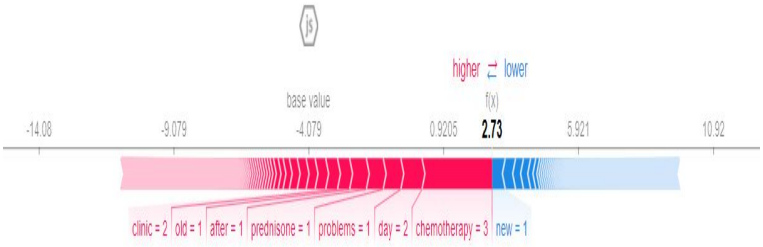

Fig. 2. SHAP explainer's example for an observation from MADE dataset for medical data classification. Where the red colours represent for EI class and blue represents Non-EI class.

LIME: This is stands for Local Interpretable Model-Agnostic Explanations.

LIME can provide a local model interpretation to understand, why the model has proposed the outcome. LIME tweak the input(by presenting and removing) to verify that, whether a outcome is changing or not. Then LIME decide how much that feature is impacting the outcome. Then it presents the results in human understandable format by exposing the black model interpretation.

LIME works more accurately for linear models compared to nonlinear functions, for the local model interpretation. It uses the linear function approximations. Still Lime gives the competitive results in order to explain the non-linear functions. It looks at the single data sample and assumes that is linear then provides the results. It is possible that a assumed linear model by LIME might not be powerful enough to explain, behavior of the original model. Non-linearity at local regions for some datasets, is more complex to explain by applying the LIME. Not being able to apply LIME in these scenario's is a significant pitfall.

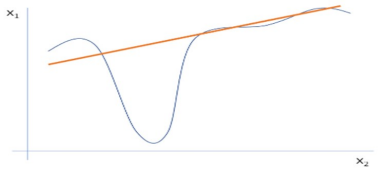

Fig. 3. LIME linear classification - the lime proposed function represented as orange color whereas blue one is actual model function, So at the particular point the lime treats the original function as linear then it evaluates the impacting features at that point[9].

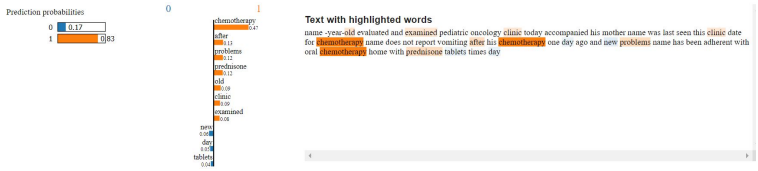

Fig. 4. LIME text classification-the lime identifies the feature which impacts the model prediction, which features impacting which category. Blue color represents the EI category whereas saffron represents the non-EI category. Thickness of the color represents high probability.

As name suggests itself, lime is mainly for local interpretation, it is not carries the potential to present the global interpretability. Lime is also a model-agnostic and it is applicable for majority of the models for explanation.

ELI5: It stands for explain like I am 5. ELI5 uses the *scikit-learn* to understand the text processing and points the words data where ever it needs. It has the several algorithms to explain the black box models. ELI5 has the potential to express all the model weights along with easy interface for using it. Firstly, ELI5 understands the model weights to understand the global performance of the model later this knowledge, ELI5 uses in the analysis for the local inputs.

Fig. 5. ELI5 probability example-it provides the observation in three ways, 1st part represents which category the observation belongs, 2nd represents how much the observation belongs to EI category, 3rd represents how much the observation belongs to non-EI category.

This explainer will give both global and local model interpretations. This is considered as Permutation importance for the global interpretation. ELI5 explainer also tries to find the feature importance to express the model activities. For the local interpretation of the model, ELI5 uses the LIME methodology only. This explainer also a model-agnostic, it has some inbuilt methods to present and explore the different models.

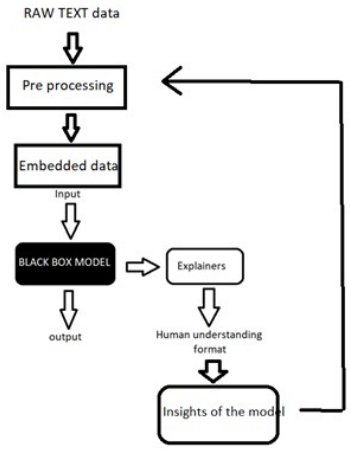

Fig. 6. Proposed architecture.

The proposed work: The explainers are using to validate the model performance based on the explainer's output, but not using the explainer's output to refine the data pre-processing. In this paper, we use explainer's results such as weights and probabilities of important features to train the model. We can consider this is a kind of reverse engineering to get the good performance of the model. As shown in the figure 6, the knowledge from the explainers will be re-used in the training of the model, in such a way of prioritising the most impacting words in the model training and removing the words causes model to choose wrong predictions. Notably we got the good results to explore in this direction of research.

Exploring the highly rated explainers with the medical data, along with word embeddings is the proposed development in this paper.

3 Experiments

Data set:
The experiments carried on mentioned medical data set which is Medical Adverse Drug Reaction(MADE) Data. This is a dataset for Classification if a sentence is ADE-related (True) or not (False) and Relation Extraction between Adverse

Drug Event and Drug. DRUG-AE.rel provides relations between drugs and adverse effects. DRUG-DOSE.rel provides relations between drugs and dosages. ADE-NEG.txt provides all sentences in the ADE corpus that DO NOT contain any drug-related adverse effects[7].

Initial steps of experiments:
We trained the linear model on the given medial data after transforming the text data into bag of words representation. Based on the accuracy, we calculated the confusion matrix which gives us correct number of predictions (True positive, True negative) and wrong number of predictions (False positive and False negative). We know that the words present in the true positive and true negative observations are the cause of rights decisions of model predictions. Same as the words with the high probabilities present in the false negative and false positive are the words causing the model to choose wrong predictions.

Every explainer takes the model and observations as inputs and estimates the feature importance and finds the probabilities of features. If we consider one observation from the data set, the figure:7 or figure :8 represents probabilities of the words for an observation. Where as the numerical number defines the strength of the feature to promote to the class and sign(-,+) represents which class it supports. So, after summing the total values of features(number of features can be defined to a explainer) if the value is in positive then the observation belongs to one class or if the value is in negative then the observation belongs to another class.

Though the model giving the good accuracy, in the context of medical data we need more accuracy. So, our intention is to find the better accuracy. We analysed the model decisioning for particular observations through multiple explainers, how the model is taking the decision, which features are getting high importance, how much probabilities are assigning to features.

3.1 Experiment-1

In this experiment, the "explainer defined stop words" are extracted from False positive and False Negative observations. the wrongly predicted observations are given to explainers and extracted high probable words along with probabilities. In those words we extracted above 0 probability from FP and below 0 probability words from FN. This implies that one of the strong category of words which caused the wrong prediction. this experiment can be replicated by choosing the other category of words.

As shown in the figure:7 and figure:8 the green boxed words only will be added to explainer defined stop words. why not all, if we add all the words from from FP and FN then there is a chance of normalizing the threshold which has no impact on the model accuracy. Considering this point, we chose one category from FN and another category from FP.

After combining all explainer defined stop words, we removed those words batch wise in the model training and verified the accuracy.

False Positives

Wrong prediction 1		Wrong prediction 2		.....	Wrong prediction n	
Feature	**Probability**	**Feature**	**Probability**		**Feature**	**Probability**
Skin	0.4356	small	0.2356		through	0.5626
Problems	0.2532	general	0.2212		was	0.3252
After	0.2109	warm	0.2209		made	0.3008
Old	0.0014	are	0.1915	.....	monitored	0.1053
Clinic	0.0001	Neck	0.1345		Consultation	0.1423
Examined	-0.0222	begin	-0.0122		during	-0.1534
new	-0.1345	heent	-0.1345		did	-0.1953
day	-0.1586	heart	-0.1580		level	-0.2134
prednisone	-0.2546	enema	-0.1643		urinary	-0.2506
evaluated	-0.2156	rash	-0.1950		he	-0.3166

Fig. 7. The boxed words from the false positive observations added to the explainer defined stop words(after sorting).

False Negative

Wrong prediction 1		Wrong prediction 2		.....	Wrong prediction n	
soft	0.3286	hours	0.2876,		refills	0.3886,
nontender	0.3178	axillary	0.2598,		today	0.3098,
and	0.2811	doxazosin	0.2145,		edema	0.2995,
weight	0.1987	protonix	0.1987,		wearing	0.1256,
good	0.0889	lasix	0.0087	.....	difficult	0.0089
bowel	-0.1003,	levoxyl	-0.0345,		right	-0.1045,
sounds	-0.1456,	cervical	-0.0954,		leg	-0.1350,
rate	-0.2581,	neurontin	-0.1945,		stable	-0.1905,
palpable	-0.2656,	received	-0.2456,		blood	-0.1996,
every	-0.2988,	his	-0.3452,		treat	-0.2492,

Fig. 8. The boxed words from the false negative observations added to the explainer defined stop words(after sorting).

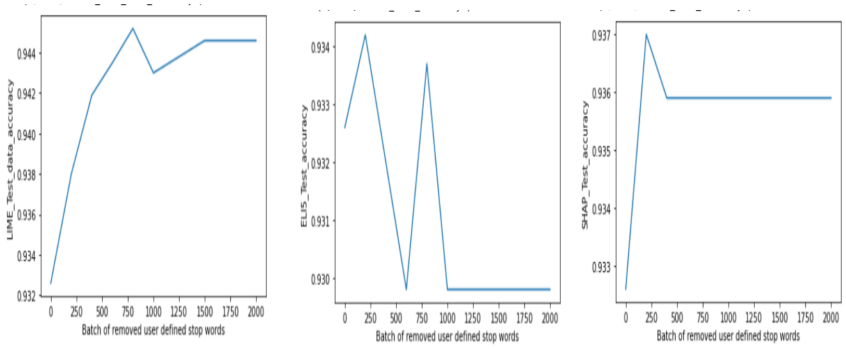

Fig. 9. the first graph represent the accuracy of test data after removing batch of stop words given by LIME explainer in pre-processing, same as 2nd and 3rd represents for ELI5 and SHAP explainers

In Figure:9,. Considerably, the lime shows increasing accuracy with compared to other explainers such as SHAP and ELI5. In that too, SHAP also showing some better accuracy progress.

3.2 Experiment-2

In this experiment, the words are extracted from the False positive and False negative parts of the model's confusion matrix. The main difference in this experiment is the explainer defined words are removed in iterative process until the there are no explainer defined stop words to remove. Considering the time complexity and computation power, the iteration process repeated until 500 words to remove.

After deep analysis of the experiment results, the accuracy is not granular observable, but precision and recall are in incremental state. One main observation is the counts FP for train and test is gradually decreasing, where as the counts of FN in train data is increasing. in the comparison of FN counts and FP counts, the model mis-classifying highly in FN category compared to FP.

The main intention of this experiment is to propose a new method to reduce a certain category of confusion matrix from the model predictions. For example in some cases we might have a requirement such as, the model should not predict wrongly in FN though it is acceptable in FP. IN such cases we need a strategy to reduce the possibility of model mis-classifying in FN category by doing this kind of experiments.

.

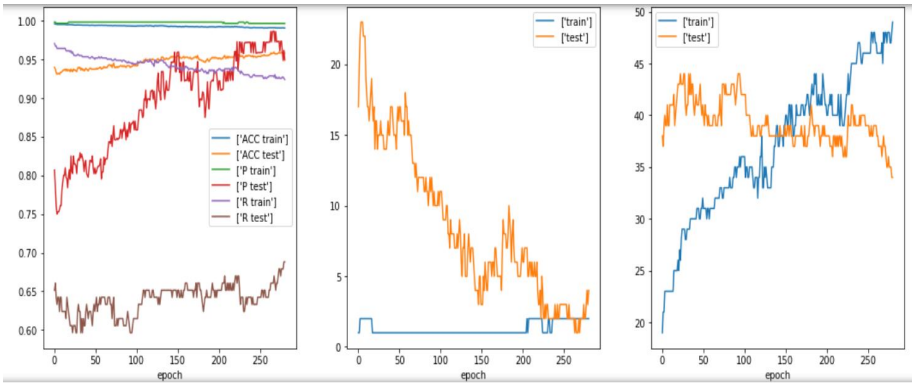

Fig. 10. Removed 250 words of explainer defined stop words - The first graphs represent the accuracy, precision and Recall of train and test data, 2nd and 3rd represents the observation's counts in FP and FN in train and test data respectively.

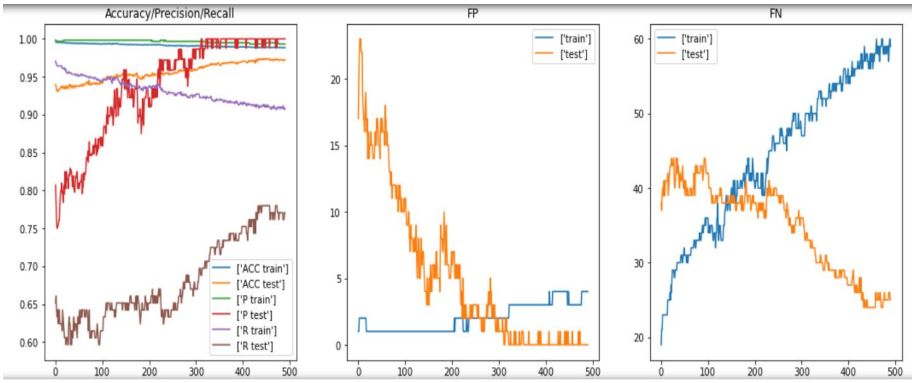

Fig. 11. Removed 500 words of explainer defined stop words - The first graphs represent the accuracy, precision and Recall of train and test data, 2nd and 3rd represents the observation's counts in FP and FN in train and test data respectively.

4 Conclusion

Explainers are playing crucial role in the model explainability to expose the block box model. By using the explainers outputs, we showcase that, there are multiple experiment possibilities to increase the model accuracy. we strongly believe that defining stop words by explainers is also a experimental direction to increase the accuracy of medical data into next level. As shown in the experiments, one of our experiments, shows that there is a accuracy increments after removing the explainer defined stop words. The strategy of defining explainer's plays a critical step in the experiments. We strongly recommend to choose proper explainer and strategy based on the type and size of the dataset to define explainer defined stop words.

5 Future Directions

As we experimented two possibilities out of many, there are other strategies also can be applied as follows based on the data.

1. We can train the model by just extracting the strong positive words, which can be extracted from the True positive observations passed through explainers and strong negative words from True negative observations passed through explainers.
2. We can experiment by just choosing the alternate to the experiment-1. which is positive values from false positive and negative values from false negative to the explainers stop words list to experiment.
3. If the requirements, there should be no False negative category in the model prediction then we can extract the explainers defined words just from false negative category, then we can minimize the number of false negative category of model predictions.

References

1. Explainable artificial intelligence - https://en.wikipedia.org/wiki/Explainable_artificial_intelligence 24 November 2021.
2. Welcome to the SHAP documentation, https://shap.readthedocs.io/en/latest/index.html 2018
3. Lucile Ter-Minassian, Shapley Values: Model-Agnostic Local Explanation Models From a Statistical Viewpoint II, Jul 07 2021
4. Local Interpretable Model-Agnostic Explanations (lime), https://lime-ml.readthedocs.io/en/latest/ 2016
5. Abhishek Sharma, Decrypting your Machine Learning model using LIME Nov 4, 2018
6. Mikhail Korobov, Konstantin Lopuhin , Welcome to ELI5's documentation! 2016-2017
7. Abhyuday Jagannatha 1, Feifan Liu 2, Weisong Liu 3 4, Hong Yu, Overview of the First Natural Language Processing Challenge for Extracting Medication, Indication, and Adverse Drug Events from Electronic Health Record Notes (MADE 1.0) Jan 2019
8. marcotcr, lime, https://github.com/marcotcr/lime -2020
9. Lars Hulstaert , https://towardsdatascience.com/understanding-model-predictions-with-lime-a582fdff3a3b -Jul 11, 2018
10. Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, https://arxiv.org/abs/1602.04938 -9 Aug 2016
11. Scott Lundberg, Su-In Lee, https://arxiv.org/abs/1705.07874 - 25 Nov 2017
12. Mikhail Korobov, Konstantin Lopuhin , Welcome to ELI5's documentation! 2016-2017
13. Collaris, D., van Wijk, J.J. Comparative evaluation of contribution-value plots for machine learning understanding. J Vis 25, 47–57 (2022). https://doi.org/10.1007/s12650-021-00776-w

14. P. Rasouli and I. C. Yu, "EXPLAN: Explaining Black-box Classifiers using Adaptive Neighborhood Generation," 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1-9, doi: 10.1109/IJCNN48605.2020.9206710.
15. TY - BOOK AU - Bataineh, Mohammad Moe PY - 2019/07/17 SP - T1 - Feature Impact for Prediction Explanation