applied sciences

MDPI

Article

Abstractive Summarizers Become Emotional on News Summarization

Vicent Ahuir [1,*], José-Ángel González [2,*], Lluís-F. Hurtado [1] and Encarna Segarra [1,3]

[1] Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València, 46022 Valencia, Spain; lhurtado@dsic.upv.es (L.-F.H.); esegarra@dsic.upv.es (E.S.)
[2] Symanto Symanto Research, C/Reina 12, 46011 Valencia, Spain
[3] Valencian Graduate School and Research Network of Artificial Intelligence (ValgrAI), Universitat Politècnica de Valéncia, 46022 Valencia, Spain
* Correspondence: viahes@dsic.upv.es (V.A.); jose.gonzalez@symanto.com (J.Á.G.)

Abstract: Emotions are central to understanding contemporary journalism; however, they are overlooked in automatic news summarization. Actually, summaries are an entry point to the source article that could favor some emotions to captivate the reader. Nevertheless, the emotional content of summarization corpora and the emotional behavior of summarization models are still unexplored. In this work, we explore the usage of established methodologies to study the emotional content of summarization corpora and the emotional behavior of summarization models. Using these methodologies, we study the emotional content of two widely used summarization corpora: CNN/DAILYMAIL and XSUM, and the capabilities of three state-of-the-art transformer-based abstractive systems for eliciting emotions in the generated summaries: BART, PEGASUS, and T5. The main significant findings are as follows: (i) emotions are persistent in the two summarization corpora, (ii) summarizers approach moderately well the emotions of the reference summaries, and (iii) more than 75% of the emotions introduced by novel words in generated summaries are present in the reference ones. The combined use of these methodologies has allowed us to conduct a satisfactory study of the emotional content in news summarization.

Keywords: news summarization; abstractive summarization; emotional content; emotional behavior

Citation: Ahuir, V.; González, J.-Á.; Hurtado, L.-F.; Segarra, E. Abstractive Summarizers Become Emotional on News Summarization. *Appl. Sci.* **2024**, *14*, 713. https://doi.org/10.3390/app14020713

Academic Editor: Valentino Santucci

Received: 30 November 2023
Revised: 3 January 2024
Accepted: 8 January 2024
Published: 15 January 2024

1. Introduction

Storytelling is an important aspect of journalism that aims to share facts or ideas in the best way to reach, captivate attention, and convince the audience. Hence, news often does not directly re-tell events, but rather gives an interpretation of those events by a human, whose feelings can often become an important part of the story's meaning [1]. Besides, there is clear evidence that using emotional cues helps to catch our attention and prolong our engagement [2]. For this reason, emotions have become an important dynamic in how news is produced and consumed, central to our understanding of journalism [3,4].

According to how online newspapers produce news articles, our entry points to a story are the headline and the summary. If they catch our attention, we will likely read the source article. Therefore, we would expect that human summarizers favor emotional content when generating summaries and headlines, potentially over/under-emphasizing some emotions compared to the source article [1]. Table 1 illustrates this with two summaries for the same article that evoke different emotions.

Few works have explored emotions under the umbrella of automatic news summarization [1], which have otherwise been considered in other domains such as dialogue or microblogging [5,6].

Nowadays, pre-trained language models are the reputable approach for developing state-of-the-art abstractive summarization systems of news articles. Their capabilities to summarize news articles have been proven, standing out in terms of phrase-overlapping

metrics like ROUGE [7], through a broad set of corpora. However, the *emotional behavior* of these systems is still unexplored. Along with other summarization aspects such as abstraction [8], faithfulness, or factuality [9], *emotional behavior* can shed light on how to develop better summarizers.

Table 1. An example of two different summaries for the same article. Using the NRC lexicon, we highlight the words that convey emotions (the emotions are listed in brackets). Phrases and emotions in blue refer to positive aspects, and those marked in red to negative aspects.

Article	Penglais Farm (Aberystwyth University) will have a total of 1000 rooms, but only 700 will be ready [anticipation] this month to welcome [joy] students. The university said developer Balfour Beatty confirmed [trust] the remaining 300 rooms will be ready [anticipation] during the 2015–2016 academic year. Balfour Beatty has been asked to comment. The unfinished [¬anticipation] rooms have not been let to students.
Summary$_1$	Hundreds of rooms at a student halls development at Aberystwyth University will not be ready [¬anticipation] for the new term.
Summary$_2$	700 rooms at Aberystwyth University will be ready [anticipation] to welcome [joy] students this month.

In this work, we explore the usage of established methodologies to study the emotional content of summarization corpora and the emotional behavior of summarization models. Using these methodologies, we carry out the first study about the emotional content of news articles and their summaries. This study is mainly based on two measures to quantify the emotional content in texts at the word level: emotion density and emotion ratio [1], and is divided into two stages. First, we study the emotional content of two widely used news summarization corpora in the literature: CNN/DAILYMAIL [10] and XSUM [11]. Second, we study the capabilities of abstractive summarizer models for eliciting emotions in the generated summaries that match the emotions introduced by humans in reference summaries. This study has been performed on three state-of-the-art transformer-based systems [12]: BART [13], PEGASUS [14], and T5 [15]. This work aims to answer the following questions: (i) what and how frequent are the emotions in documents and summaries of both corpora; (ii) how emotion densities and ratios of the generated summaries correlate with densities and ratios of the reference summaries; and (iii) whether the emotions of novel words that appear in the generated summaries but not in the source articles match emotions of their reference summary. For reproducibility purposes, the software used in this work is freely available on GitHub (https://github.com/ELiRF/EmotionsInNewsSummarization, accessed on 10 January 2024).

2. Related Work

Automatic summarization has been addressed in the literature using mainly extractive or abstractive approaches. Extractive approaches build summaries by selecting text directly from the document [16–18], while abstractive systems build the summaries by paraphrasing text from the document [19,20]. Recently, strong efforts have been made in developing abstractive systems by focusing on encoder-decoder architectures pre-trained in self-supervised ways [13–15]. One of the best-known problems of these systems is related to hallucinating content, where the models are prone to generate content in the summaries that is not directly inferable from the source document. Several works aim to reduce hallucinations or improve the factual consistency of abstractive summarizers, e.g., employing content planning [21], reinforcement learning [22], or constraining the generation [23]. Abstractive summarizers could also be guided, for instance, to work better on aggregating semantic information [8], with specific topics [24], or to represent better the keywords and relationships among the entities [25,26].

Along with hallucinations, factuality, and abstractivity, emotions are also important to be studied in summarization systems and in the corpora used to train them. Since summaries are an entry point to the source article, the emotions elicited in the summaries directly impact the perception of the users. Few works have considered emotions for

summarization in dialogue or microblog summarization [5,6], but, to our knowledge, only [1] has studied emotions in automatic news summarization. They proposed an emotion-aware news summarization system and introduced the concepts of emotion densities and ratios, which we used extensively in our work. Similarly, in our work, we use them to study salient emotions in human-written summaries of two widely used summarization corpora (CNN/DailyMail and XSUM). Different from [1], we also study the emotional behavior of abstractive summarization systems, and we do not ground emotions to predefined categories since (i) articles from the considered categories are discarded, (ii) current summarization corpora do not consider categories, and (iii) we aim to obtain global insights of emotions at newspaper-level.

Emotions have been studied out of the scope of news summarization, to understand the affective state of users in applications such as e-commerce [27], opinion analysis in social media [28,29], or healthcare [30,31]. Emotions have also been studied in the news domain to detect fake news [32] or the stance toward specific targets [33]. To our knowledge, our work is the first to analyze emotions under the umbrella of news automatic summarization to obtain insights from the emotional content of news summarization corpora and the emotional behavior of abstractive summarizers.

3. Emotional Content Measures

We aim to quantify (i) how frequent an emotion is in a text and (ii) which emotions increase/decrease their frequency in summaries compared to their frequency in articles. We base our study on the methodology introduced in [1].

Following this methodology, we assume that the presence of an emotional word in a text is enough to convey some degree of an emotion. Although this assumption oversimplifies the problem because of the inherent limitations of lexicons, such as the lack of compositionality or ambiguity, having a moderately accurate fine-grained view of emotions in texts is useful. We use the NRC lexicon [34] (version 0.92), which contains 27 k words and their associations with the eight basic emotions in Plutchik's wheel (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust). Ten thousand of these words were manually annotated through crowdsourcing, and the remaining 17 k words are Wordnet synonyms of the annotated words. We use the NRC lexicon through the NRCLEX Python package to detect words with emotions. Words from texts and the NRC lexicon are lemmatized to deal with inflections.

To measure how frequent an emotion e is in a text t, we use the emotion density defined (ED) in Equation (1).

$$ED(e,t) = \frac{count(e,t)}{|t|} \tag{1}$$

where $count(e,t)$ is the number of words in the text t that convey the emotion e following the NRC lexicon, and $|t|$ is the number of words in t. We compute the emotion density on articles, reference summaries, and generated summaries.

To quantify emotions that appear more/less frequently in a summary than in an article, we use the emotion ratio (ER). The emotion ratio of an emotion e in an article-summary pair is defined in Equation (2).

$$ER(e,a,s) = \frac{ED(e,s)}{ED(e,a)} \tag{2}$$

where a is an article and s a summary. When $ER(e,a,s) > 1$, we say the emotion e is overemphasized in the summary. On the contrary, when $ER(e,a,s) < 1$, we state that the emotion is underemphasized in the summary. Intuitively, emotions that are more frequent in reference summaries than in source articles should also be more numerous in generated summaries [1]. To measure this, we compute the emotion ratios for both reference summaries and summaries generated by abstractive summarizers.

4. Summarization Corpora

To conduct our study about emotions in news corpora and abstractive models, we choose two reference corpora in English news summarization: CNN/DAILYMAIL and XSUM. Both corpora are publicly available on the HuggingFace hub: https://huggingface.co/datasets/cnn_dailymail (CNN/DAILYMAIL, version 3.0.0, accessed on 10 January 2024), and https://huggingface.co/datasets/xsum (XSUM, accessed on 10 January 2024). Table 2 shows the number of samples and statistics for documents and summaries for both corpora.

Table 2. Statistics for the two corpora: CNN/DAILYMAIL and XSUM. From left to right: corpus size, average document, and summary length (in terms of words and sentences), and vocabulary size in document and summary.

	#docs	avg. doc. Length		avg. sum. Length		Vocabulary Size	
		Words	Sentences	Words	Sentences	Document	Summary
CNN/DAILYMAIL	311,971	693.62	38.55	49.00	3.70	839,788	231,778
XSUM	226,711	377.51	19.20	21.33	1.00	425,532	83,414

5. Emotions in Summarization Corpora

First, we study how frequently the articles and summaries contain emotional words. To this aim, Table 3 shows the percentage of articles and summaries that has at least one word of an emotion.

Table 3. Percentage of articles and summaries in both corpora containing at least one word of an emotion.

		Fear	Anger	Anticipation	Trust	Surprise	Sadness	Disgust	Joy
CNN/DM	**%Docs**	99.66	98.81	99.58	99.98	99.24	99.67	97.17	99.40
	%Summs	79.22	69.63	84.23	90.12	60.36	76.38	53.81	69.79
XSUM	**%Docs**	95.72	91.74	98.60	99.04	93.58	96.26	85.70	94.97
	%Summs	59.45	46.89	61.19	70.52	38.36	54.17	31.75	44.38

Most articles in both corpora show some emotion, and it is common to see all the emotions co-occurring (77% of articles in XSUM and 95% in CNN/DAILYMAIL have words representing all the emotions at some point). It is not so in the summaries: the percentage of summaries that elicit each emotion is lower than the percentage of articles, especially in XSUM, and it is not as frequent as in the articles where all the emotions co-occur. *Fear*, *sadness*, *anticipation*, and *trust* are the emotions that appear in a more significant number of articles and summaries.

We carried out a study of the most frequent combination of emotions in the summaries. The study shows that there are larger combinations of emotions in CNN/DAILYMAIL than in XSUM, likely because summaries are twice as long. Interestingly, summaries of CNN/DAILYMAIL are twice as long as XSUM ones, but it is four times more likely that all emotions appear in their summaries (23.08% vs. 5.68%). Of theCNN/DailyMail summaries, 52.43% are in the top-10 combinations, while, in XSUM, the top-10 combinations accumulate 29.74% of the summaries. Figures A1 and A2 of Appendix A show this study.

We found 27.8k examples in XSUM (12%) and 2.8k in CNN/DAILYMAIL (0.9%) where the reference summaries elicit at least one emotion that does not appear in the article. For both corpora, the most frequent emotions in these cases are *disgust*, *anger*, and *surprise*, and the least frequent ones are *anticipation* and *trust*. Table A1 of Appendix B shows one example from XSUM. Second, by focusing on emotion densities and ratios, we study how frequently each emotion is elicited in articles and summaries and what emotions are over/under-emphasized in the summaries. Figure 1 shows Kernel Density Estimation (KDE) plots of emotion densities and ratios for each emotion in both corpora. In these

plots, the x-axes represent values of either emotion densities or ratios, and the y-axes define the probability density function for the kernel density estimation. The figure shows that emotion densities and ratios are similarly distributed.

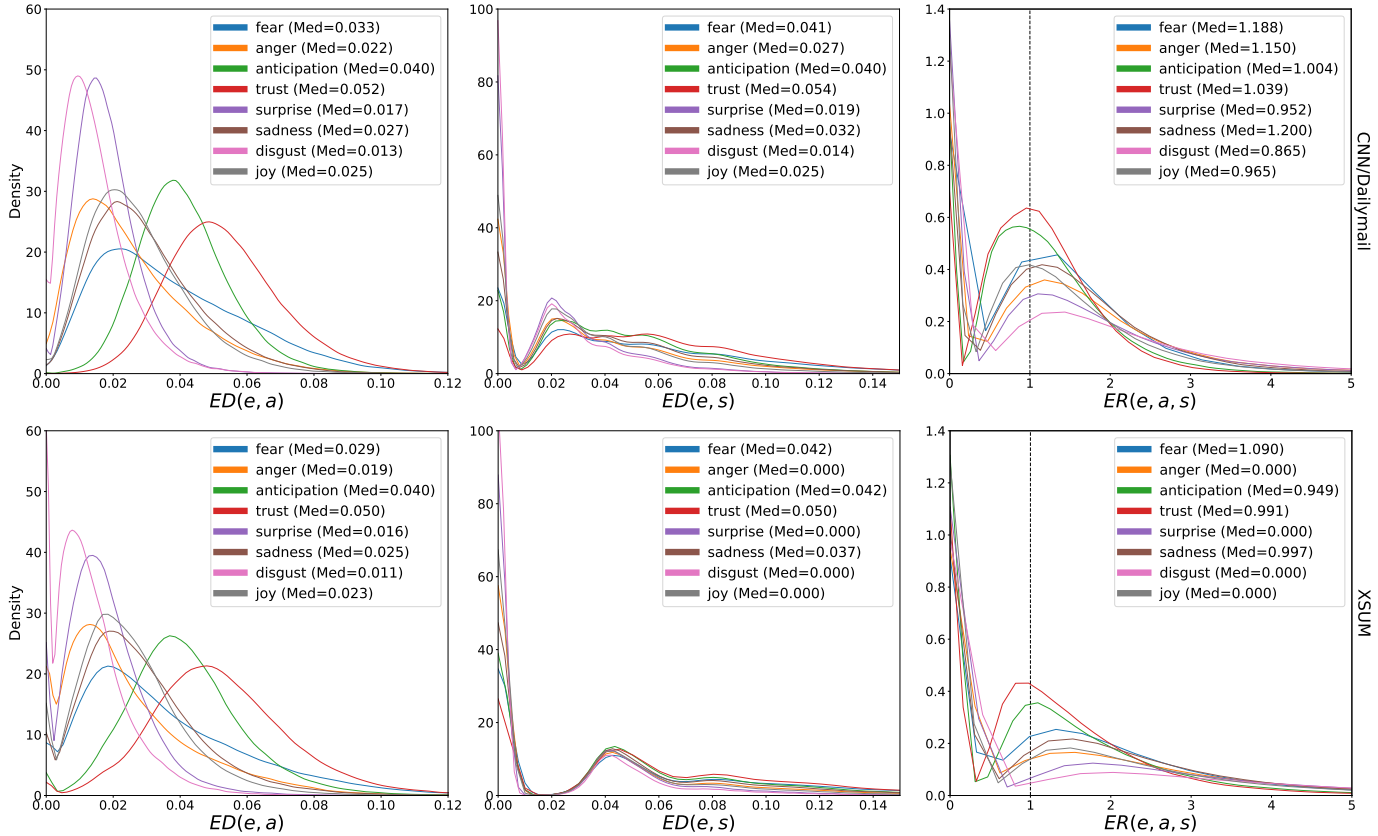

Figure 1. Density plots of emotion content measures in CNN/DAILYMAIL (top row) and XSUM (bottom row). The x-axes represent emotion densities in articles ($ED(e,a)$, left column), emotion densities in summaries ($ED(e,s)$, mid column), and emotion ratios ($ER(e,a,s)$, right column). To avoid undefined values in the emotion ratios, we discarded all those examples where $ED(e,a) = 0$. The legends include the median of each emotion.

Related to the articles (first column in Figure 1), *trust* concentrates the most significant number of articles with higher $ED(e,a)$. In contrast, *disgust* collects the most significant number of articles with lower $ED(e,a)$. The distribution of *fear* is the most skewed. Despite the differences among the $ED(e,s)$ and $ED(e,a)$ distributions, emotions in summaries (second column of Figure 1) show a similar behavior: *trust* concentrates the most significant number of summaries with higher $ED(e,s)$, and *disgust* with lower ones. In XSUM, the distributions are shifted toward higher values of $ED(e,s)$ compared to CNN/DAILYMAIL.

Regarding the ratios $ER(e,a,s)$ (third column in Figure 1), there is a tendency to overemphasize the emotion *fear* in both corpora, as suggested by the median. *Surprise*, *disgust*, and *joy* are underemphasized in both corpora. Interestingly, *disgust* is the emotion with the highest density in the tail, when $ER(e,a,s)$ is higher than $\sim$3 in CNN/DAILYMAIL and $\sim$4 in XSUM. In CNN/DAILYMAIL, the summaries tend to overemphasize emotions, especially the negative ones, while in XSUM they tend to overemphasize *fear*. In Table A2 of Appendix B, we show an example from XSUM where the emotion ratio of negative emotions is high. *Anger*, *surprise*, *disgust*, and *joy* show a median emotion ratio of 0 in XSUM. Therefore, the central tendency is not to include words with these emotions in the summary.

6. Emotions in Summarization Systems

In this section, we describe our study of the emotional behavior of three widely used state-of-the-art abstractive summarizers.

6.1. Models

We used three state-of-the-art abstractive summarization models, implemented in HuggingFace Transformers [35], as the main systems for our experimentation: BART [13], PEGASUS [14], and T5 [15]. Since the experimentation is performed with the CNN/DAILYMAIL and XSUM corpora, we used already finetuned checkpoints. For BART, we used *bart-large-cnn* (https://huggingface.co/facebook/bart-large-cnn, accessed on 10 January 2024) and *bart-large-xsum* (https://huggingface.co/facebook/bart-large-xsum, accessed on 10 January 2024). For PEGASUS, we used *pegasus-cnn_dailymail* (https://huggingface.co/google/pegasus-cnn_dailymail, accessed on 10 January 2024) and *pegasus-xsum* (https://huggingface.co/google/pegasus-xsum, accessed on 10 January 2024). Finally, for T5, due to the lack of checkpoints for these corpora in HuggingFace, we finetuned the *t5-base* (https://huggingface.co/t5-base, accessed on 10 January 2024) model with each corpus. The two T5 models have been trained with batches of 8 samples and a constant learning rate of 5×10^{-5}, using two GIGABYTE NVIDIA RTX 3090 GPUs hosted in our research laboratory. We used early stopping to stop the training after five epochs of patience on the validation loss.

We consider two baselines commonly used in the literature for completeness: LEAD and RANDOM. LEAD extracts the first sentence of the source article in XSUM and the first three sentences in CNN/DAILYMAIL. RANDOM extracts the same number of sentences as LEAD, but randomly selected from the source article. Additionally, we use an oracle to represent the best hypothetical summarization model. The oracle selects the sentence in the source article that maximizes the averaged ROUGE F_1 scores for each sentence in the reference summary.

For reproducibility, we show the results of these systems on the test sets in terms of ROUGE and BERTSCORE; measures commonly used in the literature for summarization [12]. The results are shown in Table A4 of Appendix C. The hyper-parameters used for the abstractive summarizers are shown in Table A5 of Appendix D.

6.2. Emotional Coherence and Bias

We analyze how emotion densities and emotion ratios of the generated summaries correlate with the corresponding metrics of the reference summaries. We introduce two metrics based on the Pearson correlation coefficient to this aim.

6.2.1. Emotional Coherence

Emotional coherence measures how the emotion densities for an emotion e in the generated summaries correlate with the emotion densities for e in the reference summaries. In that sense, it quantifies the strength and direction of the relation between the proportion of words with an emotion e in a generated summary and the proportion of words with that emotion in the reference summary. The emotional coherence for an emotion e is computed as the Pearson correlation between the emotion densities in the reference summaries $y = \{ED(e, s_1), \ldots, ED(e, s_N)\}$ and in the generated summaries $\hat{y} = \{ED(e, \hat{s}_1), \ldots, ED(e, \hat{s}_N)\}$. Figure 2 shows the emotional coherence between reference summaries and summaries generated by each model for all the emotions and corpora.

We observe that all the emotional coherences are higher than 0, suggesting positive relationships between the emotion densities. Abstractive models generally present a coherence higher than 0.5 in negative emotions: *fear*, *anger*, *sadness*, and *disgust*; and a coherence between 0.35 and 0.5 in the other emotions: *anticipation*, *trust*, *surprise*, and *joy*. Hence, abstractive models approximate better the emotion densities of negative emotions.

In XSUM, T5 is the abstractive model with the lowest emotional coherence and PEGASUS with the highest one. In CNN/DAILYMAIL, BART generally has a slightly higher emotional coherence than T5 and PEGASUS. All the abstractive systems show a similar emotional coherence in both corpora.

Baseline systems also show higher emotional coherence in negative emotions. However, different from abstractive ones, these systems show low emotional coherence in XSUM. LEAD shows an emotional coherence very similar to that of abstractive systems in CNN/DAILYMAIL (slightly higher for some emotions). Hence, the first sentences of the source articles keep moderately well, and similar to the abstractive models, the expected emotion densities in the summaries of CNN/DAILYMAIL. All the systems have higher emotional coherence than RANDOM in both corpora.

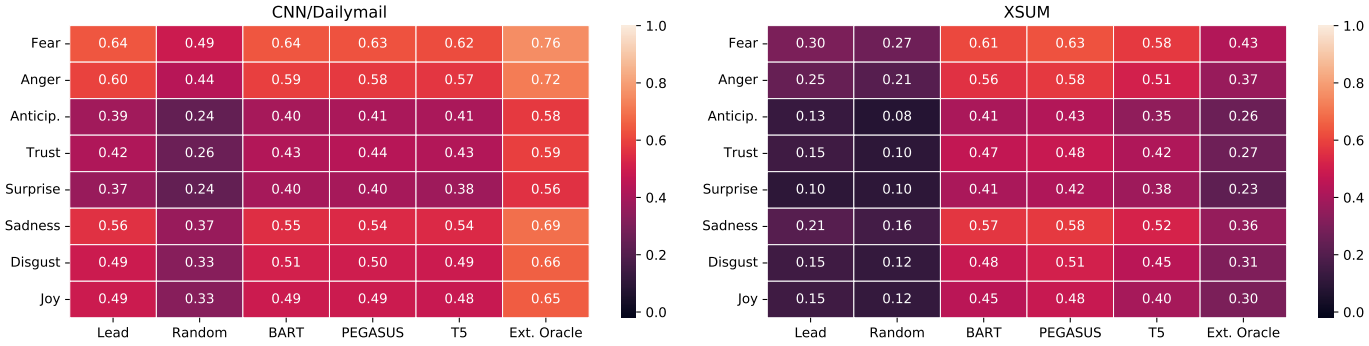

Figure 2. Emotional coherence of each model for each emotion in CNN/DAILYMAIL and XSUM. Correlations are statistically significant (p-value is 0 in all the cases).

The oracle shows the highest coherence in CNN/DAILYMAIL, suggesting that the emotional coherence of the abstractive models could be increased if they focus on better sentences from the source (in terms of ROUGE concerning the reference summary). It is not so in XSUM, where abstractive systems have higher coherence than the oracle. It suggests that focusing on the best sentences of the articles would not help to increase the emotional coherence in XSUM.

6.2.2. Emotional Bias

Emotional bias measures how the emotion ratios for an emotion e in the generated summaries correlate with the emotion ratios for e in the reference summary. Hence, it quantifies the strength and direction of the relation between the emphasis, regarding the source article, placed on an emotion e in a generated summary and the emphasis placed on that emotion in the reference summary. The emotional bias for an emotion e is computed as the Pearson correlation between the emotion ratios in the reference summaries $y = \{ER(e, s_1, a_1), \ldots, ER(e, s_N, a_N)\}$ and in the generated summaries $\hat{y} = \{ER(e, \hat{s}_1, a_1), \ldots, ER(e, \hat{s}_N, a_N)\}$. To compute the emotional bias, we discard all those examples where the emotion ratio is undefined (when $ED(e, a) = 0$).

Figure 3 shows the emotional bias between reference summaries and summaries generated by each model for all the emotions and corpora. In almost all the cases, the emotional biases are higher than 0, suggesting positive relationships between emotion ratios. The strength of the correlations is notably lower than in the emotional coherence (Figure 2). It suggests it is more difficult to approximate the emotion ratios than the emotion densities.

The abstractive systems show higher emotional bias in XSUM than in CNN/DAILYMAIL. In XSUM, T5 is the abstractive model with the lowest emotional bias. PEGASUS shows the highest emotional bias for almost all emotions in CNN/DAILYMAIL and XSUM. All the abstractive systems show, in XSUM, the lowest emotional bias for *anger* and *surprise*, and the highest emotional bias for *sadness*, *fear*, and *trust*. The emotional biases of abstractive systems are similar for all the emotions in CNN/DAILYMAIL.

Baseline models, LEAD and RANDOM, show a low emotional bias in CNN/DAILYMAIL and a negligible one (close to 0) in XSUM. The low emotional bias of LEAD indicates that the first sentences of the source articles do not show the expected emotion ratios in the summaries neither of CNN/DAILYMAIL nor Xsum. In CNN/DAILYMAIL, LEAD shows a slightly lower emotional bias than abstractive models for all the emotions, but in XSUM, the difference concerning abstractive models is high. All the systems show higher emotional bias than RANDOM.

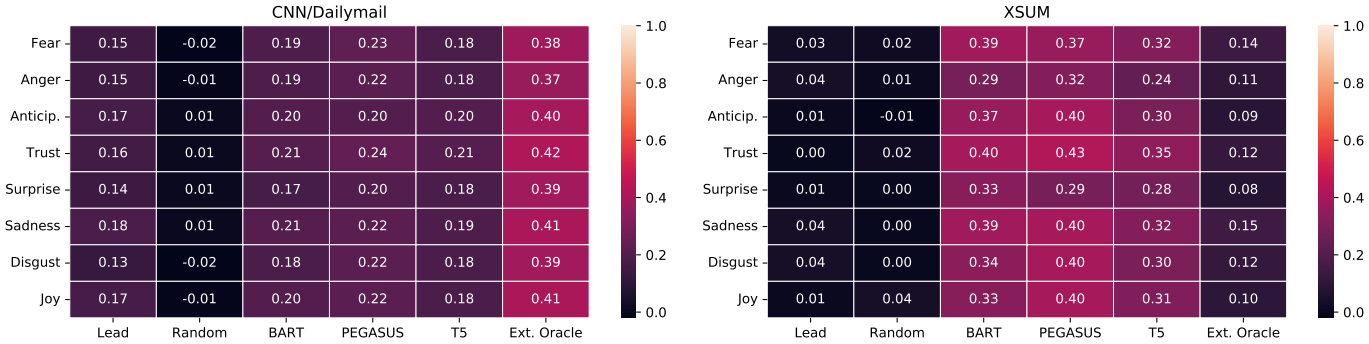

Figure 3. Emotional bias of each model for each emotion in CNN/DAILYMAIL and XSUM. Correlations are statistically significant (p-value is 0 in all the cases).

The oracle shows the highest emotional bias in CNN/DAILYMAIL but not in XSUM, where the abstractive models stand out. It could suggest again that abstractive models could increase their emotional bias in CNN/DAILYMAIL if they focus on better sentences from the source article (in terms of ROUGE with respect to the reference summary) but not in XSUM.

6.3. Emotions of Novel Words

Abstractive summarizers are moderately good at generating summary-worthy novel words that are not present in the source. These novel words could convey a set of emotions. However, whether the emotions of the novel words are those expected in the reference summary is still unclear. We study it on the test sets of CNN/DAILYMAIL and XSUM by computing the precision between the emotions of the novel words in a generated summary and all the emotions in the reference summary.

Let $\mathcal{E}_s$ be the set of emotions in a reference summary and $\mathcal{E}_{\hat{s}}$ the set of emotions of the novel words in a generated one, precision (P) for N samples is computed as shown in Equation (3); where the intersection refers to the emotions in common between those found in the novel words and the reference summary. We only consider those cases where there are novel words with emotions in the generated summary ($|\mathcal{E}_{\hat{s}}| > 0$) and the reference one has words with emotions ($|\mathcal{E}_s| > 0$).

$$\mathrm{P} = \frac{1}{N} \sum_{\forall(s,\hat{s})} \frac{|\mathcal{E}_s \cap \mathcal{E}_{\hat{s}}|}{|\mathcal{E}_{\hat{s}}|} \tag{3}$$

We also compute the recall (R) to see how many of the emotions in the reference summary are covered by the emotions of the novel words in the generated summary. Recall is computed as shown in Equation (4).

$$\mathrm{R} = \frac{1}{N} \sum_{\forall(s,\hat{s})} \frac{|\mathcal{E}_s \cap \mathcal{E}_{\hat{s}}|}{|\mathcal{E}_s|} \tag{4}$$

Table 4 shows precision and recall for each model and corpora, along with other data statistics used to compute them. Most of the novel words generated by the models have emotions that match those of the reference summaries since precision is higher than 75% in

all cases. PEGASUS is the system that shows the highest precision in both corpora. T5 has slightly higher precision than BART in CNN/DAILYMAIL, but not in XSUM.

The precision of all the models is higher in CNN/DAILYMAIL than in XSUM. The abstractive models generate more novel words in XSUM (4.9 novel words per summary) than in CNN/DAILYMAIL (0.9 novel words per summary). Then, generating more novel words will likely include more non-expected emotions. Table A3 of Appendix B shows an example from XSUM where the emotions of the novel words in a summary generated by PEGASUS do not match exactly the emotions of the reference summary.

Table 4. Precision and recall of the emotions in the novel words generated by each model, compared to the emotions of the reference summaries. The number of samples without (w/o) novel words in the generated summary and w/o emotions in the novel words of the generated summary ($|\mathcal{E}_{\hat{s}}| = 0$) are also shown. The last column indicates the number of samples finally considered in the evaluation. We also show percentages of samples in the test sets.

		Precision	Recall	Samples w/o Novel Words	Samples $\|\mathcal{E}_{\hat{s}}\| = 0$	Samples
CNNDM	**BART**	84.73	38.79	7912 (68.9%)	2499 (21.7%)	1022 (8.9%)
	PEGASUS	85.82	36.75	6394 (55.6%)	3332 (29.0%)	1707 (14.9%)
	T5	84.99	34.47	7019 (61.1%)	3090 (26.9%)	1336 (11.6%)
XSUM	**BART**	76.10	49.49	324 (2.9%)	3717 (32.8%)	6704 (59.1%)
	PEGASUS	77.73	51.01	279 (2.5%)	3685 (32.5%)	6781 (59.8%)
	T5	75.30	47.21	354 (3.1%)	3901 (34.4%)	6554 (57.8%)

Both in CNN/DAILYMAIL and XSUM, PEGASUS generates novel words in more samples than BART and T5 (lowest **Samples w/o novel words**). However, for a larger number of samples than BART and T5 in CNN/DAILYMAIL, the novel words generated by PEGASUS do not convey emotions (highest **Samples** $|\mathcal{E}_{\hat{s}}| = 0$). By contrast, PEGASUS generates emotional novel words for a slightly larger number of samples than BART and T5 in XSUM (lowest **Samples** $|\mathcal{E}_{\hat{s}}| = 0$). We notice that the models generate more novel words in XSUM than in CNN/DAILYMAIL, but the number of samples where novel words do not convey emotions is similar in both corpora.

Interestingly, the recall is between 34% and 51%, which suggests that the emotions of the novel words are enough to cover, approximately, at least a third part of the overall emotional content of the reference summaries. BART has the highest recall in CNN/DAILYMAIL and PEGASUS in XSUM. Although it is difficult to explain why, the number of emotions in the reference summaries (lower in XSUM than in CNN/DAILYMAIL) could play a big role.

Considering the overall results of the two corpora, the difference in recall is significant. We consider that it is due to the difference in the introduction of new words in both cases. The fact that the XSUM corpus is much more abstractive in nature than the CNN/DAILYMAIL means that the former incorporates a greater number of novel words, and therefore, more emotional content.

7. Discussion

We summarize the most important contributions and findings of this work in relation to the objectives stated in the introduction section.

Emotional content of summarization corpora. First, we found that 99% of articles and 70% of summaries of the studied corpora contain at least one emotion. We also found that 12% in XSUM and 0.9% in CNN/DAILYMAIL of the reference summaries elicit at least one emotion that does not appear in the article. Second, we applied two measures, emotion density, and emotion ratio, to articles and summaries of both corpora and the results that we analyzed. Related to the articles, we observed that *trust* concentrates the most significant number of articles with higher emotion densities. In contrast, *disgust* concentrates the largest number of articles with lower emotion densities. Related to emotions in summaries, we noticed a similar behavior. In XSUM, the distributions are shifted toward higher values

of emotion densities compared to CNN/DAILYMAIL. Regarding the emotion ratios, there is a tendency to overemphasize the emotion *fear* in both corpora. In CNN/DAILYMAIL, the summaries tend to overemphasize emotions, especially the negative ones, while in XSUM they tend to overemphasize *fear*.

Emotional behavior of summarization models. We introduced two new measures, *emotional coherence* and *emotional bias*, to measure how the emotion densities and ratios of generated summaries correlate with those of the reference. We found that all the emotional coherences are higher than 0, suggesting positive relationships between the emotion densities. Abstractive models generally present a coherence higher than 0.5 in negative emotions: *fear*, *anger*, *sadness*, and *disgust*; and a coherence between 0.35 and 0.5 in the other emotions. Additionally, we found a higher emotional bias in XSUM than in CNN/DAILYMAIL. In XSUM, T5 is the abstractive model with the lowest emotional bias. PEGASUS shows the highest emotional bias for almost all emotions in CNN/DAILYMAIL and XSUM. Also, we analyzed whether the novel words generated by the summarization models convey the emotions expected in their reference summaries. We observed that most of the novel words generated by the models have emotions that match those of the reference summaries. Interestingly, the recall is between 34% and 51%, which suggests that the emotions of the novel words are enough to cover, approximately, at least a third part of the emotions in the reference summaries.

Finally, we should remark that the proposed methodology is valid for studying emotions in summarization regardless of the method used to detect emotions. However, the approach used in this work presents some limitations since we assumed that the presence of an emotional word in a text is enough to convey some degree of an emotion. Although this assumption oversimplifies the problem because of the inherent limitations of lexicons, such as the lack of compositionality or ambiguity, having a moderately accurate fine-grained view of emotions in texts is helpful. Therefore, we detected emotions at the word level using lexicons, although other alternatives could exist.

8. Conclusions

We studied the prevalence of emotions in news summarization corpora, specifically, how much these emotions are emphasized in the summaries compared to the source article and the capabilities of state-of-the-art abstractive summarizers for eliciting expected emotions in the generated summaries.

A large percentage of articles and summaries in CNN/DAILYMAIL and XSUM elicit emotions, especially *fear*, *sadness*, *anticipation*, and *trust*. Our findings also suggest that reference summaries in CNN/DAILYMAIL overemphasize negative emotions, while XSUM underemphasizes all the emotions except *fear*. Abstractive summarizers approach moderately well the emotion densities in the summaries. However, they do not show the same emotional bias as human summarizers when emphasizing emotions in the summaries. Finally, we noticed that most of the novel words generated by the models convey emotions expected in the reference summaries, especially in CNN/DAILYMAIL, where the models generate few novel words.

In future work, we plan to develop news summarization models with controllable text generation driven by the emotions of the reference summaries and via prompting [36], which could produce better emotional coherence in the generated summaries and potentially, reduce undesired biases towards some emotions and stances.

Author Contributions: Conceptualization, V.A. and J.-Á.G.; methodology, J.-Á.G.; software, V.A. and J.-Á.G.; validation, V.A., L.-F.H. and E.S.; formal analysis, J.-Á.G.; investigation, J.-Á.G.; resources, L.-F.H. and E.S.; data curation, V.A.; writing—original draft preparation, V.A., J.-Á.G., L.-F.H. and E.S.; writing—review and editing, V.A., J.-Á.G., L.-F.H. and E.S.; visualization, V.A.; supervision, L.-F.H. and E.S.; funding acquisition, L.-F.H. and E.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partially supported by MCIN/AEI/10.13039/501100011033, by the "European Union and "NextGenerationEU/MRR", and by "ERDF A way of making Europe" under grants PDC2021-120846-C44 and PID2021-126061OB-C41. It is also partially supported by the Generalitat Valenciana under project CIPROM/2021/023, and by the Spanish Ministerio de Universidades under the grant FPU21/05288 for university teacher training.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: CNN/DAILYMAIL, version 3.0.0 at https://huggingface.co/datasets/cnn_dailymail/viewer/3.0.0 (accessed on 10 January 2024), and XSUM at https://huggingface.co/datasets/EdinburghNLP/xsum (accessed on 10 January 2024).

Conflicts of Interest: Author José-Ángel González was employed by the company Symanto Research. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A. 10-Top Emotions in Corpora

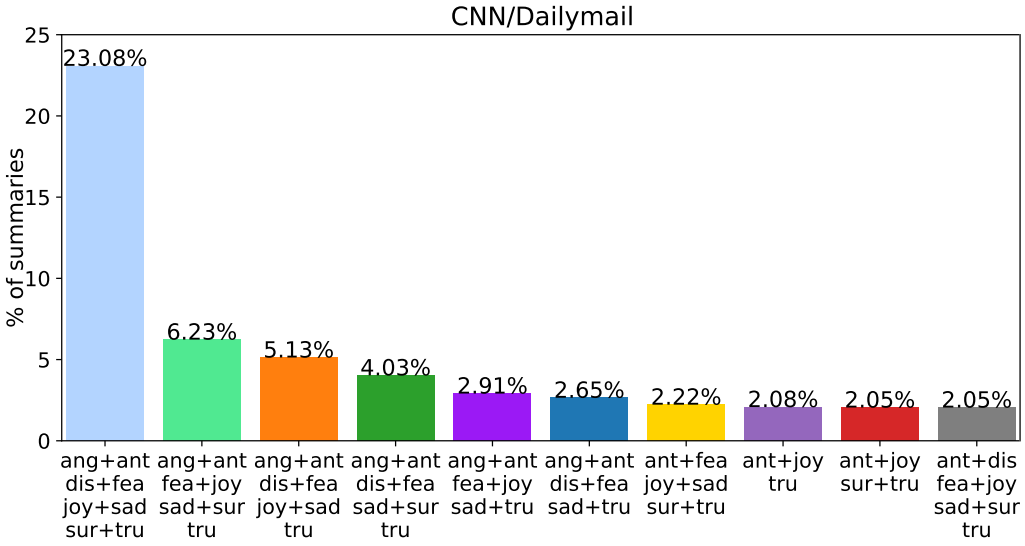

Figure A1. Ten most frequent combinations of emotions in the summaries of CNN/DAILYMAIL. Bar labels indicate the percentage of summaries in the whole corpus.

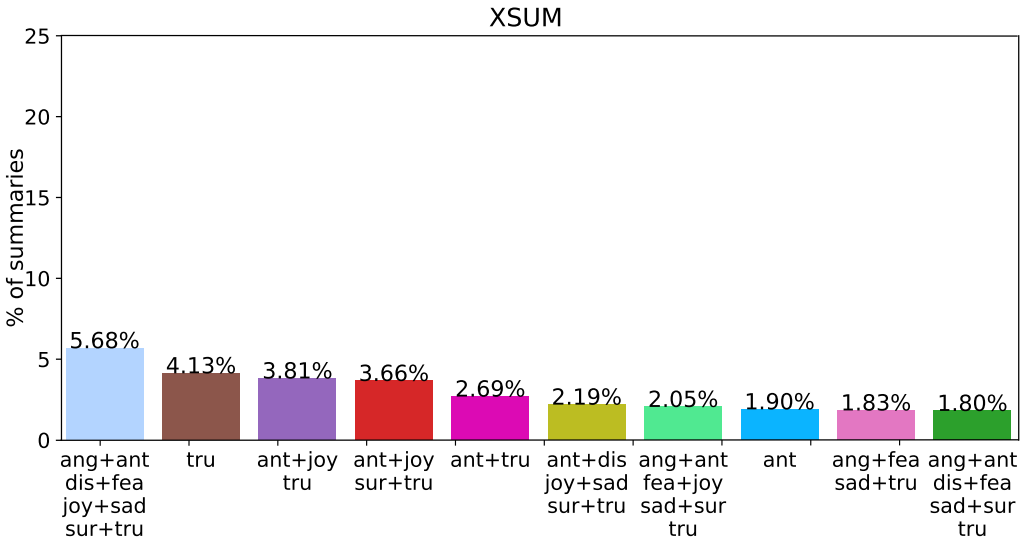

Figure A2. Ten most frequent combinations of emotions in the summaries of XSUM. Bar labels indicate the percentage of summaries in the whole corpus.

Appendix B. Examples of Emotions in News Summaries

Table A1. An example from XSUM where the emotions *fear*, *anger*, and *disgust* appear in the summary but not in the article. Bold underlined words appear in the NRC lexicon, and words in brackets are their emotions.

Article	The 29-year-**old** **[*sadness*]**, who had only joined from AC Milan three days earlier, was **accidentally** **[*surprise*]** **caught** **[*surprise*]** on the right side of their head by Lorient midfielder Didier Ndong 15 min into the match. "Jeremy was operated on under local anaesthesia on Wednesday night to sew up their ear," Bordeaux said. "We wish a speedy recovery to our player and **hope** **[*anticipation, joy, surprise, trust*]** to see them back soon."international Ndong said: "I apologise to Jeremy Menez, and to Bordeaux, and **hope** **[*anticipation, joy, surprise, trust*]** to see them back in Ligue 1 action soon. "It was completely involuntary but had **unfortunate** **[*sadness*]** consequences for him.
Reference summary	Bordeaux's France international forward Jeremy Menez **lost** **[*anger, disgust, fear, sadness, surprise*]** **part** **[*sadness*]** of their right ear in a pre-season game with FC Lorient.

Table A2. An example from XSUM where the emotion ratio, computed as shown in Equation (2), of *fear* (5.29), *anger* (7.05), and *sadness* (5.29) is higher than 5, i.e., the proportion of words with these emotions in the summary is more than 5 times higher than the proportion in the source article. Bold underlined words are those present in the NRC lexicon, and words in brackets are their emotions. Punctuation marks are not counted as words.

Article	A telehandler vehicle was used to smash a wall at the rear of the Sainsbury's Local store on Bingham Road in Cotgrave at about 04:00 BST on Monday. The **cash** **[*trust, joy, anticipation, anger, fear*]** **machine** **[*trust*]** was taken away in another vehicle described as a white vehicle, possibly an Audi, **police** **[*trust, fear*]** said. **Officers** **[*trust*]** have **sealed** **[*trust*]** off the area and have appealed for any **witnesses** **[*trust*]** to **come** **[*anticipation*]** forward. It is unclear how much of the shop has been **damaged** **[*anger, disgust, sadness*]**.
Reference summary	A **cash** **[*trust, joy, anticipation, anger, fear*]** **machine** **[*trust*]** has been **stolen** **[*anger, fear, sadness*]** in a **ram** **[*anger, anticipation*]** **raid** **[*anger, fear, surprise*]** on a Nottinghamshire supermarket

Table A3. An example from XSUM where the emotions of the novel words in a summary generated by PEGASUS do not match exactly the emotions of the reference summary (precision = 0.25). Bold underlined words are those present in the NRC lexicon, and words in brackets are their emotions. Words in **blue** are the novel words in the generated summary.

Article	The McGill's 904 service went up in flames just outside Largs on the A760 Kilbirnie Road at about 13:35 on Saturday. Emergency services attended but the driver and passengers were uninjured. A woman whose partially-sighted mother was on board later thanked the driver for keeping everyone safe. Kathleen McKenna told the BBC: "The bus started filling up with smoke. "The driver told everyone to get off as quickly as possible. He then made sure everyone was as far away as possible. "The bus was popping and banging as the fire took hold. The driver did really, really well. "The police arrived and asked if anyone needed to go to hospital but they were all fine. They just needed a cup of tea. "Police Scotland said the road was closed for a time but later re-opened. The burnt-out bus has been removed.
Reference summary	A bus driver whose vehicle **caught** **[surprise]** **fire** **[fear]** in North Ayrshire has been **praised** **[joy, trust]** after all the **passengers** **[anticipation]** were **safely** **[joy, trust]** **evacuated** **[fear]**.
Emotions in reference	surprise, fear, joy, trust, anticipation
PEGASUS summary	A bus has been **badly** **[sadness]** **damaged** **[anger, disgust, sadness]** after **catching** **[surprise]** fire in North Ayrshire.
Emotions in novel words	sadness, anger, disgust, surprise

Appendix C. ROUGE/BertScore Performance

We evaluated the quality of the summaries generated by the models in terms of ROUGE and BERTSCORE. Table A4 shows the results.

Table A4. ROUGE (R) and BERTSCORE (BS) F_1-scores for all the models and corpora.

		R1	R2	RL	BS
CNNDM	**Lead**	40.05	17.48	36.34	23.45
	Random	28.48	8.34	25.51	11.88
	BART	43.76	20.86	40.68	33.64
	PEGASUS	43.96	21.38	41.07	35.18
	T5	43.03	20.31	40.04	32.87
	Extractive oracle	52.34	30.23	48.86	39.77
XSUM	**Lead**	16.71	1.65	12.30	14.27
	Random	15.23	1.77	11.38	11.71
	BART	45.23	22.13	37.02	50.13
	PEGASUS	47.16	24.58	39.31	52.74
	T5	40.98	18.02	32.99	48.85
	Extractive oracle	29.38	8.68	22.43	22.66

Appendix D. Generation Hyperparameters

For reproducibility, we show in Table A5 the hyperparameters used for the "*generate*" method from HuggingFace Transformers. We tried to keep them similar to the original implementations [13–15].

Table A5. Hyperparameters used during generation for all models and corpora.

	Model	Length Penalty	Max Length	Min Length	N-gram Blocking	Num Beams
CNN/DailyMail	**BART**	2.0	142	56	3-g	4
	PEGASUS	0.8	128	32	No	4
	T5	2.0	142	56	3-g	6
XSUM	**BART**	1.0	62	11	3-g	6
	PEGASUS	0.6	64	No	No	6
	T5	1.0	62	11	No	6

References

1. Kennedy, A.; Kazantseva, A.; Inkpen, D.; Szpakowicz, S. Getting Emotional about News Summarization. In Proceedings of the Advances in Artificial Intelligence, Toronto, ON, Canada, 28–30 May 2012; pp. 121–132. [CrossRef]
2. Beckett, C.; Deuze, M. On the Role of Emotion in the Future of Journalism. *Soc. Media Soc.* **2016**, *2*, 3. [CrossRef]
3. Lecheler, S. The Emotional Turn in Journalism Needs to be About Audience Perceptions. *Digit. J.* **2020**, *8*, 287–291. [CrossRef]
4. Richardson, N. Journalism and Emotion. *Aust. J. Rev.* **2020**, *42*, 339–340. [CrossRef]
5. Chen, Y.; Liu, Y.; Chen, L.; Zhang, Y. DialogSum: A Real-Life Scenario Dialogue Summarization Dataset. In Proceedings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, 1–6 August 2021; pp. 5062–5074. [CrossRef]
6. Panchendrarajan, R.; Hsu, W.; Li Lee, M. Emotion-Aware Event Summarization in Microblogs. In Proceedings of the Companion Proceedings of the Web Conference 2021, Ljubljana, Slovenia, 19–23 April 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 486–494.
7. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*; Association for Computational Linguistics: Barcelona, Spain, 2004; pp. 74–81.
8. Jumel, C.; Louis, A.; Cheung, J.C.K. TESA: A Task in Entity Semantic Aggregation for Abstractive Summarization. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 8 April 2020; pp. 8031–8050. [CrossRef]
9. Maynez, J.; Narayan, S.; Bohnet, B.; McDonald, R. On Faithfulness and Factuality in Abstractive Summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 1906–1919. [CrossRef]
10. Hermann, K.M.; Kočiský, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; Blunsom, P. Teaching Machines to Read and Comprehend. In Proceedings of the NIPS'15: 28th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Volume 1, pp. 1693–1701.
11. Narayan, S.; Cohen, S.B.; Lapata, M. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 1797–1807. [CrossRef]
12. Fabbri, A.R.; Kryściński, W.; McCann, B.; Xiong, C.; Socher, R.; Radev, D. SummEval: Re-evaluating Summarization Evaluation. *Trans. Assoc. Comput. Linguist.* **2021**, *9*, 391–409. [CrossRef]

13. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 7871–7880. [CrossRef]
14. Zhang, J.; Zhao, Y.; Saleh, M.; Liu, P.J. PEGASUS: Pre-Training with Extracted Gap-Sentences for Abstractive Summarization. In Proceedings of the 37th International Conference on Machine Learning, Online, 13–18 July 2020; pp. 11328–11339.
15. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
16. Liu, Y.; Lapata, M. Text Summarization with Pretrained Encoders. 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), Hong Kong, China, 3–7 November 2019; pp. 3730–3740. [CrossRef]
17. Zhong, M.; Liu, P.; Chen, Y.; Wang, D.; Qiu, X.; Huang, X. Extractive Summarization as Text Matching. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 6197–6208. [CrossRef]
18. Mutlu, B.; Sezer, E.A.; Akcayol, M.A. Candidate sentence selection for extractive text summarization. *Inf. Process. Manag.* **2020**, *57*, 102359. [CrossRef]
19. See, A.; Liu, P.J.; Manning, C.D. Get To The Point: Summarization with Pointer-Generator Networks. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 1073–1083. [CrossRef]
20. Gehrmann, S.; Deng, Y.; Rush, A. Bottom-Up Abstractive Summarization. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 4098–4109. [CrossRef]
21. Narayan, S.; Zhao, Y.; Maynez, J.; Simões, G.; Nikolaev, V.; McDonald, R. Planning with Learned Entity Prompts for Abstractive Summarization. *Trans. Assoc. Comput. Linguist.* **2021**, *9*, 1475–1492. [CrossRef]
22. Zhang, M.; Zhou, G.; Yu, W.; Liu, W. FAR-ASS: Fact-aware reinforced abstractive sentence summarization. *Inf. Process. Manag.* **2021**, *58*, 102478. [CrossRef]
23. Zhao, Z.; Cohen, S.B.; Webber, B. Reducing Quantity Hallucinations in Abstractive Summarization. In Proceedings of the Association for Computational Linguistics: EMNLP 2020, Online, 16–20 November 2020; pp. 2237–2249. [CrossRef]
24. Belwal, R.C.; Rai, S.; Gupta, A. Text summarization using topic-based vector space model and semantic measure. *Inf. Process. Manag.* **2021**, *58*, 102536. [CrossRef]
25. Dou, Z.Y.; Liu, P.; Hayashi, H.; Jiang, Z.; Neubig, G. GSum: A General Framework for Guided Neural Abstractive Summarization. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021; pp. 4830–4842. [CrossRef]
26. Liu, S.; Cao, J.; Yang, R.; Wen, Z. Key phrase aware transformer for abstractive summarization. *Inf. Process. Manag.* **2022**, *59*, 102913. [CrossRef]
27. Bielozorov, A.; Bezbradica, M.; Helfert, M. The Role of User Emotions for Content Personalization in e-Commerce: Literature Review. In Proceedings of the HCI in Business, Government and Organizations. eCommerce and Consumer Behavior, Orlando, FL, USA, 26–31 July 2019; Nah, F.F.H., Siau, K., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 177–193.
28. Mohammad, S.; Bravo-Marquez, F.; Salameh, M.; Kiritchenko, S. SemEval-2018 Task 1: Affect in Tweets. In Proceedings of the 12th International Workshop on Semantic Evaluation, New Orleans, LA, USA, 5–6 June 2018; pp. 1–17. [CrossRef]
29. del Arco, F.M.P.; Jiménez-Zafra, S.M.; Montejo-Ráez, A.; Molina-González, M.D.; Ureña-López, L.A.; Martín-Valdivia, M.T. Overview of the EmoEvalEs task on emotion detection for Spanish at IberLEF 2021. *Proces. Del Leng. Nat.* **2021**, *67*, 155–161.
30. Muñoz, S.; Iglesias, C.A. A text classification approach to detect psychological stress combining a lexicon-based feature framework with distributional representations. *Inf. Process. Manag.* **2022**, *59*, 103011. [CrossRef]
31. Dheeraj, K.; Ramakrishnudu, T. Negative emotions detection on online mental-health related patients texts using the deep learning with MHA-BCNN model. *Expert Syst. Appl.* **2021**, *182*, 115265. [CrossRef]
32. Kumari, R.; Ashok, N.; Ghosal, T.; Ekbal, A. What the fake? Probing misinformation detection standing on the shoulder of novelty and emotion. *Inf. Process. Manag.* **2022**, *59*, 102740. [CrossRef]
33. Mascarell, L.; Ruzsics, T.; Schneebeli, C.; Schlattner, P.; Campanella, L.; Klingler, S.; Kadar, C. Stance Detection in German News Articles. In Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER), Punta Cana, Dominican Republic, 10 November 2021; pp. 66–77. [CrossRef]
34. Mohammad, S.M.; Turney, P.D. Crowdsourcing a Word-Emotion Association Lexicon. *Comput. Intell.* **2013**, *29*, 436–465. [CrossRef]
35. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 38–45.
36. Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; Neubig, G. Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.* **2023**, *55*, 195. [CrossRef]