Linguistic Summaries as Explanation Mechanism for Classification Problems

Carla Wrede[1,2][0000−0002−1809−3183], Mark H.M. Winands[1][0000−0002−0125−0824], and Anna Wilbik[1][0000−0002−1989−0301]

[1] Department of Advanced Computing Sciences (DACS), Maastricht University, Maastricht, The Netherlands
{c.wrede, m.winands, a.wilbik}@maastrichtuniversity.nl
[2] VDL Nedcar B.V., Born, The Netherlands
c.wrede@vdlnedcar.nl

Abstract. The amount and complexity of generated and collected data is rapidly growing. As a consequence, it is increasingly hard to understand the data and extract useful information. Transparency, interpretability and understandability contribute towards explainability of the data, which is crucial for the user for both efficient and effective usage of it and trust in these data-based decisions. In this paper, we investigate how linguistic summaries can serve as an explanation mechanism for classification results. Linguistic summaries are template-based, semi-natural language-like sentences that can verbalize these (classification) patterns. We develop linguistic summarizations for the classification results of two publicly available data sets and perform an initial evaluation with a small group of potential users. The preliminary results look promising.

Keywords: Explainable Artificial Intelligence · Linguistic Summaries · Classification

1 Introduction

Recent advances in information technology cause companies to discover the value of data, resulting in more and more data being gathered with the hope of analyzing it. The amount of data is beyond human cognitive capabilities and comprehension skills, for instance, healthcare data double every two years [12]. To process those data, business organizations are using many powerful data mining and knowledge discovery methods, though they still require human understanding. For this purpose, recently there is a big interest in Explainable AI methods. There is an expectation that with those means the users can understand better the machine-made recommendations.

One of the methods that help the users understand the large amounts of data are linguistic summarizations [38]. However, till now, the explanatory capabilities of linguistic summaries were investigated only to a limited extent [11]. There is a potential for these summaries as ad hoc local explanations. Moreover, they use

natural language, which is the only fully natural means of communication for human beings.

Therefore, in this paper, we present the results of a proof-of-concept experiment, in which we generate linguistic summaries to provide an insight into why a certain point may have been assigned to a certain class. We use two publicly available data sets, select subsequently a small number of random cases as test cases, generate the appropriate summaries, and show them to a group of potential users. In our questionnaire, we ask about understandability, usefulness, perceived trustworthiness and user satisfaction. The results clearly show the potential of this method for future usage as verbalized explanations.

The remainder of this paper is organized as follows. First, in Section 2 more details about the related work of both Explainable Artificial Intelligence as well as linguistic summaries are given. Afterwards, Section 3 outlines how linguistic summaries can be used to explain classification results. Next, Section 4 describes how an initial evaluation is set up and conducted. Following this, Section 5 describes the results of this evaluation. Finally, Section 6 concludes this paper and gives an outlook for future research.

2 Related Work

The following section gives a brief overview of the two research fields of Explainable AI (Section 2.1) as well as linguistic summaries (Section 2.2).

2.1 Explainable Artificial Intelligence

With the increased presence of machine-learning models, there is a need for understanding these models, which has resulted in emerging a new domain of Explainable Artificial Intelligence (XAI) [7, 1, 18]. Generally, the various XAI methods can be divided into two basic categories: model-agnostic XAI methods, which can generate explanations for any type of black-box model and model-specific XAI methods, which are designed for a particular machine-learning model. Moreover, model-agnostic methods can be further distinguished into local and global methods, depending on whether they provide an explanation for a particular data point or the whole data set.

A number of model-specific XAI methods have already been proposed, e.g., for fuzzy rule-based systems [2, 3], logical formulas [25], counterfactual facts [34, 35], knowledge representation and reasoning [8, 27, 33, 9], temporal and causal relations in Bayesian networks [19, 26, 31], and black-box machine learning algorithms [17, 24]. Regarding model-agnostic approaches, examples are Grad-CAM [30], SHAP [23], LIME [28] or DeepLIFT [32].

In [7] the authors categorize model-agnostic XAI methods in six categories, depending on the explanation means, namely: local explanations, local simplification, feature relevance, explanation by example, visualizations and text explanations. In this paper, we want to investigate to what extent linguistic summaries can serve as both local and text explanations.

2.2 Linguistic Summaries

We employ the method of a linguistic summary as proposed by Yager [40]. A linguistic summary is a template-based sentence in semi-natural language. Typically, two protoforms (templates) are used, a simple one:

$$Qy\text{'s are } P \tag{1}$$

and extended form:

$$Q\ Ry\text{'s are } P \tag{2}$$

where Q is the quantifier, e.g. *many*, *most*, P is the summarizer, i.e., a property of the object (or a set of those), R is the qualifier, i.e., a different object characterization, and y's are the objects to be summarized. For a database of cars, a linguistic summary *most cars are fast* is an example of a simple protoform summary, while *most new cars are fast* is an example of an extended protoform summary.

The basic criterion for evaluating the quality of the linguistic summary is the truth value T, also called the validity of a summary. One possibility to determine its value is to employ Zadeh's calculus [41]. In this case, the truth value for a simple and an extended protoform is calculated respectively as:

$$\mathcal{T}(Qy\text{'s are } P) = \mu_Q\left(\frac{1}{n}\left(\sum_{i=1}^{n}\mu_P(y_i)\right)\right) \tag{3}$$

$$\mathcal{T}(Q\ Ry\text{'s are } P) = \mu_Q\left(\frac{\sum_{i=1}^{n}\mu_P(y_i)\wedge\mu_R(y_i)}{\sum_{i=1}^{n}\mu_R(y_i)}\right) \tag{4}$$

where $\mu_\cdot$ is a membership function of the appropriate linguistic term. More details about linguistic summarization and different methods for the evaluation of linguistic summaries can be found in [13].

This method has been investigated by many researchers and can summarize different types of structured data: databases (cf. Kacprzyk et al. [22, 21]), time series (cf. Kacprzyk et al. [20], Castillo-Ortega et al. [10]), standardand ized texts (cf. Szczepaniak [36]), videos (cf. Anderson et al. [4–6]), sensor data (cf. Ros et al. [29], Wilbik et al. [38, 39]), web logs (cf. Zadrożny and Kacprzyk [42]) and event logs (cf. Wilbik [37, 11]).

3 Linguistic Summaries for Explaining Classification

Our idea is that we can use linguistic summaries as local explanations for a classification problem. Let us consider a data set D with features $f_1, f_2, ..., f_n$ and classes $c_1, ..., c_m$. Each feature has a pre-defined certain number of linguistic terms. Similarly, quantifiers, such as *many*, *most*, and *almost all*, are defined.

Now imagine there is a classification model M that predicts that a point x belongs to a class c_k. An explanation as proposed by us is an extended summary of the data set D of the following form:

$$QRy\text{'s are } c_k \tag{5}$$

where R is a subset of features and linguistic terms that best describe the point x connected with the conjunction *and*, e.g. "*new* and *sporty* cars". In other words, R is the subset of feature-linguistic term pairs (f_i, lt_{ij}), such that $arg\,max_j(lt_{ij}(x(f_i)))$. The summarizer is a crisp set of all elements that are of class c_k in data set D. The truth value is then calculated according to (4).

4 Experimental Setup

The following section describes the exemplary linguistic summaries that are created for two publicly available data sets, which are described in Section 4.1. Section 4.2 describes the exemplary linguistic summaries, while Section 4.3 outlines how an initial evaluation of the reception of these examples is conducted by means of a semi-structured interview.

4.1 Data Sets

The two data sets used for the creation of the linguistic summaries are well-used: the iris data set and the glass identification data set from the machine-learning repository of the University of California, Irvine (UCI) [14].

The iris data set [16] consists of 150 entries of different plants, 50 for each of the three possible types Iris Setosa, Iris Versicolor and Iris Virginica. Each flower is described by four attributes, namely the sepal length and width as well as the petal length and width. This data set is chosen because of its simplicity in terms of both the number of features as well as the low comprehension complexity.

The glass identification data set [15] in turn consists of 214 entries of glass samples, with six possible types it can be classified as. Each glass sample is described by nine different attributes, eight of them being the "weight percent in corresponding oxide" for different elements, the ninth being the refractive index of each sample. This data set is chosen because in comparison to the iris data set it contains more features and classes.

4.2 Explanation Examples

Linguistic summaries are calculated for five different exemplary data points, three from the iris data set and two from the glass data set. For three out of these five data points, an additional linguistic summary is created to allow for a preference evaluation between different types of summaries, resulting in a total of 8 linguistic summaries. The reason for this is to present the participants with a choice for preference in order to inquire what aspects of a linguistic template they put more emphasis on.

For both of these data sets, a binary, multi-class decision tree is trained in Matlab, testing out the minimum number of leaf node observations in the range of 1 to 20, resulting in an accuracy of 0.91 and 0.79 for the iris and glass data set, respectively. The linguistic summaries are based on the classification results from these models.

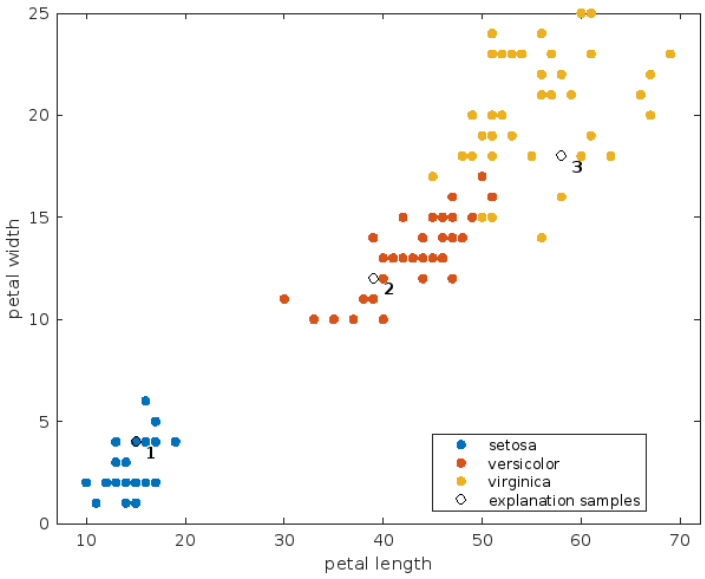

Fig. 1. The scatter plot of the iris data set and selected points to be explained

A scatter plot of the iris data set and selected points for which explanations were generated are shown in Figure 1. Also, a scatter plot of the glass data set and selected points for which explanations were generated are shown in Figure 2.

The quantifiers for the linguistic summaries can be described with multiple linguistic terms defined with open-right trapezoidal fuzzy membership functions of the following form:

$$OpenR(x, a, b) = max\left(min\left(\frac{x-a}{b-a}, 1\right), 0\right), \tag{6}$$

with x being the feature value, and a, b parameters describing the left half of the trapezoid and set by the authors.

We use six linguistic terms as quantifiers, listed here in order of increasing magnitude/meaning:

- Quite some, with $a = 0.4$ and $b = 0.45$
- At least half, with $a = 0.5$ and $b = 0.55$
- Many, with $a = 0.6$ and $b = 0.65$
- At least three quarters, with $a = 0.7$ and $b = 0.75$
- Most, with $a = 0.8$ and $b = 0.85$
- Almost all, with $a = 0.9$ and $b = 0.95$

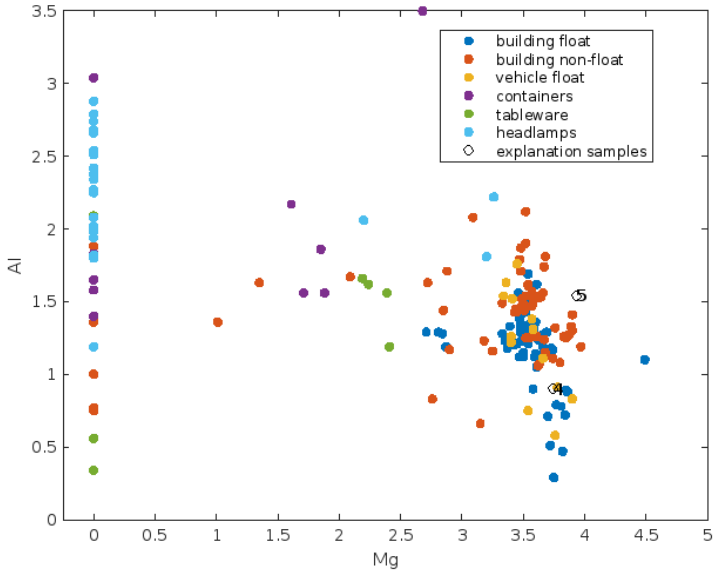

Fig. 2. The scatter plot of the glass data set and selected points to be explained

The wording is chosen purposefully as a mix between concise and ambiguous linguistic quantifiers in order to evaluate the possibly different perceptions between the two types. The parameter values have been chosen in steady decline in order to have an evenly spaced interpretation for the quantifiers.

The features of the data sets can be described with three linguistic terms defined with trapezoidal membership functions of the following form:

$$Trapezoid(x, a, b, c, d) = max\left(min\left(\frac{x-a}{b-a}, 1, \frac{d-x}{d-c}\right), 0\right), \tag{7}$$

with x being the feature value, and a, b, c, d parameters describing the four points of the trapezoid and set by the authors.

For the purpose of simplicity, we use the same three linguistic terms (*small, average, long*) to describe each (normalized) variable. Their membership functions are:

- Small/low, with $a = 0$, $b = 0$, $c = 0.2$, $d = 0.4$
- Average, with $a = 0.2$, $b = 0.4$, $c = 0.6$, $d = 0.8$
- Long/high, with $a = 0.6$, $b = 0.8$, $c = 1$, $d = 1$

A qualifier R can concern several features that are combined with a conjunction like "and", for instance *low* f_1 *and high* f_3. Again, the wording is chosen with the purpose of clearly describing the two extremes of the values for each feature, with a distinction between the two data sets of iris and glass, respectively.

Example 1 Iris data set:

"*Almost all* flowers with a *small* petal length are Iris Setosa."

This linguistic summary is chosen because of the clear meaning of the quantifier as well as the little number of features included in order to allow for a straightforward opening of the interview.

Example 2 Iris data set:

"*Most* of the flowers that have an *average* petal length and an *average* petal width are Iris Versicolor."

This linguistic summary is chosen as a transition from a concise to an ambiguous quantifier. Additionally, another feature is added to investigate if the number of features has an influence on the explanation assessment.

Example 3 Iris data set:

"*Almost all* flowers with *long* petal width are Iris Virginica."

(validity of summary: 0.73)

OR

"*Most* of the flowers with *long* petal width are Iris Virginica."

(validity of summary: 1)

For this example, two linguistic summaries are chosen in order to inquire about preference for a specific explanation. While both concise and ambiguous quantifiers were evaluated on their own before. For this example they are contrasted to investigate whether either the quantifier or the validity of a summary is regarded higher.

Example 4 Glass data set:

"*Many* of the glass particles with *high* Mg and *low* Al measurements are from float processed building windows."

(validity of summary: 0.27)

OR

"*Many* of the glass particles with an *average* refractive index, *high* Mg, *low* Al and *average* Si measurements are from float processed building windows."

(validity of summary: 0.79)

Similar to the previous example, we are interested in whether the number of features or the validity of the summary is regarded higher while keeping the quantifier constant.

Example 5 Glass data set:

"*At least half* of the glass particles with *high* Mg and *average* Al measurements are from non float processed building windows."

(validity of summary: 0.42)

OR

"*Quite some* of the glass particles with *high* Mg, *average* Al, *low* Ba and *low* Fe are from non float processed building windows."

(validity of summary: 1)

In this last example and similar to Example 3, we are again interested to see whether rank-wise similar quantifiers with their respective validity have an impact on the assessment, while also considering a difference in the number of features presented in the summary.

4.3 Evaluation Method

To evaluate the chosen linguistic summaries, a semi-structured interview is conducted. For this, nine different participants were interviewed on an individual basis. All of them can be classified as AI experts, defined by their specialization in education, holding a more than intermediate knowledge in either an AI, machine learning or algorithms study and design field, with an age range from 22 to 31. For evaluation of the examples as defined in the previous section, we decide on five selected aspects, which are judged with the help of a 5-point Likert scale:

- Understandability
- Usefulness
- Trustworthiness
- Helpfulness
- Confusedness

The interview is structured as follows:

First, an introduction to linguistic summaries and fuzzy membership functions is given. Then, the list of quantifiers without their absolute meaning is given to the participants as a ranked list from highest to lowest magnitude. It is on purpose that the participants will only receive the list of quantifiers with their relative meaning, rather than the absolute meaning in order to evaluate a possible discrepancy in perception of concise and ambiguous quantifiers.

Next, an instruction follows, indicating that there will be 5 example linguistic summaries for two different data sets. For each of these examples, they will be asked for a judgement of the previously listed five aspects, and possibly a selection of preferred summary, where applicable. Additionally, the participants are told that they will be prompted to yet undefined open-ended questions based on their chosen ranking of these five aspects.

The interview then proceeds by giving a short introduction of the iris data set, followed by the first three examples and their respective questions. Finally, the glass data set is briefly introduced as well, again followed by the last two examples and their respective questions as well.

5 Results

After having asked the participants to rate the presented linguistic summaries for the five aspects as mentioned in Section 4.3, the results are discussed in Subsections 5.1 to 5.5, respectively, as well as the preference option in Subsection 5.6. While not statistically significant due to the small sample size, these results still show an indication of how well the examples are perceived in the different aspects.

5.1 Understandability

In general, the linguistic summaries were well received, with the best summary scoring an average of 4.55 on a 5-point Likert scale and the worst summary scoring 3.77 on average. The most understandable summary was the linguistic summary from Example 1, which can be accounted towards its clear interpretation of the quantifier "almost all" for all participants and low number of features. The least understandable summary stems from Example 4, as the participants had difficulties with understanding both the reduced quantifier and the fact that none of the summaries were 100% valid. Many participants noted that they like the clear structure of the summaries, which still leads to understandable explanations when the number of features rises, as showcased by the second option in Example 5, which received an average score of 4.22 while containing 4 features in its summary.

5.2 Usefulness

The usefulness was similarly high for Examples 1 to 4, with an average score around 3.69. The least useful is from Example 5, with an average score of 3.11. The reason for Example 5 not scoring very well was accounted to the fact that the quantifiers were too small to deduce any meaning from it. Many participants even questioned the purpose of the existence of such a summary. However, the scores of this criterion are the lowest. This can be accounted to the fact the participants felt like they lacked a more clear definition of the meaning and range of both quantifiers and qualifiers. They would prefer to have more additional information to decide better on the explanation's usefulness.

5.3 Trustworthiness

Similar to the usefulness, Examples 1 to 4 scored similarly high with regards to trustworthiness, with an average score of 4.16. Just like before, Example 5 and its low quantifiers were not well perceived when it comes to how trustworthy this explanation is deemed as, resulting in the lowest average score of 3.33. Interestingly, when the validity does not reach 100% as in Example 4, it is still rated similarly high as linguistic summaries that did reach a validity of 100% (Example 1 to 3). This is due to multiple factors, one of them being the cautiousness about the high number, since "nothing ever in life is 100% sure", as one participant noted. Another factor is that some participants were missing more background information about the meaning of the qualifiers, so they could not trust that they had the same interpretation of the meaning as the underlying definition of the linguistic terms set by the authors.

5.4 Helpfulness

The most helpful summary is the explanation from Example 1, with an average rating of 4.0. Participants noted the concise and short sentence and especially

liked the used quantifier for this statement. On the contrary, they disliked the summary from Example 4 the most, with an average score of 2.89. Similar to before, the reasoning here was based on the lower validity of both options, even though one participant appreciated "the fact that it does tell you that it is not fully valid is helpful, which can be a valuable insight in combination with other explanations".

5.5 Confusedness

In line with the results concerning the understandability of the summaries, the summary from Example 1 was the least confusing, with an average score of 1.11. Again, Example 4 was the most confusing for all participants with an average score of 2.91, followed by Example 5. The reason why Example 5 was scored similarly confusing as Example 4 is because many participants did not see the purpose of a statement with such a low quantifier. A participant noted that for them, "a summary that sums up only around half the data points, if not even less, seems no better than random".

5.6 Preference Options

Focusing on the preference options, the results were clear. All of the nine participants favoured the second option in Example 4, with the main reason being the higher validity. Upon inquiry, the participants noted that even though the quantifier was of the same level, there are still many unknown factors in both sentences, as the quantifier was an ambiguous one and they lacked more definition for the linguistic representation for the qualifier, hence they relied mostly on the validity.

For the other two preferences, the results were less strong, but still highly indicative, as both second options in Examples 3 and 5 were chosen by a majority of two-thirds. For Example 3, all participants that favoured the second option mentioned the fact that even though the quantifier is smaller than in the first option, they would rather rely on a more accurate but defined explanation. Therefore, the validity has a higher relevance to them, as the first options leaves too many assumptions unanswered. For Example 5, the reasoning was similar, as the participants would rather be sure of the subpart of the data that is summarized, than "having covered more ground but with a higher uncertainty". However, multiple participants noted that while they preferred the second option and appreciated the small, but confident insight it gives, this does not mean that they find it as useful or helpful as previous explanations.

In general, it can be seen that the validity has the highest relevance when it comes to rating an explanation, as the preferred summaries always came with the greater validity of the two presented options. Upon inquiry about the features, some of the participants liked the shorter sentences that come with a reduced number of features, other participants were glad about additional specificity that comes with a higher number of features. Both parties agreed nevertheless that the structure of these explanations makes them still well understandable,

regardless of the number of features used in the sentence. However, they did note again that there should be more information about the linguistic representation of the qualifiers in order to judge it better.

6 Conclusion and Future Research

In this paper, we investigated the explanatory power of linguistic summaries for the results of two different classification problems. We collected the opinions of several potential users, which are knowledgeable about AI, concerning the perception of these types of explanations. The initial evaluation looks promising for the linguistic summaries to be used as explanations for (classification) patterns.

In general, the verbalized explanations in the form of linguistic summaries were well perceived by all participants. However, it was noted that they struggled with balancing not only the validity and the meaning of the quantifier but also the meaning of the qualifier for the features. While the concise quantifiers were easier to work with, they specifically struggled with the ambiguous ones, as they lacked a more precise indication of how many data points these quantifiers summarize, even though the ranking of all quantifiers was given to them. In order to make the explanations better, the consensus was that either only more precise quantifiers are used, or directly making use of the linguistic representation of percentiles (i.e. "about 60%" instead of "many") with a domain-appropriate division, as proposed by two of the participants. As seen by the outcome of the preference options, the participants strongly favour explanations that feature a high validity of the summary. With this, only the meaning of the quantifiers is left open to interpretation. Here, the opinions of the participants diverted. While some would prefer longer sentences, that clearly describe the range of the quantifiers more, other participants do not wish to have even longer sentences. It has to be investigated further how the quantifiers can be used better to increase the perception of the explanations. A suggestion that has been named by some of the participants is to enumerate the respective linguistic terms and feature pairs to allow for better comprehension.

One of the limitations of this research is the small set of participants that were interviewed. However, it already indicates the potential linguistic summaries have as a means of explanation. Moreover, currently the linguistic summaries presented to the users were handpicked from a set of possible summaries. In future research, more extensive user evaluation should be performed, as well as to propose a selection method for the best linguistic summary explanation. Furthermore, more complex data sets have to be investigated to see what kind of impact different types of data and feature distributions have on the user's perception of these forms of explanations.

Acknowledgements. This research has been funded by the Rijksdiesnt voor Ondernemend Nederland (RVO) in the framework of the project Green Transport Delta - Elektrificatie.

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE Access **6**, 52138–52160 (2018)
2. Alonso, J.M., Castiello, C., Magdalena, L., Mencar, C.: Explainable Fuzzy Systems - Paving the Way from Interpretable Fuzzy Systems to Explainable AI Systems. Studies in Computational Intelligence, Springer (2021), https://doi.org/10.1007/978-3-030-71098-9
3. Alonso, J.M., Ramos-Soto, A., Reiter, E., van Deemter, K.: An exploratory study on the benefits of using natural language for explaining fuzzy rule-based systems. In: Proc. of the IEEE International Conference on Fuzzy Systems (2017). https://doi.org/10.1109/FUZZ-IEEE.2017.8015489
4. Anderson, D., Luke, R.H., Keller, J.M., Skubic, M., Rantz, M., Aud, M.: Linguistic summarization of video for fall detection using voxel person and fuzzy logic. Computer Vision and Image Understanding **1**(113), 80–89 (2009)
5. Anderson, D., Luke, R.H., Keller, J.M., Skubic, M., Rantz, M., Aud, M.: Modeling human activity from voxel person using fuzzy logic. IEEE Transactions on Fuzzy Systems **1**(17), 39–49 (2009)
6. Anderson, D., Luke, R.H., Stone, E., Keller, J.M.: Segmentation and linguistic summarization of voxel environments using stereo vision and genetic algorithms. In: Proceedings IEEE International Conference on Fuzzy Systems, World Congress on Computational Intelligence. pp. 2756–2763 (2010)
7. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion **58**, 82–115 (2020)
8. Budzynska, K., Villata, S.: Argument mining. The IEEE Intelligent Informatics Bulletin **17**, 1–7 (2016)
9. Caminada, M.W., Kutlak, R., Oren, N., Vasconcelos, W.W.: Scrutable plan enactment via argumentation and natural language generation. In: Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems. pp. 1625–1626 (2014)
10. Castillo-Ortega, R., Marín, N., Sánchez, D.: Linguistic query answering on data cubes with time dimension. International Journal of Intelligent Systems **26**(10), 1002–1021 (2011)
11. Chouhan, S., Wilbik, A., Dijkman, R.: Explanation of anomalies in business process event logs with linguistic summaries. In: Proceedings of WCCI 2022 (2022)
12. Cottle, M., Hoover, W., Kanwal, S., Kohn, M., Strome, T., Treister, N.: Transforming health care through big data strategies for leveraging big data in the health care industry. Institute for Health Technology Transformation, http://ihealthtran.com/big-data-in-healthcare (2013)
13. Delgado, M., Ruiz, M.D., Sanchez, D., Vila, M.A.: Fuzzy quantification: A state of the art. Fuzzy Sets and Systems **242**, 1 – 30 (2014)
14. Dua, D., Graff, C.: UCI machine learning repository (2017), http://archive.ics.uci.edu/ml
15. Evett, I.W., Spiehler, E.J.: Rule induction in forensic science. In: Knowledge Based Systems, pp. 152–160 (1989)
16. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annals of eugenics **7**(2), 179–188 (1936)

17. Forrest, J., Sripada, S., Pang, W., Coghill, G.: Towards making NLG a voice for interpretable machine learning. In: Proc. of the International Conference on Natural Language Generation (INLG). pp. 177–182. Association for Computational Linguistics, Tilburg University, The Netherlands (2018). https://doi.org/10.18653/v1/W18-6522
18. Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.Z.: Xai—explainable artificial intelligence. Science robotics **4**(37), eaay7120 (2019)
19. Hennessy, C., Bugarin, A., Reiter, E.: Explaining Bayesian Networks in natural language: State of the art and challenges. In: Proc. of the Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI) at the International Conference on Natural Language Generation (INLG). Dublin, Ireland (2020), https://www.aclweb.org/anthology/2020.nl4xai-1.7/
20. Kacprzyk, J., Wilbik, A., Zadrożny, S.: Linguistic summarization of time series using a fuzzy quantifier driven aggregation. Fuzzy Sets and Systems **159**(12), 1485–1499 (2008)
21. Kacprzyk, J., Yager, R.R., Zadrożny, S.: Fuzzy linguistic summaries of databases for an efficient business data analysis and decision support. In: Abramowicz, W., Żurada, J. (eds.) Knowledge Discovery for Business Information Systems, pp. 129–152. Kluwer, Boston (2001)
22. Kacprzyk, J., Zadrożny, S.: Linguistic database summaries and their protoforms: toward natural language based knowledge discovery tools. Information Sciences **173**, 281–304 (2005)
23. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 4768–4777. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017). https://doi.org/10.5555/3295222.3295230
24. Mariotti, E., Alonso, J.M., Gatt, A.: Towards harnessing natural language generation to explain black-box models. In: Proc. of the Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI) at the International Conference on Natural Language Generation (INLG). Dublin, Ireland (2020), https://www.aclweb.org/anthology/2020.nl4xai-1.6/
25. Mayn, A., van Deemter, K.: Towards generating effective explanations of logical formulas: Challenges and strategies. In: Proc. of the Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI) at the International Conference on Natural Language Generation (INLG). Dublin, Ireland (2020), https://www.aclweb.org/anthology/2020.nl4xai-1.9/
26. Pereira-Fariña, M., Bugarín, A.: Content determination for natural language descriptions of predictive Bayesian Networks. In: Proc. of the Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT). pp. 784–791. Atlantis Press (2019)
27. Rago, A., Cocarascu, O., Toni, F.: Argumentation-based recommendations: Fantastic explanations and how to find them. In: Proc. of the International Joint Conference on Artificial Intelligence (IJCAI). pp. 1949–1955 (2018). https://doi.org/10.24963/ijcai.2018/269
28. Ribeiro, M., Singh, S., Guestrin, C.: "Why should i trust you?": Explaining the predictions of any classifier. pp. 97–101 (02 2016). https://doi.org/10.18653/v1/N16-3020
29. Ros, M., Pegalajar, M., Delgado, M., Vila, A., Anderson, D.T., Keller, J.M., Popescu, M.: Linguistic summarization of long-term trends for understanding change in human behavior. In: Proceedings of the IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2011. pp. 2080–2087 (2011)

30. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision **128**(2), 336–359 (Oct 2019). https://doi.org/10.1007/s11263-019-01228-7, http://dx.doi.org/10.1007/s11263-019-01228-7
31. Sevilla, J.: Explaining data using causal Bayesian Networks. In: Proc. of the Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI) at the International Conference on Natural Language Generation (INLG). Dublin, Ireland (2020), https://www.aclweb.org/anthology/2020.nl4xai-1.8/
32. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. p. 3145–3153. ICML'17, JMLR.org (2017)
33. Sierra, C., de Mántaras, R.L., Simoff, S.J.: The argumentative mediator. In: Proc. of the European Conference on Multi-Agent Systems (EUMAS) and the International Conference on Agreement Technologies (AT). pp. 439–454. Valencia, Spain (2016). https://doi.org/10.1007/978-3-319-59294-7"36
34. Stepin, I., Alonso, J.M., Catala, A., Pereira, M.: Generation and evaluation of factual and counterfactual explanations for decision trees and fuzzy rule-based classifiers. In: Proc. of the IEEE World Congress on Computational Intelligence (2020). https://doi.org/10.1109/FUZZ48607.2020.9177629
35. Stepin, I., Alonso, J.M., Catala, A., Pereira-Fariña, M.: A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. IEEE Access **9**, 11974 – 12001 (2021), https://doi.org/10.1109/ACCESS.2021.3051315
36. Szczepaniak, P.S., Ochelska, J.: Linguistic summaries of standardized documents. In: Last, M., Szczepaniak, P.S., Volkovich, Z., Kandel, A. (eds.) Advances in Web Intelligence and Data Mining, pp. 221–232. Springer Berlin Heidelberg (2006)
37. Wilbik, A., Dijkman, R.M.: Linguistic summaries of process data. In: 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). pp. 1–7 (2015). https://doi.org/10.1109/FUZZ-IEEE.2015.7337891
38. Wilbik, A., Keller, J.M., Alexander, G.L.: Linguistic summarization of sensor data for eldercare. In: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC 2011). pp. 2595–2599 (2011)
39. Wilbik, A., Keller, J.M., Bezdek, J.C.: Linguistic prototypes for data from eldercare residents. IEEE Transactions on Fuzzy Systems **22**(1), 110–123 (2013)
40. Yager, R.R.: A new approach to the summarization of data. Information Sciences **28**, 69–86 (1982)
41. Zadeh, L.A.: A computational approach to fuzzy quantifiers in natural languages. In: Computational linguistics, pp. 149–184. Elsevier (1983)
42. Zadrozny, S., Kacprzyk, J.: Summarizing the contents of web server logs: A fuzzy linguistic approach. In: Proceedings of FUZZ-IEEE 2007 (2007)