Towards Explainable AI: Assessing the Usefulness and Impact of Added Explainability Features in Legal Document Summarization

Milda Norkute, Nadja Herger, Leszek Michalak, Andrew Mulder, Sally Gao

Thomson Reuters Labs

ABSTRACT

This study tested two different approaches for adding an explainability feature to the implementation of a legal text summarization solution based on a Deep Learning (DL) model. Both approaches aimed to show the reviewers where the summary originated from by highlighting portions of the source text document. The participants had to review summaries generated by the DL model with two different types of text highlights and with no highlights at all. The study found that participants were significantly faster in completing the task with highlights based on attention scores from the DL model, but not with highlights based on a source attribution method, a model-agnostic formula that compares the source text and summary to identify overlapping language. The participants also reported increased trust in the DL model and expressed a preference for the attention highlights over the other type of highlights. This is because the attention highlights had more use cases, for example, the participants were able to use them to enrich the machine-generated summary. The findings of this study provide insights into the benefits and the challenges of selecting suitable mechanisms to provide explainability for DL models in the summarization task.

CCS CONCEPTS

• **Human Centered Computing** → Human Computer Interaction • **Computing Technologies** → Artificial Intelligence; Natural Language Processing

KEYWORDS

explainable artificial intelligence, interpretable machine learning, abstractive summarization

ACM Reference format:

Milda Norkute, Nadja Herger, Leszek Michalak, Andrew Mulder, Sally Gao. 2020. Towards Explainable AI: Assessing the Usefulness and Impact of Added Explainability Features in Legal Document Summarization. *In CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI'21 Extended Abstracts), May 8–13, 2021, Yokohama, Japan. ACM, New York, NY, USA, 7 pages*. https://doi.org/10.1145/3411763.3443441

CHI '21 Extended Abstracts, May 08–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8095-9/21/05.
https://doi.org/10.1145/3411763.3443441

1 Introduction

With recent advances in artificial intelligence (AI) and machine learning (ML), digital technology is increasingly integrating automation through algorithmic decision-making. However, balancing the powerful capabilities provided by ML with the need to design technology that people feel empowered by is a challenge. Understanding how technology may affect users is important for trusting it and feeling in control [10]. To help achieve this, machine learning algorithms need to be able to explain how they arrive at their decisions.

There has been increased attention given to interpretable, fair, accountable and transparent algorithms in the AI and ML communities [14]. In 2016, the European Union approved a data protection law known as the General Data Protection Regulation or "GDPR" [5] that includes a "right to explanation". The need for decisions made by AI to be explainable is also often present in the AI principles of the organizations that build products containing AI features [1, 6, 13]. This means that AI practitioners have to look for concrete ways to explain the decisions made by their AI models in different contexts and use cases.

One such use case is abstractive text summarization, the technique of generating a summary of a text from its main ideas, where the generated summary potentially contains new phrases and sentences that may not appear in the source text [8]. Different from extractive summarization, which refers to the process of extracting words and phrases from the text itself to create a summary, abstractive summarization closely resembles the way humans write summaries [9]. Due to its complexity, it relies on advances in Deep Learning (DL) to be successful. In this project we explored how we could increase explainability for a DL-driven abstractive text summarization model as part of a legal editorial tool.

1.1 Background

The legal editorial tool is used by a team of editors who monitor and collect new court cases and perform various editorial tasks. For example, the editors read the case and write a short summary of the allegations made by the

plaintiffs. The editors, who are trained lawyers, follow guidelines on how to write this summary. All the necessary information for this task is available in the original court case, which can range from 10 to 100 pages.

To speed up the editorial process, an AI-powered summarization model was built and integrated into the editorial tool to automatically generate an allegations summary for each case. Since this AI model has been in active use, the primary task of the editors has become to review and edit the machine-generated summaries rather than creating them from scratch based on the long input documents. However, to validate the machine-generated summaries, the editors must still review the entire court case manually. Identifying the elements of the court case that were included in the machine-generated summary is impossible without reviewing the whole case.

In this work, we implemented and studied explainability enhancements to this process. We studied two major questions: (i) Does the explainability feature reduce the time spent on validating the automated allegations summary? (ii) Does the explainability feature increase the editor's trust in the AI system?

2 Explainability for Summarization

2.1 Summarization model

To automatically generate summaries of the allegations, we used a Pointer Generator network [11] as implemented in the OpenNMT-Py toolkit [7]. The model was trained from scratch based on 800'000 court cases and associated editor-written summaries converted into digital text using Optical Character Recognition (OCR). The text was tokenized at a word level and only the first 4800 tokens were considered for model training. This is adequate for the task as allegations are typically present within the first part of the court case. The target sequences were editor-written summaries of the allegations in the court case. Apart from standard evaluation metrics like ROUGE, we measured the quality of the summarization model in a blind evaluation experiment conducted with editors. In the evaluation, we asked editors whether a summary, either generated by the model or by another editor, without them knowing which, would be publishable with minor edits. We measured that 75+/-10% of the summaries produced by the model were publishable with minor edits, compared to 88+/-10% of the summaries produced by editors from scratch. From this measurement, we concluded that the editors should review and (if needed) enhance each summary before it is published instead of pursuing a fully automated approach.

2.2 Explainability Method 1: Attention Vector Approach

The attention vector approach makes use of additional metadata produced by the Pointer Generator model in the process of generating the summary. At inference time, the model predicts the next word as part of the summary as well as the attention distribution for it. At the time of the summary generation, we are therefore able to extract an attention matrix of dimension T x S from the DL model, where S is the size of the source text and T is the length of the generated summary [2, 12] (see Figure S1 in the supplementary materials).

After obtaining a generated summary from a given source text, we computed averaged attention scores per word in the source text to end up with a vector of size 1 x S. Minimal post-processing of the attention scores were needed to use them for visualization purposes. A rolling window of 5 words for smoothing was applied, as well as min-max scaling to ensure that the minimum attention score per court document was 0 and the maximum value was 1. Attention values were then visualized through the use of opacity. We extracted the attention scores for the first 4800 tokens that were used to train the model and set the scores to 0 for the remainder of the tokens as we had no attention values available for them. We refer to this approach as '*attention highlights*' in this paper.

2.3 Explainability Method 2: Source Attribution Approach

In addition to the attention vector approach, we tested the source attribution approach. Different from the previous approach, the source attribution approach relies on a heuristic independent of the Pointer Generator model. Given the machine-generated summary, it allows us to identify one or more sentences in the source text which the summary's language primarily comes from. One or more sentences in the summary is assigned a value of 1 (= text that had influence on the summary), based on the collection of sentences that yields the highest normalized bigram precision score against the source text without excessively sacrificing recall. All remaining sentences are assigned a value of 0 (= text that had no influence on the summary). Those values can then be turned into highlights used for visualization purposes (0 = no highlights, 1= highlights). The source attribution approach was introduced as a benchmark to the attention vector approach. We refer to this approach as *'source highlights'* in this paper.

3 Research Goals

Overall, we wanted to test what effect, if any, the two explainability approaches would have on the interactions of the editors with the outputs of the summarization model. More specifically we wanted to learn if the added explainability feature would help the editors understand how the summary gets created and to trust the AI more. In addition to this, we wanted to explore if this would have any effect on the performance of the editors, namely, if the added

explainability could help them review and edit the summaries faster. If so, we wanted to find out why and how they would use the highlights. Finally, we investigated if the editors had any preferences for one of the two explainability approaches and the reasons for this.

4 Method

This was meant to be a rapid pilot study, hence why we recruited only two editors to take part in it. They were experienced editors, one male and one female, aged between 40 and 60, who worked with the editorial tool for more than 10 years. They were already familiar with the AI summarization functionality and used it for around 6 months before the study began.

Figure 1. The user interface (UI) built for testing, no highlights condition. (a) Link to the full case source text in the original PDF format. (b) A text field area to view and edit the summary generated by the model. (c) Submit button which stopped the timer and made a pop up appear where the participant could leave comments before the next case was displayed. (d) Source text viewer displaying the OCR'd source document of the case. (e) Jump to page functionality - each rectangle represents a page of the source document

The designs created to present the three editing conditions – without highlights, source highlights and attention highlights - looked very similar. Our goal was to ensure that the conditions in which the participants edited the summaries were as comparable as possible. The key difference between the editing conditions was the presence or the lack of highlighting as well as the type of highlights that the participants saw. See Figures 1, 2 and 3 for detailed illustrations and explanations of the elements in the interface for each testing condition.

A web app was built for testing purposes, which the participants used to review and edit the summaries. All responses were logged in the web app. Each participant's work was logged separately. When starting the editing task, the participant had to log in with a user identity (user ID). The user ID controlled in which of the 3 conditions the participants had to work and which summaries they had to review. We used the app to log how long it took the participants to validate, and, if needed, to edit the summary for each case. The timer started as soon as the page with the summary and the case was rendered to the screen. The end time was logged once the submit button (Figure 1c) was clicked. The participants did not know that their performance was timed, but they had been informed that diagnostic data would be captured as part of the study.

All the testing sessions were conducted remotely. At the start of each session, an introductory video call using Microsoft Teams was run with the objective of explaining the tasks and their order to the participant. Both participants were

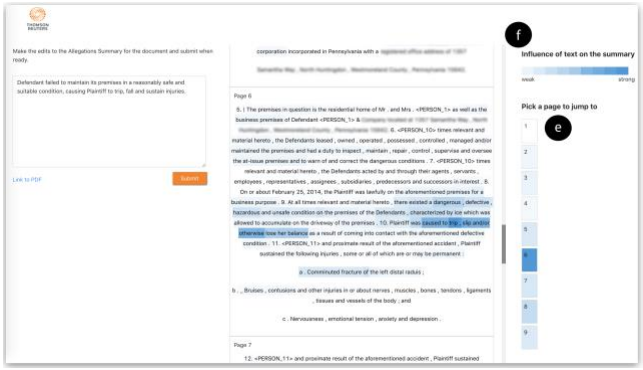

Figure 2. The UI built for testing, attention highlights condition. The highlights in the full case text viewer have many shades of blue. (e) An average attention score was computed per page and min-max scaling was applied to normalize the scores for the Jump-to-page functionality. As a result, the higher the score, the more intense the shade of the highlighting of the page in the index. (f) A scale was added to the UI to explain how to read the highlights.

instructed to edit the summaries of allegations using the web app as they would normally but were asked not to take breaks while editing the summaries. All communication during the session after the introductory video call was done using the Microsoft Teams chat functionality. The participant informed the researcher that they had completed the task by sending a chat message and then the researcher messaged the participant informing them what task should be completed next and sent the needed materials for the task. The interviews were also done through Microsoft Teams, using the video call functionality.

Most court cases enter the editorial system in a PDF format, but to perform the analysis, the documents needed to be converted to a machine-readable text format. We used an

Figure 3. The UI adjusted to display the source highlights condition. The highlights in the full case text viewer have only one shade of blue. (e) Pages in the index were either highlighted as relevant or not highlighted at all. A page was assigned a value of 1 if at least one word per page had a highlight, otherwise the page was assigned a value of 0. (f) Explanation how to interpret the highlights.

open-source OCR library to convert what is visible on the PDF into text. The OCR metadata included some layout information for each word which was used to display the source document text of the case in the web app. We included the possibility for the participants to view the source document of the case in its original PDF format to account for the possible issues with the OCR quality as we had a limited sample of cases to use for the study. However, during the study the participants reported that they used the PDFs when either the summary or highlights were unhelpful rather than when having issues with the OCR quality.

We performed two rounds of testing in total. Each round consisted of two testing sessions, one with each of the two participants. We describe each of the rounds of testing below.

In round 1, we tested whether having attention highlights would be preferable to working with no highlights. The first participant saw 11 cases with summaries (batch 1) with attention highlights and 11 cases not previously seen (batch 2) without highlights. The participant first performed the task with highlights. This was reversed for the second participant who first reviewed 11 cases with summaries from batch 1 without highlights and then 11 cases from batch 2 with attention highlights. This setup, where the participants had to review multiple summaries in one condition before switching to another, was chosen to allow the participants to experience the differences between working in two conditions. Furthermore, it enabled us to compare time needed to review each summary with and without highlights while ensuring each of the participants did not see the same case twice as this could have biased the time measurements and to account for differences in editing speeds between the participants. The two batches of cases were balanced in terms of their document length and case type (batch 1 M=16.18 pages, batch 2 M=16.64 pages). After completing each batch, the participants completed a survey. After all editing tasks were completed, an interview was conducted, where the participants were asked to reflect on their experiences when editing the summaries in different conditions.

Round 2 of testing was run three weeks after round 1 was completed. In this round, the first participant had to review 11 cases with summaries (batch 3) not previously seen in round 1 with highlights created using the source attribution approach (source highlights) and 11 cases with summaries not seen previously without highlights (batch 4). This was reversed for the second participant. Groups of cases in batches 3 and 4 were balanced in terms of their length and case type (batch 3 M=14.64 pages, batch 4 M=15.18 pages). After completing the editing task with source highlights, the participant completed the same survey that was used in round 1 of testing. Once these editing tasks (with source and no highlights) were completed, an interview was conducted to get insights on what the editing experience was like. After this, the participant was asked to review four more cases, previously seen, with attention highlights. This was done to give a reminder of what it was like to work in this setup, as the participants had seen the attention highlights three weeks prior. Finally, one additional interview was conducted, focused on comparing the two types of highlights. For an illustration of the experimental setup for both rounds of testing, see Figure S2 in the supplementary materials.

All interviews were semi-structured. The survey developed consisted of several statements taken from the system usability scale questionnaire (SUS) [4]. This was done to capture the impact of the highlights on the usability of the system. Additional statements on trust and explainability were added such as "I trusted the summaries generated by the system". The participants were asked to score these items with one of five responses that range from Strongly Agree to Strongly Disagree, the same as the SUS items.

5 Analysis

We compared the time spent working on cases between all three conditions, using the data logged in the web app while participants were completing the tasks. We also analyzed the qualitative data gathered in the interviews. The interviews were transcribed and analyzed by the researcher who also ran all of the testing sessions. Using the transcriptions, the researcher used thematic analysis [3] to identify and provide a coding schematic that was used to analyze the participants' comments. Affinity diagrams were used to cluster information generated in the coding phase to obtain additional insights. This revealed themes which were presented and discussed with the other researchers before the final themes were agreed on. The number of responses collected using the survey was too small for a more meaningful analysis, hence some of the individual items of the survey were used to

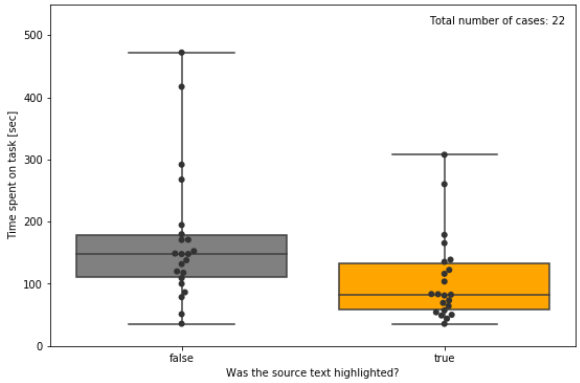

Figure 4. The distribution of time spent on each case by the participants with added explainability (orange boxplot) and the condition without any highlighting (grey boxplot). A plot of round 1 time on task comparison: attention highlight vs no highlight. Black dots represent the individual cases.

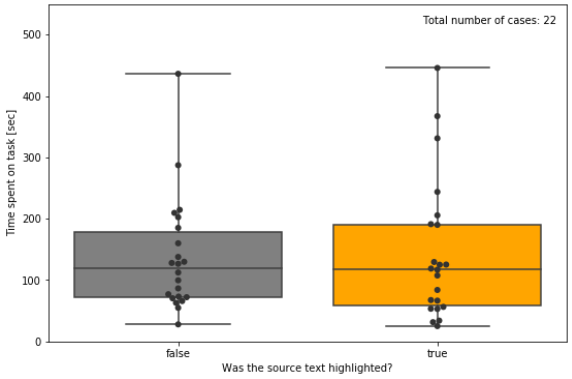

Figure 5. The distribution of time spent on each case by the participants with added explainability (orange boxplot) and the condition without any highlighting (grey boxplot). A plot of round 2 time on task comparison: source highlight vs no highlight. Black dots represent the individual cases.

validate the interview findings and are only briefly mentioned in the results section where relevant.

6 Results

6.1 Time on Task

A one-sided Wilcoxon Signed-Ranks test indicated that the difference in time required to edit the summaries with attention highlights (*N*=22, *Mdn*=82.89s) compared to the control, no highlights, condition (*N*=22, *Mdn*= 147.95s) was statistically significant (*W*=40.0, *p*=0.0018). The median percentage of time saved was 37.0% when working with attention highlights (see Figure 4). Time savings were calculated for each case using the formula (time_Highlight – time_noHighlight)/time_noHighlight. The median across those values was then reported.

The median percentage of time saved using the source highlights, calculated using the same formula as above was only 0.03%. According to a one-sided Wilcoxon Signed-Ranks test, the improvement in time saved when editing the summaries with source highlights (*N*=22, *Mdn*=117.9s) compared to the control, no highlights, condition (*N*=22, *Mdn*=119.5s) was not statistically significant (*W*=131.0, *p*=0.56), see Figure 5.

Our sample size was too small to make it worthwhile to run any statistical significance tests on whether the speed up was consistent across the two participants. Instead, we report here on some descriptive statistics. Both participants were faster when editing the summaries with attention highlights (Participant 1: *N*=11, *Mdn*=73.3s; Participant 2: *N*=11, *Mdn*=83.5s) than when working without highlights (Participant 1: N=11, Mdn=131.7s; Participant 2: N=11, Mdn=170.7s). The impact of source highlights on each participant's performance is less clear. Both participants seemed to have been slower when working with the source highlights (Participant 1: N=11, Mdn=117.0s; Participant 2: N=11, Mdn=129.5s) than when working without highlights (Participant 1: N=11, Mdn=72.2s; Participant 2: N=11, Mdn=128.0s). For a visual representation of those results we refer to Figures S3 and S4 in the supplementary materials.

6.2 Confidence and Trust

Overall, the participants reported that the attention highlights increased their confidence in the machine-generated summary. This was also reflected in the survey responses, where both participants agreed with the statement "I trusted the summaries generated by the system" for attention highlights, but stated they were neutral towards this statement for the source highlights. Seeing different shades of blue on more words in the attention highlights condition gave participants the sense that "*it* [the AI summarization system] *did look at whole case so even if it got it wrong made me trust it more".* One participant also reported: *"I saw all light blue, saw no injuries - I felt confident: if injuries were there they'd been caught by light blue*". The injuries sustained by a plaintiff are typically included in a summary, if they are the subject of the case, and the editor here was referring to the fact that the highlighting would have overlapped with a passage in case text that discussed the injuries. Participants identified parallels between their work and attention highlights. They said that this was similar to how they would approach the task – scan the whole case to identify key elements and relevant details to include in the summary.

However, the participants did not report having the impression that the whole case was checked when working

with the source highlights. This is because this method only highlighted the source text that matched the text in the summary as closely as possible. They said: *"So do I just trust that one section that was highlighted? Doesn't feel like I've done my full job as a reviewer and made sure it picked up the right section and included the full details. Sometimes I spent more time making sure that one highlighted section was correct".* Furthermore, according to the participants' observations, the source highlights were not often completely 'correct' – i.e., they did not match their expectations about the relevance of the highlighted text for the summary. Each instance of incorrect highlights reduced their trust in the system and its capability to take them to parts in the text that mattered: *"If you just show one highlight and it's wrong and it feels more wrong. I start to lose confidence"*. By contrast, with attention highlights, where they felt confident that the whole text had been looked at, mistakes seemed more forgivable to the participants. Both participants reported realizing that AI cannot always be correct as some summaries are also challenging for human writers. Based on these findings, we believe attention highlights corresponded more closely to the editor's mental model of the task.

The higher levels of trust attributed to attention highlights is likely one of the reasons why the participants were more efficient when working with attention highlights compared to source highlights. As they trusted the source highlights less, the participants felt the need to check the correctness of both the highlighting and the summary, which resulted in more time spent on the task. This was not the case when working with attention highlights.

6.3 Usage of Highlights

The source highlights were primarily used to get to the area of the document where the relevant details might be. The highlights themselves were not very useful and were therefore not used to add additional details to the summary: *"*[the source highlights] *Often didn't get me to the main point that frequently but took me close to it. They took me to the right document area. Helped me target where I was looking. But I didn't get the detail needed".* This, however, was possible with attention highlights: "[…] *the* [attention] *highlights* [light blue] *included more important information I needed to review, confirm and maybe add to summary*". It turned out that the light blue highlighting in attention highlights, indicative of moderate levels of attention by the DL model, also carried relevant information to the participants, in addition to the darkest blue, which typically showed the key information, often matching what was already included in the AI-generated summary. The participants, therefore, sometimes used the text that had been highlighted with lighter blue as additional detail to be added to the summary. In addition to this, the attention highlights also turned out to be useful even in a situation when the AI-generated summary was not satisfactory based on the editor's assessment: "*One time the summary was wrong – it said that this was a car accident case, but it actually was an insurance case. However, the highlighted text was still mostly correct and useful to me in realizing and correcting this*". Apparently in that case, the details the model had paid attention to in the source text were relevant to include in the summary and were useful for understanding what the case was about. However, the conclusion the AI system made about the type of case in the summary was incorrect and this required a correction. Therefore, overall, the attention highlights could be used by the participants in more ways beyond just being able to take the participant to a highly relevant part of the document which was also possible with source highlights.

6.4 Usage of Highlights

Both participants reported a perceived speed up when working with the attention highlights: "*the* [attention] *highlights not only made me faster … they also made the* [editing] *experience more enjoyable*". The participants were indeed significantly faster with attention highlights than when working without highlights as revealed by our analysis of the web app logs. In addition to this, both participants said that they would prefer to work with source highlights rather than with no highlights, however, both users expressed a strong preference for working with attention highlights over source highlights and found the attention highlights more useful. This was also confirmed by the responses to the survey items – for example, both participants strongly agreed with the statement "I found the features of the system useful for editing the summary" for attention highlights, as opposed to only agreeing with it for source highlights. See Figure S5 in the supplementary materials for an illustration of the responses to other survey items.

7 Conclusions

While this study was limited in sample size and ideally a similar study should be run with more participants and a larger body of sample data, we are confident that the attention highlight explainability feature that we tested had a positive impact on the editorial process. Our findings suggest attention highlights worked much better as an explainability feature for this use case than the source highlights.

Based on this, we conclude that not all methods to enable explainability are equal in the benefits they create. The attention highlights were created based on attention scores produced by the summarization model and therefore represented the model's decision making more closely than the source attribution approach which was created to be independent from the summarization model. Thus, on one hand we can say that it is important to represent the model's decision making as closely as possible for the explanation to be useful. However, we also learned that according to the participants, attention highlights worked similarly to how

they approach the task of reviewing the summaries without explainability. Thus, the explainability method chosen should ideally not only represent the decision making performed by the DL model, but also aim to match the user's mental model of the task as closely as possible.

As demonstrated by our study, a suitable explainability method can not only increase trust in the AI system, but also result in better overall work experience for the users interacting with an AI system. Furthermore, the addition of explainability features can also create measurable benefits in terms of time savings. Finally, we discovered additional use cases for the attention highlights that we did not foresee such as the participants using them to add additional details or correct the summary. This suggests that a well-chosen AI explainability feature can enrich the interactions between the AI system and the user.

7 Future Work

In addition to repeating this study at a larger scale with more participants with different levels of editorial experience and more cases to ensure the robustness of the current findings, a similar set up could also enable investigations into whether explainability features have an influence over the number of edits made and quality of the summary. This could be done by having independent reviewers evaluate the quality of the original summary and the summaries that were edited, without knowing which condition they were edited in. Furthermore, extending the current research by investigating additional AI models or post-hoc approaches which may present different explainability features could shed light to the impact of explainability on the interplay of efficiency and trust with direct implications on the business case for AI explainability.

REFERENCES

[1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier del Ser, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58: 82–115. arXiv:1910.10045. Retrieved from: https://arxiv.org/abs/1910.10045

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv:1409.0473. Retrieved from: https://arxiv.org/abs/1409.0473

[3] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. Qualitative Research in Psychology 3, 2: 77–101. DOI: https://doi.org/10.1191/1478088706qp063oa

[4] John Brooke. SUS - A quick and dirty usability scale. Retrieved 10 May, 2020 from https://hell.meiert.org/core/pdf/sus.pdf

[5] Bryce Goodman and Seth Flaxman. 2016. European Union regulations on algorithmic decision-making and a "right to explanation." DOI: https://doi.org/10.1609/aimag.v38i3.2741

[6] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. Nature Machine Intelligence 1, 9: 389–399. DOI:https://doi.org/10.1038/s42256-019-0088-2

[7] Guillaume Klein, Yoon Kim, Yuntian Deng, et al. 2018. OpenNMT: Neural Machine Translation Toolkit. Proceedings of AMTA 2018, vol. 1, 177-184. Retrieved September 1, 2020 from: https://www.aclweb.org/anthology/W18-1817.pdf

[8] Ramesh Nallapati, Bowen Zhou, Cicero Nogueira dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. arXiv:1602.06023. Retrieved from: https://arxiv.org/abs/1602.06023

[9] Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A Deep Reinforced Model for Abstractive Summarization. arXiv:1705.04304. Retrieved from: https://arxiv.org/abs/1705.04304

[10] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, 1135–1144. DOI: https://doi.org/10.1145/2939672.2939778

[11] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. arXiv:1704.04368. Retrieved from: https://arxiv.org/abs/1704.04368

[12] Abigail See. 2017. Taming Recurrent Neural Networks for Better Summarization. Retrieved September 9, 2020 from http://www.abigailsee.com/2017/04/16/taming-rnns-for-better-summarization.html

[13] Thomson Reuters. Artificial Intelligence at Thomson Reuters. Retrieved September 20, 2020 from https://www.thomsonreuters.com/en/artificial-intelligence/introduction-to-artificial-intelligence-at-thomson-reuters.html

[14] Matt Turek. 2020. Explainable Artificial Intelligence (XAI). Defense Advanced Research Projects Agency (DARPA). Retrieved September 20, 2020 from http://www.darpa.mil/program/explainable-artificial-intelligence