Counterfeit: Sentiment Improved Text Summarization

Piotr Parkitny
University of California, Berkeley
pparkitny@berkeley.edu

Abstract

I present Counterfeit, abstractive text summarization model that improves on existing approaches by adding an additional sentiment model used for text preprocessing. The sentiment model duplicates strong sentiment sentences to bias or deceive (Counterfeit) the second model into including those duplicated sentences in the summary. Counterfeit improves the performance of pretrained models ROUGE score for a new dataset. Counterfeit hyperparameters allow for precision-recall trade-off which results in the ability to create adjustable summaries that are more suitable in a practical setting.

1 Introduction

The Transformer(Vaswani et al., 2017) is the dominant architecture for natural language processing tasks. Its use in creating abstractive text summarizations has created new baseline benchmarks for state-of-the-art results.

In Counterfeit I use the Transformers library (Wolf et al., 2020) that is dedicated in supporting Transformer-based architectures and distribution of pretrained models to use two fully stacked pretrained models. The first model will be pretrained for sentence sentiment task and the second model on text summarization. Feeding results from the first models into the second to create the final text summarization. The problem that Counterfeit will solve is improving text summarization on datasets with no ground-truth available for training. In real-life application where ground-truth summaries are not available Counterfeit provides ROUGE score improvement.

The novelty of this approach is that it allows to transfer the knowledge from one model to another using the sentence duplication task. I have chosen to duplicate sentences with strong sentiment but it could be sentences with places, people, events, dates or any other sentence whose probability of making it into the summary needs to be increased.

2 Background

The problem of text summarization is well known with many existing solutions. I am going to use pretrained models that are available for download from the Transformers library. I will use 4 different text summarization models and one sentiment model.

I have decided to use two BART based models: facebook/bart-large-cnn and facebook/bart-large-xsum and two distillation models: sshleifer/distilbart-cnn-12-6 and sshleifer/distilbart-xsum-12-6. BART (Lewis et al., 2019) a denoising autoencoder for pretraining sequence-to-sequence models where the distillation (Shleifer and Rush, 2020) version is created by distilling BART and creating a smaller faster version while retaining strong performance which makes it better for practical use.

For sentiment analysis which is the first part of my stacked model I have chosen to also use a model from the Transformers library. The model is cardiffnlp/twitter-roberta-base-sentiment which is a roBERTa-base model trained on 58M tweets and finetuned for sentiment analysis with the TweetEval (Barbieri et al., 2020) benchmark. The Roberta model is a text classification model that produces three scores for each sentence it is provided. The three scores are:

- LABEL0 - [NEG] negative sentiment score for the sentence
- LABEL1 - [NEU] neutral sentiment score for the sentence
- LABEL2 - [POS] positive sentiment score for the sentence

Having a scalar metric based on NEG,NEU and POS sentence sentiment is a key in identifying sentences that should be duplicated. In Counterfeit only the NEG and POS scores are used as they identify strong sentiment sentences. Provided below are three sentences along with a score that is produced by the sentiment model.

- A bomb blast on a bus kills 12 people: NEG=0.952 NEU=0.045 POS=0.003
- This is the best assignment ever: NEG=0.002 NEU=0.007 POS=0.991
- I am going home: NEG=0.079 NEU=0.729 POS=0.191

3 Objective

The purpose behind Counterfeit is to achieve better text summarization from pretrained models by stacking them together. The baseline is created using an existing pretrained summarization model and comparing that model against itself with a stacked sentiment model on top of it.

3.1 Evaluation Metric

Counterfeit evaluation metric is the ROUGE score (Lin, 2004). ROUGE stands for Recall Oriented Understudy for Gisting Evaluation. ROUGE is a set of metrics specifically designed for evaluation of automatic summarizations. ROUGE score is composed of three values

1. F1-Score - calculated as the harmonic mean of precision and recall.
2. Precision - fraction of relevant instances among the retrieved instances. It is the accuracy of positive predictions.
3. Recall - fraction of relevant instances that were retrieved. It is the ratio of positive instance that are correctly detected.

ROUGE score has different versions. I will be using ROUGE-1, ROUGE-2, and ROUGE-L scores as those are the most popular ROUGE versions used at this time. ROUGE-1 measures the overlap of unigram between the generated summary and ground-truth summary where ROUGE-2 refers to the overlap of bigrams. ROUGE-L is the longest common subsequence-based statistics. As it stands the ROUGE score is the most common evaluation used for text summarization. It is the right evaluation since at the time is the industry standard. Scoring text summarization using the ROUGE score is what other researchers are using currently. In practice ROUGE Recall metric is the most useful metric to use but it can be easily gamed which is the reason why the F1 metric is primarily used for comparing results.

The baseline for comparison will be the ROUGE score produced by the pretrained models compared to Counterfeit results that uses the same model. It should be noted that ROUGE score is not an ideal method for comparison based on reviewing ROUGE scored examples but is the best choice.

4 Methods

4.1 Architecture

Counterfeit architecture can be described as a stacked model pipeline where we feed information thought stages and obtain the summarized text. The overall model is composed of two tasks. First create the duplicate text using the sentiment model task and then create the summarization using the summarization model task.

4.2 Stage 1:Sentiment Model Task

The sentiment model will be used to score sentences so they can be duplicated if they have a strong sentiment. The cardiffnlp/twitter-roberta-base-sentiment model available from the Hugging Face website is used for this task. Sentence Boundary Detection which is the splitting of the article into individual sentences is done using the spaCy python package. spaCy brands itself as Industrial-Strength Natural Language Processing package that is free, open-source and written in Cython. It is a very well developed and documented package.

4.3 Stage 2: Summarization Models Task

I am using four different models for text summarization available from the Hugging Face website:

- facebook/bart-large-cnn
- facebook/bart-large-xsum
- sshleifer/distilbart-cnn-12-6
- sshleifer/distilbart-xsum-12-6

The four BART based models are pretrained for text summarization using CNN(Hermann et al., 2015) or XSUM(Narayan et al., 2018) datasets.

4.4 Hyperparameters

Counterfeit has four hyperparameters that are used to fine-tune the model. Based on the very large possible pool of hyperparameters I have chosen to use random selection for selecting the hyperparameter values along with some specific choices that are used to illustrate the impact of the hyperparameter on the ROUGE score. The four parameters are:

- TPOS - Top positive sentiment sentences to duplicate
- TNEG - Top negative sentiment sentences to duplicate
- DCNT - Duplication counter, the number of times to duplicate the sentence
- BEAM - The beam parameter used by the text summarization model

5 Datasets

Four datasets are used for scoring the model. Each dataset is composed of an article and a summary. A sample of 1000 observations from each dataset is used. Figure 1 demonstrates that the overall Counterfeit improvement stabilizes at around 1000 samples and there is no real need to score on the entire set as the results would not change. It provides evidence based on the near to zero slope of the lines that 1000 sample size is sufficient, and it will not bias the model ROUGE score. Using a sample of 1000 observations is imposed due to performance limitations as it takes about 1.5 hour to run the entire 1000 samples through the model using a 16vCPU with 15GB RAM and an NVIDIA Tesla V100 GPU. Increasing the sample size would be too costly. The following dataset are used for scoring Counterfeit:

NEWSROOM[NR] (Grusky et al., 2018) is a large dataset for training and evaluating summarization systems. It contains 1.3 million articles and summaries written by authors and editors in the newsrooms of 38 major publications.

Multi-News[MN] (Fabbri et al., 2019) consists of news articles and human-written summaries of these articles from the site newser.com. Each summary is professionally written by editors and includes links to the original articles cited.

CNN-DailyMail[CNN] (Hermann et al., 2015) is an English-language dataset containing just over 300k unique news articles as written by journalists at CNN and the Daily Mail. The current version supports both extractive and abstractive summarization, though the original version was created for machine reading and comprehension and abstractive question answering.

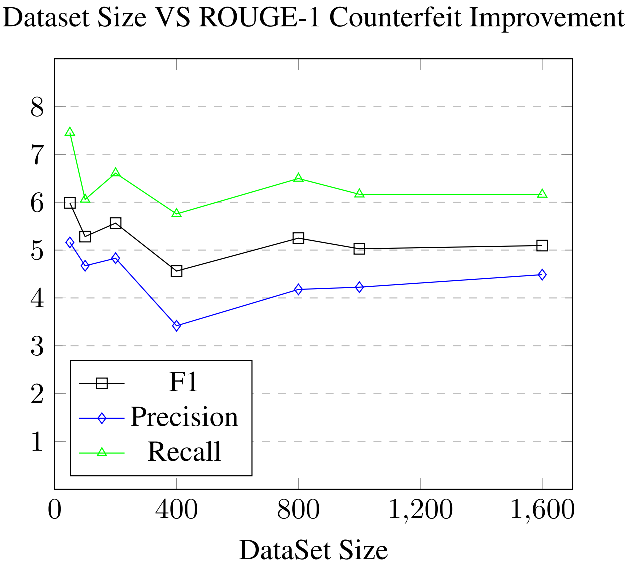

Figure 1: Counterfeit ROUGE-1 Improvement by Dataset size using NR with TPOS = TNEG = 10, DCNT = 5, BEAM = 4 and sshleifer-distilbart-xsum-12-6

6 Hyperparameters Results

Changes to hyperparameters adjust the model ROUGE score. The values reported thought-out the report are the absolute difference between the two results such as F1 score improvement from 23% to 25% would be reported as improvement of 2.

6.1 TPOS and TNEG

A key hyperparameters in the model are the TPOS and TNEG which are always kept as the same value to minimize the total count of hyperparameters. From Figure 2 we can see that increasing the TPOS and TNEG value drastically increases Counterfeit improvements over the baseline model summary.

6.2 BEAM

Precision-recall trade-off in the ROUGE score can be achieved by adjusting the BEAM hyperparameter for the summarization model. Figure 3 demonstrates such relationship for BEAM hyperparameter value from 1 to 10. We can see that as we increase the BEAM hyperparameter recall increases but precision decreases. This relationship between

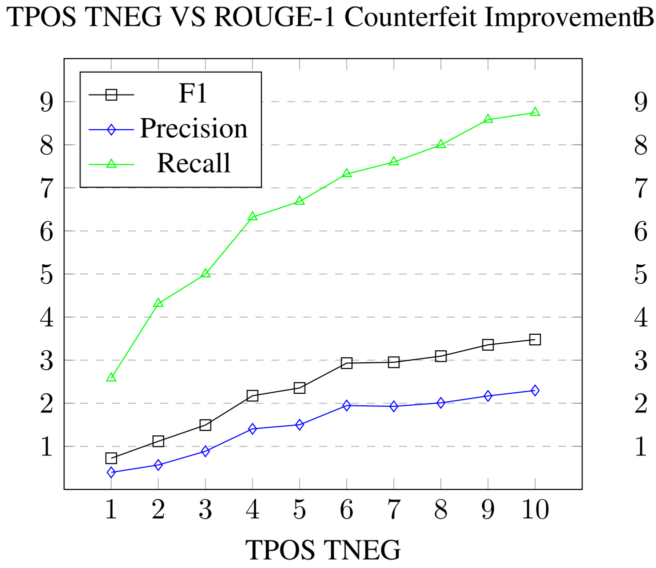

Figure 2: Coneterfeit ROUGE-1 Improvement by TPOS TNEG using NR and facebook-bart-large-cnn

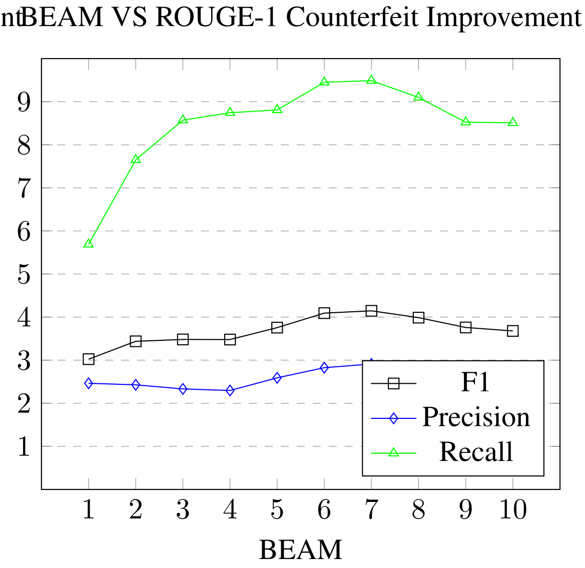

Figure 3: Counterfeit ROUGE-1 Improvement by BEAM using NR and facebook-bart-large

BEAM and recall can be adjusted based on what the type of summary is required. One major drawback of increasing the beam search value is that it increases processing time required to create the summary. When adjusting BEAM hyperparameter I have always kept the value the same in Counterfeit and baseline model and both models showed improvement however Counterfeit improved more as shown in Figure 3. Baseline model impact based on BEAM hyperparameter is shown in Figure 4. The baseline model Recall improves as we increase the BEAM hyperparameter however both F1 and precision decline.

6.2.1 DCNT

Increasing the DCNT hyperparameter increases Counterfeit ROUGE score. Figure 5 illustrates this behavior as all three of ROUGE-1 metric increase directly proportionally to DCNT increase. F1, Precision and Recall appear to stabilize at DCNT=10.

7 Results

Counterfeit ROUGE-1, ROUGE-2 and ROUGE-L metrics are higher when compared to the baseline model. Counterfeit summary does not end up with duplicated sentences in the summary which was a concern at the start of this project. Reviewing the results of the text summarization it becomes clear that generating a better summary does not mean scoring higher on the ROUGE score.

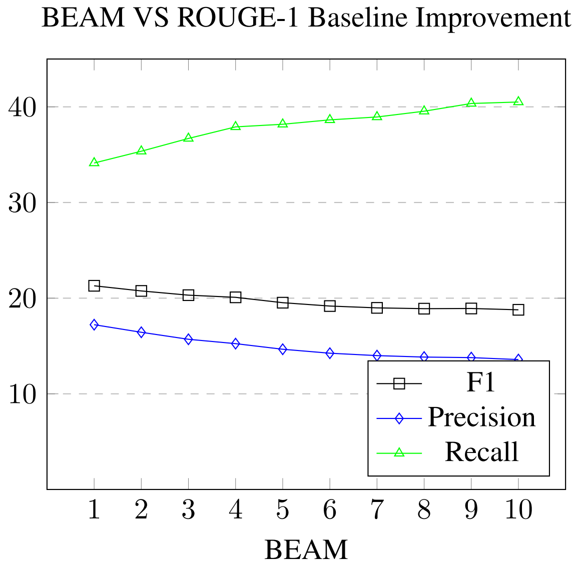

Figure 4: Baseline ROUGE-1 Improvement by BEAM using NR and facebook-bart-large

7.1 Baseline Comparison

I have compared Counterfeit on each of the data sets and provided ROUGE score results. The complete results are provided in the appendix, in total 47 comparison are provided across the hyperparameters and the different pretrained models. The NR dataset results are provided in Table 1. Counterfeit outperforms the baseline on each of the ROUGE-1 metrics. The largest increase is produced on the Recall metric.

Counterfeit improvements for the CNN data set are in Table 2. The improvements are smaller than

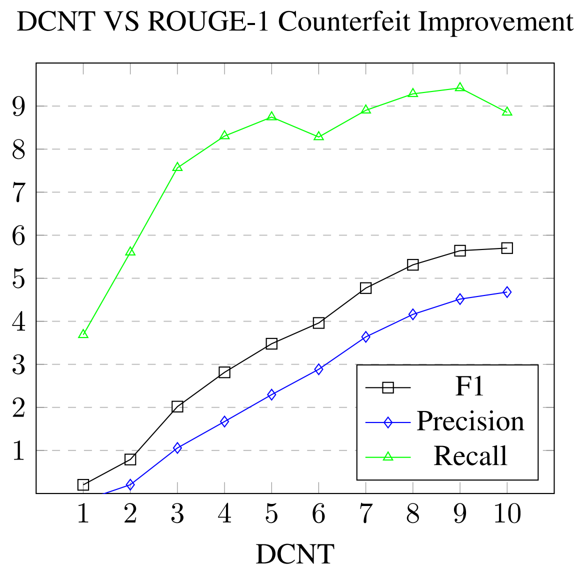

Figure 5: Counterfeit ROUGE-1 Improvement by DCNT using NR and facebook-bart-large

	Score	Baseline	Counterfeit
ROUGE-1	**F1**	18.98	23.13
	Precision	13.99	16.91
	Recall	38.94	48.43
ROUGE-2	**F1**	7.52	13.5
	Precision	5.71	10.27
	Recall	14.85	26.06
ROUGE-L	**F1**	18.59	23.52
	Precision	13.93	17.56
	Recall	34.85	44.39

Table 1: Counterfeit results for NR dataset on facebook-large-bart-cnn. Counterfeit hyperparameters TPOS=10, TNEG=10, DCNT=5, BEAM=7

the results obtained using the NR dataset but still impressive.

	Score	Baseline	Counterfeit
ROUGE-1	**F1**	20.96	24.14
	Precision	34.95	37.04
	Recall	15.75	18.97
ROUGE-2	**F1**	5.9	8.33
	Precision	10.02	12.9
	Recall	4.4	6.53
ROUGE-L	**F1**	19.3	22.32
	Precision	30.74	32.74
	Recall	14.67	17.77

Table 2: Counterfeit results for CNN dataset on facebook-bart-large-xsum. Counterfeit hyperparameters TPOS=10, TNEG=10, DCNT=5, BEAM=4

Counterfeit produces lower improvement using the MN dataset. This can be seen in Table 3. The explanation appears to be related to the ground-truth summary and hyperparameter selection. MN dataset has the longest summary and a high value of the BEAM hyperparameter should have been used. The improvement here is small, in practical terms zero.

	Score	Baseline	Counterfeit
ROUGE-1	**F1**	23.65	24.41
	Precision	48.32	47.39
	Recall	16.4	17.31
ROUGE-2	**F1**	6.73	6.8
	Precision	13.81	13.32
	Recall	4.64	4.78
ROUGE-L	**F1**	21.94	22.07
	Precision	40.41	38.87
	Recall	15.6	16.03

Table 3: Counterfeit results for MN dataset on facebook-bart-large-cnn. Counterfeit hyperparameters TPOS=10, TNEG=10, DCNT=1, BEAM=4

To be fully transparent Appendix-A contains the full list off all the tests that have been conducted. As I have presented some of the better results that Counterfeit has produced the results of all the scores are included in the appendix table. Overall, across all the tests Counterfeit produces improvement of about 2.9 for F1 ROUGE-1. The average improvement is included in Table 4. However, results for test summarization model that is matched to the dataset that it was trained on the same dataset are much lower or even negative.

	Score	Improvement
ROUGE-1	**F1**	2.90
	Precision	2.07
	Recall	5.44
ROUGE-2	**F1**	3.74
	Precision	3.27
	Recall	5.93
ROUGE-L	**F1**	3.17
	Precision	2.38
	Recall	5.31

Table 4: Counterfeit Improvement across all the tests

8 Conclusion

I have introduced Counterfeit, a stacking pre-trained model approach to text summarization which achieves higher ROUGE score when compared to baseline model on a dataset for which the model was not trained on. This approach is ideally suited to a production environment where you can't train the model due to lack of ground-truth summaries. The demonstrated model stacking approach can be expanded for text duplication to other stacked models depending on the priority of the text to be included in the summary. BEAM hyperparameter can be used for adjusting precision vs recall metric, it is extremely useful at making a longer summary as the trained model length hyperparameter is not as effective. Having a higher BEAM hyperparameter increases the length of the summary. Interestingly based on the results of the text duplication tasks the TPOS and TNEG hyperparameters can be used to adjust the summary into a more positive or negative version.

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

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.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend.

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.

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.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. 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.

Sam Shleifer and Alexander M. Rush. 2020. Pre-trained summarization distillation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Appendices

B All Results

								Baseline Model									Counterfeit									Improvement								
								Rouge-1			Rouge-2			Rouge-L			Rouge-1			Rouge-2			Rouge-L			Rouge-1			Rouge-2			Rouge-L		
#	DataSet	TPOS	TNEG	DCNT	SIZE	Summarization Model	BEAM	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R
1	cnndailymail	10	10	10	1000	facebook-bart-large-cnn	4	33.35	26.21	51.63	15.18	11.97	23.49	33.19	26.80	47.63	34.41	29.25	45.70	14.79	12.55	19.74	33.12	28.77	41.81	1.05	3.03	-5.94	-0.39	0.58	-3.75	-0.08	1.96	-5.81
2	newsroom	10	10	4	1000	facebook-bart-large-cnn	4	20.09	15.25	37.91	8.28	6.48	14.96	19.64	15.15	34.20	22.90	16.92	46.21	12.65	9.72	23.65	23.01	17.35	42.19	2.81	1.67	8.30	4.37	3.24	8.69	3.37	2.20	7.99
3	newsroom	4	4	5	1000	facebook-bart-large-cnn	4	20.09	15.25	37.91	8.28	6.48	14.96	19.64	15.15	34.20	22.26	16.65	44.23	11.86	9.26	21.99	22.39	17.05	40.48	2.17	1.41	6.32	3.58	2.78	7.03	2.74	1.90	6.28
4	newsroom	9	9	5	1000	facebook-bart-large-cnn	4	20.09	15.25	37.91	8.28	6.48	14.96	19.64	15.15	34.20	23.44	17.42	46.49	13.32	10.29	24.65	23.61	17.90	42.56	3.36	2.17	8.58	5.04	3.81	9.69	3.97	2.75	8.36
5	newsroom	3	3	5	1000	facebook-bart-large-cnn	4	20.09	15.25	37.91	8.28	6.48	14.96	19.64	15.15	34.20	21.58	16.13	42.90	10.95	8.54	20.42	21.55	16.40	39.00	1.49	0.89	5.00	2.68	2.06	5.46	1.91	1.25	4.80
6	multinews	10	10	10	1000	facebook-bart-large-cnn	4	23.65	48.32	16.40	6.73	13.81	4.64	21.94	40.41	15.60	22.81	46.68	15.81	6.20	12.85	4.26	20.76	38.29	14.76	-0.83	-1.64	-0.60	-0.54	-0.95	-0.38	-1.18	-2.12	-0.84
7	cnndailymail	10	10	10	1000	sshleifer-distilbart-xsum-12-	4	19.19	35.07	13.86	5.13	9.59	3.68	17.84	30.98	13.06	21.92	36.41	16.57	6.92	11.55	5.23	20.31	32.23	15.54	2.73	1.34	2.71	1.79	1.97	1.55	2.47	1.25	2.48
8	newsroom	1	1	5	1000	facebook-bart-large-cnn	4	20.09	15.25	37.91	8.28	6.48	14.96	19.64	15.15	34.20	20.81	15.64	40.49	9.60	7.53	17.38	20.58	15.74	36.64	0.72	0.40	2.58	1.32	1.05	2.42	0.94	0.58	2.44
9	newsroom	10	10	5	1000	facebook-bart-large-cnn	5	19.53	14.66	38.18	7.86	6.06	14.88	19.12	14.58	34.42	23.28	17.25	46.99	13.34	10.30	24.95	23.48	17.77	42.96	3.76	2.59	8.81	5.48	4.25	10.07	4.37	3.18	8.53
10	newsroom	10	10	5	1000	sshleifer-distilbart-xsum-12-	4	17.95	21.99	17.26	4.55	5.91	4.20	15.21	18.54	14.50	22.98	26.22	23.43	9.75	11.61	9.50	20.30	23.06	20.48	5.03	4.22	6.17	5.20	5.69	5.31	5.10	4.52	5.98
11	multinews	10	10	10	1000	facebook-bart-large-xsum	4	11.13	56.18	6.29	3.02	15.83	1.70	12.02	48.18	6.99	11.42	54.29	6.54	3.22	15.59	1.83	12.17	46.40	7.16	0.29	-1.89	0.25	0.20	-0.24	0.14	0.15	-1.77	0.18
12	newsroom	7	7	5	1000	facebook-bart-large-cnn	4	20.09	15.25	37.91	8.28	6.48	14.96	19.64	15.15	34.20	23.04	17.17	45.50	12.86	9.95	23.79	23.21	17.64	41.72	2.95	1.93	7.60	4.58	3.47	8.83	3.56	2.49	7.52
13	newsroom	10	10	5	50	sshleifer-distilbart-xsum-12-	4	16.64	20.64	16.01	4.19	6.28	3.64	14.14	17.56	13.30	22.63	25.80	23.47	9.96	12.27	9.68	20.34	23.31	20.88	5.99	5.16	7.46	5.76	5.98	6.03	6.20	5.75	7.58
14	newsroom	10	10	5	1000	facebook-bart-large-cnn	3	20.31	15.70	36.70	8.21	6.57	14.36	19.81	15.55	33.18	23.79	18.03	45.27	13.51	10.67	24.02	23.89	18.46	41.53	3.48	2.33	8.57	5.30	4.10	9.66	4.08	2.91	8.35
15	multinews	10	10	10	1000	sshleifer-distilbart-cnn-12-6	4	25.21	47.96	17.80	7.15	13.69	5.03	22.84	39.59	16.57	23.14	46.81	16.01	6.31	12.98	4.33	21.00	38.49	14.90	-2.07	-1.16	-1.80	-0.84	-0.70	-0.70	-1.84	-1.10	-1.67
16	newsroom	5	5	5	1000	facebook-bart-large-cnn	4	20.09	15.25	37.91	8.28	6.48	14.96	19.64	15.15	34.20	22.44	16.75	44.59	12.04	9.35	22.38	22.52	17.12	40.73	2.35	1.50	6.68	3.76	2.87	7.42	2.88	1.97	6.53
17	multinews	10	10	10	1000	sshleifer-distilbart-xsum-12-	4	10.16	56.25	5.68	2.60	15.10	1.45	11.13	48.57	6.38	10.57	54.84	5.97	2.84	15.09	1.59	11.40	46.92	6.61	0.41	-1.41	0.29	0.24	-0.02	0.15	0.27	-1.65	0.23
18	newsroom	10	10	5	1000	facebook-bart-large-cnn	10	18.78	13.58	40.51	7.67	5.69	15.87	18.55	13.64	36.40	22.46	16.20	49.01	12.98	9.72	26.14	22.93	16.88	44.90	3.68	2.62	8.51	5.31	4.04	10.27	4.37	3.24	8.50
19	newsroom	10	10	5	1000	facebook-bart-large-cnn	2	20.76	16.43	35.36	8.12	6.68	13.30	20.07	16.16	31.86	24.19	18.86	43.01	13.29	10.86	22.04	24.05	19.11	39.45	3.44	2.43	7.65	5.17	4.17	8.74	3.98	2.95	7.58
20	multinews	10	10	1	1000	facebook-bart-large-xsum	4	11.13	56.18	6.29	3.02	15.83	1.70	12.02	48.18	6.99	11.68	56.16	6.65	3.35	16.45	1.91	12.55	47.99	7.36	0.55	-0.02	0.36	0.33	0.62	0.21	0.53	-0.19	0.37
21	newsroom	10	10	3	1000	facebook-bart-large-cnn	4	20.09	15.25	37.91	8.28	6.48	14.96	19.64	15.15	34.20	22.11	16.30	45.47	11.78	9.09	22.33	22.19	16.69	41.35	2.02	1.06	7.57	3.51	2.61	7.37	2.55	1.54	7.15
22	newsroom	6	6	5	1000	facebook-bart-large-cnn	4	20.09	15.25	37.91	8.28	6.48	14.96	19.64	15.15	34.20	23.02	17.19	45.23	12.73	9.91	23.26	23.12	17.60	41.39	2.93	1.95	7.32	4.46	3.43	8.30	3.47	2.45	7.19
23	newsroom	10	10	1	1000	facebook-bart-large-cnn	4	20.09	15.25	37.91	8.28	6.48	14.96	19.64	15.15	34.20	20.29	15.08	41.59	9.22	7.20	17.63	20.03	15.11	37.44	0.20	-0.17	3.69	0.94	0.72	2.67	0.39	-0.04	3.24
24	newsroom	10	10	5	1000	facebook-bart-large-cnn	7	18.98	13.99	38.94	7.52	5.71	14.85	18.59	13.93	34.85	23.13	16.91	48.43	13.50	10.27	26.06	23.52	17.56	44.39	4.15	2.91	9.49	5.98	4.56	11.21	4.93	3.63	9.55
25	newsroom	10	10	1	1000	facebook-bart-large-xsum	4	19.11	22.14	19.32	5.73	6.84	5.61	16.22	18.71	16.20	23.01	25.52	24.22	9.93	11.50	9.94	20.08	22.22	20.86	3.90	3.38	4.90	4.20	4.66	4.33	3.86	3.51	4.65
26	newsroom	10	10	9	1000	facebook-bart-large-cnn	4	20.09	15.25	37.91	8.28	6.48	14.96	19.64	15.15	34.20	25.73	19.76	47.33	15.48	12.41	26.45	25.80	20.23	43.54	5.64	4.52	9.42	7.20	5.93	11.48	6.16	5.08	9.34
27	newsroom	10	10	6	1000	facebook-bart-large-cnn	4	20.09	15.25	37.91	8.28	6.48	14.96	19.64	15.15	34.20	24.05	18.13	46.19	13.87	10.90	24.72	24.15	18.59	42.40	3.96	2.89	8.28	5.60	4.42	9.76	4.51	3.43	8.20
28	cnndailymail	10	10	10	1000	facebook-bart-large-xsum	4	20.96	34.95	15.75	5.90	10.02	4.40	19.30	30.74	14.67	23.52	36.22	18.44	7.74	12.14	6.03	21.71	32.03	17.23	2.56	1.27	2.69	1.84	2.12	1.63	2.41	1.29	2.56
29	newsroom	10	10	5	200	sshleifer-distilbart-xsum-12-	4	17.20	20.67	16.80	4.36	5.56	4.08	14.76	17.66	14.34	22.76	25.50	23.41	9.92	12.00	9.56	20.34	22.91	20.60	5.56	4.83	6.61	5.56	6.44	5.48	5.58	5.24	6.26
30	newsroom	10	10	10	1000	facebook-bart-large-cnn	4	20.09	15.25	37.91	8.28	6.48	14.96	19.64	15.15	34.20	25.79	19.92	46.76	15.55	12.53	26.23	25.84	20.38	42.98	5.70	4.68	8.85	7.27	6.05	11.27	6.20	5.23	8.78
31	newsroom	10	10	5	1000	facebook-bart-large-cnn	4	20.09	15.25	37.91	8.28	6.48	14.96	19.64	15.15	34.20	23.57	17.54	46.65	13.47	10.45	24.83	23.74	18.04	42.71	3.48	2.30	8.74	5.20	3.97	9.87	4.10	2.88	8.51
32	newsroom	2	2	5	1000	facebook-bart-large-cnn	4	20.09	15.25	37.91	8.28	6.48	14.96	19.64	15.15	34.20	21.20	15.81	42.22	10.38	8.06	19.51	21.15	16.04	38.44	1.12	0.57	4.31	2.11	1.58	4.55	1.51	0.89	4.24
33	cnndailymail	10	10	10	1000	sshleifer-distilbart-cnn-12-6	4	35.62	28.97	50.42	16.29	13.27	23.16	35.11	29.34	46.63	34.43	30.07	43.89	14.73	12.87	18.84	33.16	29.59	40.24	-1.19	1.10	-6.53	-1.57	-0.40	-4.32	-1.95	0.25	-6.39
34	newsroom	10	10	5	1000	facebook-bart-large-cnn	1	21.29	17.23	34.13	8.25	6.86	12.85	20.78	17.15	31.09	24.32	19.69	39.82	13.11	11.18	20.02	24.16	20.07	36.54	3.02	2.46	5.69	4.86	4.32	7.17	3.38	2.93	5.45
35	newsroom	10	10	5	1000	facebook-bart-large-cnn	8	18.90	13.84	39.54	7.53	5.69	15.17	18.53	13.78	35.44	22.89	16.63	48.64	13.30	10.04	26.11	23.28	17.29	44.54	3.99	2.79	9.10	5.77	4.35	10.94	4.75	3.50	9.10
36	cnndailymail	10	10	5	1000	facebook-bart-large-xsum	4	20.96	34.95	15.75	5.90	10.02	4.40	19.30	30.74	14.67	24.14	37.04	18.97	8.33	12.90	6.53	22.32	32.74	17.77	3.17	2.09	3.21	2.43	2.88	2.13	3.02	2.00	3.10
37	newsroom	10	10	5	1000	facebook-bart-large-cnn	6	19.18	14.25	38.64	7.60	5.80	14.83	18.73	14.13	34.63	23.27	17.07	48.09	13.49	10.29	25.80	23.58	17.66	44.04	4.09	2.83	9.45	5.89	4.49	10.97	4.85	3.54	9.41
38	newsroom	10	10	5	800	sshleifer-distilbart-xsum-12-	4	17.84	21.75	17.18	4.65	5.99	4.30	15.20	18.42	14.52	23.09	25.93	23.67	9.84	11.55	9.61	20.48	22.94	20.74	5.25	4.18	6.50	5.19	5.57	5.32	5.28	4.52	6.23
39	newsroom	10	10	8	1000	facebook-bart-large-cnn	4	20.09	15.25	37.91	8.28	6.48	14.96	19.64	15.15	34.20	25.40	19.41	47.19	15.20	12.08	26.24	25.52	19.91	43.47	5.31	4.16	9.28	6.92	5.60	11.28	5.88	4.76	9.27
40	multinews	10	10	1	1000	facebook-bart-large-cnn	4	23.65	48.32	16.40	6.73	13.81	4.64	21.94	40.41	15.60	24.41	47.39	17.31	6.80	13.32	4.78	22.07	38.87	16.03	0.77	-0.93	0.91	0.07	-0.49	0.14	0.13	-1.54	0.44
41	newsroom	10	10	5	100	sshleifer-distilbart-xsum-12-	4	17.87	22.50	17.02	4.76	6.63	4.25	15.21	18.95	14.33	23.15	27.17	23.08	10.18	13.22	9.50	20.89	24.57	20.51	5.28	4.67	6.06	5.42	6.60	5.25	5.68	5.62	6.18
42	newsroom	10	10	2	1000	facebook-bart-large-cnn	4	20.09	15.25	37.91	8.28	6.48	14.96	19.64	15.15	34.20	20.88	15.45	43.51	10.31	8.02	19.93	20.88	15.70	39.50	0.79	0.20	5.60	2.04	1.54	4.97	1.24	0.55	5.30
43	newsroom	10	10	5	1000	facebook-bart-large-cnn	9	18.92	13.78	40.35	7.71	5.78	15.86	18.60	13.76	36.16	22.68	16.43	48.88	13.14	9.88	26.18	23.14	17.12	44.79	3.76	2.64	8.52	5.44	4.11	10.32	4.54	3.36	8.63
44	newsroom	10	10	5	400	sshleifer-distilbart-xsum-12-	4	17.47	21.32	16.85	4.61	5.91	4.30	14.98	18.15	14.39	22.04	24.73	22.61	9.18	11.04	8.87	19.51	21.88	19.79	4.56	3.42	5.76	4.56	5.14	4.57	4.53	3.73	5.40
45	newsroom	10	10	7	1000	facebook-bart-large-cnn	4	20.09	15.25	37.91	8.28	6.48	14.96	19.64	15.15	34.20	24.86	18.88	46.81	14.63	11.59	25.54	24.92	19.33	42.99	4.77	3.64	8.90	6.36	5.11	10.58	5.28	4.18	8.79
46	newsroom	10	10	5	1600	sshleifer-distilbart-xsum-12-	4	18.13	22.35	17.39	4.63	6.08	4.24	15.30	18.75	14.56	23.22	26.84	23.55	9.97	12.02	9.66	20.51	23.56	20.59	5.10	4.49	6.16	5.34	5.95	5.41	5.21	4.81	6.03
47	newsroom	8	8	5	1000	facebook-bart-large-cnn	4	20.09	15.25	37.91	8.28	6.48	14.96	19.64	15.15	34.20	23.18	17.26	45.90	13.04	10.10	24.13	23.33	17.71	42.00	3.09	2.01	8.00	4.77	3.62	9.17	3.68	2.56	7.80
	Total Mean							19.90	22.76	30.14	7.29	7.87	11.26	19.01	20.88	27.21	22.81	24.84	35.59	11.03	11.14	17.19	22.18	23.26	32.52	2.90	2.07	5.45	3.74	3.27	5.93	3.17	2.38	5.31

C BEAM Hyperparameter

C.1 NR, TPOS=10, TNEG=10, DCNT=5, facebook-bart-large-cnn

BEAM=1 U.S. President Obama and U.N. Secretary-General Ban Ki-moon have put climate change in the spotlight. Australian Prime Minister Tony Abbott had said he wanted the G20 meeting to focus on economic and security issues. Obama ensured climate change was front and center before he even landed in Queensland.

BEAM=2 Australian Prime Minister Tony Abbott had said he wanted the G20 meeting to focus on economic and security issues. But U.S. President Obama and U.N. Secretary-General Ban Ki-moon have put climate change squarely in the spotlight. On Wednesday, Obama announced landmark joint emissions commitments with Chinese President Xi Jinping. On Friday, in a speech at the University of Queensland in Brisbane, Obama said the United States would contribute 3 billion into the Green Climate Fund.

BEAM=10 Australian Prime Minister Tony Abbott may have thought he left global climate change off the agenda for the G20 summit in Brisbane, Australia. But U.S. President Obama and U.N. Secretary-General Ban Ki-moon have put it squarely in the spotlight through a series of actions in the past few days. Climate change has been a traditional agenda item for G20, G8 and G7 meetings in recent years, making its absence from the Brisbane gathering of world leaders noteworthy. Obama ensured climate change was front and center before he even landed in Queensland’s capital city on Friday. On Wednesday in Beijing, he announced landmark joint emissions commitments with Chinese President Xi Jinping. China agreed for the first time to peak its greenhouse gas emissions by 2030, and to dramatically scale up the use of renewable energy in its economy, to about 20 by 2020. Then on Friday, in a speech at the University of Queensland in Queensland, Obama announced that the U.s. would contribute 3 billion into the Green Climate Fund which aims to help developing countries improve their climate resilience, and transition away from fossil fuels.

D Summary Example

D.1 NR, TPOS=10, TNEG=10, DCNT=1, BEAM=4, facebook-bart-large-xsum, ID-1

Article: Police at the scene of the Tunis blast. (AAP)bomb blast on a bus transporting Tunisiaś presidential guard in central Tunis has killed at least 12 people and wounded 16 more.security source at the site of the attack said ”most of the agents who were on the bus are dead.”presidency had said the blast on the bus killed at least 14 people, but this was revised down, with the interior ministry giving the latest toll as least 12 dead.explosion, described as an ”attack” by presidential spokesman Moez Sinaoui, struck on the capitalś Mohamed V Avenue.AFP journalist reported seeing the partly burnt-out shell of the bus, with police, ambulances and fire trucks at the scene.people were in tears.bank employee working nearby reported hearing a large explosion and seeing the bus on fire.there was no immediate claim of responsibility, Tunisia has been plagued by Islamist violence since the 2011 overthrow of longtime dictator Zine El Abidine Ben Ali.attacks earlier this year claimed by ISIL terror group targeted foreigners - at the National Bardo Museum in March, killing 21 tourists and a policeman, and at a resort hotel in Sousse in June, killing 38 tourists.of members of the security forces have also been killed by extremists since 2011.

Summary: A bomb blast on a bus transporting Tunisia’s presidential guard in central Tunis on Tuesday killed at least 12 people and wounded 16, the interior ministry said.

Baseline Summary: Tunisia has been plagued by Islamist violence since the 2011 overthrow of longtime dictator Zine El Abidine Ben Ali. Two attacks earlier this year claimed by ISIL terror group targeted foreigners - at the National Bardo Museum in March, killing 21 tourists and a policeman, and at a resort hotel in Sousse in June, killing 38 tourists. Dozens of members of the security forces have also been killed by extremists since 2011.’

Counterfeit Summary: A bomb blast on a bus transporting Tunisia’s presidential guard in central Tunis has killed at least 12 people and wounded 16 more. A security source at the site of the attack said ”most of the agents who were on the bus are dead” The explosion, described as an ”attack” by presidential spokesman Moez Sinaoui, struck on the capital
ś Mohamed V Avenue. No immediate claim of responsibility, Tunisia has been plagued by Islamist violence since the 2011 overthrow of longtime dictator Zine El Abidine Ben Ali. Two attacks earlier this year claimed by ISIL terror group targeted foreigners - at the National Bardo Museum in March, killing 21 tourists and a policeman, and at a resort hotel in Sousse in June, killing 38 tourists.