# Explainable Assessment of Healthcare Articles with QA

**Anonymous EMNLP submission**

## Abstract

The healthcare domain suffers from the spread of poor quality articles on the Internet. While manual efforts exist, they are not sufficient to assess the amount of articles in circulation. The task can be automated as text classification, however, explanations for the labels are necessary for the users. While current explainable systems tackle explanation generation as summarization, we propose a new approach based on Question-Answering that allows us to generate explanations for multiple criteria using a single model. We show that this QA-based approach is competitive with current state-of-the-art systems and complements summarization-based models for explainable quality assessment.

## 1 Introduction

The Internet has become an important source of medical advice. According to Rutten et al. (2019), in 2017, 74.4% of the US population first looked for health-related information on the internet, while only 13.3% of the population first asked a physician or healthcare provider. However, poor quality reporting, including misinformation, cherry-picking, exaggerations, etc., is often present online and can be a severe threat to public health. Recent events, such as the Covid-19 pandemic, demonstrate the necessity of developing quality assessment systems for healthcare reports to limit these harms. Fortunately, websites such as HealthNewsReview[1] critically analyze medical articles to identify poor quality reporting and improve the public dialogue about healthcare. The manual review of medical news is a time-consuming task that would benefit from automated systems to scale up to the volumes needed in today's media ecosystem.

Assessing the quality of news articles has been the focus of numerous studies that tackle it as a text classification task (Louis and Nenkova, 2013; Chakraborty et al., 2016; Kryscinski et al., 2020). Text classification is well studied, but explanations for the predictions only recently started receiving attention, despite being necessary to convince the readers of such assessments. For instance, Dai et al. (2020) have built on the evaluation work conducted by the HealthNewsReview website (see Table 1) to automate article quality assessment in healthcare, but have only focused on articles classification, without providing explanations. Likewise, Wright and Augenstein (2021) have also studied exaggeration detection in healthcare as classification, but without explanations.

**Story #1511**
**Criterion 1**: Does the article adequately discuss the costs of the intervention?
Answer: Not Satisfactory
Explanation: There was no discussion of cost as there was in the competing AP story.

**Criterion 2**: Does the article adequately quantify the benefits of the treatment/test/product/procedure?
Answer: Satisfactory
Explanation: The story adequately quantified the benefits seen in the study that led to FDA approval.

**Criterion 3**: ...

Table 1: Example of an article evaluated by the HealthNewsReview website Each article is evaluated according to 10 criteria (three shown) and explanations are given to support the answers.

Previous work has formulated textual explanation generation for classification as summarization (Atanasova et al., 2020; Kotonya and Toni, 2020). However such approaches suffer from a number of shortcomings when applied to the assessment of an article based on multiple criteria. As they always output the same summary for a given input text, separate models must be trained to generate explanations for each classification label and evaluation criterion (e.g. reliability of sources, lack of information, etc.), as for the example given in Table 1.

[1] https://www.healthnewsreview.org

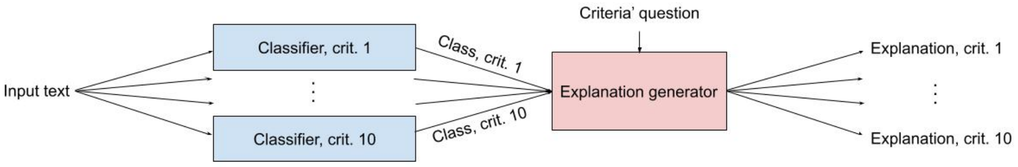

Figure 1: Pipeline for explainable quality assessment of articles

This considerably reduces the number of available training instances, because gold explanations of only one criterion at a time can be used to train each model, and it also requires developing and maintaining a model per criterion. Summarization-based models are also not appropriate to return an explanation for a label that is justified by the lack of information in the text (see criterion 1 in Table 1).

In this work, we develop an explainable quality assessment system for health news reports, and we evaluate it on the *FakeHealth* corpus (Dai et al., 2020). In addition to a classifier that makes the predictions, a QA-based model generates explanations for them by taking into consideration the definition of each evaluation criterion in the form of a question (see Table 1). This approach addresses the limitations of summarisation-based systems: it benefits from a larger training dataset, consisting of instances from all criteria and labels at once, can better generate explanations regarding the absence of information, and requires training and maintaining a single model for all criteria and labels.

We compare our approach against a summarization-based system inspired from Kotonya and Toni (2020). Our results show that both approaches are complementary and perform better in different cases. More specifically, summarization-based systems are more appropriate when relevant information is explicitly given in articles, while QA-based systems perform better when relevant information is missing. Finally, previous works used automatic metrics for evaluation, which are known to be insufficient for abstractive text generation (Kryscinski et al., 2019). For this reason, we design a human evaluation protocol to assess the fluency, consistency, and factual correctness of the explanations, and we show that automatic metrics are not appropriate for this task.

## 2 Methodology

Our approach is depicted in Figure 1: we first classify an article according to each criterion and then generate an explanation using QA, taking into account the predicted classification label. The purpose of the text classification step is to determine whether an article is satisfactory with respect to different evaluation criteria. We consider different options from the literature: logistic regression for its simplicity, BERT-based classification which is commonplace but truncates texts to 512 tokens, and a Longformer-based encoder model (Beltagy et al., 2020), which is able to deal with long input texts like those of our study. The latter is pre-trained for a large classification task on a biomedical summarization dataset, *PubMed*[2], then fine-tuned on the *FakeHealth* dataset. In line with Beltagy et al. (2020)'s recommendation, we use a classification objective that places a global attention mask on a `[CLS]` token. This token aggregates the representation of the whole text at the beginning of the input text as shown in Table 6 in Appendix C.1.

The second stage of the pipeline generates abstractive explanations for the previously predicted classes. As the QA approach takes into account the classes and the questions posed by criteria, we only need to train a single model, handling all criteria and classes. Influenced by Soni and Roberts (2020), we have chosen to work with a Longformer-based encoder-decoder that we first train on the open-domain dataset *SQuAD v2.0* (Rajpurkar et al., 2018), and then fine-tune on the *FakeHealth* dataset. For both learning steps, we use a QA-objective that applies a global attention mask on all question tokens (Beltagy et al., 2020), and we feed our model with the article, the criterion, and the class prediction. Table 6 in Appendix C.1 gives an example of the encoding of input texts and shows the global attention mask of our model. During training, we use the gold classes of articles to generate explanations, as generating post-hoc explanations for incorrectly predicted labels would not be meaningful.

Following recent previous work on explainable fact-checking in healthcare by Kotonya and Toni (2020),

[2] https://deepai.org/dataset/pubmed

we implement a baseline for the explanation generation task, based on summarization. Its summarization training objective applies a global attention mask only on the first token of the article, but it does not take into account the criterion definitions in its input. Instead, this approach requires training independent models for each class within a criterion, which results in 30 models (10 criteria × 3 classes) in the case of the *FakeHealth* dataset.

## 3 Human evaluation of explanations

Unlike previous works that assess generated text with automatic metrics, we design a human evaluation that seeks to assess four aspects of explanations: their fluency, consistency, factual correctness, and whether they are indicative of the label that they are supposed to explain. An explanation is considered fluent if it sounds natural, and consistent if it does not contradict itself, include repetitions, or information that is not mentioned in the article. The factual correctness criterion looks for incorrect facts, contradictions with respect to the article, or hallucinations. To finish, generated explanations should allow a human judge to infer the label explained.

We conducted two pilot studies in order to assess the quality of our guidelines. As reported in Table 2, Pilot 1 brought to light the ambiguity of the initial version of the guidelines, while Pilot 2 reached higher inter-annotator agreement scores. This new version of the guidelines is more detailed than the first one and provides some examples of what is expected. For instance, for all criteria, instead of asking if an explanation is fluent, the new version specifies that explanations should be rated as fluent if they sound natural and their structure is correct. Thus, the sentence “it's sunny but it's sunny” should not be considered as fluent, while “it's sunny but it's not sunny” should be considered fluent despite the contradiction, which is judged negatively under consistency. The guidelines from Pilot 2 that were used in the evaluation in Section 5 are reported in Appendix B.

## 4 Data

We evaluate our QA approach and summarization baseline on the *FakeHealth* corpus released by Dai et al. (2020). This corpus is comprised of two datasets, *HealthRelease* and *HealthStory*, both including health news articles with ratings and explanations for 10 criteria (see Table 10 in Appendix A.1). For each criterion, articles are annotated with one of three labels, *Not Satisfactory*, *Satisfactory*, and *Not Applicable*, and a textual explanation justifies the assigned label, as shown in Table 1. The label distribution across criteria is not uniform, which results in some very small classes, *Not Applicable* instances being the rarest. For example, criteria 2, 4, and 6 have at least 65 times more *Not Satisfactory* instances than *Not Applicable* ones (see Table 5 in Appendix A.2).

| | Fluency | Factual correctness | Guessed class |
|---|---|---|---|
| **Pilot 1** | -0.12 | 0.29 | 0.76 |
| **Pilot 2** | 0.46 | 0.49 | 0.58 |

Table 2: Inter-annotator agreement scores (Cohen Kappa scores) of the two pilot studies. The consistency criterion was added after Pilot 2.

| | Longformer | BERT | LogReg |
|---|---|---|---|
| Criterion 1 | **0.67** | 0.58 | 0.59 |
| Criterion 2 | **0.43** | **0.43** | 0.40 |
| Criterion 3 | 0.52 | 0.52 | 0.46 |
| Criterion 4 | **0.40** | 0.38 | 0.36 |
| Criterion 5 | 0.35 | 0.31 | **0.37** |
| Criterion 6 | **0.42** | 0.40 | 0.37 |
| Criterion 7 | 0.35 | 0.34 | 0.36 |
| Criterion 8 | **0.57** | 0.52 | 0.49 |
| Criterion 9 | **0.40** | 0.39 | 0.37 |
| Criterion 10 | **0.45** | 0.43 | 0.36 |
| **Mean** | **0.46** | 0.43 | 0.41 |

Table 3: Macro $F_1$-scores of our different classifiers for each criterion. The last row *Mean* gives the average performance of each model across criteria.

## 5 Results

### 5.1 Quality assessment per criterion

We assess the performance of our Longformer-based classifiers by comparing their macro $F_1$-scores against those of a BERT-based and a Logistic Regression models. Table 3 shows that our Longformer-based models perform the best due to their ability to encode longer texts. An analysis broken down by criterion also shows that Longformer, like all other models, performs unevenly across criteria. This suggests that some criteria are harder to handle, notably, those requiring external knowledge or subjective judgment (e.g. criterion 5 asking whether articles commit disease-mongering).

|  | **Fluency** |  | **Consistency** |  | **Factual correctness** |  | **Correct class** |  | **Count** |
|---|---|---|---|---|---|---|---|---|---|
|  | Sum. | QA | Sum. | QA | Sum. | QA | Sum. | QA |  |
| **All classes** | 74.5 | **80** | 72.5 | 72.5 | 52.5 | 52 | 85 | **86** | - |
| **Not S.** | 73.2 | **83.5** | 67 | **73.2** | 42.3 | **48.5** | 87.6 | **89.7** | 97 |
| **S.** | **79.6** | 76.3 | **80.6** | 73.1 | **63.4** | 53.8 | **86** | 82.8 | 93 |
| **Not A.** | 40 | **80** | 50 | **60** | 50 | **70** | 50 | **80** | 10 |

Table 4: Results of the evaluation of the summarization and QA-based systems per class (as percentages).

We also tried to build a single Longformer-based model handling all classes at once using a QA-approach, but it performed poorly. We suspect that its poor results are due to the discrepancy between the classification objective of the task to perform and the QA objective of the model.

### 5.2 Explanation generation

Table 4 reports the overall performance of both summarization and QA-based approaches. These results show that the QA-based approach performs better than, or as well as, the baseline system. Both approaches achieve similar performance in terms of consistency and factual correctness, but the QA approach produces explanations that are more fluent and that indicate the correct label more often. Tables 7 and 8 in Appendix C.2 provides some examples of the generated explanations.

An analysis per class (Table 4) reveals that the summarization approach performs better for the *Satisfactory* class, while the QA approach performs better for the *Not Satisfactory* and *Not Applicable* classes. This can be explained by the fact that *Satisfactory* articles include the relevant information to the criteria and require models to reuse this information to generate explanations, thus resembling summarization. On the other hand, for the *Not Satisfactory* class, models need to point out missing information and this is naturally harder for a summarization model, but easier for a QA-based one. Finally, the *Not Applicable* class suffers mainly from having very few instances for training (see Table 5 in Appendix A.2). With a single model, the QA approach is able to overcome this issue and generate better explanations.

To achieve the best performance, the previous results suggest combining both systems and use the summarization-based system for *Satisfactory* instances, and the QA-based system for all others. With this combination, 81% of explanations are fluent, 76% consistent, 57% factually correct, and 85% indicate correct labels.

### 5.3 Automatic v. human evaluation

To finish, we investigate the correlation between human judgement and automatic metrics used in previous works (Ermakova et al., 2019), including ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) scores. Table 9 in Appendix B.3 reports the correlation coefficients between all metrics. Using Kendall's Tau, we find that all these correlations are very low, at most 0.11 with ROUGE scores and 0.07 with the BLEU score. This finding was expected as most of the automatic metrics focus on word overlap, which makes it difficult to check the grammatical and syntactic correctness of explanations, as well as their factual consistency. This conclusion echoes Kryscinski et al. (2019)'s work on automatic evaluation protocols. As automatic metrics were found to be inappropriate to evaluate explanation generation systems, we only consider human evaluation to assess generated explanations.

## 6 Conclusion and discussion

In this work, we propose a new QA-based approach to generate explanations for quality assessment systems. This approach allows us to build a single model, able to generate explanations for different criteria and classes, by taking into account the questions related to criteria. We have shown that the QA-based system is competitive with the summarization-based one, and that they are complementary. Notably, the QA-based approach is more appropriate when the relevant information is not explicitly given in articles or for small classes. In addition, we have highlighted that automatic metrics, such as ROUGE, correlate very weakly with human judgment when it comes to evaluating explanation generation models.

## References

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. 2016. Stop clickbait: Detecting and preventing clickbaits in online news media. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 9–16.

Enyan Dai, Yiwei Sun, and Suhang Wang. 2020. Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. *CoRR*, abs/2002.00837.

Liana Ermakova, Jean Valère Cossu, and Josiane Mothe. 2019. A survey on evaluation of summarization methods. *Information Processing Management*, 56(5):1794–1814.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Annie Louis and Ani Nenkova. 2013. What Makes Writing Great? First Experiments on Article Quality Prediction in the Science Journalism Domain. *Transactions of the Association for Computational Linguistics*, 1:341–352.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Lila J. Finney Rutten, Kelly D. Blake, Alexandra J. Greenberg-Worisek, Summer V. Allen, Richard P. Moser, and Bradford W. Hesse. 2019. Online health information seeking among us adults: Measuring progress toward a healthy people 2020 objective. *Public Health Reports*, 134(6):617–625. PMID: 31513756.

Sarvesh Soni and Kirk Roberts. 2020. Evaluation of dataset selection for pre-training and fine-tuning transformer language models for clinical question answering. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5532–5538, Marseille, France. European Language Resources Association.

Dustin Wright and Isabelle Augenstein. 2021. Semi-supervised exaggeration detection of health science press releases.

## A *FakeHealth* dataset

### A.1 Dataset's criteria

Each article of the *FakeHealth* dataset is evaluated by at least two experts, according to ten criteria that assess diverse aspects such as “the overclaiming, missing of information, reliability of sources and conflict of interests" (Dai et al., 2020). Among them, eight criteria are common to both sources, while two are specific to *HealthRelease* and *HealthStory*. Dai et al. (2020) found zero to a minor positive correlation between the criteria, which justifies the relevance of all of them. These criteria are reported in Table 10 below.

### A.2 Class distribution

| Criterion | Not S. | S. | Not A. |
|---|---|---|---|
| **1** | 1431 | 495 | 370 |
| **2** | 1505 | 768 | **23** |
| **3** | 1413 | 717 | **166** |
| **4** | 1445 | 848 | **3** |
| **5** | **286** | 1921 | **89** |
| **6** | 1135 | 1147 | **14** |
| **7** | 1120 | 1063 | **113** |
| **8** | 538 | 1457 | 301 |
| **9** | 672 | 1543 | **81** |
| **10** | 391 | 1771 | **134** |

Table 5: Distribution of articles in each class per criterion. These numbers combine both the *HealthRelease* and *HealthStory* datasets.

# B Human evaluation

## B.1 Definition of the evaluation guidelines

To design our human evaluation protocol, we conduct two preliminary evaluation tasks. To begin with, the first task gathered three annotators who evaluated all explanations generated for the same six articles (three releases and three stories) with the baseline system for explanation generation. They were asked to determine if explanations were written in fluent English, consistent, factually correct, and which classes were suggested by explanations. This evaluation task combined both intrinsic and extrinsic methods to have a complete overview of models' performance, and we assessed to what extent annotators agreed on the evaluation task by looking at inter-annotator agreement scores computed with the Cohen Kappa score. It resulted in a high disagreement among annotators (see Table 2): annotators 1 and 2 even seemed to disagree on the fluency criterion. An in-depth exploration of their annotations revealed that they never agreed when one of them judged that an explanation was not fluent. These low inter-annotator agreement scores seem therefore to be caused by unclear guidelines.

For this reason, more detailed guidelines about the fluency and factual correctness of explanations were defined, and a second evaluation task was intended to validate them. It gathered two annotators who evaluated all explanations generated for the same five articles (two releases and three stories) with whether the baseline or the QA-based system. This second evaluation task achieved a much higher inter-annotator agreement reported in Table 2 and confirmed the new evaluation guidelines. However, the agreement score for the guessed classes slightly decreased between the first and second evaluation task. An analysis of annotations highlighted that some criteria could be ambiguous. For example, criterion 5 wonders if articles commit disease-mongering, and if they do, they should be rated as *Not Satisfactory* because it implies that they are less reliable. Consequently, a detailed description of each criterion, extracted from HealthNewsReview's website, has been given to annotators for the last evaluation task to raise all ambiguities.

## B.2 Final guidelines

Based on the outcome of the two previous evaluation tasks, annotators were asked to assess four elements for each explanation: whether it is written in fluent English, consistent, factually correct, and which class it suggests. They were given the following final guidelines:

- Fluency: Is the generated explanation written in fluent English? An explanation should be considered non-fluent if it does not sound natural or its structure is not correct (e.g. paragraphs title). Words' case (uppercase or lowercase) should not be taken into account. For example, "it's sunny but it's sunny" should be considered as non-fluent, but "it's sunny but it's not sunny" should be considered as fluent. Likewise, "intro: it's sunny, results: it's sunny, conclusion: it's sunny" should be considered as non-fluent (inappropriate structure).

- Consistency: Is the generated explanation consistent? An explanation should be considered inconsistent if it includes contradiction, repetition, extra information. For example, "it's sunny but it's sunny" should be considered as consistent, but "it's sunny but it's not sunny" should be considered as non-consistent.

- Factual correctness: Are the details (numbers, names, facts, etc.) included in the generated explanation correct? Explanations that contain incorrect facts, contradictions, or hallucinations should be evaluated as not satisfactory; but whether the factual details are related to the question or not should not be taken into consideration.

- Suggested class: According to the generated explanation, how would you classify the article? (*Not Satisfactory*, *Satisfactory*, *Not Applicable*, *Can't tell*) A *Can't tell* class has been added if generated explanations do not help classify articles. A description of what was expected for each criterion was given to annotators to raise all ambiguities. It was taken from the HealthNewsReview's website from which explanations had been extracted.

## B.3 Correlation with automatic metrics

Table 9 reports the correlation scores between human judgement and automatic metrics used in previous works (Ermakova et al., 2019), including ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) scores. Using Kendall's Tau, we find that all these correlations are very low, at most 0.11 with ROUGE scores and 0.07 with the BLEU score.

# C Model

## C.1 Model's Attention

For the Longformer model, Beltagy et al. (2020) defines different global attention mask that depends on the task we want to complete. For a classification task, the [CLS] token of input texts receives a global attention. For a QA task, the global attention mask is applied to all question tokens, while it is applied to the very first token of input texts for a summarization task. Table 6 illustrates these different attention masks.

**Question-Answering objective**

`<s>` Does the story adequately discuss the costs of the intervention? `</s>` Satisfactory `</s>` Triumph for Drug to Straighten Clenched Fingers `</s>` About one in 20 men is estimated to have Peyronie's, but figures are not precise because people with the condition tend not to discuss it publicly. ... `</s>`

**Summarization objective**

`<s>` Triumph for Drug to Straighten Clenched Fingers `</s>` About one in 20 men is estimated to have Peyronie's, but figures are not precise because people with the condition tend not to discuss it publicly. ... `</s>`

**Classification objective**

`[CLS]` Triumph for Drug to Straighten Clenched Fingers `</s>` About one in 20 men is estimated to have Peyronie's, but figures are not precise because people with the condition tend not to discuss it publicly. ... `</s>`

Table 6: Example of the inputs and global attention masks of the Longformer model for the different training objectives for HealthStory 1613, Criterion 1. The global attention is put on the tokens highlighted in grey.

## C.2 Example of models' outputs

**HealthRelease #528, Criterion 5, Satisfactory**

***Gold explanation***

There is no disease mongering in this release.

***Generated explanation – QA-based system***

There is no disease mongering in this news release. Obesity is a serious health issue, and the release provides some context about the prevalence of obesity in the US.

***Generated explanation – baseline system***

There is no disease mongering in the news release. The release provides some context about the prevalence and severity of obesity and its consequences.

Table 7: Examples of explanations generated with our QA-based and summarization-based baseline systems.

**HealthStory #1619, Criterion 1, Satisfactory**

***Gold explanation***

The story notes that a 5.29-ounce tube of the product retails for $39.99 at drugstores.

***Generated explanation – QA-based system***

The story notes that the cream is available at Sephora.com for $69.99. The story could have done a better job of comparing the cost of the cream to other products on the market.

***Generated explanation – baseline system***

The story states that the cream "takes the redness out of new stretch marks." That's good enough for the story, but it would have been better if the story would have mentioned the cost.

Table 8: Examples of explanations generated with our QA-based and summarization-based baseline systems.

| | **Fluency** | **Consistency** | **Factual correctness** | **Correct class** | **Can't tell class** |
|---|---|---|---|---|---|
| ROUGE-1 Precision | 0.09 | 0.04 | 0.03 | 0.08 | -0.07 |
| ROUGE-1 Recall | -0.02 | -0.08 | -0.05 | -0.04 | -0.00 |
| ROUGE-1 F1 | 0.01 | -0.05 | -0.01 | 0.00 | -0.04 |
| ROUGE-2 Precision | 0.08 | 0.05 | 0.04 | 0.09 | **-0.11** |
| ROUGE-2 Recall | 0.04 | -0.02 | -0.01 | 0.04 | -0.09 |
| ROUGE-2 F1 | 0.06 | 0.01 | 0.01 | 0.07 | **-0.11** |
| ROUGE-L Precision | 0.10 | 0.08 | 0.05 | 0.09 | -0.09 |
| ROUGE-L Recall | 0.01 | -0.04 | -0.03 | -0.01 | -0.03 |
| ROUGE-L F1 | 0.06 | 0.03 | 0.02 | 0.06 | -0.08 |
| BLEU | -0.01 | -0.07 | -0.04 | -0.01 | -0.03 |
| Length ratio | 0.09 | 0.08 | 0.05 | 0.08 | -0.06 |
| Cosine similarity | 0.08 | -0.01 | 0.03 | 0.06 | -0.05 |
| Euclidean distance | -0.04 | 0.01 | -0.04 | -0.02 | 0.03 |

Table 9: Correlation between human and automatic evaluation metrics (Kendall Tau correlation coefficient).

| **Criterion** | **Question** |
|---|---|
| **Criterion 1** | Does it adequately discuss the costs of the intervention? |
| **Criterion 2** | Does it adequately quantify the benefits of the treatment/test/product/procedure? |
| **Criterion 3** | Does it adequately explain/quantify the harms of the intervention? |
| **Criterion 4** | Does it seem to grasp the quality of the evidence? |
| **Criterion 5** | Does it commit disease-mongering? |
| **Criterion 6** | Does the story use independent sources and identify conflicts of interest? / Does the news release identify funding sources & disclose conflicts of interest? |
| **Criterion 7** | Does it compare the new approach with existing alternatives? |
| **Criterion 8** | Does it establish the availability of the treatment/test/product/procedure? |
| **Criterion 9** | Does it establish the true novelty of the approach? |
| **Criterion 10** | Does the story appear to rely solely or largely on a news release? / Does the news release include unjustifiable, sensational language, including in the quotes of researchers? |

Table 10: Datasets' criteria.