

Using Sentence Embeddings and Semantic Similarity for Seeking Consensus when Assessing Trustworthy AI

Dennis Vetter¹, Jesmin Jahan Tithi², Magnus Westerlund³, Roberto V. Zicari^{3,4} and Gemma Roig¹

¹Goethe University Frankfurt, 60629 Frankfurt am Main, Germany

²Intel Labs, Santa Clara, CA 95054, United States

³Arcada University of Applied Sciences, 00550 Helsinki, Finland

⁴Seoul National University, Seoul 08826, South Korea

Abstract

Assessing the trustworthiness of artificial intelligence systems requires knowledge from many different disciplines. These disciplines do not necessarily share concepts between them and might use words with different meanings, or even use the same words differently. Additionally, experts from different disciplines might not be aware of specialized terms readily used in other disciplines. Therefore, a core challenge of the assessment process is to identify when experts from different disciplines talk about the same problem but use different terminologies. In other words, the problem is to group problem descriptions (a.k.a. issues) with the same semantic meaning but described using slightly different terminologies.

In this work, we show how we employed recent advances in natural language processing, namely sentence embeddings and semantic textual similarity, to support this identification process and to bridge communication gaps in interdisciplinary teams of experts assessing the trustworthiness of an artificial intelligence system used in healthcare.

Keywords

Sentence Embedding, Semantic Similarity, Natural Language Processing, Trustworthy Artificial Intelligence

1. Introduction

The design, development and implementation of artificial intelligence (AI) systems requires knowledge from many different disciplines to be successful. Therefore, the teams involved in AI projects are often interdisciplinary to provide knowledge of all the relevant areas. Each area of expertise comes with its own specialized language, terms, definitions and jargon that can make communication between experts from different fields challenging, as they do not necessarily share the same concepts and may use the same words to mean something different [1]. Additionally, often time, people from one field might not be familiar with specialized terms used in another field. For example, an AI engineer might know the meaning of the terms


1st International Workshop on Imagining the AI Landscape After the AI Act (In conjunction with The first International Conference on Hybrid Human-Artificial Intelligence), June 13, 2022, Amsterdam, The Netherlands

✉ vetter@em.uni-frankfurt.de (D. Vetter); roig@cs.uni-frankfurt.de (G. Roig)

🆔 0000-0002-5977-5535 (D. Vetter); 0000-0002-6439-8076 (G. Roig)



© 2022 Copyright 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)

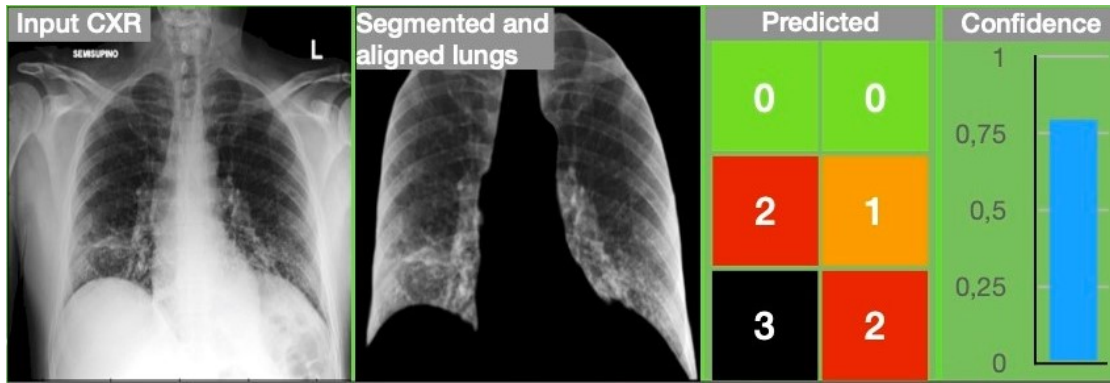


Figure 1: Schematic overview of the AI solution with the sub-tasks segmentation, alignment and Brixia score estimation. For the Brixia score, the lung is separated into 6 regions and each is rated with a number from 0 (no damage) to 3 (high damage) [6]. Image modified from [5].

“precision and recall” whereas a healthcare professional may know the word “prognosis” which the AI engineer might not know about.

One practical example where the interdisciplinary nature of communication shows up is the case where a team of interdisciplinary experts assesses an AI system for its trustworthiness [2, 1, 3, 4]. The stakeholders performing the assessment need to be aware of possible differences in the meaning of specialized terms so that they can understand each other properly. This requires them to cooperate with each other to work on a common vocabulary [2, 1].

In this paper, we show how recent advances in the AI domain of natural language processing (NLP) can be used to support this process. Concretely, we apply it in the assessment of the trustworthiness of an AI system developed to evaluate the degree of lung damage in COVID-19 patients from their chest X-ray (CXR) images. Italian researchers developed the AI system in early 2020 to support the radiologists of a local hospital during the drastically rising cases of COVID-19 that overwhelmed the hospital system [5]. The goal of the system was to provide the radiologists with a qualified second opinion so they can work more confidently, faster, and with fewer mistakes.

The assessed AI system consists of multiple neural networks, one for each of the following sub-tasks: (1) segmentation of the CXR image into lung and background, (2) alignment of the image, and (3) estimation of the semi-quantitative Brixia score. For the Brixia score, the lung is separated into six regions and each region is assigned a number between 0 (no damage) and 3 (highly damaged). This separation into different areas and scoring based on a pre-defined set of values allows for efficient communication between radiologists [6]. A schematic view of the tasks performed by the AI system is given in Fig. 1.

To train the networks, the researchers collected a large dataset of CXR images and annotations by either one radiologist (used for training) or the consensus of multiple radiologists (used for evaluation). Their results show that the AI system is performing equally well as an average human radiologist [5].

The assessment of the above AI system [5] used the Z-Inspection® process described by Zicari et al. [2], which is a holistic approach and includes participation of the entire community of

key stakeholders. For assessing trustworthiness, Z-Inspection[®] builds on the *Ethics Guidelines for Trustworthy AI* by the European Commission’s High-Level Expert Group on AI with the four ethical principles of (i) respect for human autonomy, (ii) prevention of harm, (iii) fairness, and (iv) explicability, which are implemented through the seven key requirements of (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) societal and environmental well-being, and (7) accountability [3].

Part of the assessment is to use socio-technical scenarios [7, 8] to identify different potential issues (ethical, legal, technical, etc) with the system, based on interviews with the whole team, the developers, other stakeholders, and additional materials such as academic papers, source code, datasets that are available. To achieve a disciplinary depth, the group of stakeholders is split into working groups (WGs) according to the different backgrounds of the participants. Each of these WGs then describes what potential issues/problems/tensions (conflicts between two or more desirable goals) they see with the system. This is followed by the *mapping* step, where the issues are structured and connected to the ethical principles and key requirements that they are conflicting with [2]. The goal of this mapping step is to have a description of the issues in “structural ethical terms” [9]. The output of this mapping is then used for *consolidation* where a group-based consensus is reached regarding which issues can be combined and which issues are redundant. This consolidation allows to distill the most critical issues identified about the system; the consolidated statement is then reported to the system’s developers and stakeholders, along with recommendations on possible steps to mitigate the issues or lower their impact. A schematic illustration of this process can be found in Fig. 2.

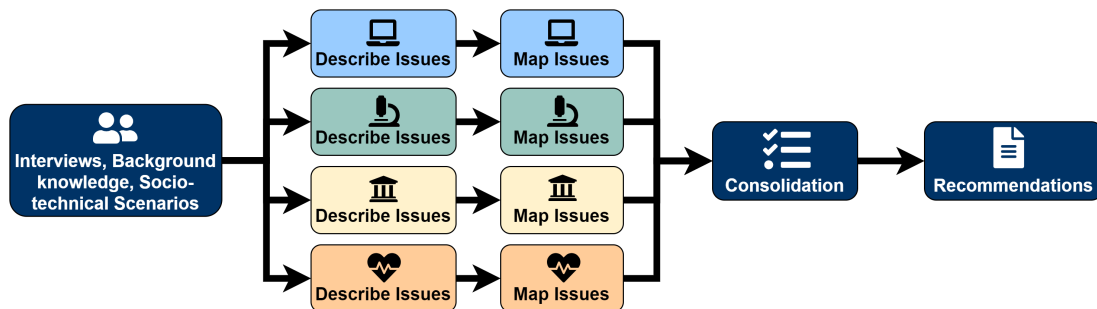


Figure 2: Schematic illustration of the mapping process. First step is to build a common knowledge base to develop socio-technical scenarios. Then the group is separated into WGs, according to the different backgrounds. The results of the WGs are combined in the consolidation step, based on which recommendations to the stakeholders are made. Adapted from [2].

For the assessment, the team consisted of a large number of participants from many different disciplines who described the issues they identified in their own language and jargon from their fields. This resulted in a large number of issues, sometimes talking about similar things from slightly different perspectives using different terminology. According to the participants, the large number made manual consolidation as described in [2, 9] both intellectually challenging and labor-intensive. They described the main difficulty as identifying which issues could be combined. From reports of previous assessments [9, 10, 11], it was considered highly likely that

different issues could be combined as they describe the same tension, but the final number of tensions, as well as the number of issues per tension, were infeasible to estimate.

To help in the consolidation process, we decided to use modern text analysis methods to lift semantic meaning from the text based on the concept of Semantic Textual Similarity (STS) [12]. In NLP, STS is the task of determining the overlap in meaning between texts. The goal of STS is to provide a numerical score where high values indicate that two texts have similar meanings and low values indicate that their meanings are different [12]. In this context, the task of identifying issues that describe the same conflict can be seen as identifying and clustering groups of issues that share high STS scores.

We make the following contributions:

- we show how NLP models can facilitate communication between experts from different domains in trustworthy AI assessment process,
- we present and evaluate two different approaches of STS to group related issues identified by multidisciplinary teams of experts: 1) a clustering-based approach 2) a graph-based approach, both of these use deep learning based STS computation underneath for scoring
- we show that the graph-based approach works comparably well to clustering, while not requiring tuning of hyperparameters.

2. Method

2.1. Word Embeddings and Sentence Embeddings

Currently, the best performing systems for STS are using deep learning-based embeddings. The basic type of embeddings are word embeddings. In word embeddings, a deep neural network is used to map a word into a fixed dimensional vector space. This mapping is done in a way that captures the meaning of the word so that words with similar meanings have similar vector representations, and analogies in word meanings can be approximated by mathematical operations. As an example, with the analogy “king is to queen as man is to woman” the encoding emb_X in the vector space should fulfill the equation $emb_{king} - emb_{queen} \approx emb_{man} - emb_{woman}$ [13, 14, 15].

Sentence embeddings are extensions of word embeddings to complete sentences. Again, deep neural networks are used to map the sentence into a high-dimensional vector space, so that the vector representation also captures the meaning of the sentence [16, 17, 18].

2.2. Measuring Semantic Textual Similarity - STS

After training word or sentence embeddings, the semantic textual similarity of words or sentences is computed from the similarity of their vector representations. A popular metric for this is the cosine similarity. For two words or sentences A and B , this is defined as the cosine of the angle θ between their vector representations emb_A and emb_B :

$$similarity(A, B) := \cos(\theta) = \frac{emb_A^T \cdot emb_B}{\|emb_A\| \cdot \|emb_B\|} \quad (1)$$

The computation of STS scores from embeddings is widely used for a variety of tasks such as checking if similar questions were already asked in a forum or the identification of different topics in large text corpora [18].

2.3. Identifying groups of similar issues

For the identification of groups of similar issues, we compared two approaches: 1) clustering-based and 2) graph-based.

Cluster-based group identification. Separating a set of objects into groups such that objects in the same group have higher similarity and objects in different groups have a lower similarity is the description of a classical clustering problem. Good clustering is best achieved through an iterative process with four key steps: (1) feature selection, (2) cluster identification, (3) cluster validation, and (4) result interpretation. Validation and interpretation are especially important, as algorithms used for cluster identification can always find a division of the objects, but judging whether the division is appropriate and useful, or if a different division should be produced is a decision to be made by the user [19].

In our use-case, feature extraction is performed by creating sentence embeddings that map the English text to a high-dimensional vector. An essential strength of this approach is that it allows us to use raw sentences and does not require any preprocessing. This makes the approach straightforward, especially when compared to other approaches where high-quality results may require extensive preprocessing pipelines and tuning [20, 21]. The following step is to perform dimensionality reduction, as clustering algorithms are known to have problems when working with high-dimensional vectors. We used UMAP [22] to map the high-dimensional embedding vectors to lower dimensions, such that most of the relevant local and global structures in the data are preserved [22]. Compared with other popular dimensionality reduction techniques, UMAP preserves more of the global and local structure of the data than PCA [22], while also producing more compact and better separated clusters than t-SNE [22, 23], which makes it well suited to our task.

The next step is to iteratively use a clustering algorithm and verify and interpret the resulting clusters until a satisfactory result is found. With this approach, the different clusters correspond to the different groups of issues with high similarity. Fig. 3 in the next section shows the output of this approach.

Graph-based group identification. Another approach that works well with data with a similarity measure is spectral clustering [24]. For spectral clustering, the data is arranged in a weighted, fully connected graph, which is called the similarity graph. In the similarity graph, each node corresponds to a data point and the weight of the edge between two nodes to the similarity of the two associated data points. This allows to reformulate the clustering problem into a graph partitioning problem, where the edges between partitions have low weights [24]. A popular variation of the similarity graph is the *k-nearest neighbor graph*. With this variation, a node N_I is connected to another node N_J , if N_J is among the k nearest neighbors of N_I [24].

Applied to the use-case considered here, the nodes in the similarity graph correspond to issues, and the weight of the edge between two nodes corresponds to the cosine similarity

of their embeddings. To simplify the resulting graph, we apply the *1-nearest-neighbor graph* variation, meaning that each node is only connected to the node of it’s most similar issue. With this construction, we found that the similarity graph consists of multiple weakly connected components, groups of connected nodes with no connections between nodes from different groups. This simplified the spectral clustering task to identifying the weakly connected components, which in turn provide the separation into groups of issues with high similarity. In addition, we use the PageRank algorithm [25] to assign importance to each of the nodes, based on the connected nodes and their respective importance. The idea behind this is that nodes with many incoming edges are more important and often better better represent an underlying issue compared to nodes with only one incoming edge. Fig. 4 in the next section shows the output of this approach, the outputs of the two approaches will be compared in the next section.

3. Experiments

In this section, we present the dataset and a subjective evaluation of the results of the two approaches. Code to reproduce our findings is available on GitHub¹.

3.1. Dataset

The dataset was made available to us by the authors of the use-case [26], it contains the issues as described by the different expert WGs in a tabular form. Each issue has the following information: an ID, WG name, a title, and a description. The title is a short summary of the issue, while the description provides additional context; the sentence embedding is computed from a concatenation of both. An example issue is listed in Table 1.

Table 1

Example issue with ID, WG, Title, and Description.

ID	E2
WG	ethics / healthcare
Title	Not all patients may benefit equally from the tool.
Description	The adoption of the system may lead to different care standards for different patient groups.

In total, the dataset consists of 58 issues described by 51 experts in the six working groups: *technical*, *social*, *ethics*, *ethics / healthcare*, *radiologists*, and *healthcare*. Table 2 gives a summary of size and issues described by each WG.

3.2. Evaluation of different sentence embeddings

Computations of sentence embeddings are central to our approaches, as this step implicitly defines the similarities between the issues. It is therefore important to use a well-performing NLP model for this task, for which deep neural networks are state-of-the-art [18]. The implementation

¹<https://github.com/dennisrv/iail2022>

Table 2

Size and number of issues per WG

WG	Members	Issues
technical	21	23
social	5	9
ethics	3	4
ethics / healthcare	4	8
radiologists	3	5
healthcare	15	9
total	51	58

provided by Reimers et al. [18] makes it possible to use a number of different large, pre-trained networks. For our use case, the *all-mpnet-base-v2* network produced the best results. This network uses MPNet, a transformer architecture with 12 layers, 12 attention heads and a hidden size of 768 [27], which was then fine-tuned for general purpose textual similarity tasks using a dataset with over one billion sentence pairs [28]. In general, we could observe a correlation between the subjective quality of the embeddings and the average performance of the network on several NLP tasks, consistent with the findings in [28]. With this network, the sentence embeddings are a 768-dimensional real-valued vector.

3.3. Results of the cluster-based approach

The clustering-based approach required some tuning of parameters to achieve a good result. The best results were achieved with a two-step dimensionality reduction with UMAP, which first reduced the 768 dimensions of the sentence embeddings to 15 and then to 2. Following this step, we performed a clustering with the HDBSCAN algorithm [29], as this algorithm can find a good number of clusters from the data and does not need the desired number of clusters as an input parameters. Fig. 3 shows the results of the clustering approach.

The result contained 12 groups of issues with most of them containing issues from different WGs. As expected, most of the groups were rather small with 3-5 issues and one larger group containing 9 issues. Through manual inspection, we found that most of the group assignments were reasonable, and only few cases of wrongly assigned issues were found. An example of this is that the issue *Transparency would seem to be enhanced if others could have access to the system* was clustered with issues that were about concerns regarding data safety and privacy.

The strength of this approach is that the low-dimensional mapping with UMAP enables a 2D visualization of the clusters and their relative positions. It is therefore possible to identify cases where a manual inspection could identify that both clusters might be about the same topic, as these clusters will be closer to each other. An example of this are clusters 2 and 3 (bottom left) in Fig. 3. These clusters are thematically related; cluster 2 contains issues about privacy in the dataset, while cluster 3 contains issues about data safety and access to the dataset.

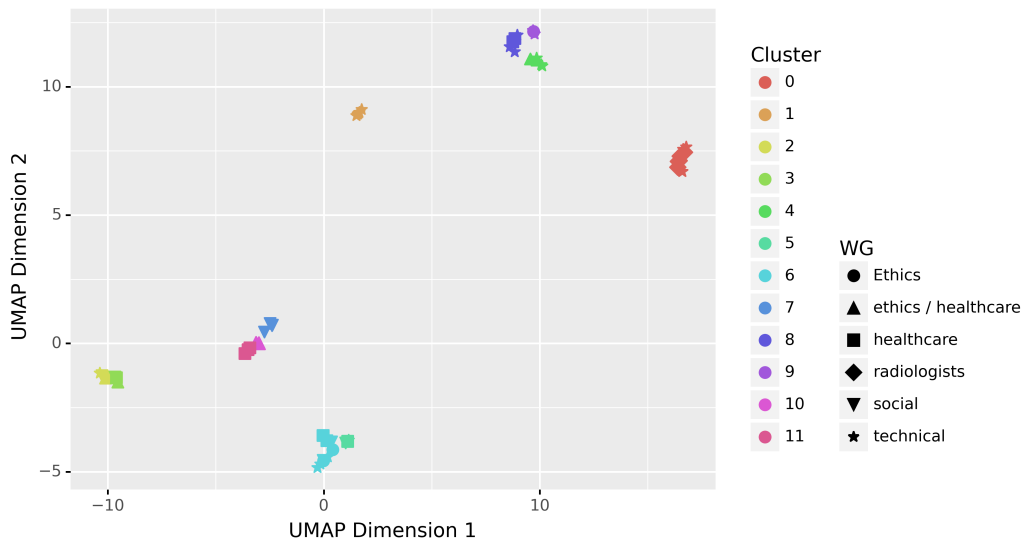


Figure 3: Clustering of issues after using UMAP for a mapping to 2 dimensions that preserves most of the relevant local and global structure in the data.

3.4. Results of the graph-based approach

Our special construction of the similarity graph and the following simplification of the spectral clustering task to the identification of weakly connected components made it possible to us to omit a pre-specification of the number of clusters, a common input parameter for spectral clustering algorithms [24]. Instead, the number of clusters emerged naturally as the number of weakly connected components.

In Fig. 4 we show the results of the graph-based approach. This approach identified 11 groups of issues (i.e., clusters), with more equally distributed sizes compared to the cluster-based approach. Most of the groups also contained issues from at least 2 different WGs. While the result of this approach and the clustering-based approach were not identical, a manual inspection confirmed that it also produced a reasonable grouping of issues.

The strength of this approach is that it does not require tuning. In addition, nodes with high importance were generally found to capture group content well, which facilitated the manual review. An example of this can be seen later in Table 3 where the top issue is the most important and also captures the problem at a more general level.

3.5. Comparison of the approaches

Comparing the two approaches, we could observe that the cluster-based approach seemed to prefer grouping the issues in smaller, more specific groups, such as *concerns about stakeholder inclusion* (4 issues) and *concerns on patient benefits* (3 issues). With the graph-based approach, it was found more likely to combine issues from multiple smaller clusters into one larger group, such as *inclusion of and benefit for patients unclear* (9 issues). The differences in size of produced groups are highlighted in Fig. 5a. Additionally, we found the graph-based approach more

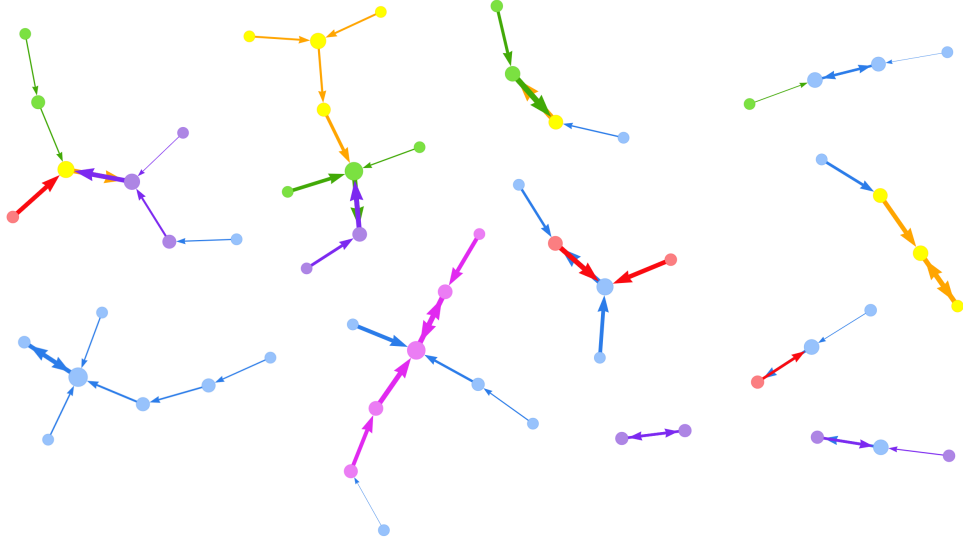


Figure 4: Example of the similarity graph constructed from the issues identified by the experts. The color of a node corresponds to the WG describing the issue. The thickness of the edges is proportional to the similarity of connected issues, the size of nodes is proportional to the importance of the associated issues.

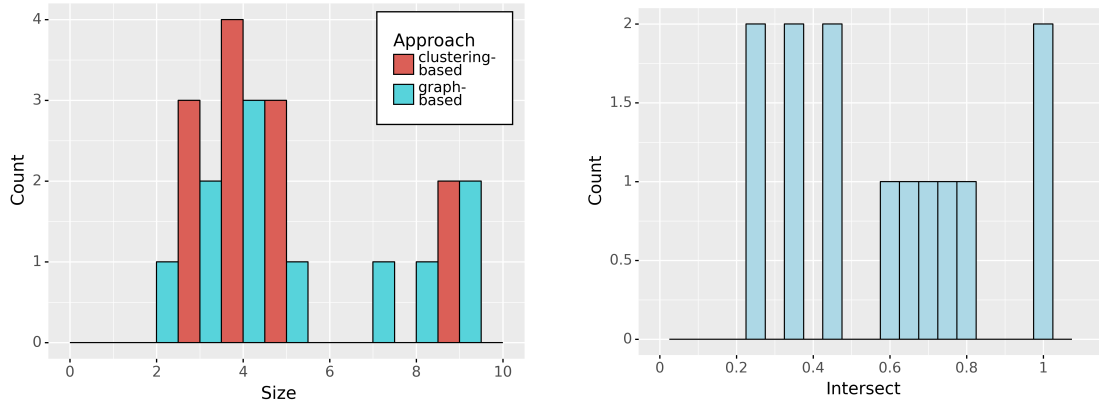
likely to assign issues to groups where we could see no clear connection, although with low importance and therefore easy to identify. Contrary, the clustering-based approach subjectively produced less inappropriate groupings, but the lack of an importance within the cluster made the issues that don't belong more difficult to identify.

Furthermore, we could also observe that the two approaches agreed on which issues belonged to the same group in many cases. While often there was no complete agreement, there was still a high overlap between the assigned groups, as shown in Fig.5b. For this purpose we computed the overlap of sets of issues s_A and s_B as

$$overlap(s_A, s_B) = \frac{|s_A \cap s_B|}{\max(|s_A|, |s_B|)} \quad (2)$$

An example of a group of issues that both approaches agreed on can be seen in Table 3.

For the current assessment [26], we only used the results of the graph-based approach for a pre-screening of the issue groupings during the consolidation phase. However, we plan to use a combination of both clustering approaches for future assessments, as both provide slightly different perspectives and, therefore, are a good start for the discussion between participants. We should also note that in some cases, it was not immediately apparent to the participants whether issues talk about the same problem or not; this could only be solved via discussion and group-consensus.



(a) Sizes of groups identified by the different approaches. (b) Overlap with the most similar group identified by the other approach.

Figure 5: Histogram of group sizes (a) and overlap with the most similar group identified by the other approach (b).

Table 3

Issues that belong to the same group in both approaches. The issues are ordered according to their importance as assigned by the graph-based approach (descending).

WG	Description
technical	The dataset used for training is likely not representative for the general population it is currently used on
ethics	[..] there is no way to know whether diverse demographics receive disparate treatment.
technical	The model is trained on a particular set of devices and software, undermining the reliability in different scenarios and context.

3.6. Limitations

While we found the two approaches to produce sufficient results for our purpose, we could not verify them with data from additional use-cases, as such data was not readily available. In addition, we observed cases where the sentence embeddings put too much importance on single words or phrases. For example, the issues “*Transparency would seem to be enhanced if others could have access to the system*” and “*There is a [data safety] concern if data and software engineers have access to the system and others outside of the medical profession*” were assigned to the same group. While the issues have different meanings, the overlap in words used was sufficient to let them appear “similar enough” to the sentence embedding network. Another occurrence was that all issues containing the word “Score” were grouped together, where some of them were later manually assigned to other groups.

Our proposed solution is to have an iterative process in which discussions with the stakeholders about the results of the grouping are conducted, and to use this approach as a support tool only, and not one that gives the definite answer.

4. Conclusions

Sentence embeddings and semantic textual similarity can be a useful tool for a Trustworthy AI self-assessment to help an interdisciplinary team of experts and stakeholders with identifying possible risks related to the use of an AI system. Our approach was used in practice in a complex use-case with over 50 experts. The approach was used to support initial expert discussions and help build group consensus in a situation where a large number of participants in the assessment made manual consolidation very time-consuming and cumbersome. Participants described it as too demanding for one person to be aware of everyone else's work, making it difficult to find consensus. Instead, our analytical method helped by providing experts with an initial descriptive measure to start the consolidation discussion. Since both modeling approaches presented provided an initial result of sufficient and similar quality, we cannot say that one approach is clearly superior to the other. However, the main advantage of both approaches is that they provide an initial grouping of issues. This initial grouping made it much easier to understand the different questions and helped the experts to get a broad picture of the work done by other groups. Because the groupings of questions share a common semantic topic, it was also easier to identify errors in the algorithmic approach and to identify groupings of questions that might belong together.

To summarize, in the eyes of the participants, the main strength of our method was that it improved their ability to effectively participate in the communication and focus on contributing to the assessment process. In future assessments, we plan to further validate this approach for consolidation and to investigate with a panel of stakeholders which of the two approaches is more effective for finding consensus.

Acknowledgments

DV received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 101016233 (PERISCOPE), and from the European Union's Connecting Europe Facility program under grant agreement no. INEA/CEF/ICT/A2020/2276680 (xAIM). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

- [1] J. Whittlestone, R. Nyrup, A. Alexandrova, K. Dihal, S. Cave, Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research, Nuffield Foundation, London, 2019. URL: <https://www.nuffieldfoundation.org/wp-content/uploads/2019/02/Ethical-and-Societal-Implications-of-Data-and-AI-report-Nuffield-Foundat.pdf>.
- [2] R. V. Zicari, J. Brodersen, J. Brusseau, B. Düdler, T. Eichhorn, T. Ivanov, G. Kararigas, P. Kringen, M. McCullough, F. Möslein, N. Mushtaq, G. Roig, N. Stürtz, K. Tolle, J. J. Tithi, I. van Halem, M. Westerlund, Z-Inspection®: A Process to Assess Trustworthy AI, IEEE Transactions on Technology and Society 2 (2021) 83–97. doi:10.1109/TTS.2021.3066209, conference Name: IEEE Transactions on Technology and Society.

- [3] High-Level Expert Group on Artificial Intelligence, Ethics guidelines for trustworthy AI, Text, European Commission, 2019. URL: <https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1>.
- [4] High-Level Expert Group on Artificial Intelligence, Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment, Text, European Commission, 2020. URL: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342.
- [5] A. Signoroni, M. Savardi, S. Benini, N. Adami, R. Leonardi, P. Gibellini, F. Vaccher, M. Ravanelli, A. Borghesi, R. Maroldi, D. Farina, BS-Net: Learning COVID-19 pneumonia severity on a large chest X-ray dataset, *Medical Image Analysis* 71 (2021) 102046. URL: <https://www.sciencedirect.com/science/article/pii/S136184152100092X>. doi:10.1016/j.media.2021.102046.
- [6] A. Borghesi, A. Zigliani, S. Golemi, N. Carapella, P. Maculotti, D. Farina, R. Maroldi, Chest X-ray severity index as a predictor of in-hospital mortality in coronavirus disease 2019: A study of 302 patients from Italy, *International Journal of Infectious Diseases* 96 (2020) 291–293. URL: <https://linkinghub.elsevier.com/retrieve/pii/S1201971220303283>. doi:10.1016/j.ijid.2020.05.021.
- [7] J. Leikas, R. Koivisto, N. Gotcheva, Ethical Framework for Designing Autonomous Intelligent Systems, *Journal of Open Innovation: Technology, Market, and Complexity* 5 (2019) 18. URL: <https://www.mdpi.com/2199-8531/5/1/18>. doi:10.3390/joitmc5010018, number: 1 Publisher: Multidisciplinary Digital Publishing Institute.
- [8] F. Lucivero, Ethical Assessments of Emerging Technologies: Appraising the moral plausibility of technological visions, number 15 in *The International Library of Ethics, Law and Technology*, 1st ed., 2016 ed., Springer International Publishing : Imprint: Springer, Cham, 2016. doi:10.1007/978-3-319-23282-9.
- [9] J. Brusseau, What a Philosopher Learned at an AI Ethics Evaluation, *AI Ethics Journal* 1 (2020). URL: <https://www.aiethicsjournal.org/10-47289-aij20201214>. doi:10.47289/AIEJ20201214.
- [10] R. V. Zicari, J. Brusseau, S. N. Blomberg, H. C. Christensen, M. Coffee, M. B. Ganapini, S. Gerke, T. K. Gilbert, E. Hickman, E. Hildt, S. Holm, U. Kühne, V. I. Madai, W. Osika, A. Spezzatti, E. Schnebel, J. J. Tithi, D. Vetter, M. Westerlund, R. Wurth, J. Amann, V. Antun, V. Beretta, F. Bruneault, E. Campano, B. Düdder, A. Gallucci, E. Goffi, C. B. Haase, T. Hagendorff, P. Kringen, F. Möslein, D. Ottenheimer, M. Ozols, L. Palazzani, M. Petrin, K. Tafur, J. Tørresen, H. Volland, G. Kararigas, On Assessing Trustworthy AI in Healthcare. Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls, *Frontiers in Human Dynamics* 3 (2021) 30. URL: <https://www.frontiersin.org/article/10.3389/fhumd.2021.673104>. doi:10.3389/fhumd.2021.673104.
- [11] R. V. Zicari, S. Ahmed, J. Amann, S. A. Braun, J. Brodersen, F. Bruneault, J. Brusseau, E. Campano, M. Coffee, A. Dengel, B. Düdder, A. Gallucci, T. K. Gilbert, P. Gottfrois, E. Goffi, C. B. Haase, T. Hagendorff, E. Hickman, E. Hildt, S. Holm, P. Kringen, U. Kühne, A. Lucieri, V. I. Madai, P. A. Moreno-Sánchez, O. Medlicott, M. Ozols, E. Schnebel, A. Spezzatti, J. J. Tithi, S. Umbrello, D. Vetter, H. Volland, M. Westerlund, R. Wurth, Co-Design of a Trustworthy AI System in Healthcare: Deep Learning Based Skin Lesion Classifier, *Frontiers in Human Dynamics* 3 (2021) 40. URL: <https://www.frontiersin.org/article/10.3389/fhumd.2021.688152>. doi:10.3389/fhumd.2021.688152.

- [12] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, L. Specia, SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1–14. URL: <http://aclweb.org/anthology/S17-2001>. doi:10.18653/v1/S17-2001.
- [13] J. Pennington, R. Socher, C. Manning, Glove: Global Vectors for Word Representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543. URL: <http://aclweb.org/anthology/D14-1162>. doi:10.3115/v1/D14-1162.
- [14] T. Mikolov, W.-t. Yih, G. Zweig, Linguistic regularities in continuous space word representations, in: Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies, 2013, pp. 746–751.
- [15] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed Representations of Words and Phrases and their Compositionality, in: Advances in Neural Information Processing Systems, volume 26, Curran Associates, Inc., 2013. URL: <https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html>.
- [16] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, A. Bordes, Supervised Learning of Universal Sentence Representations from Natural Language Inference Data, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 670–680. URL: <https://aclanthology.org/D17-1070>. doi:10.18653/v1/D17-1070.
- [17] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, R. Kurzweil, Universal Sentence Encoder, arXiv:1803.11175 [cs] (2018). URL: <http://arxiv.org/abs/1803.11175>, arXiv: 1803.11175.
- [18] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, arXiv:1908.10084 [cs] (2019). URL: <http://arxiv.org/abs/1908.10084>, arXiv: 1908.10084.
- [19] R. Xu, D. C. Wunsch, Survey of Clustering Algorithms, IEEE TRANSACTIONS ON NEURAL NETWORKS 16 (2005) 35.
- [20] V. Srividhya, R. Anitha, Evaluating preprocessing techniques in text categorization, International journal of computer science and application 47 (2010) 49–51. URL: http://sinhgad.edu/ijcsa-2012/pdppapers/1_11.pdf.
- [21] D. S. Vijayarani, J. Ilamathi, Nithya, Preprocessing Techniques for Text Mining - An Overview, International Journal of Computer Science & Communication Networks 5 (2015) 11.
- [22] L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, arXiv:1802.03426 [cs, stat] (2020). URL: <http://arxiv.org/abs/1802.03426>, arXiv: 1802.03426.
- [23] D. Kobak, G. C. Linderman, Initialization is critical for preserving global data structure in both t-SNE and UMAP, Nature Biotechnology 39 (2021) 156–157. URL: <https://www.nature.com/articles/s41587-020-00809-z>. doi:10.1038/s41587-020-00809-z, number: 2 Publisher: Nature Publishing Group.
- [24] U. von Luxburg, A tutorial on spectral clustering, Statistics and Computing 17

- (2007) 395–416. URL: <http://link.springer.com/10.1007/s11222-007-9033-z>. doi:10.1007/s11222-007-9033-z.
- [25] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web., Technical Report 1999-66, Stanford InfoLab, 1999. URL: <http://ilpubs.stanford.edu:8090/422/>, backup Publisher: Stanford InfoLab.
- [26] H. Allahabadi, J. Amann, I. Balot, A. Beretta, C. Binkley, J. Bozenhard, F. Bruneault, J. Brusseau, S. Candemir, L. A. Cappellini, S. Chakraborty, N. Cherciu, C. Cociancig, M. Coffee, I. Ek, L. Espinosa-Leal, D. Farina, G. Fieux-Castagnet, T. Frauenfelder, A. Gallucci, G. Giuliani, A. Golda, I. van Halem, E. Hildt, S. Holm, G. Kararigas, S. A. Krier, U. Kühne, F. Lizzi, V. I. Madai, A. F. Markus, S. Masis, E. Wiinblad Mathez, F. Mureddu, E. Neri, W. Osika, M. Ozols, C. Panigutti, B. Parent, F. Pratesi, P. A. Moreno-Sánchez, G. Sartor, M. Savardi, A. Signoroni, H. Sormunen, A. Spezzatti, A. Srivastava, A. F. Stephansen, B. T. Lau, J. J. Tithi, J. Tuominen, S. Umbrello, F. Vaccher, D. Vetter, M. Westerlund, R. Wurth, R. V. Zicari, Assessing Trustworthy AI in times of COVID-19. Deep Learning for predicting a multi-regional score conveying the degree of lung compromise in COVID-19 patients., Preliminary manuscript made available by the authors (2022).
- [27] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, MPNet: Masked and Permuted Pre-training for Language Understanding, arXiv:2004.09297 [cs] (2020). URL: <http://arxiv.org/abs/2004.09297>, arXiv: 2004.09297.
- [28] N. Reimers, Pretrained Models – Sentence-Transformers documentation, no date. URL: https://www.sbert.net/docs/pretrained_models.html, accessed: 2022-01-11.
- [29] L. McInnes, J. Healy, S. Astels, hdbscan: Hierarchical density based clustering, The Journal of Open Source Software 2 (2017) 205. URL: <http://joss.theoj.org/papers/10.21105/joss.00205>. doi:10.21105/joss.00205.