DivHSK: Diverse Headline Generation using Self-Attention based Keyword Selection

Venkatesh E, Kaushal Kumar Maurya,
Deepak Kumar and **Maunendra Sankar Desarkar**
Indian Institute of Technology Hyderabad, India
{venkateshelangovan.tce, deepak.soe.cusat}@gmail.com,
cs18resch11003@iith.ac.in, maunendra@cse.iith.ac.in

Abstract

Diverse headline generation is an NLP task where given a news article, the goal is to generate multiple headlines that are true to the content of the article, but are different among themselves. This task aims to exhibit and exploit semantically similar one-to-many relationships between a source news article and multiple target headlines. Towards this, we propose a novel model called DivHSK. It has two components: KeySelect for selecting the important keywords, and SeqGen, for finally generating the multiple diverse headlines. In KeySelect, we cluster the self-attention heads of the last layer of the pre-trained encoder and select the most-attentive *theme* and *general* keywords from the source article. Then, cluster-specific keyword sets guide the SeqGen, a pre-trained encoder-decoder model, to generate diverse yet semantically similar headlines. The proposed model consistently outperformed existing literature and our strong baselines and emerged as a state-of-the-art model. Additionally, We have also created a high-quality multi-reference headline dataset from news articles[1].

1 Introduction

Generating diverse and semantically similar multiple outputs in natural language generation (NLG) is an important and challenging task (Tevet and Berant, 2021). The traditional single headline generation task is formulated as a sequence-to-sequence learning problem and has been extensively studied for more than a decade now (Banko et al., 2000; Zajic et al., 2002; Dorr et al., 2003; Lopyrev, 2015; Takase et al., 2016; Gavrilov et al., 2019). Recently, researchers are also interested towards diverse output sequence generation tasks. This falls into the one-to-many generation category and is being studied for multiple tasks such as paraphrase generation (Yu et al., 2021; Gupta et al., 2018), machine translation (Shen et al., 2019), question generation (Shen et al., 2022) and summarization (Cho et al., 2019). In this work, we consider the problem of generating diverse headlines given a single news article. Diverse headlines present the theme of the article in semantically related yet lexically different short sentences, which may attract different sets of audiences and increase the consumption of the news.

The existing approaches for diverse sequence generation mostly diversify the decoding steps through alternative search algorithms (Vijayakumar et al., 2018; Fan et al., 2018) or mixture decoder approaches (Shen et al., 2019; Maurya and Desarkar, 2020) where different decoders generate difference output sequences. Recently, Cho et al. (2019) proposed a two-stage modeling involving a *diversification stage* to extract diversifying attributes and a *generation stage* to guide the encoder-decoder model for diverse generations. The diversifying attributes are keywords extracted from the input text with the expectation-maximization algorithm. They consider text summarization and question-generation tasks. In similar lines, Yu et al. (2022) leverage external knowledge graph, i.e., ConceptNet (Speer et al., 2017) to extract diverse yet relevant keywords at *diversification stage* and generate diverse common sense reasoning texts. These models are not directly applicable for diverse headline generation tasks because the headlines are mostly oriented toward a single common theme (event, person, etc.) in a short sentence, and these models distract the semantics of generated headlines. Our empirical experiments (Section-5) validate this point. Liu et al. (2020) used manually extracted keywords with a multi-source transformer for diverse headline generation. The model is not scalable to other datasets/tasks because keyword extraction requires a human annotator. Unlike these, we used an automated self-attention-based approach to obtain the most attentive keywords from the article

[1]Our code and dataset are available at https://github.com/kaushal0494/DivHSK

Findings of the Association for Computational Linguistics: ACL 2023, pages 1879–1891

automatically.

To overcome the limitations of the existing models, we propose DIVHSK, a simple yet effective model for diverse headline generation using a self-attention-based keyword selection. The model has two modules/components: (a) KEYSELECT - a pre-trained encoder model to extract diversifying attributes i.e. *theme* and *general* keywords from input news article and (b) SEQGEN - a regular pre-trained encoder-decoder architecture guided by diversifying attributes for generating multiple diverse yet semantically similar headlines.

Overall, our main contributions are as follows: (1) We propose a novel model DIVHSK- Diverse Headline Generation using Self Attention based Keyword Selection to generate diverse yet semantically similar headlines. (2) We release a high quality MRHEAD: *Multi-Reference Headline Dataset* for diverse headline generation task. (3) The performance of the proposed model is compared with several strong baselines using both automated and human evaluation metrics.

2 Problem Formulation

Given a news article, the goal is to generate *semantically similar*, *grammatically coherent*, *fluent* and *diverse* headlines. Formally, given a news article x, the goal is to model the conditional distribution for k target outputs $p(y_k|x)$ with valid mappings $x \rightarrow y_1, \ldots, x \rightarrow y_k$ where $\{y_1, y_2, \ldots, y_k\}$ should be diverse. Here we consider $k = 3$, i.e., the task is to generate three diverse headlines.

3 Methodology

The proposed DIVHSK model has two components (1) pre-trained encoder, i.e., KEYSELECT and (2) regular pre-trained encoder-decoder, i.e., SEQGEN. As per Liu et al. (2020), multiple headlines should convey the common theme, differing on a lexical level and the headline tokens should be uniformly distributed across the source article. Towards these goals, in KEYSELECT, we first cluster the encoders' last-layer self-attention heads to find the most attentive keywords for each cluster from the input news article. We observe that: (a) all the clusters have a few most-attentive common keywords called as *theme* and (b) cluster-specific most attentive keywords called as *general* (i.e., non-theme) keywords. We combine *theme* with cluster-specific *general* keywords to create diversifying attributes. For each of the k clusters, there is a corresponding diversifying attribute. Table-4, in Appendix, presents a few sample themes and general keywords.

The input news article, theme, and general keywords (from diversifying attributes) are concatenated with [SEP] tokens to create modified input for the SEQGEN module. In this way, different cluster leads to generate diverse headlines. The theme and general keywords in the cluster lead to semantically similar and theme-oriented headlines. For pre-trained encoder and pre-trained encoder-decoder models, we use the 'encoder of T5-base' (Raffel et al., 2020) and T5-base checkpoints, respectively. See Figure 1 for an overview of the proposed model. More details about each component are given below:

3.1 KEYSELECT: Keyword Selection Module

3.1.1 Self-Attention Heads Clustering

We take a pre-trained encoder model with l self-attention heads $h_1, h_2, \ldots, h_l$ from the last layer. Each self-attention head h_i usually focuses on different parts of the inputs text (Peng et al., 2020). We group these heads into k clusters $C = \{c_1, c_2, \ldots, c_k\}$; so each cluster has $g = \frac{l}{k}$ heads. Here we cluster the heads in a sequential manner. Next, we identify the m most-attentive keywords (not BPE) from each head. As one keyword may get high attention values from multiple heads, it may result in overlap among the keyword sets obtained from each head. Consequently, we get a maximum of $g * m$ keywords from each cluster. Stop-words/function-words are not considered in keyword sets.

We have clustered the multiple heads of multi-head attention of the last-hidden layer in a sequential manner. The adoption of this approach can be justified from two perspectives. Firstly, during the pre-training phase of a language model, the weights of each head within the multi-head attention mechanism are initialized with random values. Over the course of pre-training, these weights undergo the process of learning to acquire diverse values. The different heads aim to focus on different parts of the input and provide a diverse view, which is suitable for diverse keyword selection. Secondly, the proposed model is trained end-to-end, and the weights of the KEYSELECT module are consistently updated rather than being fixed. Moreover, the target headlines associated with different heads (clusters) are different. Therefore, during back-propagation, the different heads learn to focus on the keywords

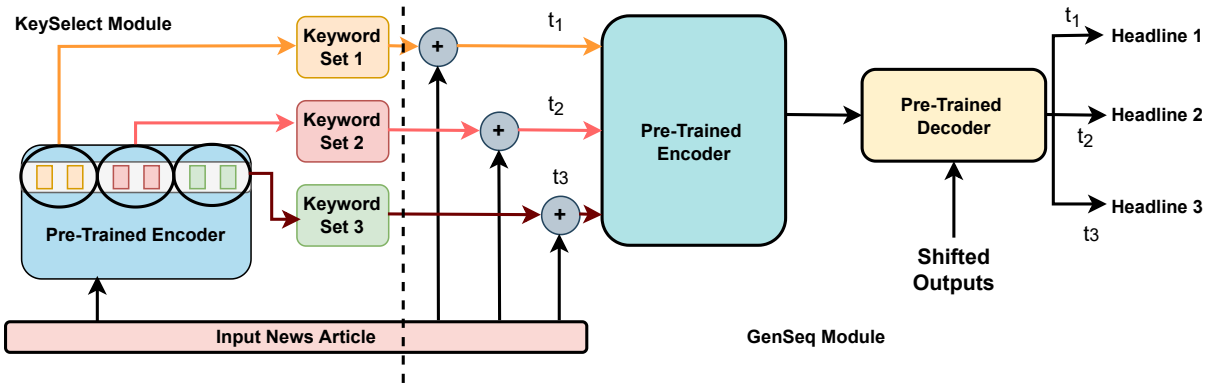


Figure 1: Overview of proposed *DivHSK* model. Where time-steps $t_1 > t_2 > t_3$.

relevant to their respective target reference headlines. Based on these points, we conclude that clustering heads in any order does not have a significant impact, and we choose a simple sequential manner for the clustering of the attention heads.

3.1.2 Creating Diversifying Attributes

Suppose the total number of keywords to guide the SEQGEN module is n. We keep r keywords as theme keywords and the remaining $n - r$ as general keywords. The r keywords are the most-attentive common keywords across all c clusters. The rest of the $n - r$ keywords are the most-attentive non-overlapping keywords specific to individual clusters c_i. These n keywords form the diversifying attributes $K_{c_i}^{guide}$ for cluster c_i. r is a hyper-parameter and its value can be determined empirically. In case r common keywords can not be found[2], then we can take the available r' common keywords that can be found, and the remaining $n - r'$ keywords can be taken from the individual clusters. See Algorithm-B in Appendix for more details.

3.2 SEQGEN: Pre-trained Seq2Seq Module

The diversifying attributes $K_{c_i}^{guide}$ are concatenated with the source article x as:

theme-keywords [SEP] *general*-keywords [SEP] *article*

to form the extended article $x_{c_i}^{e}$. Each cluster corresponds to specific attributes, resulting in different extended articles. We fine-tune a pre-trained encoder-decoder model with an extended article and a corresponding headline. Additionally, we employed word-mover distance (WMD; Kusner et al. (2015)) between predicted (h_p) and reference (h_r) headlines token ids, as an additional component in the loss function to control the diversity with λ. Finally, the KEYSELECT and SEQGEN modules are trained in end-to-end manner to minimize the loss L as:

$$L = \sum_{i=1}^{c}(1-\lambda)(-logP_\theta(y_i|x_i^e)) + \lambda(\text{WMD}(h_{pi}, h_{ri})) \quad (1)$$

4 Experimental Setup

4.1 Dataset

One of the essential elements of the proposed work is the inclusion of multiple reference headlines for each news article. Specifically, each example in the dataset will consist of a quadruple in the following format: <`article, headline-1, headline-2, headline-3`>. However, the proposed approach can be easily extended to a single reference setup. Towards this, we have created a dataset that we refer to as MRHEAD: Multi-Reference Headline.

•DataSet Collection: To create the dataset, first, we scrape news articles and their headlines from Inshorts (`https://www.inshorts.com/`) news website and add them to a seed set. Articles under *'All News'* category, i.e., politics, sports, technology, etc. were considered. Next, we identify news articles from other public news websites that are semantically similar to the articles in the seed set, and also note their headlines against the corresponding article in the seed set. To find semantically similar news articles we use sentence-BERT (Reimers and Gurevych, 2019) and cosine-similarity scores. Then, human annotators verify the dataset content and remove the poor-quality headlines. Following this process, we obtained 3012 articles each with at least three parallel headlines. We split the data into training, validation, and test splits of sizes 2330, 100, and 582 respectively. Dataset creation, human verification, and other statistics are reported in Appendix-A.

[2]We have not encountered any scenario where the theme keywords are not present in one or more clusters.

4.2 Baselines

We have meticulously chosen six baseline models for our experimentation and analysis. Our extensive observations have revealed that single-output generation models, such as text-summarization/headline generation models, do not perform well in multi-output generation settings. The primary issue with such multiple generated outputs is their lack of lexical diversity. Therefore, we have selected three literature baselines: Mixture-Decoder (MixD; Shen et al. (2019)), Mixture Content Selector (MixCS; Cho et al. (2019)), and Knowledge Graph Experts (MoKGE; Yu et al. (2022)). Additionally, we have designed three robust baselines based on diverse search algorithms and with modified loss functions: T5+DSA (diverse search algorithm), T5+WMD (Kusner et al., 2015), and T5+Avg-Loss. More details about these baselines are provided in Appendix-C.

4.3 Evaluation Metrics

We use four automated evaluation metrics that rely on a lexical and semantic match in a one-to-many evaluation setup, as, for a given generation there are three reference headlines. We consider BLEU-4 (BLEU; Papineni et al. (2002)) and ROUGE-L (Lin, 2004) metrics as lexical-match metrics, and BERTScore (Zhang et al., 2020) and BARTScore (Yuan et al., 2021) as semantic match based metrics. To measure the diversity among the generated headlines, we use Pairwise-BLEU (self/P-BLEU; Ott et al. (2018)) metric similar to Shen et al. (2019). As stated by Shen et al. (2019), there is always a trade-off between performance and diversity, i.e., if the generated headlines are correct but similar, then the performance (BLEU and ROUGE-L scores) will be high due to large lexical overlap but the diversity will be low (high P-BLEU) and vice-versa. Towards this concern, we consider the harmonic mean (HMean) between $(1 - \text{PBLEU})$ and BLEU as a *combined* evaluation metric. For more certainty about model performance, we also conducted the human evaluation with four metrics, i.e., *Fluency (Flu)*, *Relatedness (Rel)*, *Correctness (Corr)* and *Diversity* similar to (Cho et al., 2019). To manage the load on evaluators, we selected three baseline models for human evaluation. Two of the models were the best-performing (according to HMean) competitor models from literature (MixCS and MoKGE), and the other one was T5-Avg-Loss, the best-performing baseline model designed by us. We randomly selected 50 generated headlines from the baselines and the proposed DIVHSK model as a human evaluation sample. Further, we employ two sets of annotators for human evaluation to avoid any biased evaluation. For *diversity* we asked an absolute evaluation score on a scale of 1 (lowest) to 5 (highest) and for other metrics a comparative evaluation. See more details about human evaluation guidelines in Appendix-D.

5 Results and Discussions

5.1 Diversity vs. Accuracy Trade-off

Table-1 displays the automated evaluation scores obtained for various baselines and the proposed DIVHSK models. The mixture decoder model, which employs multiple decoders, achieves the highest BLEU and ROUGE-L scores. However, the high P-BLEU score for this model indicates low diversity in the generated headlines, defeating the purpose of having multiple decoders. Similar observations are noted for the T5+DSA model. Additionally, the high scores obtained for BERTScore and BARTScore metrics suggest that the DIVHSK model exhibits superior semantic similarity with the reference headlines. This is one of the key constraints that ensure the generated outputs are semantically coherent. The ideal model should obtain reasonable BLEU and ROUGE-L scores, high BERTScore and BARTScore (high semantic similarity), low P-BLEU (high diversity), and high HMean scores. The proposed DIVHSK model satisfies these ideal conditions and emerges as a state-of-the-art model. The necessary ablation experimental results are added in Table-5.

5.2 Comparison with State-of-the-Art

We have compared the performances of DIVHSK with MixD, MixCS, and MoKGE, which are state-of-the-art literature models. Although these models perform well for other tasks, they exhibit poor performance for the diverse headline generation task. As discussed in Section 1, recent models like MoKGE perform poorly for diverse headline generation tasks due to the inclusion of tokens/keywords from the knowledge graph that may not align with the headline's theme and distract the learning process. Overall, it is evident from the performances of MixCS and MoKGE that existing text summarization models do not perform well for headline generation tasks. This could be due to the fact that summaries are generally long, while headlines are

Model	Headline 1(⇑)					Headline 2(⇑)					Headline 3(⇑)					P-BLEU (⇓)
	BLEU	R-L	BES	BAS	HMean	BLEU	R-L	BES	BAS	HMean	BLEU	R-L	BES	BAS	HMean	
T5+DSA	22.83	**0.342**	67.21	61.43	0.525	22.97	0.345	67.89	61.26	0.525	25.39	0.346	67.57	61.88	0.525	0.734
T5+WMD	14.60	0.346	64.11	57.83	0.529	16.37	0.353	64.91	57.32	0.530	14.81	0.346	64.23	58.08	0.529	0.730
T5+Avg-Loss	12.07	0.310	61.31	56.44	0.637	12.02	0.308	62.11	56.03	0.637	11.06	0.306	61.72	56.95	0.636	0.672
MixD	**23.18**	0.322	**71.64**	68.52	0.320	**25.63**	**0.349**	71.91	68.28	0.320	**25.84**	**0.351**	71.43	68.88	0.320	0.838
MixCS	14.01	0.242	64.12	56.98	0.347	15.62	0.245	64.83	56.73	0.347	16.77	0.241	64.36	57.22	0.348	0.824
MoKGE	8.94	0.185	57.32	51.11	0.571	12.44	0.208	57.74	50.67	0.576	7.87	0.163	57.34	50.82	0.568	0.705
DIVHSK	16.83	0.289	71.56	**69.01**	**0.690**	17.95	0.295	**72.03**	**68.66**	**0.691**	17.72	0.295	**71.55**	**69.98**	**0.690**	**0.647**

Table 1: Automated evaluation results of the models. Where R-L, BES and BAS indicate ROUGE-L, BERTScore and BARTScore metrics, respectively. Additionally, *HMean* indicates the harmonic mean between p-BLEU and BLEU metrics. High HMean and low P-BLEU desirable.

	DivHSK Vs T5+Avg-Loss			DivHSK Vs MixCS			DivHSK Vs MoKGE		
	Win	Lose	Tie	Win	Lose	Tie	Win	Lose	Tie
Annotator Set 1									
Flu	**44.0**	34.0	22.0	**48.0**	30.0	20.0	**52.0**	36.0	12.0
Rel	26.0	20.0	**54.0**	**48.0**	32.0	20.0	**42.0**	22.0	36.0
Corr	**38.0**	24.0	**38.0**	**38.0**	26.0	36.0	**42.0**	32.0	26.0
Annotator Set 2									
Flu	**38.0**	32.0	30.0	**42.0**	40.0	18.0	**50.0**	42.0	8.0
Rel	28.0	26.0	**46.0**	34.0	30.0	**36.0**	**48.0**	28.0	24.0
Corr	**38.0**	32.0	30.0	**42.0**	34.0	24.0	**48.0**	28.0	24.0

Table 2: Comparative human evaluation results for propose DIVHSK vs baselines models. All the scores are reported in percentage (%).

Model	Diversity(⇑)	
	Annotators set-1	*Annotators set-2*
T5 + Avg-Loss	3.12	3.06
MixCS	2.74	2.56
MoKGE	3.08	2.96
DIVHSK	**3.60**	**3.72**

Table 3: Absolute average human evaluation diversity scores.

short and more focused. The models fail to adapt to these settings.

5.3 Human Evaluation Results

For more reliable evaluation, we also conducted human evaluation and results are reported in Tables 2 and 3. For *Fluency*, *Relatedness* and *Correctness* metrics, the DIVHSK model most of the time either wins or ends up with tie versus all considered baselines. Similar trends are observed across both the annotator sets. The human evaluation scores correlate well with automated evaluation scores. The average absolute diversity scores are reported in Table-3 and it is found that generated text are more diverse for proposed DIVHSK model. Considering decent automated and human evaluation scores, we conclude that our model performs reasonably well and outperforms the other methods consistently.

5.4 Effect of n and r Parameters

In Figure 2, we investigate the effect of varying the values of n (the total number of selected keywords) and r (the number of theme keywords) on the performance of the DIVHSK model. As n and r increase, we observe a decrease in the P-BLEU scores, indicating an increase in diversity (headlines are lexically diverse). However, the BLEU

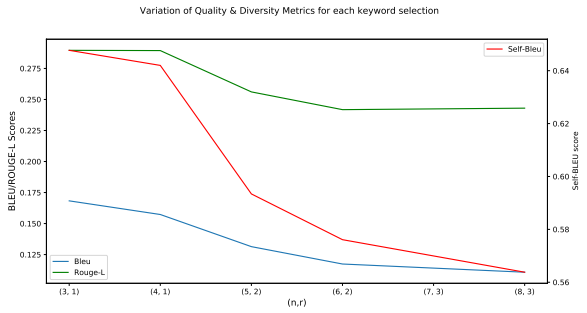


Figure 2: Trend of BLEU, ROUGE-L and P-BLEU scores for n and r values.

and ROUGE-L scores also decrease due to high diversity as these metrics are based on lexical matching. Therefore, the optimal values of n and r are important to maintain the diversity and performance trade-off.

6 Conclusion

In this work, We present a novel task and dataset for diverse headline generation. We also propose a strong neural architecture for the task. The model, referred to as DIVHSK, uses self-attention-based clustering to create diversifying attributes that guide the pre-trained encoder-decoder model to generate diverse headlines. We empirically demonstrate that the DIVHSK consistently outperforms all baseline models on both automated and human evaluation metrics, while maintaining diversity as a key criterion.

Limitations

- We are unable to test the proposed model's performance on other datasets due to the unavailability of public multi-reference headline generation datasets.
- Our dataset is created over a period of 6 months and contains around 3000 examples. Although there are several commonly used benchmark datasets with a similar number of examples: e.g., $\mathcal{R}^4\mathcal{C}$ reading comprehension dataset (6.4K examples) (Inoue et al.,

2020), FIRE-LID (3357 examples), IIITH-NER (3084 examples) datasets in GLUECoS benchmark (Khanuja et al., 2020), WNLI (634 examples), RTE (2500 examples) and MRPC (3700 examples) datasets in GLUE benchmark (Wang et al., 2018), NOPE Corpus (around 2.7K examples) (Parrish et al., 2021), we believe that it will be better to have a larger dataset for this challenging task. We plan to create a larger version of the dataset in future work.

References

Michele Banko, Vibhu O. Mittal, and Michael J. Witbrock. 2000. Headline generation based on statistical translation. In *Association for Computational Linguistics*, ACL '00, page 318–325, USA.

Jaemin Cho, Minjoon Seo, and Hannaneh Hajishirzi. 2019. Mixture content selection for diverse sequence generation. 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 3121–3131, Hong Kong, China. Association for Computational Linguistics.

Bonnie Dorr, David Zajic, and Richard Schwartz. 2003. Hedge trimmer: A parse-and-trim approach to headline generation. In *Association for Computational Linguistics*, HLT-NAACL-DUC '03, page 1–8, USA.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Daniil Gavrilov, Pavel Kalaidin, and Valentin Malykh. 2019. Self-attentive model for headline generation. In *ECIR*.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2018. A deep generative framework for paraphrase generation. In *AAAI Press*.

Benjamin D Horne, Sara Khedr, and Sibel Adali. 2018. Sampling the news producers: A large news and feature data set for the study of the complex media landscape. In *Twelfth International AAAI Conference on Web and Social Media*.

Naoya Inoue, Pontus Stenetorp, and Kentaro Inui. 2020. R4C: A benchmark for evaluating RC systems to get the right answer for the right reason. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6740–6750, Online. Association for Computational Linguistics.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online. Association for Computational Linguistics.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 957–966. JMLR.org.

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.

Dayiheng Liu, Yeyun Gong, Yu Yan, Jie Fu, Bo Shao, Daxin Jiang, Jiancheng Lv, and Nan Duan. 2020. Diverse, controllable, and keyphrase-aware: A corpus and method for news multi-headline generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6241–6250, Online. Association for Computational Linguistics.

Konstantin Lopyrev. 2015. Generating news headlines with recurrent neural networks. *ArXiv*, abs/1512.01712.

Kaushal Kumar Maurya and Maunendra Sankar Desarkar. 2020. Learning to distract: A hierarchical multi-decoder network for automated generation of long distractors for multiple-choice questions for reading comprehension. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1115–1124.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. *ArXiv*, abs/1803.00047.

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

Alicia Parrish, Sebastian Schuster, Alex Warstadt, Omar Agha, Soo-Hwan Lee, Zhuoye Zhao, Samuel R Bowman, and Tal Linzen. 2021. Nope: A corpus of naturally-occurring presuppositions in english. *Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL)*.

Hao Peng, Roy Schwartz, Dianqi Li, and Noah A. Smith. 2020. A mixture of h - 1 heads is better than h heads. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6566–6577, Online. Association for Computational Linguistics.

Lianhui Qin, Lemao Liu, Wei Bi, Yan Wang, Xiaojiang Liu, Zhiting Hu, Hai Zhao, and Shuming Shi. 2018. Automatic article commenting: the task and dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 151–156.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Tianxiao Shen, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2019. Mixture models for diverse machine translation: Tricks of the trade. *ArXiv*, abs/1902.07816.

Xinyao Shen, Jiangjie Chen, Jiaze Chen, Chun Zeng, and Yanghua Xiao. 2022. Diversified query generation guided by knowledge graph. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, page 897–907, New York, NY, USA. Association for Computing Machinery.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Sho Takase, Jun Suzuki, Naoaki Okazaki, Tsutomu Hirao, and Masaaki Nagata. 2016. Neural headline generation on Abstract Meaning Representation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1054–1059, Austin, Texas. Association for Computational Linguistics.

Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346, Online. Association for Computational Linguistics.

Ashwin Vijayakumar, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Wenhao Yu, Chenguang Zhu, Lianhui Qin, Zhihan Zhang, Tong Zhao, and Meng Jiang. 2022. Diversifying content generation for commonsense reasoning with mixture of knowledge graph experts. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1896–1906. Association for Computational Linguistics.

Wenhao Yu, Chenguang Zhu, Tong Zhao, Zhichun Guo, and Meng Jiang. 2021. Sentence-permuted paragraph generation. 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 5051–5062. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

David Zajic, Bonnie Dorr, and Richard Schwartz. 2002. Automatic headline generation for newspaper stories. In *Workshop on automatic summarization*, pages 78–85.

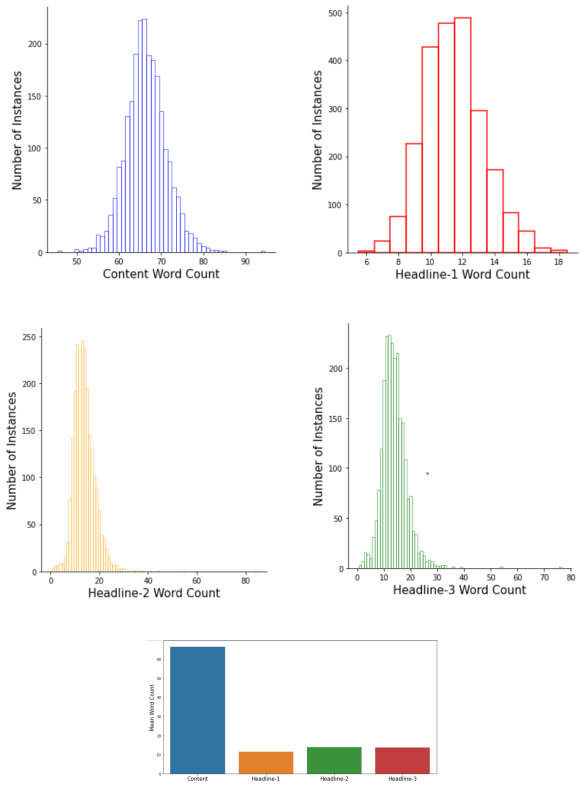
Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A MRHEAD Dataset Creation Strategy[3]

One of the key requirements of our work is to have multiple reference headlines for a news article i.e., <`article, headline-1, headline-2, headline-3`>[4]. Towards this requirement, we have created a dataset MRHEAD: *Multi-Reference Headline Dataset*. First, we scrape news articles and their headlines from Inshorts (`https://www.`

[3]We will publicly release dataset, code, model checkpoints and generated text

[4]Nevertheless, the proposed approach can be easily extended to single reference setup with modification in the loss function.



(a) Mean word count of content and all headlines

Figure 3: The various statistics of the dataset are presented. Here the news article is referred as content.

inshorts.com/) news website. We keep all news categories in consideration. Once the headline from Inshorts is collected, we try to collect multiple similar headlines from other news websites with following steps:

- Make a google search with news headline text as the search query.
- Parse the google search response and retrieve the list of URLs from the search result.
- From the URL list obtained, remove URLs that belong to Wikipedia, Facebook, Twitter, etc.
- Remove URLs that correspond to docx, pdf or ppt files.
- Make a HTTP call to the remaining URLs. Retrieve similar headlines by parsing the response.

Next, we use Sentence-BERT (Reimers and Gurevych, 2019) to get the similarity scores and pick two headlines from the list of similar headlines based on similarity scores. Therefore, each entry in our dataset consists of 4 features: <article, headline-1, headline-2, headline-3>. Further, we ask human annotators to verify the quality of the dataset and filter/modify the records accordingly. This exercise carried out over a period of 6 months resulted in around 3000 records in total. The available data was split into 2330, 100, and 582 samples of training, validation and test splits respectively. The dataset statistics are shown in figure 3. Table 4 displays a few samples from our dataset.

As part of the dataset, we have released the URLs to news articles (these articles are already in the public domain) and the reference headlines. Sharing of the urls/news articles is done in several existing datasets, e.g. NELA2017 dataset (Horne et al., 2018), Article Commenting Dataset (Qin et al., 2018).

B KeySelect Module

Algorithm 1 Keyword Selection Algorithm

Require: l self-attention heads $h_1, h_2, \ldots, h_l$
Require: c clusters $c_1, c_2, \ldots, c_c$
Require: m, n, r: Keyword-Selection hyper-parameters
Initialize $g = \frac{l}{c}$
for i $\in \{0, c-1\}$ **do**
    Assign g heads (h_{ig+1}-$h_{(i+1)g)}$) to the cluster c_i
    Initialize set $w_i \leftarrow \emptyset$ to store the keywords of c_i
    for each h_j in c_i **do**
        Select top m attentive words from h_j and update the set w_i
    end for
    # c_i will contain at most $g * m$ keywords
end for
for i $\in \{0, c-1\}$ **do**
    Select r (or r') theme keywords from overlapping keywords across c clusters based on attention scores
    Select $n - r$ (or $n - r'$) general keywords from non-overlapping keywords specific to the cluster c_i based on attention scores
    Cluster c_i have corresponding diverse keywords set K_i^{guide} of size n
end for
Use K^{guide} consists list of selected keywords for c clusters in *SeqGen* module

C Baselines

We compare the proposed model performance with three literature and three other strong baselines. Details of the baselines are mentioned below:

1. **Mixture-decoder:** In the mixture decoder (Shen et al., 2019) approach, three different decoders are used to generate the diverse headlines. Each decoder is trained with a different headline and we take the average cross-entropy loss for the particular news article.

S.No.	Example-1	Example-2
News Article	Days after a 20-year-old B.Com student was found unconscious with her hands and legs tied up on the outskirts of Andhra Pradesh, Vizianagaram town police said the incident was staged.The woman left her hostel to meet a male friend. After her brother inquired about her at the hostel she staged the incident to convince her family she was kidnapped.	A video showing a Chandigarh female traffic police constable holding her baby in her arms while on duty has gone viral on social media. The constable Priyanka was reportedly pulled up for not reporting to work at 8 am following which she took her baby to work. The clip was captured near the roundabout of Chandigarh's Sector 15:23 on Friday.
Headline 1	20-yr-old Andhra woman found with hands, legs tied staged 'kidnap': Police	Chandigarh traffic constable reports for duty with baby in arms; video goes viral
Headline 2	Andhra woman found 'unconscious' had staged 'kidnap' say police	Video of Chandigarh cop holding baby while on duty goes viral: The Tribune India
Headline 3	Kidnapping victim found tied up in backseat after police stop wrong way driver in Olympia	Video of a Woman Traffic Constable Holding Baby on Duty Goes Viral Netizens Demand Free Daycare for Cops

Table 4: Sample examples from MR-Head dataset

News	Keyword Set 1	Keyword Set 2	Keyword Set 3	Theme Keyword
Actress Raveena Tandon who will be making her digital **debut** with the **crime** **thriller** series Aranyak said that her kids are **excited** to see her on OTT. She added My **kids**...tell me Mom you re going to be on **Netflix** It s a cool thing for them. Speaking about her character as a **cop** in the series Raveena said She has incredible strength.	thriller cop Netflix	crime excited Netflix	debut kids Netflix	Netflix
China filed the **highest** number of patent applications globally in 2020 **retaining** its top **position** for the **second** consecutive year the UN s World Intellectual Property Organization WIPO said. **China** filed 68,720 applications last year while the US filed 59,230. In 2019 **China** had **replaced** the US as the **top** patent application filer for the first time in over four decades.	second position China	highest retaining China	top replaced China	China

Figure 4: Examples to illustrate theme and general keywords selected with KEYSELECT module. Here, the general keywords set is a subset of the keyword set.

2. **Mixture Content Selector:** In MixCS (Cho et al., 2019), the authors introduced a selection module SELECTOR to perform the diversification process. The SELECTOR module generates three different sets of keywords which were concatenated by input news and fed into the standard encoder-decoder model for headline generation.

3. **Experts(MoKGE):** In MoKGE (Yu et al., 2022) approach, apart from keyword extraction from the input news, the authors leverage the use of knowledge graph, i.e., ConceptNet (Speer et al., 2017) to extract the diverse set of keywords to guide an encoder-decoder model to generate diverse headlines.

4. **T5+ DSA (Diverse Search Algorithm):** We fine-tune the T5-base checkpoint to return the three sequences with a combination of top-k and top-p sampling.

5. **T5+WMD (Word Mover Distance):** Similar to T5+ DSA but additionally we added WMD along with standard cross-entropy loss. The loss function is given as follows.

$$L = (1 - \lambda) \times L_{CE} + \lambda \times L_{WMD} \quad (2)$$

$$L_{WMD} = WMD(h_p, h_r) \quad (3)$$

Here, L_{CE} indicates the standard cross-entropy loss, and L_{WMD} indicates word mover distance as a loss, where h_p and h_r are predicted and reference headlines. λ is a hyperparameter. For the best-performing model, λ is *0.5*.

6. **T5+Avg Loss:** Similar to T5+DSA, but additionally the final loss is an average cross-entropy loss for the same news article. The loss function is given as follows.

$$L = \frac{L_{1CE} + L_{2CE} + L_{3CE}}{3} \quad (4)$$

The losses L_{1CE}, L_{2CE} and L_{3CE} are calculated with respect to each headline-1, headline-2 and headline-3 respectively.

D Human Evaluation Setup

We conducted a human evaluation with four metrics i.e., *Fluency (Flu)*, *Relatedness (Rel)*, *Correctness (Corr)*, and *Diversity*. *Fluency* measures how fluent and grammatical the generated text is. *Relatedness* indicates how much the generated outputs are in the context with input(s), *Correctness* measures semantics and meaningfulness. Finally, *Diversity* measures how diverse the generated headlines are.

A human evaluation task was conducted to compare the results of our proposed model with baselines. The evaluations were carried out by 20 human evaluators, each of whom held at least a Masters's degree and possessed a good knowledge of the English language. We selected 50 input news articles randomly from the dataset and generated three headlines for each article using the selected models. For each input, we randomly selected the k^{th} generated headline ($k \sim 1, 2, 3$) from the models (both baselines and proposed). For example, if $k = 2$, we selected the second generated headline from the proposed model as well as from all the other baselines. This process was repeated for all 50 input news articles. For the first task, the dataset consists of 3-tuples containing the news article, headline from the proposed model, and headline from the baseline model. The annotators were asked to provide relative scores based on fluency, relatedness, and correctness between the two headlines. They were given three options (0, 1, 2), where 1 indicated that headline-1 was better, 2 indicated that headline-2 was better, and 0 indicated a tie. The annotators were not informed about the baseline and proposed model results.

The second task aims to ensure the diversity of generated headlines. Similar to the first task, we selected 50 samples from the proposed model and other baselines for the same news articles. The dataset consists of a news article and three headlines. The annotators were asked to provide diversity scores ranging from 1 to 5, where 1 indicated headlines with the least diversity or unacceptable quality and 5 indicated diverse headlines along with good quality.

E Implementation Detail

In our proposed model, we utilized pre-trained weights of the T5-base encoder for the pre-trained encoder used in the KEYSELECT module during training. The model was trained for 20 epochs, and the best checkpoint was selected based on the validation loss. We used $l = 12$ self-attention heads from the pre-trained encoder of the *KeySelect* module. As we aimed to generate three diverse headlines, we set $c = 3$, which implies $g = 4$. The optimal values for our best-performing model were $m = 10$, $n = 3$, $r = 2$, and $\lambda = 0.5$. The total number of parameters was 3×10^8. We utilized the Adam optimizer technique with a learning rate of $1e - 4$. During the test phase, we used the combination of Top-K and Top-p sampling decoding strategies, where $K = 50$ and $p = 0.95$. The batch size was 32. We implemented all the models using PyTorch (Hugging-face). Model training was performed on a V100, 32GB single GPU.

F Ablation Study

We conducted an ablation study to analyze the effect of different model components on the performance of our proposed model. The experimental results are presented in Table-5. First, we added a plug-and-play module called WordNet (Fellbaum, 1998) to our model, which is used to obtain related keywords from the input text. Specifically, if n keywords are extracted from the input text in a cluster c_i, then the final set of keywords after using the WordNet module would be at least 2n keywords for that particular cluster c_i. However, in this experiment, we observed a significant drop in quality across all generated headlines. Next, we experimented with removing the Word Mover Distance component from the loss function and observed a drop in performance in terms of BLEU and P-BLEU scores compared to our proposed DIVHSK model. We also experimented with different values of the hyperparameter λ used in the loss function and found that our proposed model outperforms all other variations of the model. Overall, the ablation study demonstrates the importance of the different model components in achieving the best performance for headline generation.

G Model Generated Headlines

In this section, we present the results generated by our proposed model, along with the results of baseline models. The generated headlines, along with input news and reference headline, are tabulated in Tables 6 and 7.

Experiments	Headline-1(⇑)		Headline-2 (⇑)		Headline-3 (⇑)		P-BLEU (⇓)
	BLEU	ROUGE-L	BLEU	ROUGE-L	BLEU	ROUGE-L	
DivHSK without WMD	15.10	0.2552	14.55	0.2419	15.88	0.2541	**0.6488**
DivHSK with WordNet	15.05	0.2671	14.71	0.2673	14.62	0.2699	0.6087
DivHSK Model (λ= 0.1)	14.39	0.2763	13.97	0.2795	13.45	0.2722	0.5897
DivHSK Model (λ = 0.2)	15.31	0.2864	15.12	0.2824	16.31	0.2882	0.6211
DivHSK Model(Ours) (λ = 0.5)	**16.83**	**0.2896**	**17.95**	**0.2954**	**17.72**	**0.2955**	0.6477

Table 5: Different ablation experiments that provide clarification for model design choices.

News	Actress Raveena Tandon who will be making her digital debut with the crime thriller series Aranyak said that her kids are excited to see her on OTT. She added My kids...tell me Mom you re going to be on Netflix It s a cool thing for them. Speaking about her character as a cop in the series Raveena said She has incredible strength.		
Reference Headlines	My kids feel it's a cool thing to be on OTT: Raveena on her digital debut	Raveena Tandon on her digital debut with Aranyak My kids feel it's a cool thing to be on Netflix	My kids feel being on Netflix is a cool thing Raveena Tandon on digital debut with Aranyak
Model	**Generated Headline 1**	**Generated Headline 2**	**Generated Headline 3**
Mixture Selector	My kids are excited to see me on Netflix: Raveena Tandon	My kids are excited to see me on Netflix: Raveena	My kids are excited to see me on OTT: Raveena
MoKGE	Raveena Tandon Says Her Kids Are Excited To See Her On Netflix	My kids are excited to see me on OTT: Raveena	My kids are excited to see her on Netflix
T5-Avg	Tell me mom you're going to be on Netflix it's a cool thing for kids: Raveena	Tell me mom you're going to be on Netflix it's cool for kids: Raveena	Tell me mom you're going to be on Netflix it's a cool thing for kids, Raveena
Mixture Decoder	My kids are excited: Raveena on making digital debut in 'Aranyak'	Kids excited to see me on Netflix: Raveena on 'Aranyak': Tandon	Kids excited to see me on Netflix: Raveena on making digital debut with 'Aranyak'
Ours	Actress Raveena to play as cop in a thriller on Netflix	I am super excited for my kids to see me on Netflix: Raveena	Mom is to be on Netflix. It's a cool thing for kids: Raveena on her OTT debut

Table 6: Sample generated headlines with different baselines and proposed model

News	China filed the highest number of patent applications globally in 2020 retaining its top position for the second consecutive year the UN s World Intellectual Property Organization WIPO said. China filed 68,720 applications last year while the US filed 59,230. In 2019 China had replaced the US as the top patent application filer for the first time in over four decades.		
Reference Headlines	China files highest patents globally for 2nd year in a row: UN	China becomes world's top patent filer after four decades with US on top	China extends lead over U.S. in global patents filings U.N. says
Model	**Generated Headline 1**	**Generated Headline 2**	**Generated Headline 3**
Mixture Selector	China tops the list of top patent filers for 2nd consecutive year	China files highest number of patent applications globally for 2nd consecutive year	China files highest number of patent applications globally for 2nd consecutive year
MoKGE	China tops the list of top patent filers globally in 2020	China retains top spot for 2nd consecutive year: UN	China tops the list of world's top patent exporters in 2020
T5-Avg	China files highest number of patent applications globally in 2020 retains top position	China files highest number of patent applications globally in 2020 retains top position: UN	china files highest number of patent applications in 2020 retains top position: UN says
Mixture Decoder	China retains top ranking in 2020, file the highest patent applications globally	China retains top position in 2020, filed highest number of patent applications	China retains top position in 2020, filed highest number of patent applications
Ours	China retains top position in global patent filings for second consecutive year.	China files highest number of patents globally in 2020, retains top spot: UN	China replaces US as top patent applicant: UN

Table 7: Sample generated headlines with different baselines and proposed model

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
After the conclusion as stated in the call for the main conference paper.

☒ A2. Did you discuss any potential risks of your work?
We use clean dataset after human verification and validation.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

section 3

☑ B1. Did you cite the creators of artifacts you used?
section 1 and 4

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
I have used all the publicly available artifacts which don't have any research restrictions.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
section 1 and 4

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
section 4 and appendix A

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
section 1

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
section 4 and appendix A

C ☑ Did you run computational experiments?

Section 4 and Appendix E

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix E

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
section 4 and appendix E

☒ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Due to the large set of experiments and computationally constrained

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix E

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 4 and A

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
The dataset is created by two of the co-authors and they were well aware of the risk and other details. They have considered the expected policies.

☒ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
The dataset is created by two of the co-authors.

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
The dataset comprises of data points (news articles) available in the public domain. The urls are part of the dataset.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The created data source is reviewed by the multiple stockholders

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
The dataset is created by two of the co-authors.