

Explainable Classification of Medical Documents Through a Text-to-Text Transformer

Mihai Horia Popescu¹, Kevin Roitero¹ and Vincenzo Della Mea¹

¹Dept. of Mathematics, Computer Science and Physics, University of Udine, Udine, Italy

Abstract

Death certificates are important medical records which are collected for the purpose of public healthcare and statistics by multiple organizations around the globe. Due to their importance, those certificates are compiled by experienced medical practitioner according to a standard defined by the World Health Organization including rules to select an underlying cause of death (UCOD). For this reason, the coding of death certificates is a slow and costly process. To overcome these issues, the scientific community proposed deep learning approaches to perform such a task. Despite those systems achieve high accuracy scores (close to 1), their complexity makes the obscure to the final user, making it unfeasible the adoption as a decision support system.

In this paper, we propose a model based on text-to-text transformers which is able to provide a UCOD as well as to generate a human-readable explanation for its classification. We compare the proposed approach to state-of-the-art interpretable rule-based systems.

Keywords

deep learning, XAI, automated coding, medical documentation, generative model


1. Introduction

Traditionally, natural language processing (NLP) applications have been built on techniques that are natively explainable. Such techniques are generally referred to as “white box” techniques, and are mainly implemented using rule-based heuristics, decision trees, hidden Markov models, etc. [1]. Recent advances in Deep Learning (DL), a “black box” machine learning technique, have dramatically improved Neural Network (NNs) accuracy and increasingly gained interest from stakeholders. As a result, DL became the dominant approach in NLP and have seen wide adoption in a large amount of applications [2, 3]. Such a popularity of DL based approaches have been pursued by focusing merely on effectiveness on such a systems and thus resulting in effective models lacking of interpretability. Hence, concerns have been raised on the adoption of such black box methodologies in specific sensitive applications such as healthcare, decision making, and finance, in which settings it is fundamental to rely on interpretable models [4, 5]. As a result, for sensitive domains and real-world decision-making systems, the mere effectiveness of the system is not enough; those systems also need to support the reliability of the produced result and

HC@AIxIA 2022: 1st AIxIA Workshop on Artificial Intelligence For Healthcare, November 28 – December 2, 2022, Udine, It

✉ mihaihoria.popescu@uniud.it (M. H. Popescu); kevin.roitero@uniud.it (K. Roitero); vincenzo.dellamea@uniud.it (V. Della Mea)

ORCID 0000-0003-3378-0368 (M. H. Popescu); 0000-0002-9191-3280 (K. Roitero); 0000-0002-0144-3802 (V. Della Mea)

 © 2022 Copyright ©2022 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

thus provide a feedback e.g., in the form of a confidence score or a human-readable explanation to inform the final user if the produced result is likely to be correct and/or trustworthy or to explain the rationale behind the model decisions [6]. For these reasons, in recent times we observed an increase of interest from the community to develop and improve methods for the interpretability of DL models, especially towards the generation of human-readable explanations generated using explainable artificial intelligence (XAI) models [1, 3, 7].

Recently, many works have been developed to produce natural language explanations for DL systems [8, 9]. While diverse approaches that can be used to generate explanation exist, most of these methods can be categorized as producing post-hoc explanations. Such kind of techniques target models that are not interpretable by design and are used to enhance the interpretability of the underlying model choices [7].

In this paper we propose a methodology able to generate a human readable explanation for the predictions produced by a model designed in the context of select the underlying cause of death from death certificates with, which achieves very high accuracy scores (close to 1) [10, 11, 12], but it is not being adopted in practice due to its lack of interoperability.

2. Background and Related Work

In general, XAI approaches can be categorized from different perspectives: local versus global [13], transparent models versus post-hoc explainability [7], based on XAI goals (such as trustworthiness, causality, transferability, etc.) [7]. Our work is based on local and post-hoc explainability, given that from the explanations generated it is possible only to understand the reason for the predicted UCOD. We have identified two major goals that the users may desire and which those explanations can support; *trustworthiness* and *informativeness*.

Different approaches to enhance interpretability exists in the literature. Ribeiro et al. [14] studied the explainability of a model's predictions using feature importance-based explanations. Other approaches, such as the one proposed by Camburu et al. [15], first generate a free-form natural language explanation, then use such an explanations to infer the classification prediction. Similarly, Brand et al. [16, 17] shown that one can jointly predict and generate an explanation for classifying the veracity of statements. From a different perspective, some other works used the confidence of the model as a reliability measure for the correctness of the predictions by computing a calibrated confidence score [6]. Finally, other works such as the one proposed by Agarwal et al. [18] leveraged alternative measures like the variance of gradients to measure model reliability and instance difficulty.

3. Data

3.1. The Death Certificate

The death certificate is the main source of mortality data. Such data is supposed to be collected in compliance with the standard death certificate format defined in [19] and [20].

The death certificate contains: administrative details, a part called Frame A, and a part called Frame B. Frame A is used to record the sequence of events leading directly to death, and may

contain conditions that do not belong to the sequence but their presence contributed to death. Conversely, Frame B contains additional health conditions, such as previous surgery, mode of death, or place of occurrence. It should be noted that while Frame A contains the textual expression of conditions as filled by physicians, their corresponding ICD-10 codes are generally provided by experts coders. The coded version of the certificate is the format used for the selection of the UCOD.

The UCOD is the most important information extracted from mortality data, and it is used for statistical comparison and public health data. It is defined as '*(a) the disease or injury which initiated the train of morbid events leading directly to death, or (b) the circumstances of the accident or violence which produced the fatal injury*' [19]. The UCOD is selected according to the coding rules defined in the reference guide. The chosen code is usually one of the conditions present in the chains reported by the certifying doctor in Frame A.

3.2. Generation of Ground Truth Explanations

The system used for the generation of the gold explanations is called DORIS [21], a prototype rule-based system for mortality coding-based ICD-10 and ICD-11. Those rules can be subdivided into 2 categories; selection and modification rules. Currently, the system fully supports 18 out of 38 selection rules, and about 95% of the modification rules. The remaining rules are only partially implemented. The system was evaluated on datasets for both ICD-10 and ICD-11. DORIS is unable to code 8.2% of the total certificates and has an accuracy of 78% for ICD-10 [21].

The explanation generated by DORIS describes the coding instructions used to reach the selection of the UCOD and the conditions on which the rule is applied. In Table 1 we have presented two cases of explanations used by DORIS for two coding instructions and the associated description used in the reference guide. To select the UCOD multiple coding instructions may be used, as a result the explanations are concatenated.

3.3. Data Source and Preparation

The death certificates data files were collected from the U.S. National Center for Health Statistics (NCHS)¹. The dataset contains a total of 12,919,268 records for the years 2014–2017 including administrative data, coded conditions for frames A and B, and the UCOD that we used as ground truth. From the full dataset, we extracted 510,000 records for which the rule-based system presented in the Section 3.2 was able to correctly select the UCOD. The data then have been pre-processed to select only the data needed for our experiment. For this task, we choose to use the sex and age features from the administrative data and conditions from the Frame A. The dataset has been split into three smaller parts using randomization and stratified sampling by target UCOD. For the train set, we have selected 400,000 records, 100,000 records for the test set, and the remaining 10,000 certificates for the validation set. The dataset contains the same records, dataset split, and reverse coding format used for the NLP model used for the selection of the underlying cause of death using reverse coding as proposed by Della Mea et al. [10] and detailed in the following.

¹https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm

Table 1
Example explanation generated by DORIS.

| Coding instruction | Explanation used | Rule description |
|---------------------------|--|---|
| SP1 | Malignant neoplasm of prostate is the unique condition reported in the certificate and is the new tentative starting point (TUC). | If there is only one condition reported on the certificate, this is the new TUC. |
| SP2 | Unspecified injury of head is the first condition reported on the single used line, which is selected as the new tentative starting point. | If only one line is used but multiple conditions are preset, select the first condition as the new TUC. |

As input, the model proposed takes the version encoded as text of the certificates. Since the certificates do not have the original textual conditions present, we had to reverse the work done by coders because it brings the certificate back to text. The certificate encoded as text needs to encode both administrative data and conditions. The administrative data were put in an explicit form (e.g., Female, 39y old). Each line is encoded with the title entity, while for multiple codes per line, the titles are merged using "or" expression and the entire line go between parentheses. The sequence of lines then is concatenated with the expression "due to", where Part 2 if present, is concatenated using "in the context of" between the last line of Part 1 and Part 2.

4. Methods

4.1. Generating Explanations

We develop and train our models by relying on both the PyTorch² and HuggingFace³ frameworks. The experiments have been carried out on a Linux server equipped with 16x Intel(R) Core(TM) i7-10700 CPU @ 2.90GHz, 70GB of RAM, and 2x Nvidia Geforce RTX 3090 GPUs. We make the trained model available to the community.⁴

T5 [22] is a transformer based model trained on a mixture of both supervised and unsupervised tasks (i.e., summarization, translation, etc.) [22, Appendix Section]. In this work, we rely on the T5-base model⁵, which is a 220 million parameters model composed of an encoder-decoder stack involving 12 blocks, each of those implementing a self-attention mechanism, an encoder-decoder attention one, and a feed forward network.

Many available transformer-based architectures leverage separate transformer models for either discriminative (e.g., classification) or generative (e.g., text-generation) tasks. As opposed to this approach, we take inspiration from E-BART [16, 17], a model designed in the context

²<https://pytorch.org/>

³<https://huggingface.co/>

⁴To request access to the model, send an email to the paper authors.

⁵<https://huggingface.co/t5-base>

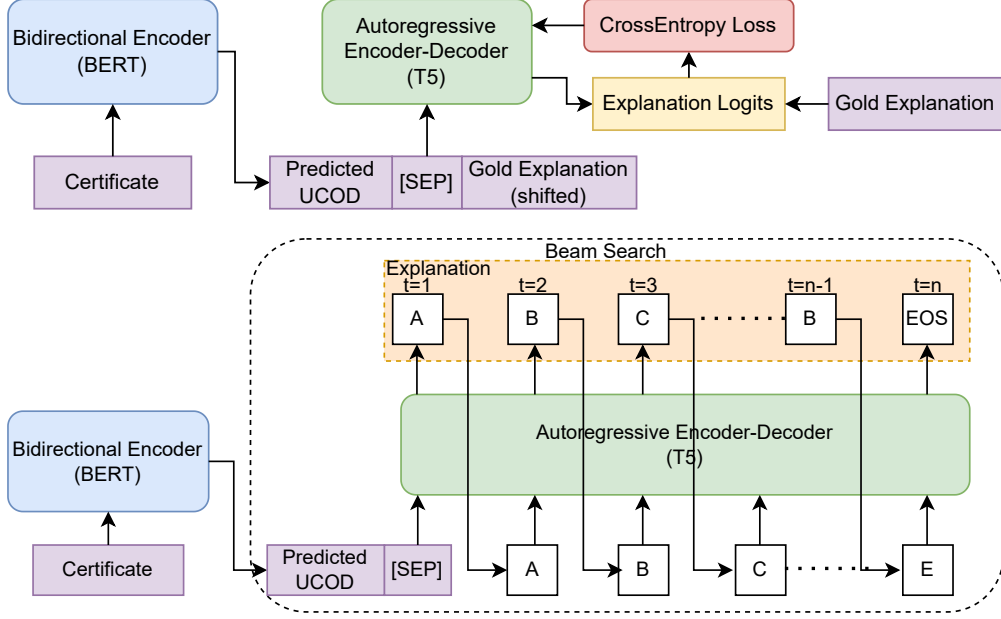


Figure 1: Model training (above) and inference (below).

of misinformation and veracity assessment to perform a discriminative task (i.e., classify the truthfulness of statements) and a generative one (i.e., generate a human-readable explanation for the former step) at the same time. In a similar fashion, we develop a model which is capable of classifying the UCOD of a certificate and generate a human-readable explanation for such a task. The model training and inference phases are detailed in the following, and summarized in Figure 1. Given that the focus of this work is on the generated explanations, in the following we omit the description of the discriminative model (which is anyway a standard BERT-based model equipped with a classification head) and we focus only on the generative one.

To generate the explanations, the model takes in input, separated by the [SEP] token, the death certificate encoded as text described in Section 3.3, the string generated as a report by the rule based system presented in Section 3.2, and the UCOD, the code predicted by the discriminative model representing the underlying cause of death. The model is then trained in a causal fashion, thus trained to auto-regressively predict the gold explanation (shifted), which is the description generated by the rule based system detailed in previous sections. The model loss is computed by considering the conventional multi-class cross-entropy loss function, where the number of classes is equal to the size of the vocabulary, defined as

$$\mathcal{L} = -\frac{1}{B} \sum_{b=1}^B \sum_{k=1}^{|V|} y_k^b \log(\hat{y}_k^b)$$

where b is the batch and B the batch size, $|V|$ is the vocabulary size, y is the true token to be predicted by the model, and \hat{y}_k is the output probability distribution over the vocabulary at each time-step.

Table 2

Effectiveness of the model on the generated explanations.

| Dataset | Rouge-1 | | | Rouge-2 | | | Rouge-L | | |
|---------------|---------|--------|--------|---------|--------|--------|---------|--------|--------|
| | Prec | Rec | F1 | Prec | Rec | F1 | Prec | Rec | F1 |
| CDC-Test 100K | 0.9988 | 0.9985 | 0.9986 | 0.9983 | 0.9980 | 0.9981 | 0.9986 | 0.9983 | 0.9983 |

During the inference step, the model generates the text by leveraging beam search, thus generating the explanation token-by-token by feeding the input tokens via the cross-attention layers to the decoder, and then auto-regressively generating the decoder output. To optimize the generation process, we set the early stopping parameter to the value of true so that the beam generation is stopped when all beam hypotheses reach the EOS token. Experimentally, we found that such generation procedure is suitable for the task, and generates relevant explanations for each input string, thus we found no need to implement constrained search techniques or try alternatives to beam search. For the same reason, we always select the output sequence with the highest likelihood as computed by the model.

4.2. Metrics

We evaluate the generated summaries using the Rouge score [23], a recall-oriented measure designed to compare a generated textual summary to an ideal one, usually generated by a human [24, 25]. More in detail, Rouge-N denoted an n-gram metric between a candidate summary and the reference summary. In this work we consider Rouge-1 (uni-gram based metric), Rouge-2 (bi-gram based), and Rouge-L, that is computed by considering the Longest Common Subsequence (LCS). More in detail, Rouge precision is defined as the number of overlapping n-grams between the candidate and the reference summary divided by the number of n-grams in the candidate summary, Rouge recall is defined as the number of overlapping n-grams between the candidate and the reference summary divided by the number of n-grams in the reference summary, and Rouge F1 is the harmonic mean of precision and recall.

5. Results and Discussion

Table 2 shows the Rouge scores for the considered datasets. As we can see, we have reached an overall score near to 1 for all the evaluated n-grams. Each has a high score of precision and recall and F1 values. The recall shows that in almost all cases the n-grams in the gold explanation are also present in the generated explanation, while the precision shows that almost all the n-grams in the generated explanations are present in the reference explanation. Comparing the n-grams proposed, the bi-grams (Rogue-2 score) has the lowest F1 score with a value of 0.9981.

Since the overall scores are very high, most of the generated explanations have a perfect match with the gold explanation. For the remaining cases we also perform a qualitative analysis of the generated explanations, by comparing them to the rule-based system, considering the structure of the rule, the conditions involved and terminology. Table 3 shows the certificate, as well as the gold and generated explanation for a sample of the instance present in the datasets.

As we can see from the table, the explanation generated is not fully correct. In particular, the error occurs in both cases on the obvious causes selection, where multiple causes are obvious causes of the TUC, but the generated description was not able to identify one. In the first case the explanation lead to an error for the selection of the UCOD, while in the second did not influence the final result. In all the cases where the description was incorrect, we have noticed that the terminology used was always consistent. The rules structure was correctly applied, and the conditions used were always consistent with those of the certificate. The errors were mainly for SP6 obvious causes and M1 special instructions, where categories were not recognized as part of the rule. While those cases are not recognized, is most likely that they were not part of the training set.

6. Conclusions

We have presented a system that is able to enhance the interpretability of a classification model by generating explanations using as reference a rule-based system. The model was not only able to generate appropriate explanations consistently (about 0.998 F1 score), but it was able to correctly learn and use the structure of the rules and their terminology. The proposed model has the ability to predict the UCOD, since the last sentence always specifies the category suggested; this feature is very important since the rule-based system used to learn the explanations do not reach the same accuracy of the classification model, and the suggested UCOD of the explanation can be used to crosscheck the classification model UCOD to understand when the explanation is likely to be incorrect. Some limitations of this preliminary experiment comes from the dataset used. In fact, for this experiment, we have used a dataset as big as the preliminary evaluation of the classification model, while the certificates used needed to be encoded as text as a reverse encoding from coded conditions. Those limitations come from a lack of certificates with natural textual conditions, which were still sufficient to evaluate the feasibility of this approach.

This paper opens for plenty of future work. More in detail, for future experimentation and evaluation, we plan to retrieve and use a dataset with original textual conditions, that is where plain text is available natively. Furthermore, we plan to evaluate and compare the generated explanations with the death certificates for which DORIS fails, that is where DORIS is not able to correctly predict the underlying cause of death; extrapolating from the results discussed in this paper, we expect those explanation to be well structured, with an incorrect rule applied, but this needs to be proven with further experimentation. We also plan to use an extend the dataset for the training and evaluation phases by employing the full dataset used by the model that selects the UCOD [11], also by including the cases originated after 2017, where recent cases of COVID-19 may have changed the data distribution and patterns. Finally, we plan to extend the training dataset for generating the explanations with hand written explanations generated by trained experts to be able to compare and support cases where DORIS is not able to correctly identify the UCOD.

Table 3

Qualitative evaluation of the generated explanations for the datasets.

| Certificate | Explanation | Generated Explanation |
|--|---|---|
| Male, 79y old: Unspecified dementia in the context of (Other specified degenerative diseases of nervous system or Alzheimer disease, unspecified or Cerebrovascular disease, unspecified or Mental and behavioural disorders due to use of tobacco: Unspecified mental and behavioural disorder). code: G318 | Unspecified dementia is the first condition reported on the single used line, which is selected as the new tentative starting point. Other specified degenerative diseases of nervous system, Alzheimer disease, unspecified, Cerebrovascular disease, unspecified (HIV) is an obvious cause of the current TUC Unspecified dementia (Infectious Diseases). Other specified degenerative diseases of nervous system is the underlying cause of death. | Unspecified dementia is the first condition reported on the single used line, which is selected as the new tentative starting point. Alzheimer disease, unspecified, Cerebrovascular disease, unspecified (HIV) is an obvious cause of the current TUC Unspecified dementia (Infectious Diseases). Alzheimer disease, unspecified is the underlying cause of death. |
| Male, 69y old: Cardiac arrest, unspecified due to Atherosclerotic heart disease due to Essential (primary) hypertension in the context of (Respiratory failure, unspecified or Subarachnoid haemorrhage, unspecified or Subdural haemorrhage (acute)(nontraumatic) or Epilepsy, unspecified or Immobility). code: I251 | Essential (primary) hypertension present on last used line causes all the conditions reported on the lines above, which is selected as the new tentative starting point (TUC). Special instruction apply where Essential (primary) hypertension is with mention of Atherosclerotic heart disease, Subarachnoid haemorrhage, unspecified, Nontraumatic subdural haemorrhage. Atherosclerotic heart disease is selected using priority order. Atherosclerotic heart disease is the underlying cause of death. | Essential (primary) hypertension present on last used line causes all the conditions reported on the lines above, which is selected as the new tentative starting point (TUC). Special instruction apply where Essential (primary) hypertension is with mention of Atherosclerotic heart disease, Nontraumatic subdural haemorrhage. Atherosclerotic heart disease is selected using priority order. Atherosclerotic heart disease is the underlying cause of death. |
| Male, 54y old: Cardiomegaly due to Atherosclerotic heart disease due to Myocarditis, unspecified due to Generalized and unspecified atherosclerosis in the context of Atherosclerotic cardiovascular disease, so described. code: I251 | Generalized and unspecified atherosclerosis present on last used line causes all the conditions reported on the lines above, which is selected as the new tentative starting point (TUC). Special instruction apply where Generalized and unspecified atherosclerosis is with mention of Atherosclerotic heart disease, Myocarditis, unspecified, Atherosclerotic cardiovascular disease, so described. Myocarditis, unspecified is selected using priority order. Special instruction apply where Myocarditis, unspecified is with mention of Atherosclerotic heart disease, Atherosclerotic cardiovascular disease, so described. Atherosclerotic heart disease is selected using priority order. Atherosclerotic heart disease is the underlying cause of death. | Generalized and unspecified atherosclerosis present on last used line causes all the conditions reported on the lines above, which is selected as the new tentative starting point (TUC). Special instruction apply where Generalized and unspecified atherosclerosis is with mention of Atherosclerotic heart disease, Atherosclerotic cardiovascular disease, so described. Atherosclerotic heart disease is selected using priority order. Atherosclerotic heart disease is the underlying cause of death. |

References

- [1] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, 2019. URL: <https://arxiv.org/abs/1910.10045>. doi:10.48550/ARXIV.1910.10045.
- [2] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, P. Sen, A survey of the state of explainable ai for natural language processing, arXiv (2020). URL: <https://arxiv.org/abs/2010.00711>. doi:10.48550/ARXIV.2010.00711.
- [3] J. Yu, A. I. Cristea, A. Harit, Z. Sun, O. T. Aduragba, L. Shi, N. A. Moubayed, Interaction: A generative xai framework for natural language inference explanations, 2022. URL: <https://arxiv.org/abs/2209.01061>. doi:10.48550/ARXIV.2209.01061.
- [4] R. McAllister, Y. Gal, A. Kendall, M. van der Wilk, A. Shah, R. Cipolla, A. Weller, Concrete problems for autonomous vehicle safety: Advantages of bayesian deep learning, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, 2017, pp. 4745–4753. URL: <https://doi.org/10.24963/ijcai.2017/661>. doi:10.24963/ijcai.2017/661.
- [5] R. Challen, J. Denny, M. Pitt, L. Gompels, T. Edwards, K. Tsaneva-Atanasova, Artificial intelligence, bias and clinical safety, *BMJ Quality & Safety* 28 (2019) 231–237. URL: <https://qualitysafety.bmj.com/content/28/3/231>. doi:10.1136/bmjqs-2018-008370.
- [6] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, 2017. URL: <https://arxiv.org/abs/1706.04599>. doi:10.48550/ARXIV.1706.04599.
- [7] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, 2019. URL: <https://arxiv.org/abs/1910.10045>. doi:10.48550/ARXIV.1910.10045.
- [8] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, M. Rohrbach, Multimodal explanations: Justifying decisions and pointing to the evidence, 2018. URL: <https://arxiv.org/abs/1802.08129>. doi:10.48550/ARXIV.1802.08129.
- [9] S. Kumar, P. Talukdar, Nile : Natural language inference with faithful natural language explanations, 2020. URL: <https://arxiv.org/abs/2005.12116>. doi:10.48550/ARXIV.2005.12116.
- [10] V. Della Mea, M. H. Popescu, K. Roitero, Underlying cause of death identification from death certificates using reverse coding to text and a nlp based deep learning approach, *Informatics in Medicine Unlocked* 21 (2020) 100456. URL: <https://www.sciencedirect.com/science/article/pii/S2352914820306067>. doi:<https://doi.org/10.1016/j.imu.2020.100456>.
- [11] K. Roitero, B. Portelli, M. H. Popescu, V. D. Mea, Dilbert: Cheap embeddings for disease related medical nlp, *IEEE Access* 9 (2021) 159714–159723. doi:10.1109/ACCESS.2021.3131386.
- [12] M. H. Popescu, K. Roitero, S. Travasci, V. Della Mea, Automatic assignment of icd-10 codes to diagnostic texts using transformers based techniques, in: 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI), 2021, pp. 188–192. doi:10.1109/ICHI52183.2021.00037.
- [13] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of

- methods for explaining black box models, *ACM Comput. Surv.* 51 (2018). URL: <https://doi.org/10.1145/3236009>. doi:10.1145/3236009.
- [14] M. T. Ribeiro, S. Singh, C. Guestrin, "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, KDD '16*, Association for Computing Machinery, New York, NY, USA, 2016, p. 1135–1144. URL: <https://doi.org/10.1145/2939672.2939778>. doi:10.1145/2939672.2939778.
- [15] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, P. Blunsom, e-snli: Natural language inference with natural language explanations, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, volume 31, Curran Associates, Inc., 2018. URL: <https://proceedings.neurips.cc/paper/2018/file/4c7a167bb329bd92580a99ce422d6fa6-Paper.pdf>.
- [16] E. Brand, K. Roitero, M. Soprano, G. Demartini, E-bart: Jointly predicting and explaining truthfulness., in: *TTO*, 2021, pp. 18–27.
- [17] E. Brand, K. Roitero, M. Soprano, A. Rahimi, G. Demartini, A neural model to jointly predict and explain truthfulness of statements, *J. Data and Information Quality* (2022). URL: <https://doi.org/10.1145/3546917>. doi:10.1145/3546917.
- [18] C. Agarwal, D. D'souza, S. Hooker, Estimating example difficulty using variance of gradients, 2020. URL: <https://arxiv.org/abs/2008.11600>. doi:10.48550/ARXIV.2008.11600.
- [19] World Health Organization, International statistical classification of diseases and related health problems, 10th revision, Volume 2, https://icd.who.int/browse10/Content/statichtml/ICD10Volume2_en_2016.pdf, 2016. [Online; accessed 21-September-2022].
- [20] World Health Organization, International statistical classification of diseases and related health problems, 11th revision, <https://icd.who.int/en>, 2022. [Online; accessed 21-September-2022].
- [21] M. H. Popescu, C. Celik, V. Della Mea, R. Jakob, Preliminary validation of a rule-based system for mortality coding using ICD-11, *Stud. Health Technol. Inform.* 294 (2022) 679–683.
- [22] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, et al., Exploring the limits of transfer learning with a unified text-to-text transformer., *Journal of Machine Learning Research* 21 (2020) 1–67.
- [23] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: *Text summarization branches out*, 2004, pp. 74–81.
- [24] C.-Y. Lin, F. Och, Looking for a few good metrics: Rouge and its evaluation, in: *Ntcir workshop*, 2004.
- [25] F. Liu, Y. Liu, Correlation between rouge and human evaluation of extractive meeting summaries, in: *Proceedings of ACL-08: HLT, short papers*, 2008, pp. 201–204.