Multi Document Summarization Evaluation in the Presence of Damaging Content

Avshalom Manevich[*1], **David Carmel**[2,3], **Nachshon Cohen**[2], **Elad Kravi**[2] **and Ori Shapira**[2]

[1]Bar-Ilan University [2]Amazon [3]Technion - Israel Institute of Technology
avshalomman@gmail.com
{dacarmel, nachshon, ekravi, orishap}@amazon.com

Abstract

In the Multi-document summarization (MDS) task, a summary is produced for a given set of documents. A recent line of research introduced the concept of *damaging documents*, denoting documents that should not be exposed to readers due to various reasons. In the presence of damaging documents, a summarizer is ideally expected to exclude damaging content in its output. Existing metrics evaluate a summary based on aspects such as relevance and consistency with the source documents. We propose to additionally measure the ability of MDS systems to properly handle damaging documents in their input set. To that end, we offer two novel metrics based on lexical similarity and language model likelihood. A set of experiments demonstrates the effectiveness of our metrics in measuring the ability of MDS systems to summarize a set of documents while eliminating damaging content from their summaries.[1]

1 Introduction

In the Multi-Document Summarization (MDS) task a summary is generated for a given set of documents. The goal is for the summary to contain the most relevant information within the input documents. MDS was demonstrated in various domains including news articles (Over and Yen, 2004; Fabbri et al., 2019), Wikipedia pages (Ghalandari et al., 2020; Liu et al., 2018), and product reviews (Bražinskas et al., 2020b; Brazinskas et al., 2021; Shapira and Levy, 2020), including either an extractive approach, where the summary consists of spans from the input documents, or an abstractive approach, where the summary is synthetically generated.

Recent studies (Carmel et al., 2022; Sauchuk et al., 2022; Giorgi et al., 2022) have defined the concept of Damaging Documents—documents containing sensitive content that should not be exposed to readers; for example, toxic or offensive content. Furthermore, these works have emphasized the challenge of MDS in a real-world setting, where imperfect filtering algorithms fail to prevent damaging documents from sneaking into the summarizer's input set. MDS systems are, therefore, required to summarize the input documents without exposing any of the damaging content in the output.

MDS systems are typically evaluated using reference-based metrics by comparing the system summary to one or more reference summaries. A commonly used metric is *ROUGE* (Lin, 2004), but other metrics (e.g., Zhang et al., 2019; Zhao et al., 2019; Peyrard et al., 2017) are also being used, some of them do not require a reference summary. The various metrics capture a summary's quality-aspects including relevance, coherence, faithfulness and more (Fabbri et al., 2021). However, to the best of our knowledge, there is no metric that captures how 'contaminated' the summary is, i.e., the amount of damaging content it contains.

The goal of this study is to develop a metric that can appraise the amount of damaging content in a summary. To do so, the notion of 'damaging content' must first be defined. This is challenging since we rely in this work on external annotations, at the document level, which mark damaging documents (e.g. spam reviews). However, a damaging document may contain both damaging and legitimate content, and the legitimate content should be reflected in the summary. Therefore, when damaging documents appear in the input, the summarizer's task is to differentiate between legitimate and damaging content, and the evaluation task is to measure its success. Consider, for example, the summary presented in Figure 1, generated from some restaurant's negative reviews. While it accurately portrays the negative feedback on food quality and poor service, it also undesirably ex-

[*] Work done as an intern at Amazon.

[1]Resources available at https://github.com/avshalomman/mds-damaging-eval.

Summary: *This place used to be pretty good, but over the past few years it has gone downhill. The food is always undercooked, and the service has gone down hill. The owner is always sitting in the corner drunk, but does not give a shit that he's run this place into the ground. Stay away unless you are looking to get drunk and draw far away.*

Figure 1: A summary generated by the AdaSum system (Brazinskas et al., 2022) for negative restaurant reviews in the *D-Restaurants* dataset (Rayana and Akoglu, 2015). Damaging content is marked in red.

poses offensive and toxic language in the reviews that must be avoided.

Another challenge that we face is finding the right balance between measuring the consistency of the summary with the input, and the lack of damaging content. By aggressive filtering, a summarizer can reduce the damaging content in the output, with the cost of less consistency with the input. In contrast, by focusing on consistency, we may preserve a lot of damaging content in the summary, especially in the presence of a lot of damaging content in the input. The metric should take into consideration these two aspects when evaluating the summary quality.

In this work we propose two novel metrics that measure the capability of MDS systems to reduce the amount of damaging content in their output. The core idea behind our novelty is to add a penalty score denoting the dependency of the summary on damaging documents in the input. The metrics are based on two different approaches: The first extends the *ROUGE* metric (Lin, 2004) to also consider lexical similarity of the summary to the damaging documents. The second extends the *PMI* metric (Tam et al., 2022) to also consider a language model likelihood of the summary, conditioned on the damaging documents.

We challenge our evaluation metrics in the domain of customer reviews, where the task is to summarize negative reviews, some of them labeled as spam. Experimental results reveal that the suggested metrics can successfully assess the amount of damaging content in a summary. Additionally, we experiment in a real world setting where the input data is first filtered by a spam classifier, prior to summarization. In such a setting, choosing a threshold for classification introduces a tradeoff, since with a higher threshold many spam documents will be filtered out, at the cost of filtering legitimate documents as well. Our results demonstrate that the proposed metrics are beneficial in setting the optimal threshold for classification.

Contributions. The main contributions of this work are:

- Presenting the challenges of evaluating MDS systems in the presence of damaging content.
- Proposing two new metrics that measure the extent to which MDS systems prevent damaging content in their output. The metrics are grounded on lexical similarity and likelihood-based approaches.
- Conducting experiments with multiple MDS systems over two customer review datasets, demonstrating the effectiveness of the proposed metrics in measuring summary contamination.

2 Background and Related Work

The task of multi-document summarization (MDS) – generating a summary from a set of source documents – has been researched extensively, with accumulating works that introduce benchmarks and evaluation metrics. Traditionally, the typical evaluation protocol estimates a summary's quality by measuring its relevance to the source documents, i.e., how salient the information in the summary is. This is done by comparing the summary with one or more corresponding reference summaries, representing the most salient information within the document sources. Some widely used reference-based metrics include ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005) and BERTScore (Zhang* et al., 2020). Other metrics attempt to measure summary quality *without* reference summaries, when they are unavailable (e.g., Gao et al., 2020; Wu et al., 2020; Chen et al., 2021). One family of reference-free metrics uses language models to predict a summary's probability given the source documents (Vasilyev et al., 2020; Egan et al., 2022).

In addition to relevance, other factors must be taken into consideration in order to assess a summary's quality. Early DUC[2] benchmarks defined readability and content responsiveness criteria (Dang, 2005). More recently, Kryscinski et al. (2019) listed the four main quality dimensions as

[2] https://duc.nist.gov

1) relevance, the salience of summary content with respect to the source, 2) coherence, the linguistic validity of the overall summary structure, 3) fluency, the linguistic quality of individual sentences in the summary, and 4) consistency, the faithfulness of the summary to the source, i.e., whether the summary adheres to the source documents' information. Consistency is often measured by comparing the system summary to the source (Gabriel et al., 2021), and many such metrics have been recently proposed (e.g., Fabbri et al., 2022; Li et al., 2022).

In our work, we propose an additional evaluation dimension for the cases where certain source documents are damaging. Examples of damaging documents include offensive, spam, adult content for children, or classified documents for non-certified users (Carmel et al., 2022). While the presence of damaging documents is a known issue in information retrieval tasks (Cormack and Lynam, 2005; Clarke et al., 2009; Hussain et al., 2020), this is not the case in summarization tasks. Multi-document summarization datasets deliberately prepare document sets that are assumed to be fully legitimate. However, in real-life scenarios, the presence of damaging content cannot be ignored and therefore summarizers are expected not to expose it in their summary. Generally, generation of toxic content, a specific kind of damaging content, has recently gained some attention – as posed in the HELM framework (Liang et al., 2022).

In this work we are interested in the ability of a summarizer to disregard damaging pieces of text in the input. Traditional unsupervised summarization methods (Erkan and Radev, 2004; Verma and Om, 2019), initially treat all parts of the input equally, and may hence fail to omit damaging content. Meanwhile, supervised systems (Zhang et al., 2020; Brazinskas et al., 2022; Xiao et al., 2022) are faithful to the examples they were trained on, and can therefore generate outputs with certain kinds of content or style. For example, systems may be implicitly trained to output only non-harmful content. Other systems use a pipeline method, where certain text spans are filtered out as a first step (Lebanoff et al., 2018; Dong et al., 2021). Such an approach can attempt to leave out damaging content if desired. Our proposed evaluation framework aims to assess a summarizer's ability to remain consistent with allowed source documents and penalizes it for being consistent with forbidden source documents. In §5 we analyze this ability on the three described summarizing approaches.

Specifically, we demonstrate our evaluation framework in the domain of customer reviews, where the task is to summarize a set of reviews, partly consisting of spam reviews. Due to the clear potential utility of customer review summarization, a rich line of research has addressed this task (e.g., Angelidis and Lapata, 2018; Chu and Liu, 2019; Bražinskas et al., 2020b; Oved and Levy, 2021; Brazinskas et al., 2022), however, to the best of our knowledge, we are the first to view the task in the presence of spam reviews. Most of the more recent works on review summarization evaluate on product and service review datasets (Bražinskas et al., 2020a). These datasets contain summaries per sets of eight reviews, enabling the use of common summarization evaluation metrics, as described above. While most works focus on evaluating summary relevance, Oved and Levy (2021) emphasized the tendency of opinion summarizers to disappoint with quality issues concerning consistency and coherence. We address a more specific type of consistency-checking with our proposed evaluation scheme.

3 Evaluating MDS in the Presence of Damaging Content

Given a set of documents $\mathcal{D}$, the task of MDS is to generate a summary S of $\mathcal{D}$, such that S contains the most salient information in $\mathcal{D}$ (relevance), and only information from $\mathcal{D}$ (consistency). In this work, we further assume that $\mathcal{D}$ can be partitioned into two subsets, *unknown* to a summarization system. One subset consists of the legitimate documents $\mathcal{L}$, and the other of the damaging documents $\mathcal{B}$, where $\mathcal{L} \cup \mathcal{B} = \mathcal{D}$ and $\mathcal{L} \cap \mathcal{B} = \emptyset$. In this setting, a summary should be evaluated for its consistency with $\mathcal{L}$ and its avoidance of damaging content in $\mathcal{B}$. As is common in summarization, where consistency is typically measured by approximating the content overlap between the summary and source document(s) (Maynez et al., 2020), in our case "consistency" and "avoidance" can be captured through the content overlap between the summary and the two marked subsets of documents.

In the rest of this section we present two metrics that exemplify the desired evaluation. The metrics measure a summary's consistency with the legitimate documents, penalizing for its consistency with damaging documents. The first metric is based on *ROUGE* (Lin, 2004) and the second is based on

PMI (Tam et al., 2022).

3.1 Penalizing-ROUGE

ROUGE. The *ROUGE* family of metrics measures lexical similarity between texts by determining word-level overlap. It is commonly used to evaluate the performance of summarization systems by comparing system summaries to their corresponding reference summaries.

In this work we focus on *ROUGE*$_N$ variants, which measure N-gram overlap between two texts, $N \in \{1, 2\}$. Moreover, since *ROUGE* does not effectively measure consistency when used with *reference summaries* (Maynez et al., 2020), we compare a summary to its *source documents* to better approximate the summary's consistency. Given a text T, we denote by T_N the set of N-grams in T. Then, *ROUGE*$_N$ is defined as follows:

$$\begin{aligned} \text{Prec}_N(S, \mathcal{D}) &= \tfrac{1}{|\mathcal{D}|} \textstyle\sum_{D \in \mathcal{D}} |S_N \cap D_N| / |S_N| \\ \text{Recall}_N(S, \mathcal{D}) &= \tfrac{1}{|\mathcal{D}|} \textstyle\sum_{D \in \mathcal{D}} |S_N \cap D_N| / |D_N| \\ \textit{ROUGE}_N(S, \mathcal{D}) &= \frac{2 * \text{Prec}_N(S, \mathcal{D}) * \text{Recall}_{\mathcal{N}}(\mathcal{S}, \mathcal{D})}{\text{Prec}_N(S, \mathcal{D}) + \text{Recall}_{\mathcal{N}}(\mathcal{S}, \mathcal{D})} \end{aligned} \quad (1)$$

ROUGE$_N(S, \mathcal{D})$ approximates the consistency of summary S with document set $\mathcal{D}$ by measuring the average precision and recall of N-gram overlap across all documents in $\mathcal{D}$, and computing their harmonic mean.

Penalizing-ROUGE. In the presence of damaging content, we would like to account for how much a summary is consistent with the legitimate documents, and refrains from consistency with the damaging documents. In this regard, we consider any N-gram that appears in a legitimate document valid, regardless of whether it appears in a damaging document. Therefore, we define the set of damaging N-grams to be $B_N = \cup_{D \in \mathcal{B}} D_N \setminus \cup_{D \in \mathcal{L}} D_N$. The contribution of the damaging fraction of the summary S to its precision, with respect to the source documents, is $\text{D-Prec}_N(S, \mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{B \in \mathcal{B}} |S_N \cap B_N| / |S_N|$.

Ideally, we would like to subtract the damaging precision from the legitimate precision, accounting for how much the summary is consistent with the legitimate documents, but not with the damaging documents. However, when the damaging precision is greater than the legitimate precision the resulting value is negative. In such a case the harmonic mean would be undefined since the operation is intended for positive values only. To alleviate this issue, we first define the penalized precision (PP) as:

$$\text{PP}_N(S, \mathcal{L}, \mathcal{B}) = \text{Prec}_N(S, \mathcal{L}) - \text{D-Prec}_N(S, \mathcal{B}) \quad (2)$$

Then, we define Penalizing-ROUGE as follows:

$$\textit{P-ROUGE}_N(S, \mathcal{L}, \mathcal{B}) = \begin{cases} \frac{2 * \text{PP}_N(S, \mathcal{L}, \mathcal{B}) * \text{Recall}_N(S, \mathcal{L})}{\text{PP}_N(S, \mathcal{L}, \mathcal{B}) + \text{Recall}_N(S, \mathcal{L})} & \text{if } PP_N(S, \mathcal{L}, \mathcal{B}) > 0 \\ PP_N(S, \mathcal{L}, \mathcal{B}) & \text{otherwise} \end{cases} \quad (3)$$

We note that *P-ROUGE*$_n$ is in the range of $[-1, 1]$.[3]

3.2 Penalizing-PMI

Contrary to *ROUGE* which was not initially designed for measuring consistency, there has been a line of research dedicated to measuring summary consistency. In one such work, Tam et al. (2022) introduced a metric based on Pointwise Mutual Information (*PMI*) that uses a large language model to compute the probability of a summary given the source documents. The main hypothesis is that a good summary, especially one that is consistent with its source, should appear as a natural complementary extension to the source documents, while a bad summary has a lower likelihood to complement the documents. We first assume the existence of a general-purpose pre-trained language model, LM,[4] which is capable of estimating $P_{LM}(t_{i+1}|t_1, \ldots, t_i)$—the probability of a token t_{i+1} given a sequence of tokens $(t_1, \ldots, t_i)$. The Mean Log Likelihood (MLL) of a summary S, given source documents $\mathcal{D}$ is defined as $MLL_{LM}(S|\mathcal{D}) := \frac{1}{T} \sum_{i=1}^{T} \log P_{LM}(t_i|\mathcal{D}, t_1, \ldots, t_{i-1})$, where T is the number of tokens in S, and documents in $\mathcal{D}$ are concatenated in an arbitrary order, followed by the summary prefix $(t_1..t_{i-1})$. In the following, we remove dependency on *LM* for clarity. The *PMI* metric is then defined as

$$PMI(S; \mathcal{D}) = MLL(S|\mathcal{D}) - MLL(S|\emptyset) \quad (4)$$

which factors out the inherent probability of the summary text, independently of $\mathcal{D}$.

Next, we would like to evaluate the unique contribution of damaging documents to the summary. In our use case, we would like to erase the contribution of legitimate documents to the summary. Pointwise Conditional Mutual Information (*PCMI*) is an information theory metric

[3] The penalized precision PP_N lies within $[-1, 1]$ since it is a subtraction of two values in $[0, 1]$, and the recall is in $[0, 1]$.

[4] We used the GPT-Neo LM for our metric implementation.

that measures the information in the text that is uniquely attributable to a specific variable, and can be computed using a language model as well (Paranjape and Manning, 2021). For our use case, given a summary S, a set of legitimate documents $\mathcal{L}$ and a set of damaging documents $\mathcal{B}$, we define $PCMI(S;\mathcal{B}|\mathcal{L}) = PMI(S;\mathcal{B}\cup\mathcal{L}) - PMI(S;\mathcal{L})$. This measurement represents the contribution of the damaging documents in $\mathcal{B}$ to the summary S, while factoring out the contribution of the legitimate documents $\mathcal{L}$.

Finally, we propose the Penalizing-PMI (*P-PMI*) metric, which measures *PMI* between the summary and the legitimate documents, and subtracts the contribution of the damaging documents. Specifically, we define

$$P\text{-}PMI(S;\mathcal{L};\mathcal{B}) = PMI(S;\mathcal{L}) - PCMI(S;\mathcal{B}|\mathcal{L}) \quad (5)$$

We remark that Equation 5 is similar in spirit to the damaging precision defined in Equation 2. We start by measuring the summary's dependency on the legitimate documents, and then subtract the dependency on the damaging document while excluding the legitimate content. In the *P-PMI* case, dependency is measured by the summary's likelihood conditioned on a given input and not via lexical matching, but the core concept is the same.

4 Experimental Setup for Metric Assessment

Our next step is to assess the proposed metrics for MDS in the presence of damaging content. For that, we aim to answer the following research questions:

- How sensitive are the metrics to the amount of damaging content in the source documents?
- Can the metrics distinguish between summaries according to the amount of damaging content in their output?
- What is the level of agreement between the suggested metrics?

To answer these questions we evaluate our proposed metrics over several MDS systems on real-world datasets containing damaging documents. Specifically, we experiment with two customer review datasets that contain a substantial amount of spam reviews – to be considered as damaging documents. The rest of the section includes a description of our experimental setup including the summarization systems and the datasets. §5 discusses the experiments conducted to address the above questions.

4.1 Summarization Systems

In our experiments, we apply the proposed metrics on the following MDS systems:

AdaSum (Brazinskas et al., 2022). An abstractive review summarization system, employing a BART (Lewis et al., 2019) model with a self-supervised adapter. AdaSum was pre-trained on a large unlabeled corpora of customer reviews, and then was fine-tuned on a small annotated dataset. It yields state-of-the-art *ROUGE* scores for review summarization over several common benchmarks.

CopyCat (Bražinskas et al., 2020b). An abstractive, unsupervised summarization system which uses a hierarchical continuous latent representation of products and individual reviews.

COOP-BiMeanVAE (Iso et al., 2021). An approach for generating abstractive summaries from review representations encoded in a latent space. It consists of searching for a convex combination of latent review vectors that maximizes word overlap between the reviews and the generated summary. We denote this system as **COOP** for the rest of the paper.

LexRank (Erkan and Radev, 2004). LexRank is an extractive, unsupervised system. It uses a graph composed of the documents' sentences as nodes. Edges represent the *tf-idf* similarity scores between the nodes. A summary is formed by selecting sentences based on a graph centrality measure.

System	$ROUGE_1$	$ROUGE_2$
AdaSum	0.398	0.108
CopyCat	0.320	0.058
COOP	0.366	0.072
LexRank	0.287	0.055

Table 1: Self-reported $ROUGE_1$ and $ROUGE_2$ scores of the tested summarization systems on a product review summarization benchmark (Bražinskas et al., 2020a).

As a point of reference, Table 1 presents the $ROUGE_1$ and $ROUGE_2$ scores of the different systems on the Amazon product review summarization benchmark (Bražinskas et al., 2020a). System summaries were evaluated against crowdsourced reference summaries. With respect to these two metrics, there is a large gap in favor of the AdaSum system.

4.2 Review Datasets

In our experiments, we focus on the task of summarizing negative reviews, with 1 or 2 stars (on a

1-5 Likert scale). It is known from previous studies (Bražinskas et al., 2021) that negative reviews are notoriously harder for automatic systems to summarize. Part of this complexity follows from the fact that spam or toxic content is prevalent in negative reviews, thus making it harder to create a summary that captures the negative attitude while avoiding the damaging content (see Figure 1).

We use two datasets for assessing the proposed metrics. The first dataset, *D-Products*, consists of 3.7K products from the Amazon.com website, each accompanied with some reviews marked with a binary spam label.[5] The second dataset, *D-Restaurants*, consists of restaurants and their reviews, annotated with spam labels denoting fake/suspicious reviews (Rayana and Akoglu, 2015).[6] We denote by *entity* a product in the *D-Products* dataset and a restaurant in the *D-Restaurants* dataset, specifying the granularity on which reviews are summarized. For the rest of this section we use the terms *spam* and *damaging* interchangeably.

In order to balance the data, we selected entities having enough legitimate and damaging reviews in order to evaluate summarization systems under different configurations. Hence, only entities with at least 4 damaging reviews were selected. For the qualifying entities we sampled the same number of legitimate reviews, creating balanced datasets.[7] Additional information about the size of each dataset is provided in Table 2.[8]

Dataset	Entity Type	# Entities	# Reviews
D-Products	product	3,743	44,464
D-Restaurants	restaurant	896	11,134

Table 2: Review dataset statistics. Number of damaging and legitimate reviews per entity is balanced.

4.3 Varying Portions of Damaging Documents

The purpose of our metrics is to detect the amount of damaging content within a summary. We make the assumption that if a summarizer is exposed to more damaging documents in the input, its summary will contain more damaging content as well. Therefore, it is desirable for our metric to distinguish between different proportions of damaging documents that the summarizer gets as input. To this end, we create dataset subsets A_x, where for each entity, a portion of x reviews are spam and a portion of $1-x$ reviews are legitimate. A_0 denotes that only legitimate reviews are considered per entity, while a A_1 denotes that only spam reviews are considered. Overall, five dataset subsets with varying spam portions were constructed $(A_0, A_{0.33}, A_{0.5}, A_{0.67}, A_1)$. We ran our summarization systems over these datasets, generating $92,775$ summaries overall.

5 Experiments for Metric Assessment

5.1 Sensitivity to Damaging Content

In the first experiment we evaluate the metrics' sensitivity to damaging content. For that, we measure their ability to rank a pair of review summaries, according to the amount of damaging content they contain. We assume that for any MDS, the more spam in the input, the more spam we can expect in the summary.

We experiment with the vanilla metrics (*ROUGE* and *PMI*) and their penalizing versions (*P-ROUGE* and *P-PMI*). For the vanilla metrics, we measure the summary consistency with the legitimate reviews, i.e. applying Equations 1 and 4 using $\mathcal{L}$ as the reference set. For each entity, we generate $\binom{5}{2}$ pairs of summaries, (s_x, s_y), based each on a different spam ratio; s_x is generated for the corresponding reviews in A_x, and s_y for the corresponding reviews in A_y. For example, the pair (s_0, s_1) denotes that one summary is generated from the entity's ham (i.e., legitimate) reviews while the other from the entity's spam reviews. Given such a pair of summaries, we rank them based on the metric scores and observe if their ranking matches the expected ranking according to the spam portion of their corresponding subsets ("oracle order"), i.e., according to the natural order between x and y. We expect the penalizing metrics to excel in agreement with the oracle order as they are designed to capture the existence of damaging content in the summary. The average accuracy of matching orders over all pairs, in each dataset, are depicted in Table 3.

It can be seen that the accuracy varies among systems and metrics; the most prominent results are for AdaSum and LexRank, while CopyCat and COOP systems yield much lower accuracy. In all the examined cases (excluding CopyCat on *D-*

[5]The *D-Products* dataset does not represent the Amazon customer reviews corpus, Specifically, spam reviews contained within it are not part of that corpus.

[6]Based on publicly available data.

[7]We consider only reviews with 10 to 150 words.

[8]The final *D-Restaurants* dataset can be reproduced at https://github.com/avshalomman/mds-damaging-eval.

Dataset	System	*P-PMI*	*PMI*	*P-ROUGE*$_1$	*ROUGE*$_1$	*P-ROUGE*$_2$	*ROUGE*$_2$
D-Products	AdaSum	**82.44**	77.59	**84.95**	82.16	**84.57**	80.8
	LexRank	**83.39**	81.82	**82.97**	80.5	**85.17**	83.86
	COOP	**57.79**	55.35	**75.96**	74.02	**67.77**	62.23
	CopyCat	50.9	**52.24**	**61.32**	60.4	**55.55**	53.49
D-Restaurants	AdaSum	**80.48**	72.67	**83.68**	80.05	**83.7**	78.54
	LexRank	**84.77**	83.67	**84.79**	81.01	**86.79**	86.05
	COOP	**57.53**	53.64	**75.1**	72.65	**66.22**	61.59
	CopyCat	**54.98**	53.55	**59.95**	57.24	**55.85***	53.65*

Table 3: Ranking accuracy of the different metrics, measured by matching with the oracle order. **Bold** indicates better results between the vanilla metric and its penalizing variant. * marks a *non*-statistically significant difference (McNemar test of homogeneity with p-value < 0.05).

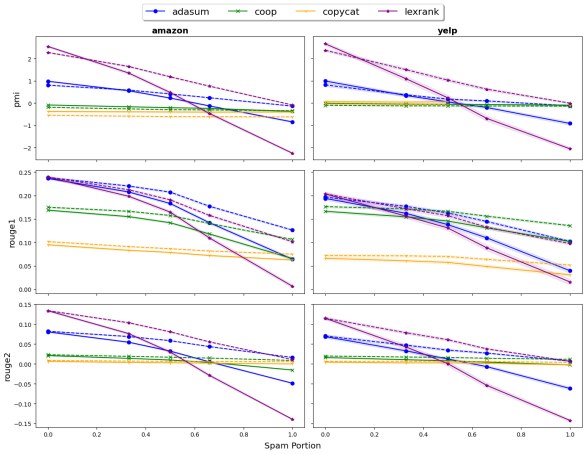

Figure 2: Mean values of vanilla (dashed line) and penalizing (solid line) metrics as a function of spam portion, for the *D-Products* (left) and *D-Restaurants* (right) datasets. The sleeve around the lines denotes a 95% confidence-interval of each line.

Products) the penalizing metric outperforms the vanilla version, indicating that it better captures the occurrence of damaging content in the generated summary. *P-ROUGE*$_1$ and *P-ROUGE*$_2$ outperform *P-PMI* for most systems and across datasets, suggesting that spam is more easily detected through text overlap rather than by a contextual likelihood measurement.

We further compare the metrics sensitivity for varying spam portions in the input. Figure 2 shows the metrics' value, averaged over all summaries in each of the subsets, as spam portion increases. The top row presents *PMI* (dashed line) and *P-PMI* (solid line) scores, as a function of the spam portion, for the two datasets. The two bottom rows present *ROUGE* and *P-ROUGE* scores for unigrams and bigrams, respectively.

It can be clearly seen that for all metrics LexRank and AdaSum scores decrease significantly as spam portion grows. Recall that the vanilla metrics measure consistency with the legitimate documents only, therefore a decrease in these metric values is observed when spam portion increases. However, the larger decrease is observed for the penalizing metrics, indicating their superior sensitivity to damaging content in the summaries.

An interesting phenomenon can be observed for COOP and CopyCat systems. *ROUGE*$_1$ and *P-ROUGE*$_1$ degrade for these systems as spam portion increases, similarly to their behavior for AdaSum and LexRank (although with a less steep slope). On the other hand, *ROUGE*$_2$ and the *PMI* are stable for these systems, unaffected by the increase in spam portion. There is a possibility that these systems are almost indifferent to spam, as changes in the amount of damaging content are barely noticed.

Another interesting observation is that while *P-ROUGE* is always less than or equal to *ROUGE* (by definition), *P-PMI* is higher than *PMI*, across all systems and datasets, when no spam documents appear in the input (leftmost point on the graphs). In such a case, when damaging documents are considered in the reference set, while there is no damaging content in a summary, the likelihood of such a summary S^* degrades when conditioned on the damaging documents, $PMI(S^*; B \cup \mathcal{L}) \leq PMI(S^*; \mathcal{L})$. It follows that $PCMI(S^*; \mathcal{B}|\mathcal{L}) \leq 0$, and therefore, $P\text{-}PMI(S^*; B \cup \mathcal{L}) = PMI(S^*; \mathcal{L}) - PCMI(S^*; \mathcal{B}|\mathcal{L}) \geq PMI(S^*; B \cup \mathcal{L})$.

Finally, we calculate the Pearson correlation between the metric scores for all summaries of all systems, over all subsets with varying spam portions;

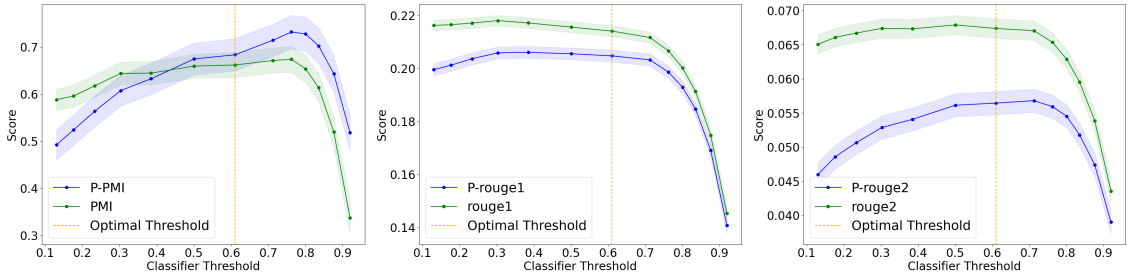

Figure 3: Filter-then-summarize pipeline results for different spam ratios, as determined by the different classification thresholds. Sleeve around the lines denotes 95% confidence interval.

Dataset	Metric	*PMI*	*P-PMI*
D-Products	$ROUGE_1$	**0.54**	0.46
	$P\text{-}ROUGE_1$	0.51	**0.57**
	$ROUGE_2$	**0.68**	0.56
	$P\text{-}ROUGE_2$	0.51	**0.70**
D-Restaurants	$ROUGE_1$	**0.41**	0.37
	$P\text{-}ROUGE_1$	0.44	**0.52**
	$ROUGE_2$	**0.68**	0.55
	$P\text{-}ROUGE_2$	0.52	**0.74**

Table 4: Pearson correlation between the different metrics over the *D-Products* and *D-Restaurants* datasets. All correlation results are statistically significant (p-value < 0.01). **Bold** denotes the highest value with respect to a given pair of metrics (penalizing and vanilla versions).

results are depicted in Table 4. Penalizing metrics are better correlated among themselves; this also holds for vanilla metrics. Moreover, the correlation among penalizing versions is higher than among vanilla versions, indicating that the evaluation criteria taken by *ROUGE* and *PMI* are less consistent than those taken by their penalizing versions.

5.2 Filter Then Summarize

A common practical approach to prevent damaging content from seeping into a summary is to filter it out at a preliminary step (Sauchuk et al., 2022) ('filter-then-summarize'). Filtration can be executed by employing a classifier that detects damaging content, as we examine next.

In this experiment, we focus on AdaSum since it provided more stable results in previous experiments. For spam filtering, we train a distilbert-base-uncased[9] classifier (learning rate: $2e-5$, batch size: 32, weight decay: 0.01). We took entities not in the *D-Products* and *D-Restaurants* datasets with balanced ratio of spam/not spam reviews and split to train/test sets in a $(0.95, 0.05)$ ratio. Overall $100K$ reviews, from each dataset source, were selected for the classifier training task.

Choosing a threshold for the classifier introduces a tradeoff, since a higher threshold will filter out more spam, at the cost of filtering out more legitimate documents as well. We focus on the *D-Products* dataset for which the spam classifier performs reasonably well (AUC = 0.85). We examine 13 threshold values overall, including the optimal, chosen according to Youden's J Statistic (Ruopp et al., 2008), and 12 additional values, using equally-spaced samples from the optimal threshold value on the ROC curve. For each threshold we apply the spam classifier, summarize the remaining documents, and evaluate the resulting summary by the penalizing and non-penalizing metrics. Results are depicted in Figure 3.

As the classification threshold increases, more data (spam and ham) is being filtered out from the input set. For *PMI*, on the left, the penalizing metrics score are lower than the vanilla score due to the large amount of spam. When the threshold exceeds 0.8, all metrics crash due to low recall of the generated summaries. For smaller threshold values, the vanilla metrics are indifferent to filtering and remain stable as the threshold grows (in particular $ROUGE_1$ and $ROUGE_2$). In contrast, the penalizing versions show improvement with more filtering, up to a point where performance starts to degrade. Their peak is close to Youden's J Statistic (though not the same), denoting their usefulness in searching for the optimal filtering threshold value.

6 Conclusion

In this work we raise the need to evaluate MDS systems for their capability of preventing damaging content in their summaries. We propose two

[9] huggingface.co/distilbert-base-uncased

metrics, one based on *ROUGE* and the other on the *PMI* metric. Through a set of experiments conducted over customer review datasets and several summarization systems, we demonstrate that our metrics are capable of ranking different summarizers according to how they treat damaging content. Importantly, the proposed metrics and the increased awareness to damaging content should motivate further research on optimizing MDS in the presence of damaging documents.

7 Limitations

This paper conducts extensive experiments with two datasets. The *D-Restaurants* dataset is a subset of a publicly available dataset and will be released to the community for reproduction and for further research. On the other hand, the *D-Products* dataset was exposed to us for this particular research only, and unfortunately cannot be publicised according to its terms of use.

Both metrics proposed in the paper take a specific approach of deducting consistency with damaging documents from consistency with legitimate document. There are likely other approaches that can approximate the task objectives, and we hope the community continues to explore such directions.

Finally, we do not methodologically evaluate the metrics. Future work calls for a meta-evaluation framework, including high quality benchmarks dedicated to our task. This would enable a more accurate appraisal of metrics, including new ones to be brought forth in future research.

8 Ethical Considerations

There is a high degree of sensitive information in the data used in this study, including offensive and toxic language. The metrics proposed in the work which measure the contamination level of generated summaries, can provide a first barrier to using such data. However, it is important to note that these metrics are not involved in the summarization task per se and therefore cannot avoid the appearance of damaging content in systems' summaries, only to attest on it.

References

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing Opinions: Aspect Extraction Meets Sentiment Prediction and They Are Both Weakly Supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020a. Few-Shot Learning for Opinion Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4119–4135, Online. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020b. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169, Online. Association for Computational Linguistics.

Arthur Brazinskas, Mirella Lapata, and Ivan Titov. 2021. Learning opinion summarizers by selecting informative reviews. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 9424–9442. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2021. Learning opinion summarizers by selecting informative reviews. *arXiv preprint arXiv:2109.04325*.

Arthur Brazinskas, Ramesh Nallapati, Mohit Bansal, and Markus Dreyer. 2022. Efficient few-shot fine-tuning for opinion summarization. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1509–1523, Seattle, United States. Association for Computational Linguistics.

David Carmel, Nachshon Cohen, Amir Ingber, and Elad Kravi. 2022. IR evaluation and learning in the presence of forbidden documents. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 556–566. ACM.

Wang Chen, Piji Li, and Irwin King. 2021. A Training-free and Reference-free Summarization Evaluation Metric via Centrality-weighted Relevance and Self-referenced Redundancy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 404–414, Online. Association for Computational Linguistics.

Eric Chu and Peter Liu. 2019. MeanSum: A Neural Model for Unsupervised Multi-Document Abstractive Summarization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1223–1232. PMLR.

Charles Clarke, Nick Craswell, and Ian Soboroff. 2009. Overview of the TREC 2009 Web Track. In *Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009*, Gaithersburg, Maryland.

Gordon Cormack and Thomas Lynam. 2005. TREC 2005 Spam Track Overview. In *Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005*, Gaithersburg, Maryland.

Hoa Trang Dang. 2005. Overview of DUC 2005. In *Proceedings of the Document Understanding Conference*, volume 2005, pages 1–12.

Luobing Dong, Meghana N. Satpute, Weili Wu, and Ding-Zhu Du. 2021. Two-Phase Multidocument Summarization Through Content-Attention-Based Subtopic Detection. *IEEE Transactions on Computational Social Systems*, 8(6):1379–1392.

Nicholas Egan, Oleg Vasilyev, and John Bohannon. 2022. Play the Shannon Game with Language Models: A Human-Free Approach to Summary Evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10599–10607.

Gunes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-Based Factual Consistency Evaluation for Summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Alexander R Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. *arXiv preprint arXiv:1906.01749*.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. GO FIGURE: A Meta Evaluation of Factuality in Summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 478–487, Online. Association for Computational Linguistics.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards New Frontiers in Unsupervised Evaluation Metrics for Multi-Document Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.

Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. 2020. A large-scale multi-document summarization dataset from the wikipedia current events portal. *arXiv preprint arXiv:2005.10070*.

John Giorgi, Luca Soldaini, Bo Wang, Gary Bader, Kyle Lo, Lucy Lu Wang, and Arman Cohan. 2022. Exploring the challenges of open domain multi-document summarization. *arXiv preprint arXiv:2212.10526*.

Naveed Hussain, Hamid Turab Mirza, Ibrar Hussain, Faiza Iqbal, and Imran Memon. 2020. Spam Review Detection Using the Linguistic and Spammer Behavioral Methods. *IEEE Access*, 8:53801–53816.

Hayate Iso, Xiaolan Wang, Yoshihiko Suhara, Stefanos Angelidis, and Wang-Chiew Tan. 2021. Convex Aggregation for Opinion Summarization. In *Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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.

Logan Lebanoff, Kaiqiang Song, and Fei Liu. 2018. Adapting the Neural Encoder-Decoder Framework from Single to Multi-Document Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4131–4141, Brussels, Belgium. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Yiyang Li, Lei Li, Qing Yang, Marina Litvak, Natalia Vanetik, Dingxin Hu, Yuze Li, Yanquan Zhou, Dongliang Xu, and Xuanyu Zhang. 2022. Just ClozE! A Fast and Simple Method for Evaluating the Factual Consistency in Abstractive Summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic Evaluation of Language Models.

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.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Nadav Oved and Ran Levy. 2021. PASS: Perturb-and-Select Summarizer for Product Reviews. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 351–365, Online. Association for Computational Linguistics.

Paul Over and James Yen. 2004. An introduction to duc-2004. *National Institute of Standards and Technology*.

Ashwin Paranjape and Christopher Manning. 2021. Human-like informative conversations: Better acknowledgements using conditional mutual information. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 768–781, Online. Association for Computational Linguistics.

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 74–84.

Shebuti Rayana and Leman Akoglu. 2015. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 985–994, New York, NY, USA. Association for Computing Machinery.

Marcus D Ruopp, Neil J Perkins, Brian W Whitcomb, and Enrique F Schisterman. 2008. Youden index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50(3):419–430.

Artsiom Sauchuk, James Thorne, Alon Halevy, Nicola Tonellotto, and Fabrizio Silvestri. 2022. On the role of relevance in natural language processing tasks. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1785–1789.

Ori Shapira and Ran Levy. 2020. Massive multi-document summarization of product reviews with weak supervision. *arXiv preprint arXiv:2007.11348*.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2022. Evaluating the Factual Consistency of Large Language Models Through Summarization.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the BLANC: Human-free quality estimation of document summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20, Online. Association for Computational Linguistics.

Pradeepika Verma and Hari Om. 2019. MCRMR: Maximum coverage and relevancy with minimal redundancy based multi-document summarization. *Expert Systems with Applications*, 120:43–56.

Hanlu Wu, Tengfei Ma, Lingfei Wu, Tariro Manyumwa, and Shouling Ji. 2020. Unsupervised Reference-Free Summary Quality Evaluation via Contrastive Learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3612–3621, Online. Association for Computational Linguistics.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. *CoRR*, abs/1904.09675.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. *CoRR*, abs/1909.02622.