

Topic-aware Multimodal Summarization

Sourajit Mukherjee¹ Anubhav Jangra¹ Sriparna Saha¹ Adam Jatowt²

¹Indian Institute of Technology Patna, India

{mailsourajit25, anubhav0603, sriparnasaha}@gmail.com

²University of Innsbruck, Austria

adam.jatowt@uibk.ac.at

Abstract

Multimodal Summarization (MS) has attracted research interest in the past few years due to the ease with which users perceive multimodal summaries. It is important for MS models to consider the topic a given target content belongs to. In the current paper, we propose a topic-aware MS system which performs two tasks simultaneously: differentiating the images into "on-topic" and "off-topic" categories and further utilizing the "on-topic" images to generate multimodal summaries. The hypothesis is that, the proposed topic similarity classifier will help in generating better multimodal summary by focusing on important components of images and text which are specific to a particular topic. To develop the topic similarity classifier, we have augmented the existing popular MS data set, MSMO, with similar "on-topic" and dissimilar "off-topic" images for each sample. Our experimental results establish that the focus on "on-topic" features helps in generating topic-aware multimodal summaries, which outperforms the state of the art approach by 1.7% in ROUGE-L metric.

1 Introduction

Due to the continuous growth of multimedia content, users often look for ways to read and go through only the crucial information content, and to avoid redundancy as much as possible. To cater to this need for concise information availability, automatic summarization systems are the need of the hour.

Extensive research works have produced summaries of a single modality like text (Gambhir and Gupta, 2017; Jangra et al., 2020a) or video (Apostolidis et al., 2021). However, researchers have also demonstrated that users are more satisfied with multimodal summaries than uni-modal summaries (Zhu et al., 2018). Thus, generating output summaries of different modalities like text and images makes sense. Images play essential roles in help-

ing users understand the text and make the summary more attractive, contextualized, and complete. Topic information is crucial for correctly identifying pictures and text as a part of a multimodal summary. However, existing work (Jangra et al., 2021a) in the field of multimodal summarization has not yet utilized the sample's topic information to improve the multimodal summary quality.

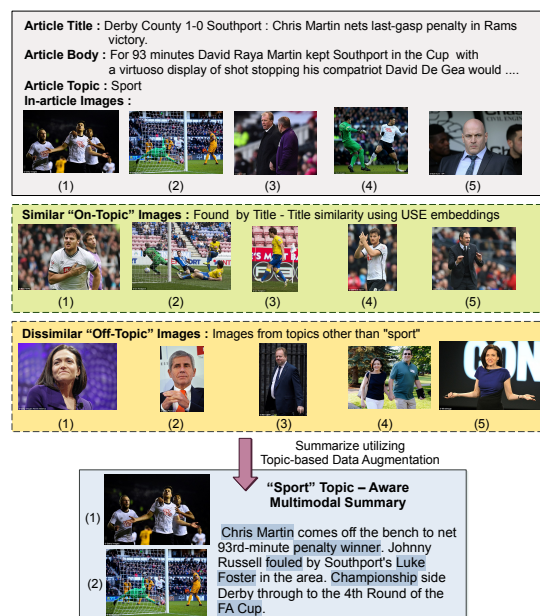


Figure 1: In this example we select two images to be part of the pictorial summary. Using topic information, i.e., "sport", the model can decide that the images (1) and (2) are highly related to "sport" topic as compared to others, and hence increase their probability of being used in pictorial part of the final summary.

In this paper, we introduce *Topic-aware Multimodal Summarization (TMS)* where multimodal summaries consisting of texts and images are generated by also focusing on topic-centric information. Incorporating topic information of the source content aids the summarization process because the generated multimodal summary also considers the key elements of that topic. For example the

summary of an article in the "sport" topic should highlight how a player scored a goal in a football match as a part of the text summary, and the images of the player as a part of the image summary (as shown in Fig. 1). In contrast, the multimodal summary of any article belonging to the "travel" topic should highlight details of the place mentioned, and the image summary should showcase the images of that place. Thus different topics may require different kinds of focus.

In our experiments we have investigated the following research objectives: i) the significance of the topic similarity classifier with respect to the combination of different modalities, ii) the impact of using similar "on-topic" image feature vectors instead of zero-padded vectors for samples having limited number of in-article images and iii) the comparison with respect to the existing state-of-the-art technique.

The key contributions of our work are as follows:

1. To the best of our knowledge, this is the first study where topic information is integrated with multimodal summary generation to improve the performance.
2. The existing MSMO data set is augmented with "on-topic" and "off-topic" images to perform an auxiliary task of topic similarity identification from images.¹
3. A multi-task learning approach is proposed which solves simultaneously the two tasks:
 - classification of in-article images into "on-topic" and "off-topic" categories
 - generation of multimodal summary.

The first task is our auxiliary task to extract more useful features from image and text modalities which in turn can help in generating better multimodal summaries.

2 Related Works

Multimodal Summarization has gained in popularity in the recent years due to the enhanced quality and user experience it is able to offer. Jangra et al. (2021a) provided an overview of the recent developments in the field of Multimodal Summarization. The summarization process can produce a single

modality output (Chen and Zhuge, 2018; Palaskar et al., 2019; Li et al., 2018; Khullar and Arora, 2020) or a multi-modal output (Zhu et al., 2018; Jangra et al., 2020b,c, 2021b). In our work we focus on the latter case, and we have considered the MSMO model (Zhu et al., 2018) as the baseline which produces multi-modal output summary in the form of text and images. One of the recent works inspired by the MSMO model is (Zhu et al., 2020); however, unlike our model, it does not focus on producing topic-aware summaries. Zhu et al. (2020) also used an extended version of the MSMO dataset for training its image selection module in a supervised fashion. In contrast, we have used an unsupervised approach similar to MSMO for training our model's image selection module. Training our model using the extended dataset used by Zhu et al. (2020) might produce better results in the future.

Recently, Transformer-based models like MTMS (Ye et al., 2021) and CtnR (Zhang et al., 2021) were developed based on the MSMO dataset. However, these models either produce text-only summaries using multimodal input or use different input parameter sizes (max. encoder length, decoder length, max. number of images, etc.) for training the model. Thus because of these factors, we have not considered these Transformer-based models as a baseline for comparison. Furthermore, MTMS also uses 80% of the test data for fine-tuning its model with the image-saliency-based loss. Our model does not require using any segment of the test data for training purposes.

Multi-task learning (MTL) involves sharing representations between related tasks which helps in achieving better performance in the target task. Earlier MTL has been used for producing both textual (Nishino et al., 2019; Isonuma et al., 2017) and multimodal summaries (Zhao et al., 2016). Taking inspirations from these works we have used MTL for making our summarization model topic-aware.

Producing multimodal summaries, which relate well with the topic they belong to, helps users get a better understanding of the actual content. Earlier, research has been done in developing systems that produce multimodal summaries related to a specific topics like sports (Tjondronegoro et al., 2011; Sanabria et al., 2019), movies (Evangelopoulos et al., 2013) or E-commerce (Li et al., 2020) but those works have used datasets which are specific

¹The extended dataset and our model's code is available at github.com/maillsourajit25/Topic-Aware-Multimodal-Summarization

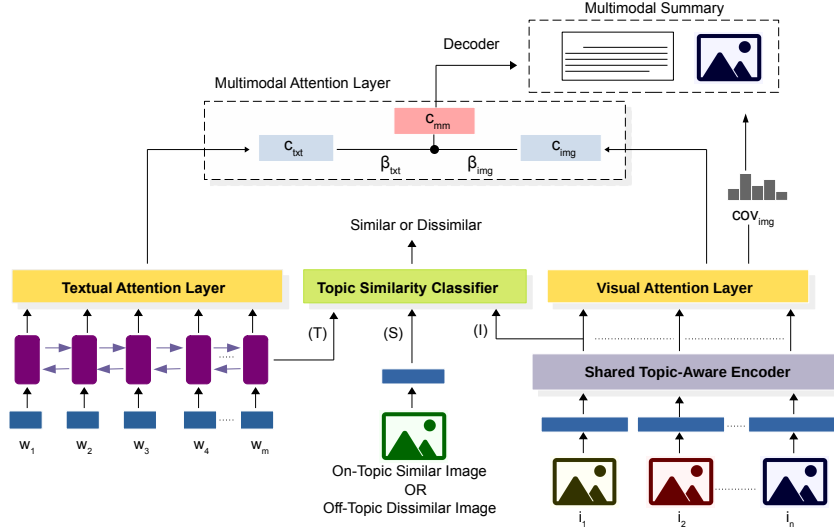


Figure 2: Proposed Architecture. The label (T) is the text encoder’s last time-step output, (I) is the projected in-article image feature vector and (S) is the corresponding similar/dissimilar image feature vector, to be passed as input into the topic similarity classifier.

to a single particular topic. We are the first to introduce a model that not only can produce topic-aware multimodal summaries, but is also trained using a topic-generic dataset.

3 Our Model

3.1 Problem Definition

TMS task is defined as follows: Given a multimodal input $\{T, I\} \in D$, where T is a text article having W words, I is the set of in-article images and D is the topic of the article, the task is to create a multimodal topic-aware summary $\{T', I'\}$ highly related to the topic D and reflecting the content of $\{T, I\}$. The textual summary T' is composed of W' words such that $|W'| < |W|$. The pictorial summary I' such that $|I'| \leq |I|$ represents the set of recommended images extracted from I .

3.2 Model Architecture

Our model is a multi-task learning model that is trained to perform summarization as well as topic similarity identification. It is composed of a Bi-directional LSTM based text encoder for encoding the textual part of the input and a unidirectional LSTM based summary decoder. The image part of the input is encoded using a VGG19-based (Simonyan and Zisserman, 2015) image encoder which has been pretrained on ImageNet dataset (Deng et al., 2009). Zhu et al. (2018) passed the encoded images through a projection layer to project

them into same dimension as text. We have re-defined the image projection layer as the **shared topic-aware encoder** because it is now shared between the classifier and the MSMO model. Previous researches (Zhu et al., 2020; Li et al., 2018) have shown that global features are more effective compared to local features. Hence in this paper, we have extracted the 4,096 dimensional global features of the pre-softmax fully-connected layer denoted by g . These vectors are projected into the same dimension of textual context vector, using the following equation: $g^* = W_I^2(W_I^1 g + b_I^1) + b_I^2$, where W_I^1, b_I^1, W_I^2 and b_I^2 are trainable parameters. The output of the shared topic-aware encoder is branched off into two directions as shown in Fig. 2. One part is passed into the topic similarity classifier (discussed in Sec. 3.3) for topic similarity identification and the other one to the visual attention layer for the summarization task.

Next, the textual context vector c_{txt}^t is computed from the textual attention layer (Bahdanau et al., 2016; Luong et al., 2015), and c_{img}^t from visual attention layer (Li et al., 2018). These context vectors are then passed to the multimodal attention layer (Zhu et al., 2018) which combines the visual and textual attentions together to produce the multimodal context vector, c_{mm}^t , given as the weighted sum of c_{txt}^t and c_{img}^t . During decoding, the summary decoder takes as input the previously predicted word and c_{mm}^t to predict the next word. Further, in order to prevent repeated attention, we

compute textual and visual coverage vectors, cov_{txt}^t and cov_{img}^t , as the sum of the respective attention weights over the previous decoding steps.

Our summary decoder is based on Pointer Generator Network (PGN) (See et al., 2017). It can decide whether to generate words from a fixed vocabulary or rather to copy words from the source while constructing the summary for a given input text. Finally, the loss at a time step t is given as the summation of the negative log likelihood of the target word, w_t , and the textual coverage loss $L_{txt}^{cov} = \sum_i \min(\alpha_i^t, cov_i^t)$ and the visual coverage loss $L_{img}^{cov} = \sum_j \min(\alpha_j^t, cov_{img,j}^t)$, where α_i^t and α_j^t are textual and visual attention weights, respectively.

$$L_t = -\log p_{w_t} + L_{txt}^{cov} + L_{img}^{cov} \quad (1)$$

Image Decoding: The visual coverage scores cov_{img}^t for every image at the last decoding timestep, are used to select the most relevant images representing the pictorial summary of the source. A higher coverage score indicates greater relevance.

3.3 Topic Similarity Classifier (TSC)

The topic similarity classifier helps the model to also consider the topic information while calculating attention for both image and text. The target output for the classifier is labelled as (topic) "similar" when similar "on-topic" images are passed as input while it is labelled as (topic) "dissimilar" when "off-topic" images are passed as input into the classifier (as shown in Fig. 2). The other inputs to the classifier are the text encoder's last time step output and the in-article image features. The topic similarity classifier is used only during training the model. During testing, the trained weights of the shared topic-aware encoder and the text encoder help the model in extracting topic-centric information. Thus, the classifier performs an auxiliary task of topic similarity classification during training that should aid the shared topic-aware encoder and the text encoder to learn and extract topic-related information during the encoding process. This would have impact on the visual and textual attention layers as now the model will provide more attention on images and text which are more related to the topic that the article belongs to. The classifier is defined as follows:

$$O_{TSC}^{sim} = \sigma(W_{txt}h_{txt} + W_s h_{img}^{sim} + W_{img}h_{img}) \quad (2)$$

$$O_{TSC}^{dissim} = \sigma(W_{txt}h_{txt} + W_s h_{img}^{dissim} + W_{img}h_{img}) \quad (3)$$

where W_{txt} , W_s and W_{img} are trainable parameters having dimensions $\mathbb{R}^{1 \times d_{enc}}$, $\mathbb{R}^{1 \times 4096}$ and $\mathbb{R}^{1 \times d_{enc}}$. Here d_{enc} denotes the dimension of the Bi-LSTM based text encoder. O_{TSC}^{sim} and O_{TSC}^{dissim} denote the classification outputs of the classifier when we pass as input the VGG19-based feature vectors, h_{img}^{sim} , and h_{img}^{dissim} , of the similar and dissimilar images, respectively. h_{txt} denotes the hidden state output for the last time step of the text encoder. h_{img} represents the projected feature vectors of the in-article images obtained after passing through the shared topic-aware encoder. The classifier loss is defined as follows:

$$L_{TSC} = BCELoss([O_{TSC}^{sim}, O_{TSC}^{dissim}], [y^{sim}, y^{dissim}]) \quad (4)$$

where $BCELoss$ refers to binary cross-entropy loss. y^{sim} and y^{dissim} refer to the true labels for the classifier. Finally, the total loss for our model, with λ_{TSC} as classifier weight is computed as:

$$L = -\log p_{w_t} + L_{txt}^{cov} + L_{img}^{cov} + \lambda_{TSC}L_{TSC} \quad (5)$$

4 Experimental Settings

4.1 Dataset

The MSMO dataset (Zhu et al., 2018) is the only large-scale dataset best-suited for the task defined in Sec. 3.1. It was originally constructed using news articles collected from the *Daily Mail* website. It contains 293,965 samples in the train set, 10,355 samples in the validation and 10,261 samples in the test set. Each sample contains a multi-sentence news article (720 tokens on average), the set of multiple image and caption pairs (6 pairs on average) and the manually-written² multi-sentence highlights of each article (70 tokens on average). Furthermore, every multi-sentence article has a title and a body. We have considered the body of every article as the source text. To train our model using the topic similarity classifier, we need a similar "on-topic" image and a dissimilar "off-topic" image for every in-article image of the train set. To cater to this need for training our model, we have augmented the training set of the MSMO dataset.

Proposed Dataset Augmentation: For augmentation, we first determined the "topic" of each sample (or news article) from its URL. The URL path contains the name of the category or the topic to which the article belongs. The URL path also includes the name of the sub-topic of the article. Although the sub-topic is a better representation of an article's

²Created by *Daily Mail* (<http://www.dailymail.co.uk>).

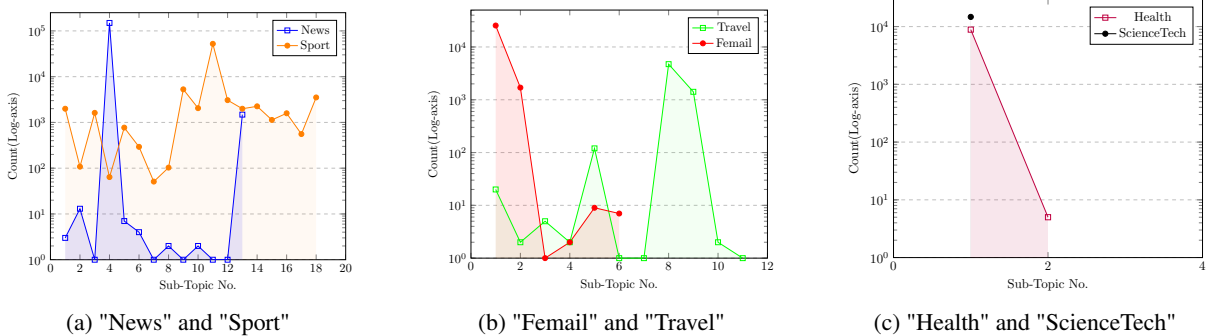


Figure 3: Sub-Topic Count Distributions for different Topics. For the "sport" topic, sub-topics having count smaller than 50 are not shown in the plot.

category, we found an uneven count distribution of the number of samples belonging to each sub-topic (as shown in Fig. 3). Even the total number of sub-topics for different topics varied from being as high as 87 for the "sport" topic to just 1 for the "ScienceTech" topic. The irregular sub-topic count and uneven count distribution would pose a difficulty in finding similar images for every sample. Hence we decided to find topic-wise similar/dissimilar images for every sample instead of doing it sub-topic-wise.

We generated Universal Sentence Encoder (USE) embeddings (Cer et al., 2018) for the title of every article and grouped all samples into their respective topics. For topics having less than 5,000 samples, we grouped them under the "Others" category. To find similar images for a sample belonging to a certain topic, we compared its USE-based title embeddings with the title embeddings of 20,000³ other randomly chosen samples belonging to the same topic by comparing their cosine similarity scores. The samples with the highest cosine-similarity scores were chosen and at most 10 images from these articles were selected as "On-Topic" - similar images. Table 1 discusses the details of the topic-wise cosine similarity scores in the augmented dataset.

Furthermore, we have also extracted the publication dates of the articles so that, in the future, we could use this augmented dataset to find similar images using temporal similarity. Temporal search can provide an alternate means of finding similar images faster by searching for similar articles within a specific time range before or after the target article's publication date (assuming that similar news articles get published often on consecutive days). Hence articles missing timestamp in-

³Limited number of comparisons were done to reduce the search space for generating similar images faster.

Topic	Sample Count (SC)	Mean Sim. Score	Sim. Score Std. Dev.	SC in Test Set
News	150551	0.535	0.073	5105
Sport	79098	0.671	0.098	2605
Femail	26983	0.550	0.083	971
Travel	6261	0.515	0.077	162
ScienceTech	14592	0.552	0.097	408
Health	8815	0.534	0.074	334
Others	6920	0.523	0.117	266

Table 1: Augmented Dataset: Similarity score statistics

formation were not considered, leading to 293,220 samples in the train set. We focused however on the title-based search in our work as it seemed more intuitive concerning our model architecture.

For finding dissimilar images for a sample belonging to a certain topic, we randomly picked 10 images from samples belonging to a different topic. In all the experiments, at most 10 in-article images were considered per sample such that for each in-article image, only one similar and one dissimilar image were taken during the classification task (Sec. 3.3).

4.2 Compared Methods

To evaluate the performance of our model, we have compared its performance with the following baselines:

- **MSMO (ATG, ATL HAN):** Zhu et al. (2018) proposed the ATG, ATL and HAN models, which uses the global, local and hierarchical image features, respectively, for the multi-modal abstractive summarization task.
- **GuideRank (GR):** We have also considered an extractive summarization baseline GuideRank (Li et al., 2016, 2017) which employs LexRank (Erkan and Radev, 2011) along with a guidance mechanism. In this approach, the captions rank the accompanying sentences based on relatedness. After using GR to estab-

lish the ranks of the sentences and captions, we remove sentences that satisfy the minimum length requirement as a text summary by the text’s rating. Next, we pick the image whose caption ranks first among the captions to obtain the visual summary.

The previous researches (Zhu et al., 2018; Li et al., 2018) have already established that global image features perform better than local and hierarchical features in the multimodal summarization task. Hence we consider only the ATG model as a baseline for comparing the topic-wise results (Sec. 5.2) and human-evaluation results (Sec. 5.3).

We have performed experiments testing the following models:

- **TSC-MSMO-TIS:** This model consists of inputs (T), (I) and (S) as shown in Fig. 2.
- **TSC-MSMO-IS:** We have fed only (I) and (S) as input into the TSC.
- **TSC-MSMO-TS:** We have used (T) and (S) as input into the TSC.
- **TSC-MSMO-SIMPAD-TIS:** We have kept the inputs to the classifier unchanged, but if any sample has less than 10 images, then instead of padding with zero vector we replaced those with the similar "on-topic" images. All of these architecture changes were done during training the models.

4.3 Hyper-parameters and Evaluation Metrics

For training our 0.5M parameter models, we have considered 400 textual tokens and ten images per sample. Our models were trained for 255,000 iterations (around 13 epochs for a batch size of 16) without considering coverage loss, followed by coverage loss for extra 45,000 iterations. We have considered a vocabulary size of 50000 tokens. Early stopping was used by observing the running average of the loss on the validation set. For decoding our summaries, we have used a beam-search decoder with a beam length of 4. During decoding, we considered the maximum size of decoded tokens to be 120 and the minimum as 35. Our rest of the hyper-parameters regarding learning rate, word-embedding dimensions and LSTM-hidden unit dimensions are as reported in See et al. (2017). Although we have considered $\lambda_{TSC} = 1$, for all our experiments but to study its impact, we have experimented with different weights for our best-performing model (Sec. 5.1).

Model	R-1	R-2	R-L	IP
ATG (Zhu et al., 2018)	40.63	18.12	37.53	59.28
ATL (Zhu et al., 2018)	40.86	18.27	37.75	62.44
HAN (Zhu et al., 2018)	40.82	18.30	37.70	61.83
GR (Li et al., 2016)	37.13	15.03	30.21	61.70
TSC-MSMO-IS	41.03	18.77	37.90	63.93
TSC-MSMO-TS	41.0	18.70	37.87	63.8
TSC-MSMO-TIS	40.79	18.55	37.65	64.13
TSC-MSMO-SIMPAD-TIS	41.42	19.06	38.17	63.81

Table 2: Results on the Test set. We skipped the articles for which no relevant image labels were available resulting into evaluation of 9,851 articles from the test set.

λ_{TSC}	R-1	R-2	R-L	IP
0	40.63	18.12	37.53	59.28
0.5	40.79	18.57	36.67	64.45
1	41.42	19.06	38.17	63.81
1.5	40.95	18.68	37.92	63.99

Table 3: Impact of changing classifier weight (λ_{TSC}) on TSC-MSMO-SIMPAD-TIS model’s performance.

For evaluation of the textual summaries we have considered ROUGE (Lin, 2004). The official ROUGE script is used to report all of our ROUGE scores. For assessing the images recommended by the model as the pictorial summary we have used Image Precision (IP) defined by Zhu et al. (2018).

5 Quantitative Analysis

5.1 Overall Results

From Table 2, we can see that all the different model variants have outperformed the baselines. The TSC-MSMO-SIMPAD-TIS model has performed well in ROUGE-related metrics but did not perform as well as the TSC-MSMO-TIS in the IP metric. Using similar image feature vectors instead of zero-padded vectors during training has helped the SIMPAD model gain better textual understanding through the multimodal attention layers. However, using zero padded image vectors during testing did not support the model score well in the IP metric. Furthermore, the improved performance of the TSC-MSMO-IS in both ROUGE and IP metrics compared to the other non-SIMPAD variants supports the conclusion that the classifier works well when only image features are passed for classification.

The TSC-MSMO-TIS model also shows a marginal drop in the ROUGE-L score. The reason behind it may be that passing both textual and image features make it difficult for the classifier to decide whether to focus more on improving the "image" encoder or the "textual" encoder since the

target labels (similar/dissimilar) of the classifier depend on the augmented "images". The improved performances of the non-*TIS*-based models also suggest the benefit of experimenting with the *SIMPAD* versions of those models in the future.

To verify the effectiveness of *TSC*, we have experimented by adjusting its weight λ_{TSC} , as shown in Table 3. Although for $\lambda_{TSC} = 0.5$, we get a higher IP value, but a reduced weight on the classifier decreases the ROUGE score. Hence $\lambda_{TSC} = 1$ is a better choice giving good values for both IP and ROUGE metrics.

Model	Femail			
	R-1	R-2	R-L	IP
ATG (Zhu et al., 2018)	37.52	15.49	33.92	46.09
TSC-MSMO-IS	36.74	14.81	33.16	46.26
TSC-MSMO-TS	36.99	15.06	33.55	46.45
TSC-MSMO-TIS	36.69	14.84	33.16	46.05
TSC-MSMO-SIMPAD-TIS	37.50	15.37	33.85	46.01

Table 4: Topic-wise results on the test set for the "Femail" topic.

Model	Others			
	R-1	R-2	R-L	IP
ATG (Zhu et al., 2018)	34.13	13.92	31.02	54.52
TSC-MSMO-IS	32.74	12.53	29.68	53.71
TSC-MSMO-TS	32.49	12.22	29.27	53.73
TSC-MSMO-TIS	31.89	11.81	29.05	55.02
TSC-MSMO-SIMPAD-TIS	32.80	12.54	29.50	54.15

Table 5: Topic-wise results on the test set for the "Others" topic.

Model	Health			
	R-1	R-2	R-L	IP
ATG (Zhu et al., 2018)	42.04	19.89	39.15	82.82
TSC-MSMO-IS	40.66	18.78	37.68	83.18
TSC-MSMO-TS	41.50	19.31	38.49	83.89
TSC-MSMO-TIS	41.36	19.24	38.33	82.49
TSC-MSMO-SIMPAD-TIS	41.69	19.25	38.55	83.28

Table 6: Topic-wise results on the test set for the "Health" topic.

5.2 Topic-wise Results

As a part of the experimental analysis, we have also computed results for different topics by different models. Except for "Femail"⁴ (Table 4), "Health" (Table 6) and "Others" (Table 5) topics, the majority of the variants of the proposed model outperform the *MSMO-ATG* model in topic-wise results. A possible reason behind the poor performance of our model is that the topics "Femail" and "Health" consist of a high count of samples (the high spikes

⁴A topic in DailyMail that covers news related to fashion, shopping, etc.

Model	News			
	R-1	R-2	R-L	IP
ATG (Zhu et al., 2018)	44.55	21.61	41.20	63.1
TSC-MSMO-IS	44.42	21.63	41.15	63.75
TSC-MSMO-TS	44.51	21.64	41.21	63.8
TSC-MSMO-TIS	44.25	21.44	40.93	64.15
TSC-MSMO-SIMPAD-TIS	44.78	21.91	41.39	63.89

Table 7: Topic-wise results on the test set for the "News" topic.

Model	Sport			
	R-1	R-2	R-L	IP
ATG (Zhu et al., 2018)	37.11	15.16	34.30	68.52
TSC-MSMO-IS	37.10	15.21	34.35	68.57
TSC-MSMO-TS	36.7	14.85	33.96	68.11
TSC-MSMO-TIS	36.49	14.75	33.76	68.51
TSC-MSMO-SIMPAD-TIS	37.43	15.44	34.54	68.05

Table 8: Topic-wise results on the test set for the "Sport" topic.

shown in Fig. 3b and 3c) belonging to a "Others" sub-topic. The "Others" sub-topic indicates a collection of multiple sub-topics within a topic. Furthermore, the "Others" topic being composed of numerous topics, implicitly contains various sub-topics. Multiple sub-topics make it difficult for our title-based similarity search to find good quality similar images for the classifier hence leading to poor performance.

Major improvements are seen for "News" (Table 7), "Sports" (Table 8), "Travel" (Table 9), and "ScienceTech" (Table 10) topics. Even though the highest spike in the sub-topic sample count plot for the "news" topic (Fig. 3a) corresponds to the "Others" sub-topic, our model still performed well. The reason is that the high sample count within the "news" topic (as shown in Table 1) helped our title-based similarity find good-quality images for the classification task. Moreover, it is also observed that the models *TSC-MSMO-TS* and *TSC-MSMO-SIMPAD-TIS* have performed poorly only for the "Travel" topic as compared to other topics due to smaller training data available for the travel topic as it could be seen from Table 1. Thus data abundance in a particular topic plays an important role in improving the performance of our model.

5.3 Human Evaluation

We describe in this section the results of human evaluation of our proposed approach. For this, we employed three graduate student annotators to evaluate the multi-modal summaries produced by our best-performing model. We chose 100 random articles from the test set for the evaluation task. We then asked the annotators to judge the multi-modal

Model	Travel			
	R-1	R-2	R-L	IP
ATG (Zhu et al., 2018)	35.71	15.97	32.69	51.76
TSC-MSMO-IS	35.73	15.54	32.71	53.44
TSC-MSMO-TS	35.12	15.26	32.53	50.58
TSC-MSMO-TIS	36.53	16.35	33.62	55.66
TSC-MSMO-SIMPAD-TIS	36	16.37	33.17	52.27

Table 9: Topic-wise results on the test set for the "Travel" topic.

Model	ScienceTech			
	R-1	R-2	R-L	IP
ATG (Zhu et al., 2018)	41.29	20.23	38.41	72.94
TSC-MSMO-IS	41.65	20.61	38.77	73.53
TSC-MSMO-TS	41.42	20.15	38.48	72.8
TSC-MSMO-TIS	41.73	20.39	38.70	73.92
TSC-MSMO-SIMPAD-TIS	41.67	20.52	38.69	73.07

Table 10: Topic-wise results on the test set for the "ScienceTech" topic.

summaries based on the following criteria: (1) *Coverage*: where the model-generated textual summary is compared with the actual textual summary to check if the major points are adequately covered. (2) *Grammar*: where we investigate whether the model-generated textual summary is semantically correct. (3) *Topic-Aware-Text*: where we analyze whether the model-generated textual summary follows the suitable writing style such that it reflects the topic it belongs to. For example, a "sport"-topic summary should cover player names, whereas a "ScienceTech" topic summary should explain scientific facts using scientific terms. (4) *Topic-Aware-Image*: with this measure, we check whether the images selected by our model reflect the topic or not. For example, a "sport" topic pictorial summary should select pictures of players rather than spectators watching the game, whereas a "ScienceTech" topic pictorial summary should highlight the scientific event correctly.

It is difficult to judge the topic-aware criterion for samples belonging to topics like "news" or "others" due to its multiple sub-topics. So, the annotators were instructed to judge the topic-aware summary quality based on the "topic" they could determine from the title of the sample. For each evaluation criterion, the annotators were instructed

Model	Coverage	Grammar	TA-Text	TA-Image
ATG (Zhu et al., 2018)	3.71	4.42	4.12	4.09
TSC-MSMO-SIMPAD-TIS	3.83	4.38	4.33	4.47

Table 11: Human evaluation results. Here TA denotes Topic-Aware.

Topic : Sport	
Human Written Summary	A young Huddersfield Town fan wrote a charming letter to director Sean Jarvis. The boy, called Adam, found a £5 note at the John Smith's Stadium on Saturday. He did n't keep the cash, instead choosing to send it to club director Sean Jarvis. Adam pencilled a note to Jarvis asking if Aaron Mooy could be given the money. Jarvis shared a picture of the letter on Twitter, describing it as 'Pure class'. Midfielder Mooy also tweeted, writing: 'I would love to meet you Adam'. Mooy scored as Adam and his dad saw Huddersfield beat Manchester United 2-1.
MSMO - ATG Summary	Adam wrote a charming letter to huddersfield town director sean jarvis . Huddersfield's shock victory saw them rise to 11th in the premier league table . Three points at anfield would see them leapfrog the reds .
TSC-MSMO-IS	Adam wrote a charming letter to huddersfield town director sean jarvis . the young fan attended huddersfield 's 2-1 win over manchester united . adam has been taught not to hold on to what is not his , so he sent the money .
TSC-MSMO-TS	A boy called adam wrote a charming letter to huddersfield town director . the young fan attended huddersfield 's famous 2-1 win over manchester united . Adam has been taught not to hold on to what is not his , so he sent the money to jarvis and suggested that manager david wagner give it to mooy .
TSC-MSMO-TIS	A boy called adam wrote a charming letter to huddersfield town director . the young fan attended huddersfield 's famous 2-1 win over manchester united . Adam has been taught not to hold on to what is not his so he sent the money .
TSC-MSMO-SIMPAD-TIS Summary	A boy called adam wrote a charming letter to sean jarvis after finding a £ 5 note . The young fan attended huddersfield 's famous 2-1 win over manchester united . Adam has been taught not to hold on to what is not his , so he sent the money .
Images	

Figure 4: Topic: "Sport" Example Summaries comparison with Baseline. The circular colour codes corresponding to each category is used to represent the image selected as pictorial summary. The Green textual highlights refer to well-summarized content covering the major points of actual summary (highlighted in blue). The yellow highlights indicate extractive textual output.

to give a score from 1 (minimum) to 5 (maximum) to multimodal summaries.

As it can be seen from Table 11 there is a 3.2% increase in "Coverage" score, 5.1% increase in Topic-Aware-Text score, and 9.2% increase in Topic-aware-Image scores. The higher ratings of the topic-aware metrics and the rise in coverage-related metrics indicate that topic-awareness helps cover the major points discussed in the article. A minor decrease (0.9%) in the "grammar" related score can be due to some punctuation errors.

6 Qualitative Analysis

As shown in Fig. 4⁵, the textual summary produced by the baseline ATG model could not cover the content related to sending the letter well. In contrast, our best-performing model *TSC-MSMO-SIMPAD-TIS* captured the details of why the boy had sent the letter and gave insights into the match scores due to focusing on "sport"-topic-related features. Our *TSC-MSMO-SIMPAD-TIS* model selected the image of the letter (Image No. (1) in Fig. 4) as

⁵More examples are shown in Appendix A.1

part of the pictorial summary. Although the letter image was not in the human-annotated pictures list, its selection complements our textual summary well. The generated multimodal output indicates that the model maintains a balance between topic awareness and content relevance while producing the output. The balanced output may be because the classifier and the other summarizing components were given equal weights in the final loss function. The images chosen by the other *TSC*-variants and the baseline *ATG* were also not bad. In the given example, 4 images were chosen as part of the pictorial summary.

A significant limitation of our work, is the highlighted extractive textual summaries (Fig. 4) that resulted from using a PGN-based decoder. However, there are few extractive elements in the human written summaries, as seen from the blue highlighted text. Thus, the model learns this extractive behavior from the training data itself. Another limitation, as seen from the topic-wise results (Sec. 5.2), is the dependence on data size for producing good quality output. The presence of low-sample-count sub-topics further adds to the problem of finding good "On-topic"-similar images for the classifier, thus leading to a deterioration in the quality of the summary produced by our model.

7 Conclusion and Future Study

Multimodal summaries help users absorb rich multimedia knowledge by generating brief and pertinent summaries. Adding topic information helps our model learn the different representation styles of various topics resulting in better quality summaries. The improvement in ROUGE and IP scores in the overall test set and the topic-wise segments for all our experiments indicate that making the model learn topic-related information helps produce better quality multimodal summaries. Furthermore, our experiments also established that using similar image features instead of the zero-padded vectors for samples having lesser in-article images does help in producing better summaries.

In a future study, we can find similar images using temporal information already present in our augmented dataset. Exploration of other techniques like comparing image-image, image-caption, or image-title embeddings for finding similar photos can also be done. A novel dataset can be created with lesser sub-topics and well-defined topics for studying our topic-based summarization technique.

7.1 Acknowledgement

Dr. Sriparna Saha gratefully acknowledges the Young Faculty Research Fellowship (YFRF) Award, supported by Visvesvaraya Ph.D. Scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia) for carrying out this research.

References

- Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. 2021. Video summarization using deep neural networks: A survey. *Proceedings of the IEEE*, 109(11):1838–1863.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. [Neural machine translation by jointly learning to align and translate](#).
- Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. [Universal sentence encoder](#).
- Jingqiang Chen and Hai Zhuge. 2018. Extractive text-image summarization using multi-modal rnn. In *2018 14th International Conference on Semantics, Knowledge and Grids (SKG)*, pages 245–248, Guangzhou, China. IEEE, IEEE.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. [Imagenet: A large-scale hierarchical image database](#). In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255, Miami, Florida, USA. IEEE Computer Society.
- Günes Erkan and Dragomir R. Radev. 2011. [Lexrank: Graph-based lexical centrality as salience in text summarization](#). *CoRR*, abs/1109.2128.
- Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Raptatzikos, Georgios Skoumas, and Yannis Avrithis. 2013. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 15:1553–1568.
- Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1):1–66.
- Masaru Isonuma, Toru Fujino, Junichiro Mori, Yutaka Matsuo, and Ichiro Sakata. 2017. [Extractive summarization using multi-task learning with document classification](#). In *Proceedings of the 2017 Conference on*

- Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2101–2110, Copenhagen, Denmark. Association for Computational Linguistics.
- Anubhav Jangra, Raghav Jain, Vaibhav Mavi, Sriparna Saha, and Pushpak Bhattacharyya. 2020a. [Semantic extractor-paraphraser based abstractive summarization](#). In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 191–199, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLP AI).
- Anubhav Jangra, Adam Jatowt, Mohammed Hasanuzzaman, and Sriparna Saha. 2020b. [Text-image-video summary generation using joint integer linear programming](#). In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II*, volume 12036 of *Lecture Notes in Computer Science*, pages 190–198. Springer.
- Anubhav Jangra, Adam Jatowt, Sriparna Saha, and Mohammad Hasanuzzaman. 2021a. [A survey on multi-modal summarization](#). *CoRR*, abs/2109.05199.
- Anubhav Jangra, Sriparna Saha, Adam Jatowt, and Mohammad Hasanuzzaman. 2020c. [Multi-modal summary generation using multi-objective optimization](#). In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1745–1748. ACM.
- Anubhav Jangra, Sriparna Saha, Adam Jatowt, and Mohammad Hasanuzzaman. 2021b. [Multi-modal supplementary-complementary summarization using multi-objective optimization](#). In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 818–828. ACM.
- Aman Khullar and Udit Arora. 2020. [MAST: multi-modal abstractive summarization with trimodal hierarchical attention](#). *CoRR*, abs/2010.08021.
- Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. [Aspect-aware multimodal summarization for chinese e-commerce products](#). In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8188–8195, New York, NY, USA. AAAI Press.
- Haoran Li, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2016. [Guiderank: A guided ranking graph model for multilingual multi-document summarization](#). In *Natural Language Understanding and Intelligent Applications - 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2-6, 2016, Proceedings*, volume 10102 of *Lecture Notes in Computer Science*, pages 608–620. Springer.
- Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, and Chengqing Zong. 2018. [Multi-modal sentence summarization with modality attention and image filtering](#). In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4152–4158, Stockholm, Sweden. ijcai.org.
- Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2017. [Multi-modal summarization for asynchronous collection of text, image, audio and video](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1092–1102, Copenhagen, Denmark. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. [Effective approaches to attention-based neural machine translation](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421, Lisbon, Portugal. The Association for Computational Linguistics.
- Toru Nishino, Shotaro Misawa, Ryuji Kano, Tomoki Taniguchi, Yasuhide Miura, and Tomoko Ohkuma. 2019. [Keeping consistency of sentence generation and document classification with multi-task learning](#). 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3193–3203, Hong Kong, China. Association for Computational Linguistics.
- Shruti Palaskar, Jindrich Libovický, Spandana Gella, and Florian Metze. 2019. [Multimodal abstractive summarization for how2 videos](#).
- Melissa Sanabria, Frédéric Precioso, and Thomas Menguy. 2019. [A deep architecture for multimodal summarization of soccer games](#). In *Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports*, pages 16–24.
- Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. [Get to the point: Summarization with pointer-generator networks](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

- Karen Simonyan and Andrew Zisserman. 2015. [Very deep convolutional networks for large-scale image recognition](#).
- Dian Tjondronegoro, Xiaohui Tao, Johannes Sasongko, and Cher Han Lau. 2011. [Multi-modal summarization of key events and top players in sports tournament videos](#). In *IEEE Workshop on Applications of Computer Vision (WACV 2011), 5-7 January 2011, Kona, HI, USA*, pages 471–478, Kona, HI, USA. IEEE Computer Society.
- X. Ye, Z. Yue, R. Liu, and Q. Lu. 2021. [Mtms: A fact-corrected summarization model based on multitask learning and multimodal fusion](#). In *2021 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, pages 238–247, Los Alamitos, CA, USA. IEEE Computer Society.
- Chenxi Zhang, Zijian Zhang, Jiangfeng Li, Qin Liu, and Hongming Zhu. 2021. [Ctr: Compress-then-reconstruct approach for multimodal abstractive summarization](#). In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.
- Sicheng Zhao, Hongxun Yao, Sendong Zhao, Xuesong Jiang, and Xiaolei Jiang. 2016. Multi-modal microblog classification via multi-task learning. *Multi-media Tools and Applications*, 75(15):8921–8938.
- Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. [MSMO: multimodal summarization with multimodal output](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4154–4164, Brussels, Belgium. Association for Computational Linguistics.
- Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020. [Multimodal summarization with guidance of multimodal reference](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9749–9756.

A Appendices

A.1 Sample Summaries


Topic : Health	
Human Written Summary ●	Since 2006 , nearly half of all food contamination warnings in California have been for lead in candy , according to a new study. Almost all of the contaminated candies have been imported , mainly from Mexico , China and India. In the wake of the Flint , Michigan water crisis , the study authors advocate for vigilance in identifying lead contamination and protecting children
MSMO – ATG Summary ●	The university of california , san francisco study reports that since the state passed a law on testing and monitoring candy in 2006 . As many as 10,000 children get lead poisoning in california each year , according to the study . Recalling the flint , michigan water crisis , the study's author urges consumers to be mindful and watchful for lead contamination .
TSC-MSMO-SIMPAD-TIS Summary ●	Lead in candy has accounted for 42 percent of food contamination warnings in California since 2006 , a study found . The university of california , San Francisco study reports that since the state passed a law on testing and monitoring candy in 2006 . There have been more reports issued warning about lead in sweet treats -- mostly imported ones -- than for any other contamination .
Images	

Figure 5: **Example 1:** Comparison between Multimodal summary generated by our best performing model TSC-MSMO-SIMPAD-TIS and the baseline for a sample belonging to "Health" topic. Only one image is part of the sample, and it is selected as the pictorial summary

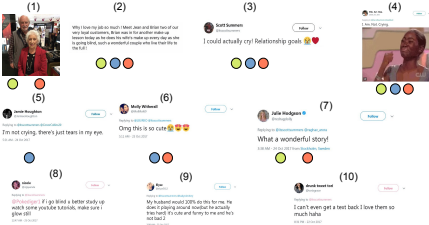
Topic : Femail	
Human Written Summary ●	Jean and Brian are regulars at a local beauty store. Jean is going blind , so Brian goes for make-up lessons. He is learning how to do her cosmetics so he can help her when she can no longer do it herself. The image has been widely shared on social media with many proclaiming the pair to be " couple goals "
MSMO – ATG Summary ●	Couple Jean and Brian are regulars at one local make-up shop. They go to the store together so brian can take make-up lessons . The identities and location of the man and woman are unknown, but their devoted bond has melted hearts.
TSC-MSMO-SIMPAD-TIS Summary ●	Couple Jean and Brian are regulars at one local make-up shop , but not because they're eager to get their hands on all the latest products or test new beauty . Rather , the two go to the store together so brian can take make-up lessons , and learn to put his wife 's face on before she goes blind and can no longer do it herself . The identities and location of the man and woman are unknown , but their devoted bond has melted hearts universally.
Images	

Figure 6: **Example 2:** Comparison between Multimodal summary generated by our best performing model TSC-MSMO-SIMPAD-TIS and the baseline for a sample belonging to "Femail" topic. Only one image is part of the sample, and it is selected as the pictorial summary

As shown in the 1st example (Fig. 5) our TSC-MSMO-SIMPAD-TIS model's summary stated the exact percentage of lead contamination (42%). In contrast, the human summary has stated: "nearly half" to explain the lead contamination rate. The ATG model covered facts regarding the number of children affected each year. However, it missed the detail that the "imported"-candies were mostly contaminated and should be avoided. This fact was covered well by our proposed model's textual

summary.

In the 2nd example (Fig. 6), the textual summary produced by our TSC-MSMO-SIMPAD-TIS model was able to capture the significant reason why the couple went to the make-up-shop. The reason that Brian's wife would be going blind was not covered in the textual summary by ATG model. Further in the pictorial summary, although the ATG model chose good images of the tweets but missed the picture of the couple (Image no. (1) in Fig. 6), which our model chose.