CONTENTS

Using Landmarks for Explaining Entity Matching Models

Andrea Baraldi, Francesco Del Buono, Matteo Paganelli, Francesco Guerra
UNIMORE-DIEF
Modena, Italy
firstname.lastname@unimore.it,andrea.baraldi96@unimore.it

ABSTRACT

The state of the art approaches for performing Entity Matching (EM) rely on machine & deep learning models for inferring pairs of matching / non-matching entities. Although the experimental evaluations demonstrate that these approaches are effective, their adoption in real scenarios is limited by the fact that they are difficult to interpret. Explainable AI systems have been recently proposed for complementing deep learning approaches. Their application to the scenario offered by EM is still new and requires to address the specificity of this task, characterized by particular dataset schemas, describing a pair of entities, and imbalanced classes.

This paper introduces Landmark Explanation, a generic and extensible framework that extends the capabilities of a post-hoc perturbation-based explainer over the EM scenario. Landmark Explanation generates perturbations that take advantage of the particular schemas of the EM datasets, thus generating explanations more accurate and more *interesting* for the users than the ones generated by competing approaches.

1 INTRODUCTION

Despite the effort put in the past 30 years, Entity Matching (EM), the task that identifies data items that refer to the same real-world entity, is still an open challenge. State of the art approaches (e.g., DeepER [7], DeepMatcher [12], DITTO [10], and many others [2, 19]), based on Machine Learning (ML) and Deep Learning (DL) models, have been demonstrated to be effective in the experimental datasets. Nevertheless, their adoption in real business scenarios is hampered by several factors, including the need for large amounts of training data, the need for expert users for the configuration of their hyper-parameters and the inability to easily interpret how the models make their decisions.

Explaining the behavior of ML and DL models is now a challenging research topic [5]. Its application to EM could facilitate the adoption of EM techniques in business scenarios. An improved ability to interpret the models would increase 1) user confidence in the adoption of ML and DL techniques, 2) the ability to debug erroneous behaviors and diagnose unexpected results, and 3) improve the functionality of the approaches. Moreover, it would decrease the need for domain experts to evaluate the effectiveness of EM approaches, task that is typically executed through manual, expensive, and time-consuming processes.

Although several explanation systems have already been proposed in the literature (e.g., LIME [14], Shapley [8], Anchor [15], and Skater[1]) , their application to EM tasks is not straightforward and only few approaches have partially addressed it [4, 6, 11, 17].

[1]https://github.com/oracle/Skater

left_name	right_name	left_description	right_description	left_price	right_price
sony digital camera with lens kit dslra200w	nikon digital camera leather case 5811	sony alpha digital slr camera with lens kit dslra200w 10.2 megapixels	leather black	849.99	7.99

Figure 1: Example of EM record to explain.

The main motivation is that EM is conceived by ML and DL systems as a binary classification problem, where the class shows if the pairs of entities described in the dataset records are matching, and the dataset entry is composed of pairs of attributes describing the same feature of different entities. This structure is "unusual" in ML and DL, where the records conversely describe single evidence. Moreover, the datasets are usually imbalanced: the number of records belonging to the matching class is far less than the non matching ones. Finally, the attributes describing the same features of different entities have close statistical distributions (or close word distributions in case of categorical attributes) even when they refer to different entities.

In this paper, we present Landmark Explanation a system for explaining EM model predictions, that extends the capabilities of a post-hoc perturbation-based local explainer over this specific scenario. Post-hoc perturbation-based explainers analyze the records to explain and build a surrogate linear model where the features are the tokens (e.g., the words in case of textual attributes) composing the attribute values. The explanation is directly generated from the surrogate model: its linear coefficients represent the importance of the tokens. This model is trained with synthetic data, generated via a two-step approach where the values of the records to explain are properly altered (in the perturbation phase) and then passed to the original model to get their class (in the reconstruction phase).

Example 1.1. Figure 1 shows an example of a record describing a pair of entities. A suffix is added to the attribute names to show which entity they are describing. The application of a DL based model to the record (e.g. DeepMatcher) let us know that the entities in the record do not refer to the same real-world entity. This is evident for a human person, being clear from the attributes that the left entity is a digital camera and the right entity is a leather case. But the EM model is not able to explain the reasons for its choice. This is the task for an explainer which takes the record to explain, transforms it into tokens (we create a token for each space-separated term), generates a number of perturbations (typically performed by casually dropping tokens), passes each perturbation to the model to get the class, and exploits the so generated dataset to train a linear model. The coefficients associated to the linear model are the ones explaining the behavior of the EM model on the target record. The tokens `sony`, `lens`, `dsira200w` can be considered by an explainer as an evidence of the fact that the record is describing a non-matching entity.

The direct application of a perturbation mechanism based on token removals is not effective for the dataset used in EM. The reason is that removing random tokens is likely to affect both the

entities represented by the dataset item. The generated synthetic records may then contain *null perturbations* where the same tokens referring to the different entities are removed. Moreover, since the EM datasets are largely imbalanced, perturbations frequently lead to records belonging to the non-matching class. To solve this issue, Mojito [4] introduces the "COPY" perturbation mechanism, where attribute values describing one of the entities in a record are substituted to the corresponding attribute values of the second entity. The aim is to introduce a perturbation that increases the match probability between pairs of entities. The aim is to create records representing matching entities. But duplicating entire attribute values does not allow the approach to discriminate among the tokens that, thanks to the copy, will provide the same contribution in the explanation.

Landmark Explanation addresses these issues by introducing two main innovations. The first is the generation of two explanations for each dataset entry, each one explaining the model decision from the perspective of one of the two entities described in the record. These explanations are generated by selecting one of the entities constituting a dataset entry in turn as a landmark. The landmark is preserved from the perturbation, which is subjected to the other entity (the varying entity). The second is a mechanism for computing explanations for records belonging to non-matching classes. Before the perturbation, we inject additional tokens extracted from the landmark entity into the varying entity. The perturbation of the varying entity with injected tokens produces a set of synthetic entities. These will all be concatenated with the landmark entity to generate the synthetic EM dataset used to train the surrogate model. The idea is to contrast the asymmetric nature of the problem: an explanation of a matching pair is always composed of "interesting" tokens since they express the reason why the entities have been considered as matching. The same does not happen for non-matching entities, since non-matching entities have many reasons to be different. So it is difficult to generate an explanation with interesting tokens for non-matching pairs. Therefore the problem is to generate the most interesting explanations. These are the ones involving tokens from one entity that if used to describe the second entity would have brought the EM model to classify the record as matching. Thanks to the injection described above non-matching pairs are pushed to be match and the resulting explanation will be more interesting.

Example 1.2. To explain the inference of an EM model applied to the record in Figure 1 , Landmark Explanation generates two explanations. The top 3 tokens generated by the explanation with the left entity as a landmark are *leather*, *nikon* and *5811*. These tokens are the ones that best differentiate entities. This means that if the left entity were described by these tokens, the record would probably be classified in the matching class by the EM model. The top 3 tokens generated by the second explanation with the right entity as a landmark are *dsira200w*, *lens* and *849.99*.

We evaluate Landmark Explanation coupled with LIME. The results of the experiments show that the explanations generated outperform the ones of the competing approaches in accuracy and "interest" for the users.

Summarizing, the main contributions of this paper are: (1) the introduction of Landmark Explanation, a tool that extends the capability of a generic post-hoc perturbation-based explainer to generate accurate local explanations of EM models; (2) the realization of an extensive experimentation of Landmark Explanation coupled with the LIME explainer [14] to demonstrate the effectiveness and the quality of the approach in comparison with different competitor systems.

The rest of the paper is organized as follows. Section 2 introduces some related work. Section 3 introduces our approach that is evaluated in Section 4. Finally, in Section 5 we sketch out conclusion and future work.

2 RELATED WORK

Explaining AI. The interpretation of machine learning techniques represents a hot topic and two main approaches for its resolution can be identified [5]. On the one hand, there are intrinsically interpretable models, such as decision trees, rule-based and linear models, which rely on structures that can be directly interpreted by humans. On the other hand, there are techniques that analyze the behavior of black-box machine learning methods via a second intermediate model built from the first. These post-hoc interpretation methods are model-agnostic (i.e. they are applicable to any ML / DL model), however they provide less faithful explanations than intrinsically interpretable models.

Regardless of the explanation technique adopted, it is further possible to distinguish between global and local interpretations [5]. In the first case, the entire functioning of an ML / DL model is examined, while in the second its behavior is studied only locally (i.e. by explaining its logic on individual predictions).

The main exponent of the category of local post-hoc interpretation techniques is LIME [14], which exploits an interpretable linear surrogate model (e.g. Lasso) to evaluate the behavior of the original model in the neighborhood of a specific data instance. It will be used in our experiments, and an extension of it is Anchor [15], which generates explanations based on if-then rules. Some examples of global explanation systems are BRL [9] and Skater[2]. Similar techniques are *permutation feature importance* and *drop-column importance* [1], which can be used to detect the global relevance of features in any model.

In this paper we focus exclusively on local post-hoc interpretation techniques (for simplicity in the rest of the paper they will also be identified as generic "explanation systems") and we propose their adaptation, through Landmark Explanation, to the Entity Matching problem.

Explainable Entity Matching. Entity matching, that is the task that identifies the records that refer to the same real-world entity in multiple datasets, represents one of the main steps of data integration and has been under study for several years. Many techniques have been proposed: from the more traditional rule-based approaches to the most modern machine learning and deep learning methods. Some examples of the first category are [16, 18]. They are intrinsically interpretable, however, the identification of the most effective set of matching rules is a complex and non-trivial task [13].

Recently, several approaches based on Deep Learning have proved particularly effective in solving this task. Some examples are DeepER [7], DeepMatcher [12], DITTO [10] and many others [2, 19]. In addition to requiring a significant amount of annotated data and a complex configuration, the main problem with these systems is the inability to interpret their behavior, affecting their usability in business environments [3].

This motivated the realization of several studies on the use of interpretation techniques in the entity matching area [11, 17], and

[2]https://github.com/oracle/Skater

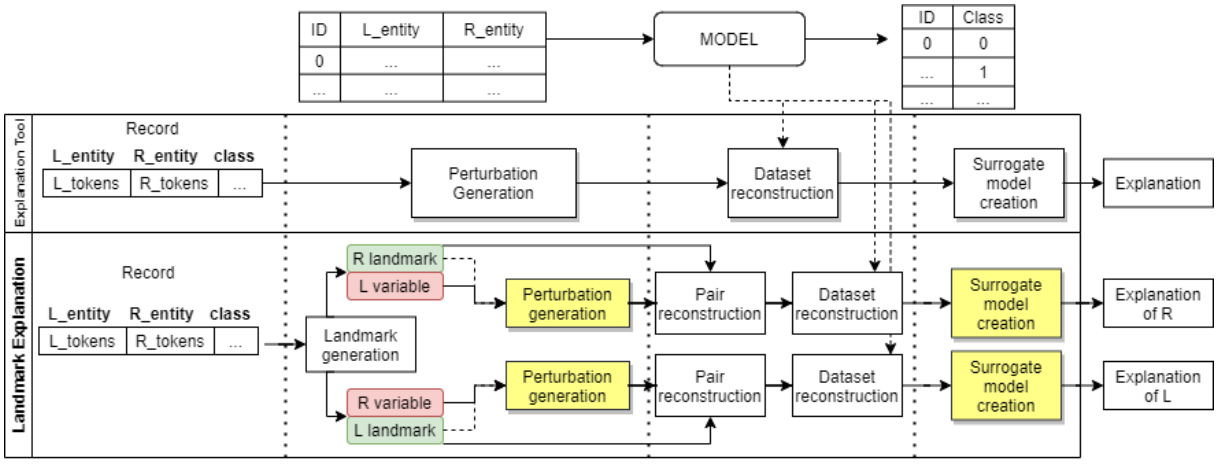

Figure 2: A generic post-hoc perturbation based explanation system (at the top of the image) compared with its extension with the Landmark Explanation **Framework (at the bottom).**

tools, like Mojito [4] and Explainer [6], have been proposed. ExplainER provides a unified interface for applying well-known interpretation techniques (e.g., LIME, Shapley, Anchor, and Skater) in the EM scenario. Mojito adapts LIME for the explanation of single EM predictions and represents the work closer to our approach. It extends LIME in two ways: 1) it exploits the subdivision of EM data into attributes, 2) it introduces a new form of data perturbation, called LIME-COPY[3], which allows generating match elements starting from non-match elements. Unlike Landmark Explanation, Mojito treats attributes atomically, distributing its impact equally to its constituent tokens. Furthermore, Landmark Explanation analyzes the diversified impact that the same token can generate depending on the entity considered as a landmark for the explanation.

3 THE LANDMARK EXPLANATION APPROACH

Landmark Explanation is a generic and extensible framework that can extend a generic local post-hoc and model-agnostic perturbation based explanation systems to the interpretation of EM model predictions. The main assumption of these generic systems is that the prediction of a model computed on a given instance can be approximated by a linear function of the predictions calculated in the neighborhood of that input instance. Their functional architecture is shown at the top of Figure 2 and can be schematized in 3 main blocks: the component for the *Perturbation generation*, for the *Dataset reconstruction* and for the *Surrogate Model Creation*.

The *Perturbation generation* component takes the record to analyze as input and generates a number of perturbations from it. These perturbations constitute a new features space, defined in the neighborhood of the record, which, once integrated by the *Dataset reconstruction* component with the predictions of the original model, can be used by the *Surrogate Model creation* component to infer its behavior in the locality of the record.

Landmark Explanation extends the generic explanation system to improve its effectiveness on datasets representing EMs. In particular, as shown at the bottom, of Figure 2, Landmark Explanation adds the *Landmark generation* component before the perturbation. This component generates for the record to explain two representations, where only one of the two entities composing the item will be subject to the perturbation and the other one will be kept fixed as a landmark. This constitutes the input for the *Perturbation generation* component that will be called twice, once for each representation. In this way, each perturbation obtained with this process will involve the attributes of one entity. The tokens of the second entity (the landmark entity) will be added by the *Pair reconstruction* component before the *Dataset reconstruction*. This allows Landmark Explanation to perturb the information of one entity at a time while preserving the pairwise structure of the EM data. A perturbation of the input entity pair is generated by varying only one of the two entities while preserving the pairwise structure of the EM data.

Note that the component performs differently when the record to explain is referring to a non-matching class. In this case, the tokens of both entities are concatenated into the varying entity and passed to the *Perturbation generation* component. The behavior of the *Pair reconstruction* component does not change, by concatenating after the perturbation the contribution of the landmark entity. This mechanism has been implemented to contrast the dataset imbalance and to generate explanations that can be more interesting for the users since based on a richer set of tokens. Finally, as for the generic explanation system, the synthetic dataset just created is used to train a linear model whose coefficients constitutes the explanation. These coefficients can be positive or negative thus indicating which tokens should be added (the one with a positive score) and which should be removed (the one with negative score) to create a description that is closed to the reference entity.

The yellow-shadowed components in the Figure are the ones provided by the explanation system we are extending. In our experiments, these components are provided by the LIME explainer. Since Landmark Explanation conceives them as black box modules, other explanations systems can be easily coupled with our approach.

3.1 Landmark Explanation components

Landmark Explanation is composed of three main components for the *Landmark generation*, the *Pairs reconstruction* and the *Dataset reconstruction*.

Landmark generation component. The goal of this component is to generate input for the *Perturbation generation* component. A

[3]In Section 4 we refer to this technique as Mojito Copy to emphasize that this technique is part of the Mojito tool.

Dataset	Type	Datasets	Size	% Match
S-BR	Structured	BeerAdvo-RateBeer	450	15.11
S-IA		iTunes-Amazon	539	24.49
S-FZ		Fodors-Zagats	946	11.63
S-DA		DBLP-ACM	12,363	17.96
S-DG		DBLP-GoogleScholar	28,707	18.63
S-AG		Amazon-Google	11,460	10.18
S-WA		Walmart-Amazon	10,242	9.39
T-AB	Textual	Abt-Buy	9,575	10.74
D-IA	Dirty	iTunes-Amazon	539	24.49
D-DA		DBLP-ACM	12,363	17.96
D-DG		DBLP-GoogleScholar	28,707	18.63
D-WA		Walmart-Amazon	10,242	9.39

Table 1: Magellan Benchmark

Tokenizer is firstly needed to transform the dataset entry in a format suitable for generating meaningful explanation. We implement a tokenization mechanism similar to the one adopted in other systems as Mojito [4] that preserves the structure of the pair of entities described in the record. A token is generated for each space-separated term in the attribute values. A prefix is introduced to each token to indicate the attribute where the original value is located in the entity schema. The prefix enumerates the tokens, to manage multiple occurrences of the same word in an attribute value.

After the tokenization, Landmark Explanation implements two mechanisms for performing this task. With the *single-entity generation*, the tokens composing the entities are separated and the perturbation component is called twice, each time with the element of a different entity. The output for each execution are the tokens of one entity (the landmark) and a number of perturbations for the second entity. This technique generates a perturbation that highlights the differences of one entity with respect to the other. It is then particularly effective when the record to explain is belonging to the matching class.

With the *double-entity generation*, the perturbation component receives as input the tokens of an artificial entity created by concatenating, for each attribute, the tokens of both the entities. The output for each execution are then the tokens of one entity (the landmark) and a number of perturbations of the artificial entity created by the concatenation. The idea of this technique is to generate the perturbation of a more extensive set of tokens (obtained by the union of the tokens of both the entities) that is effective for generating explanations for records classified as non-matching items.

Pair reconstruction component. The component receives as input the landmark entity and a number of perturbations of the tokens of the varying entity and "reconstructs" the corresponding pairs of entities (one for each perturbation). The prefixes introduced by the *Tokenizer* are exploited for this purpose and removed from the generated records.

Dataset reconstruction component. This component generates the synthetic dataset to be used for training the surrogate linear model. This is obtained by passing each pair of entities reconstructed by the previous component to the original EM model for getting the predicted class.

4 EXPERIMENTAL EVALUATION

We evaluated the explanations generated by Landmark Explanation according to two main perspectives: their reliability in representing the EM Model (in Section 4.2) and the "quality" of the explanation provided (in Section 4.3).

(a) Matching label.

	Single		Double		LIME	
	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE
S-BR	**0.923**	**0.121**	0.796	0.136	0.830	0.147
S-IA	**0.940**	**0.226**	0.793	0.251	0.847	0.240
S-FZ	**0.934**	**0.228**	0.841	0.237	0.865	0.236
S-DA	0.887	0.171	**0.894**	**0.164**	0.573	0.337
S-DG	**0.836**	**0.196**	0.823	**0.196**	0.757	0.200
S-AG	0.896	**0.074**	**0.903**	0.112	0.698	0.148
S-WA	**0.954**	**0.071**	0.928	0.115	0.659	0.228
T-AB	**0.908**	**0.066**	0.854	0.146	0.758	0.118
D-IA	0.899	**0.090**	**0.975**	0.112	0.780	0.156
D-DA	0.942	**0.030**	**0.979**	0.041	0.940	**0.025**
D-D	0.929	**0.107**	**0.963**	0.152	0.891	0.115
D-WA	**0.916**	**0.045**	0.901	0.090	0.813	0.074

(b) Non-matching label.

	Single		Double		LIME		Mojito Copy	
	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE
S-BR	0.747	0.092	**0.927**	**0.037**	0.843	0.100	0.011	0.369
S-IA	0.669	0.248	**0.736**	**0.127**	0.624	0.267	0.022	0.569
S-FZ	0.811	0.188	0.853	**0.134**	**0.953**	0.189	0.032	0.681
S-DA	0.975	**0.021**	0.590	0.287	**0.985**	0.066	0.005	0.574
S-DG	0.895	**0.086**	0.660	0.306	**0.935**	0.107	0.005	0.504
S-AG	0.835	0.107	0.895	**0.056**	**0.905**	0.097	0.010	0.445
S-WA	**0.990**	**0.028**	0.955	0.217	0.890	0.352	0.000	0.746
T-AB	**0.860**	0.076	0.680	**0.047**	0.795	0.092	0.045	0.328
D-IA	**0.874**	**0.019**	0.291	0.070	0.390	0.129	0.242	0.191
D-DA	0.615	0.071	0.300	**0.027**	**0.690**	0.036	0.010	0.173
D-DG	0.540	0.305	0.375	**0.118**	**0.640**	0.235	0.040	0.437
D-WA	0.500	0.184	**0.785**	**0.078**	0.500	0.192	0.005	0.380

Table 2: Token-based evaluation.

4.1 Experimental setup

We run the experiments on a VM deployed on Google Cloud with 12 GB of RAM, GPU K80, and Intel(R) Xeon(R) CPU @ 2.30GHz.

Dataset and Model. The EM model explained in the experiments is a Logistic Regression Classifier. We experimented Landmark Explanation against the datasets provided by the Magellan library[4] which is considered as a standard benchmark for the evaluation of EM tasks. The datasets are listed in Table1, where the size and the percentage of records representing matching entities are shown. The records in all datasets represent pairs of entities described with the same attributes. A label is provided to express if the record represents a matching / non-matching pair of entities. In the experiments, we sampled 100 records per label and we computed their explanations. Note that all records are sampled when the dataset contains less than 100 records (see for example the dataset S-BR which contains only 68 records labeled as matching entity).

4.2 Reliability of the explanations

The goal of the experiment is to evaluate the reliability of the explanations generated by Landmark Explanation in interpreting the behavior of an EM model through single predictions. An explanation is considered reliable if it is able to consistently recognize the importance of the features with the EM model. To evaluate this, we performed two kinds of experiments, one analyzing the weights assigned by Landmark Explanation to the *tokens* it generates, the second the weights assigned by the EM model to the dataset *attributes*.

4.2.1 Token-based evaluation. Through this first kind of experiment, we evaluate if the weights assigned by Landmark Explanation to the tokens generate a surrogate model consistent with the EM model. We performed an experiment that is similar

[4]https://github.com/anhaidgroup/deepmatcher/blob/master/Datasets.md

to the one proposed in the evaluation of LIME: 25% of tokens are randomly selected and removed from the record to explain, defining a new item. We then compared the probability score obtained passing the new item to the EM model with the one of the original record, where we have subtracted the sum of the coefficients associated with the removed tokens. If the explanation model correctly represents the EM model these two values should be close. We repeated the experiment 100 times for each class (see the beginning of Section 4), and we measured the performance obtained by means of two metrics: the mean average error (MAE) and the accuracy on the predicted class. We performed the experiments for all datasets and testing all techniques for generating the perturbations as reported in Table 2. Note that column LIME shows the results obtained with LIME / Mojito Drop[5] with the same setting. Non-matching settings also include a comparison with the Mojito Copy technique, which has been designed for this kind of record.

Discussion. Table 2a shows that Landmark Explanation, applied to records labeled as matching entity, performs better than LIME in the datasets when the perturbation is generated with the single-entity technique (it obtains better accuracy in all datasets and low MAE in 11/12 datasets). The double-entity generation technique performs slightly worse: in 9/12 it obtains better accuracy and in 6/12 lower MAE). Nevertheless, the scores, when worst, are very close to LIME. Note that in some datasets there is some small contradiction between accuracy and MAE scores computed for the same dataset. For example, we observe that Landmark Explanation applied to the S-BR dataset with the double-generation configuration has a better MAE score than LIME/Mojito Drop. This is not the same for the accuracy value, where LIME performs better. This is motivated by the fact that the probability scores generated by the model are close to the decision threshold (fixed to 0.5). Then, small fluctuations in the surrogate model can generate mismatches in the class predicted by the explained model for the records, even if the EM model and the surrogate model are very close. Table 2b shows the accuracy and the MAE obtained analyzing records referring to non-matching labels. In this scenario, the double entity perturbation obtains the best scores with an accuracy better than LIME/Mojito Drop in 4/12 datasets and a lower MAE in 10/12 datasets. The reason is due to the effect of the duplicated tokens inserted in the dataset used for training the surrogate model. These tokens, being similar to the ones of the entities described in the record, push the EM model to classify the record towards a matching label even in the case of an imbalanced dataset. By using the same tokens, the entities will likely be considered by the model as similar. Conversely, the perturbations generated by LIME / Mojito drop are subsets of the original record, which, in this case, was classified as non-match. By removing tokens from descriptions of entities classified as non-matching, the probability score of the EM model usually decreases and it is unlikely to obtain descriptions of entities that a classifier evaluates as matching. If we pushed the decision threshold to 0.4 (instead of 0.5), Landmark Explanation would obtain a better performance than LIME/Mojito drop in 10/12 datasets.

Note that the copying technique introduced by Mojito to manage records associated with non-matching labels does not show high performance. The reason is that Mojito generates a perturbation by duplicating entire attributes.

The result of this operation is that the tokens of the replaced attribute have the same weights, thus decreasing the performance.

[5]the Mojito Drop technique implements the LIME approach

(a) Matching label.

	Single	Double	LIME
S-BR	1.000	1.000	1.000
S-IA	0.261	**0.538**	0.495
S-FZ	**0.592**	**0.592**	0.143
S-DA	**0.520**	**0.520**	0.200
S-DG	1.000	1.000	1.000
S-AG	**1.000**	0.545	0.545
S-WA	0.901	**1.000**	0.544
T-AB	0.545	0.545	0.545
D-IA	0.892	**0.939**	0.848
D-DA	1.000	1.000	1.000
D-DG	1.000	1.000	1.000
D-WA	0.526	**0.681**	0.526

(b) Non-matching label.

	Single	Double	LIME	Mojito Copy
S-BR	0.733	**1.000**	**1.000**	**1.000**
S-IA	0.312	0.538	0.687	**0.756**
S-FZ	0.333	0.518	**0.864**	0.414
S-DA	0.200	**1.000**	0.200	0.520
S-DG	**1.000**	0.520	**1.000**	0.333
S-AG	**1.000**	0.545	0.545	0.545
S-WA	0.573	**1.000**	0.872	**1.000**
T-AB	0.545	0.545	0.545	**1.000**
D-IA	0.925	0.899	0.776	**0.939**
D-DA	0.813	**1.000**	**1.000**	**1.000**
D-DG	**1.000**	**1.000**	**1.000**	0.813
D-WA	0.681	0.681	0.681	0.681

Table 3: Attribute-based evaluation (weighted Kendall measure applied on the ranked list of attributed as generated by the EM and the surrogate model).

Lesson learned. The surrogate model built by Landmark Explanation with the single-entity perturbation is an accurate representation of the EM model for records representing matching pairs of entities. The model built with the double-entity perturbation is an accurate representation of the EM model for record representing non-matching pairs of entities.

4.2.2 Attribute-based evaluation. The attribute-based evaluation proceeds in the opposite direction: it starts from the internal structure of the EM model and evaluates if the weights it gives to the attributes are close to the ones we can derive from the tokens obtained by Landmark Explanation. For this reason, we have analyzed the weights given to the dataset attributes by the Logistic Regression model used as EM model in the experiments and ranked the attributes according to their absolute values. We have done a similar operation with the surrogate model, where the weights of the attributes have been computed by summing the absolute weights of their composing tokens. The idea is that the order of the attributes computed on the basis of their weights should be the same in both models. In Table 3, we measured the correlation computed by applying the weighted Kendall tau correlation measure, between the ranked list of attributes of the EM and surrogate model.

Discussion. Table 3a shows the experiments on the records representing matching entity pairs. The correlation scores achieved by Landmark Explanation with the double-entity perturbation approach are better or equal to the ones achieved by LIME/Mojito for all dataset. Table 3b shows the experiments on the records representing non-matching entity pairs. In this case, the single-entity configuration obtained better/equal results than LIME/Mojito Drop in 7/12 datasets (4/12 against Mojito Copy); the double-entity configuration obtained better/equal results than LIME/Mojito drop in 9/12 datasets (the same against Mojito Copy). Note that Mojito Copy, that has been explicitly designed for non-matching entities, performs better than LIME/Mojito Drop in 5/12 datasets only and equal/close to LIME/Mojito Drop in 4/12 datasets and worst in the remaining 3 datasets.

Lesson learned. Landmark Explanation creates surrogate models that maintain a relative importance of the attributes similar to the ones of the EM model to explain.

4.3 Quality of the explanations

To introduce this experiment, let us consider an application that aims to provide the explanation for a record labeled as a non-matching entity. The tokens of non-matching are "less polarized":

(a) Matching label.

	Single	Double	LIME
S-BR	0.643	0.593	**0.686**
S-IA	0.652	0.404	**0.702**
S-FZ	0.606	0.447	**0.612**
S-DA	**1.000**	0.940	0.965
S-DG	0.660	0.610	**0.925**
S-AG	0.955	0.800	**0.990**
S-WA	**1.000**	0.785	0.870
T-AB	0.985	0.575	**0.995**
D-IA	**0.561**	0.278	0.311
D-DA	0.695	0.715	**0.800**
D-DG	0.635	0.530	**0.735**
D-WA	**0.915**	0.545	0.880

(b) Non-matching label.

	Single	Double	LIME	Mojito Copy
S-BR	0.298	**0.927**	0.331	0.011
S-IA	0.545	**0.736**	0.393	0.000
S-FZ	0.079	**0.853**	0.047	0.000
S-DA	0.000	0.030	0.000	0.005
S-DG	0.020	**0.545**	0.020	0.000
S-AG	0.075	**0.895**	0.070	0.010
S-WA	0.015	**0.955**	0.000	0.000
T-AB	0.305	**0.680**	0.340	0.045
D-IA	**0.670**	0.291	0.379	0.027
D-DA	0.205	**0.300**	0.125	0.000
D-DG	0.200	**0.375**	0.160	0.030
D-WA	0.190	**0.785**	0.130	0.005

Table 4: Evaluation of the interest associated to the computed explanations.

there are many reasons to be dissimilar for two entities. For this reason, it is easy for this application to say why two entities do not match, since there is plenty of tokens that do not match between the entities in the record. Nevertheless, the explanation would be more interesting if the tokens returned would be the ones changing class of the record from non-matching to matching. In other words, we claim that an interesting explanation for non-matching entities should return the tokens that, if shared by the second entity, would make the record classified as matching.

In this section, we describe the evaluations we performed to evaluate the aforementioned situation. The experiments are similar to the first experiments described in Section 4.2, but in this case we select the tokens to remove. For sake of completeness, we performed a similar experiment with records classified as matches even if this evaluation is less meaningful. When the record is associated with a matching label, we remove all positive tokens (all tokens that contribute to the decision). The negative tokens are removed when the label represents a non-matching record. In Table 4 to evaluate the experiment we measure the *interest*, which is the accuracy computed on the records where the removal of the tokens was able to generate a change in the label.

Discussion. Table 4a shows that Landmark Explanation is good but slightly worse than LIME in terms of interest, when the records are labeled as matching class. This happens even if the surrogate model is really accurate (the MAE score is the lowest for all experiments with the single-entity configuration). The problem is that in most of the cases, even removing all tokens, the explanation created by Landmark Explanation belongs to the same class as before the token removal. Note that if we set a decision threshold to 0.4, our approach has the best results in all datasets. Table 4b shows that the explanations of non-matching entities generated by Landmark Explanation in the setting double-entity outperform the ones of Lime/Mojito Drop and Mojito Copy.

Lesson learned. Landmark Explanation generates interesting explanations, and the perturbation made with the double-entity generation technique effectively increases "the interest" of non-matching record explanations.

5 CONCLUSION

This paper introduces Landmark Explanation a tool that makes a post-hoc perturbation-based explainer able to deal with ML and DL models describing EM datasets. The approach has been experimented coupled with the LIME explainer, which is one of the most used state of the art approaches. The results show that the explanations generated by Landmark Explanation outperform the ones generated by the competing approaches in accuracy. Moreover, the explanations generated by Landmark Explanation have been experimented to be "more interesting" for the users.

Future work includes the study of techniques for summarizing the explanations to facilitate the interpretation of the EM model as an whole.

ACKNOWLEDGEMENT

This work was partially funded by SBDIO I4.0 (https://www.sbdioi40.it/), an industrial research project funded by the POR FESR Emilia Romagna 2014-2020 as part of the Smart Specialization Strategy (S3).

REFERENCES

[1] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
[2] Ursin Brunner and Kurt Stockinger. 2020. Entity Matching with Transformer Architectures - A Step Forward in Data Integration. In *EDBT*. OpenProceedings.org, 463–473.
[3] Zhaoqiang Chen, Qun Chen, Boyi Hou, Zhanhuai Li, and Guoliang Li. 2020. Towards Interpretable and Learnable Risk Analysis for Entity Resolution. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 1165–1180.
[4] Vincenzo Di Cicco, Donatella Firmani, Nick Koudas, Paolo Merialdo, and Divesh Srivastava. 2019. Interpreting deep learning models for entity resolution: an experience report using LIME. In *aiDM@SIGMOD*. ACM, 8:1–8:4.
[5] Mengnan Du, Ninghao Liu, and Xia Hu. 2020. Techniques for interpretable machine learning. *Commun. ACM* 63, 1 (2020), 68–77.
[6] Amr Ebaid, Saravanan Thirumuruganathan, Walid G Aref, Ahmed Elmagarmid, and Mourad Ouzzani. 2019. Explainer: Entity resolution explanations. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 2000–2003.
[7] Muhammad Ebraheem, Saravanan Thirumuruganathan, Shafiq R. Joty, Mourad Ouzzani, and Nan Tang. 2018. Distributed Representations of Tuples for Entity Resolution. *Proc. VLDB Endow.* 11, 11 (2018), 1454–1467.
[8] Amirata Ghorbani and James Y. Zou. 2019. Data Shapley: Equitable Valuation of Data for Machine Learning. In *ICML (Proceedings of Machine Learning Research)*, Vol. 97. PMLR, 2242–2251.
[9] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, David Madigan, et al. 2015. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics* 9, 3 (2015), 1350–1371.
[10] Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang-Chiew Tan. 2020. Deep Entity Matching with Pre-Trained Language Models. *Proc. VLDB Endow.* 14, 1 (Sept. 2020), 50–60. https://doi.org/10.14778/3421424.3421431
[11] Xiaolan Wang Laura Haas Alexandra Meliou. 2018. Explaining Data Integration. *Data Engineering* (2018), 47.
[12] Sidharth Mudgal, Han Li, Theodoros Rekatsinas, AnHai Doan, Youngchoon Park, Ganesh Krishnan, Rohit Deep, Esteban Arcaute, and Vijay Raghavendra. 2018. Deep Learning for Entity Matching: A Design Space Exploration. In *SIGMOD Conference*. ACM, 19–34.
[13] Matteo Paganelli, Paolo Sottovia, Francesco Guerra, and Yannis Velegrakis. 2019. TuneR: Fine Tuning of Rule-based Entity Matchers. In *CIKM*. ACM, 2945–2948.
[14] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.
[15] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. In *AAAI*. AAAI Press, 1527–1535.
[16] Rohit Singh, Venkata Vamsikrishna Meduri, Ahmed K. Elmagarmid, Samuel Madden, Paolo Papotti, Jorge-Arnulfo Quiané-Ruiz, Armando Solar-Lezama, and Nan Tang. 2017. Synthesizing Entity Matching Rules by Examples. *Proc. VLDB Endow.* 11, 2 (2017), 189–202.
[17] Saravanan Thirumuruganathan, Mourad Ouzzani, and Nan Tang. 2019. Explaining Entity Resolution Predictions: Where are we and What needs to be done?. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*. 1–6.
[18] Jiannan Wang, Guoliang Li, Jeffrey Xu Yu, and Jianhua Feng. 2011. Entity Matching: How Similar Is Similar. *PVLDB* (2011).
[19] Chen Zhao and Yeye He. 2019. Auto-EM: End-to-end Fuzzy Entity-Matching using Pre-trained Deep Models and Transfer Learning. In *WWW*. ACM, 2413–2424.