Affective Feedback Synthesis Towards Multimodal Text and Image Data

PUNEET KUMAR, Indian Institute of Technology Roorkee, India
GAURAV BHATT, Indian Institute of Technology Hyderabad, India
OMKAR INGLE, Indian Institute of Technology Roorkee, India
DAKSH GOYAL, National Institute of Technology Karnataka, India
BALASUBRAMANIAN RAMAN, Indian Institute of Technology Roorkee, India

In this article, we have defined a novel task of affective feedback synthesis that generates feedback for input text and corresponding images in a way similar to humans responding to multimodal data. A feedback synthesis system has been proposed and trained using ground-truth human comments along with image–text input. We have also constructed a large-scale dataset consisting of images, text, Twitter user comments, and the number of likes for the comments by crawling news articles through Twitter feeds. The proposed system extracts textual features using a transformer-based textual encoder. The visual features have been extracted using a Faster region-based convolutional neural networks model. The textual and visual features have been concatenated to construct multimodal features that the decoder uses to synthesize the feedback. We have compared the results of the proposed system with baseline models using quantitative and qualitative measures. The synthesized feedbacks have been analyzed using automatic and human evaluation. They have been found to be semantically similar to the ground-truth comments and relevant to the given text–image input.

CCS Concepts: • Computing methodologies → Natural language generation; Machine learning; • Information systems → Multimedia and multimodal retrieval; Sentiment analysis;

Additional Key Words and Phrases: Affective computing, feedback synthesis, multimodal input, dataset construction, context vector

ACM Reference format:
Puneet Kumar, Gaurav Bhatt, Omkar Ingle, Daksh Goyal, and Balasubramanian Raman. 2023. Affective Feedback Synthesis Towards Multimodal Text and Image Data. ACM Trans. Multimedia Comput. Commun. Appl. 19, 6, Article 190 (May 2023), 23 pages.
https://doi.org/10.1145/3589186

1 INTRODUCTION

Multimodal data processing has emerged as an important sub-domain in Artificial Intelligence (AI) research due to the fast growth of multimedia data in the last few years [33]. One of the goals of AI

The Ministry of Education INDIA has supported this research through grant no. 1-3146198040.

Authors’ addresses: P. Kumar, O. Ingle, and B. Raman, Indian Institute of Technology Roorkee, India, 247667; emails: pkumar99@cs.iitr.ac.in, iuttam@mt.iitr.ac.in, bala@cs.iitr.ac.in; G. Bhatt, Indian Institute of Technology Hyderabad, India, 502285; email: gauravbhatt@iith.ac.in; D. Goyal, National Institute of Technology Karnataka, India, 575025; email: dakshgoyal.171cv111@nitk.edu.in.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
1551-6857/2023/05-ART190 $15.00
https://doi.org/10.1145/3589186

ACM Trans. Multimedia Comput. Commun. Appl., Vol. 19, No. 6, Article 190. Publication date: May 2023.
is to enable machines to respond to multimodal data in the same manner as humans. A response could be feedback, answer, caption, vocal signal, facial reaction, bodily gesture, and more. [42]. Humans portray different emotions through various modalities, among which text and images are known to contain human emotions and intentions most effectively [36]. Machines and systems capable of generating affective feedback to multimodal text and image data could be very useful [13]. Here, the terms affective and human-like are used in the sense that the feedback synthesis system should be able to synthesize feedback to a multimodal input in a way that humans do.

The ability to synthesize feedback to multimodal data could be helpful in various applications, such as determining user response to products, social behavior analysis, evaluating multimodal educational content (e.g., slides, blogs, and books), and predicting user engagement in advertisements [13]. Multimodal feedback synthesis systems can also be used to predict emotions that multimedia contents would induce in users and, hence, predict the success of the contents [39]. Likewise, the success of advertisements and educational content can also be predicted upfront by gauging the kind of feedback that users are likely to have to them [4]. Moreover, the analysis of multimodal data generated by patients can help predict their mental states [37].

Among the existing research tasks, multimodal summarization and dialog generation are related to the affective feedback synthesis task introduced in our work [14, 17, 25]. The text and image summarization approaches use the information from a single modality whereas multimodal summarization leverages the information from multiple modalities to fine-tune the summary [65]. However, most multimodal summarization approaches do not use actual human responses to train their models. Likewise, textual and visual dialog generation models are also not trained on human responses. They also suffer from the problem of short and uninformative dialog generation [58]. Moreover, affective feedback generation can be abstract and more difficult compared with the aforementioned tasks. For instance, identifying sarcasm is a common challenge in affective feedback generation that can alter the conveyed sentiment. To this end, the proposed multimodal feedback synthesis system has been trained using actual human comments along with text and image inputs. The relevance of the comments is decided based on their number of likes (upvotes). We have incorporated textual and visual features that enable the proposed system to generate informative feedback considering the textual and visual context of the inputs.

In this article, we have proposed a novel task to synthesize affective feedback to text–image inputs, which is illustrated in Figure 1. The idea is to build a model that could generate contextually relevant feedback as a new modality from two given modalities, that is, news images and text. Affective feedback can be considered a special type of summary produced by humans based on their state of mind induced on processing the emotional context of the inputs [38]. Affective feedback aligns with the sentiment being expressed and is potentially useful in analyzing human behavior. We have also constructed a large-scale dataset, the IIT Roorkee Multimodal Feedback (IIT-R MMFeed) dataset, for multimodal feedback synthesis. For simplicity, the IIT-R MMFeed dataset is referred to as the MMFeed dataset in this article. The MMFeed dataset has been constructed by crawling news articles through Twitter feeds. It contains images, text, human comments, and the number of likes for the comments.

The proposed system contains textual and visual encoders (as shown in Figure 3) to which text and images are given as the inputs. The textual encoder uses a multi-headed self-attention-based transformer [55] and extracts textual features in the form of textual context vectors. The visual encoder uses a Faster Region–Based Convolutional Neural Network (R-CNN) model [49] and extracts visual features in the form of a visual context vector. Each text encoder output vector is concatenated with a visual context vector and passed through a feedforward network that outputs multimodal contextual vectors with the same dimension as the text encoder vectors. The decoder block takes textual context vector, multimodal context vectors, and ground-truth comments as
inputs and generates the feedback as output. The synthesized feedbacks are evaluated using qualitative and quantitative methods for their relevance with the ground-truth comments and inputs.

Apart from defining a novel task to generate feedback to multimodal input and constructing a large-scale dataset for multimodal feedback synthesis, this article’s contributions include the proposal of a multimodal feedback synthesis system. It involves the extraction of textual and visual features, synthesis of the feedbacks according to the multimodal context, and language-agnostic similarity evaluation of the synthesized feedbacks as compared with the multimodality inputs and ground-truth comments. The proposed task involves a major technical challenge to create the alignment between textual and visual modalities to bridge the semantic gap between these two modalities. This alignment is achieved by multimodal fusion. The feedforward layer in the multimodal fusion block has been trained against the loss back-propagated by the decoder module, which maximizes the capture of semantic similarity while concatenating two context vectors. This process is very challenging compared with concatenation-based information fusion approaches.

The important contributions made in this article are as follows:

- The novel task has been defined to generate feedback to multimodal input containing images and text similar to feedback of humans. The task aims to synthesize a new modality: feedback that is contextually relevant to the two given modalities, news images and text.
- A large-scale dataset, the IIT-R MMFeed dataset, has been constructed for multimodal feedback synthesis by crawling news articles through Twitter feeds. It contains images, text, and Twitter user comments, along with the number of likes (upvotes) for each comment.
- A multimodal feedback synthesis system has been proposed. Textual and visual encoders extract textual and visual features using a text transformer and a Faster R-CNN model. The decoder synthesizes the feedback according to the multimodal context that combines textual and visual features.

The organization of the remainder of the article is as follows. The research tasks related to affective feedback synthesis are surveyed in Section 2. Section 3 describes the dataset construction process and the proposed system. The experiments and results are discussed in Section 4 and
Section 5, respectively. Our major conclusions are drawn in Section 6 along with future research directions.

2 RELATED WORK
This section surveys the research advances related to the task of affective feedback synthesis introduced in this article.

2.1 Text Summarization
There are two major types of text summarization approaches: abstractive and extractive summarization [12, 14]. Extractive summarization is focused on extracting useful words from the given text. In this context, Narayan et al. [40] worked on neural extractive summarization and utilized side information such as the title and captions of the news images. Abstractive summarization expresses the main content of a given text using different words instead of creating the summary by selecting the words or sentences from the text. Abstractive summarization of sentences has been explored through neural attention-based models [51]. Extractive summarization is limited by the vocabulary set of the input text, whereas abstractive summarization approaches are not able to generate factually consistent and human-like summaries [26]. Moreover, text summarization does not leverage the information from multiple modalities to fine-tune the summary.

2.2 Image Summarization
Summarization of a scene from an image collection is an existing research problem [54]. In this context, Pan et al. [43] proposed an approach to generate a visual summarization for a set of images. They considered social attractiveness features such as image quality and aesthetics while summarizing the images. Samani et al. [52] considered semantic features along with social attractiveness features to summarize the images in a context-sensitive manner. Image summarization approaches consider only visual information to produce the summary of the contents, whereas the multimedia contents often contain a combination of visual and textual information [29]. Moreover, the output of these approaches is in visual form only, which is more challenging to infer the emotion and context-related information as compared with textual and multimodal outputs [27]. The limitations of text summarization and image summarization of not leveraging the information from multiple modalities are addressed by multimodal summarization approaches [65].

2.3 Multimodal Summarization
In the context of multimodal summarization, Chen and Zhuge [7] used recurrent networks to generate a summary from text and images. They performed simultaneous summarization of images and text documents and aligned the sentences and images to generate the summaries. Zhu et al. [65] used a pointer generator network for multimodal summarization instead of performing manual alignment of the text and images. They picked the most relevant image among the input images for a data sample and selected the essential keywords from the text inputs by performing extractive summarization. Summarization of multimedia news has also been explored using extractive summarization approaches [8]. In an attempt to utilize videos instead of images for multimedia news articles, Zellers et al. [61] implemented self-attention to automatically choose a suitable video frame based on the semantic meaning of the article. They modeled the article’s semantic meaning along with the input video jointly. In another work, Shang et al. [53] considered the time-related information for multimodal video summarization. They implemented a time-aware transformer model along with an attention mechanism. The attention mechanism enabled them to attend the inputs differently depending on the timing of the input and the timing related information learnt in this way was used while summarizing the videos.
Multimodal processing faces the challenges of either missing out on specific modalities or biasing the results on a particular modality [42]. Zhu et al. used Multimodal reference [66] to handle the modality-bias problem in multimodal summarization. They used the multimodal reference for guidance, designed an objective function, and proposed a novel evaluation metric based on the joint multimodal representation that considered the loss of image selection and summary generation. Context-aware summarization techniques have also been explored for multimodal summarization. In this direction, Li et al. [31] incorporated aspect coverage and corresponding use cases for various product categories. This research did not consider human comments to the multimodal input data while training their summarization models. The proposed system is trained on human-generated comments along with multimodal inputs containing text and images. It generates the feedback as a new modality from two input modalities: images and text.

2.4 Textual Dialog

Textual dialog systems consider text input questions and generate text output responses in the form of a dialog. In this context, Zhou et al. [63] developed chatting agents that maintain and use a memory of emotional keywords. In another work, Xu et al. [60] worked on conversation modelings for dialog generation, whereas Zhao et al. [62] used conditional variational autoencoders to understand the diverse use cases for dialog generation models. Researchers have also worked on increasing the relevance, diversity, and originality of generation results [59]. In this context, Gu et al. [17] trained a conditional generative network for dialog modeling. They modeled data distribution by training a generative adversarial network considering multiple possible probability distributions of the topics and sentiments in the latent variable space. In an attempt to incorporate human-ness in the generated dialogs, Zhou and Wang [64] used conditional autoencoders and included emojis in the synthesized responses. Most of the real-life multimedia context is expressed through multiple modalities in which combining the complementary information from various modalities helps in understanding the underlying emotional context effectively [21]. However, textual dialog systems consider only the textual context while generating the response to input questions. There is a need to consider the corresponding visual context as well, which has given rise to the development of visual dialog systems, as explored in the next section.

2.5 Visual Dialog

Visual dialog (VisDial) is another closely related task to Multimodal Affective Feedback Synthesis, which aims to generate dialogs about visual input contents. In this context, Kang et al. [25] proposed a Dual Attention Network (DAN) to resolve visual references between given images and dialog history. They implemented multi-head attention to learn the relationships between the given question and the dialog history and bottom-up attention to model the image features and output dialogs’ representations. In a similar work, Chen et al. [6] implemented dual-channel reasoning to learn the context from the images and dialog history together. The dual-channel reasoning that they proposed enabled them to learn rich semantic representations of the questions compared with the single-channel reasoning approach followed by DAN. In another work, Niu et al. [41] used recursive attention for finding the image component referred to by a particular text entity. Their work was extended by Park et al. [45] for multi-view settings of recursive attention. Multimodal VisDial synthesis has also been explored for audio-visual data [1, 20]. To effectively model the visual features, Jiang et al. [24] implemented a region-based graph attention network to learn question-aware relationships of the input images and dialog history-aware question features. This inspired us to incorporate attention-based visual feature extraction in our proposed work.

The existing VisDial methods are not able to solve our problem of affective feedback synthesis. We have empirically found that they could not synthesize meaningful feedback, and the responses
are limited to a few words. Little work has been carried out in this context. Further, multimodal summarization and VisDial systems are prone to the modality-bias problem — they tend to consider one of the input modalities more than the others [66].

### 2.6 Visual Storytelling

Visual storytelling or visual narrative involves using visual media to describe a narrative [22, 23]. It uses photos, videos, graphics, and text to describe a story’s context. In this context, both visual and textual modalities are used simultaneously, and sometimes one modality is used to generate the other. For instance, researchers have used image description to synthesize natural text and captions [18]. The synthesized text can be considered a response to the input image. Further work has been done on text-to-image generation guided by emotions. In the context of text-guided emotion generation, Cho et al. [9] proposed text-generation models to generate text labels based on multimodal inputs. In another work, Li et al. [32] used recurrent networks and adversarial learning to generate a story from given images and optimize it enforcing emotion-related information. Though these visual storytelling approaches indirectly induce emotions in the reader’s mind, they do not aim to explicitly generate a response to the multimodal input.

### 2.7 Evaluation of Machine Synthesized Sentences

The quality of machine-synthesized sentences — such as translations, text summaries, image captions, and dialogues — is subjective. Various quantitative metrics based on recall, precision, and sensitivity have also been used. For example, Bhandari et al. [3] and Zhu et al. [66] used a recall-oriented metric, ROUGE, to evaluate machine-synthesized visual dialogs against textual reference sentences. They also evaluated their output against the reference description of input images using precision value. Ranking-based evaluation metrics, such as Recall@k [50] and Mean Reciprocal Rank [10] have also been utilized to rank machine-synthesized sentences against a set of reference sentences.

As discovered from the literature surveyed earlier, synthesis of affect during multimodal summarization has not been explored to its full potential. Most of the methods discussed earlier synthesized a textual response to multimodal data; however, they did not use actual human responses to train their models. An adequate dataset for multimodal feedback synthesis is also not available. With that as an inspiration, a method to synthesize affective feedback to multimodal data has been proposed in this work. A large-scale dataset has also been constructed for multimodal feedback synthesis.

Multimodal summarization and VisDial systems are prone to the modality-bias problem in which they tend to consider one of the input modalities more than the others [66]. Subjective human evaluation is helpful to determine whether the output is relevant to both the input modalities, that is, text and images [28]. For an exhaustive evaluation of the feedback synthesized by the proposed system, we have implemented five metrics based on recall, precision, and sensitivity (BLEU Score, ROUGE, Meteor, CIDEr, and SPICE) and two ranking-based metrics (Recall@k and Mean Reciprocal Rank). Human evaluation has also been carried out to subjectively evaluate the relevance of the synthesized feedback with the input text, images, and ground-truth comments.

### 3 PROPOSED SYSTEM

This article proposes a multimodal system to synthesize feedback for given text and image data similar to the way that humans do it. The formulation of the proposed task has been described as follows, along with the construction of the MMFeed dataset and proposed affective feedback synthesis system’s architecture.
Table 1. MMFeed Dataset’s Parameters and Their Values

| Parameter               | Value          |
|-------------------------|----------------|
| No. of news articles    | 9,479          |
| No. of samples          | 77,790         |
| Avg. comments per article | 8.21           |
| Avg. no. of likes per comment | 1.51         |
| Avg. length of news text | 611 words      |
| Avg. length of comments | 15.71 words    |

Fig. 2. MMFeed dataset’s distribution as per number of comments (denoted as “c”).

3.1 Problem Formulation

Given a multimodal input $M = \{T, I\}$, where $T = \{t_1, t_2, \ldots t_m\}$ is the news text and $I$ is an image (where $m$ denotes the length of the news text sequence), the proposed system generates affective (i.e., human-like) feedback $F = \{f_1, f_2, \ldots f_k\}$, where $k$ denotes the length of the feedback sequence. The problem is to generate a new modality: feedback from two given modalities, that is, news images and text. The synthesized feedback is considered to be affective or human-like in the sense that the proposed system synthesizes them in a way that humans do.

3.2 Dataset Construction

A large-scale dataset, the IIT Roorkee Multimodal Feedback (IIT-R MMFeed) dataset, has been constructed for multimodal feedback synthesis. The procedure to compile and preprocess the dataset is presented in Appendix A. It was collected by crawling news articles from corresponding Twitter handles (such as “TIME,” “CNN,” “NYTimes,” “BBCBreaking,” etc.) using NLTK\(^1\) and newspaper3k\(^2\) libraries and Tweepy API.\(^3\) The MMFeed dataset consists of 77,790 samples collected through 9,479 Tweets containing images, text, and user comments, along with the number of likes (upvotes) for each comment. Table 1 contains various parameters of the dataset while example data instances and the dataset’s distribution as per the number of comments for each image can be found in Table 2 and Figure 2, respectively.

The MMFeed dataset stands out from the existing datasets, such as the New York Times Articles & Comments (2020) dataset,\(^11\) as it contains the images as well as the number of likes (upvotes), which are helpful to evaluate the relevance of the comments in the context of the input images and text. The MMFeed dataset includes data from various genres, such as sports, politics, and current affairs, which can be utilized for the proposed feedback synthesis system’s robust training. As human users have generated comments in response to input text and images, they can be considered the combined representation of the textual and visual contexts. Moreover, the number of likes associated with the comments can be considered to denote the comments’ relevance to the input images and text. The proposed system has been trained on the input text and images and evaluated against the comments.

3.3 Proposed System

The proposed system’s overall architecture is described in Figure 3 and elaborated in the following sections. The first two blocks are textual and visual encoders that work in parallel. Input text and images are given as the input to these blocks, respectively. The output of the textual encoder is a time series of vectors (textual context vectors, $z^*$), and the output of the visual encoder

---

\(^1\)https://nltk.org/.
\(^2\)https://newspaper.readthedocs.io/.
\(^3\)https://docs.tweepy.org/en/stable/.
Table 2. A Few Samples from the MMFeed Dataset

| Title                                                                 | Text                                                                                                                                                                                                 | Image                                                                 | Comment                                                                                     | Likes |
|-----------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------|--------------------------------------------------------------------------------------------|-------|
| Cruise carnival ship crashed while docking at a port in Mexico on Friday | One carnival cruise ship crashed into another while trying to dock at a port in Mexico on Friday leaving six injured passengers a damaged hull and a handful of expletiveladen social media videos in its wake the crash happened as the cruise ship was maneuvering to dock... | ![Image](image1.png)                                                    | And this is just one reason why I do not go on cruise ships.                              | 27    |
| COVID: Police and protesters clash during Dutch curfew demo            | Riot police in the Netherlands have clashed with protesters. Officers used water cannon and tear gas. They had gathered in defiance of a new curfew. Some protesters threw fireworks, looted supermarkets... | ![Image](image2.png)                                                    | A lot of people seem to be throwing fire on the actual problem.                            | 7     |
| Fully vaccinated people can travel in US without tests or quarantines, as long as they remain masked, CDC says... | People who are fully vaccinated against covid can travel freely in the us as long as they remain masked on planes buses and trains the centers for disease control and prevention announced friday it is unclear how much impact the new guidance will have people are already traveling... | ![Image](image3.png)                                                    | I can that is amazing usually my crippling social anxiety keeps me from doing this but now i am cured. | 13    |
| Health workers, stuck in the snow, administer corona-virus vaccines   | Public health workers in Oregon were driving back from a coronavirus vaccination site when they got stuck in the snow. They then walked from car to car, giving shots to drivers with vaccine doses... | ![Image](image4.png)                                                    | Thank goodness for resourceful and kind people.                                          | 13    |
| China reports its highest Covid-19 infections since March             | China recorded 57 local Covid-19 cases on Sunday, the highest number the country has seen since it brought the coronavirus largely under control in March, as per the Health Commission... | ![Image](image5.png)                                                    | Less than one hundred. It is nothing compared to the number in America.                   | 17    |
| Maryland-bound train derailed, two cars fall into Potomac River       | Two freight cars fell into the Potomac River near Harpers Ferry, WV. The cause of the derailment remains unclear and is under investigation, CSX said. The railway is working to swiftly to clean the area... | ![Image](image6.png)                                                    | Looks like the train cars fell through the pedestrian bridge.                            | 5     |

Here, “likes” are the number of likes (upvotes) for the corresponding comment, denoting its relevance to the input text and image.

is a single vector (visual context vector, g*). Each text encoder output vector is concatenated with g* and passed through a feedforward network that outputs another time series of vectors (multimodal contextual vectors, y*) with the same dimension as the text encoder vectors. Next is the decoder block for which z* and y* are the two inputs. During training, the ground-truth comments are provided as the third input against which the model is trained. During testing, feedback is synthesized as an output. The similarity module evaluates the similarity between the input comments and synthesized feedback.

### 3.3.1 Text Encoder

Encoding of textual data is achieved by a multi-headed self-attention-based transformer [55]. It converts the embeddings of textual tokens into vectors using the embedding layer. The encoder block includes a self-attention layer and a feedforward layer. Its sublayers (i.e., feedforward and self-attention layers) have residual connections around them, and each layer is followed by a normalization layer. To account for the order of tokens in input sequence position embedding is applied. **Textual Attention**: Three vectors — Q (query), K (key), and V (value) — are obtained by multiplying the encoder’s i<sup>th</sup> input vector with three weight matrices W(Q), W(K), W(V) and used to find the output of the self-attention layer z<sub>i</sub> at position i. This process is repeated for all input vectors to get attention head matrix z as per Equation (1). The attention head and key vector’s dimension are...
Fig. 3. The schematic architecture of the proposed system. Here, three textual encoder blocks and three decoder blocks are used along with a visual encoder network. The blue-shaded area shows the frozen part of the network. The dotted arrows represent the residual connections. The similarity module generates the similarity score, which denotes the significance of the synthesized feedback.

denoted by \( \{z_1, z_2, \ldots, z_n\} \) (where \( n \) is the length of input sequence) and \( d_k \). \( W(Q), W(K), W(V) \) are randomly initialized vectors that are trained during the process.

\[
z = \text{Attention}(Q, K, V) = \operatorname{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V, \tag{1}\]

where \( Q, K, \) and \( V \) are the matrices packed with all of the queries, keys, and values, respectively, whereas \( K^T \) denotes the Transpose of the matrix \( K \) and \( d_k \) is the scaling factor.

We have calculated multiple attention heads that provide multiple representation subspaces to the attention layer. Each of the attention heads is calculated with a separate set of Query, Key, Values weight matrices. Then, these attention heads are concatenated and multiplied by the weight matrix \( W_O \) to generate an intermediate vector \( z' \) that has the same dimension as the single attention head, which is passed through the feedforward layer to generate the textual context vector, \( z^* \). The keys, values, and queries are linearly projected \( h \) times to \( d_k, d_v, \) and \( d_k \) dimensions. As shown in Equation (2), the attention function is operated on the projected keys, values, and queries simultaneously to get \( d_v \)-dimensional output.

\[
\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_O
\]

\[
\text{head}_i = \text{Attention} \left( QW_i^Q, KW_i^K, VW_i^V \right), \tag{2}\]

where \( h \) denotes the number of attention heads whereas \( W_i^Q, W_i^K, W_i^V \), and \( W_O \) are the parameter matrices with the projections of queries, keys, values, and output, respectively.

The encoder stack has a fully connected feedforward network for each layer. The input layer of this network is of 256 dimensions, which is the same as the dimension of the output layer, whereas the hidden layers are of 512 dimensions. The network’s output for input \( x \) is computed as per Equation (3).

\[
\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2, \tag{3}\]

where \( x \) is the input, \( b_1 \) and \( b_2 \) are bias terms, and \( W_1 \) and \( W_2 \) are the weight matrices.
3.3.2 Image Encoder. We have used a pretrained Faster R-CNN [49] model to extract the visual features. The input image is fed to a series of convolutional layers to obtain the feature maps. Then, the Region Proposal Network (RPN) runs a sliding window of size $n \times n$ through them to generate the anchor boxes. The total number of anchor boxes for a feature map is $W_i \times H_i \times K_i$, where $H_i$ and $W_i$ denote the height and width of a feature map, respectively, and $K_i$ is the number of anchors for each position. An objectiveness score defined as a binary class label (of being an object or not) is given to each anchor box based on the Intersection Over Union (IoU). This score can be either positive or negative (Equation (4)). The anchors, which are neither positive nor negative, do not contribute to the training. The anchors are then passed into the classification (object classification) and regression (object localization) layers, which output the classified boxes in the image.

$$\text{Objectiveness Score} = \begin{cases} \text{Positive;} & \text{IoU} > 0.7 \\ \text{Negative;} & \text{IoU} < 0.3 \end{cases}$$ (4)

Here, 1,601 classes were used for the anchor box. The model is trained to classify the bounding boxes. The bounding boxes and the corresponding features are then concatenated to form a global feature vector that is then passed to the Visual Attention Network (VAN).

The rationale behind choosing a Faster R-CNN for object detection during visual feature extraction is governed by its speed and applicability in our use case. The Faster R-CNN model is computationally Faster in extracting the visual features than the other models of its family, R-CNN [16], and Fast R-CNN [15] that incorporated an ROI (Region of Interest) pooling layer to speed up the conventional R-CNN model. The Faster R-CNN replaces the conventional selective search-based region proposal with an RPN. The reason why Faster R-CNN was used over the much Faster You only look once (YOLO) model [47] is due to the inefficiency of YOLO while identifying small objects. We could not afford to lose any such information in our image data. The YOLO model also struggles in identifying objects with skewed aspect ratios, whereas this is not the case with the Faster R-CNN model. Moreover, the proposed task of feedback synthesis does not need to be real time. Finally, the Faster R-CNN model was chosen for the implementation.

Visual Attention: The VAN computes the visual context vector $g^*$, which is the final output of attention for state $s$ computed using Equation (5), where $g$ is the global feature vector, $c^*_i$ denotes the compatibility score for state $s$ and box $i$, whereas $b^*_i$ and $a^*_i$ denote the feature vector and attention vector for state $s$ and box $i$. The state $s$ denotes any particular hidden state of our neural network at any given step.

$$c^*_i = (b^*_i)^T . g;$$
$$a^*_i = \frac{\exp(c^*_i)}{\sum^n_j \exp(c^*_j)};$$
$$g^* = \sum^n_{i=1} (a^*_i . b^*_i)$$ (5)

3.3.3 Multimodal Fusion Block. Generation of multimodal context is achieved by combining textual context and visual context in a Multimodal Fusion block. The output of the text encoder (textual context vectors, $z^*$) and the output of the visual encoder (visual context vector, $g^*$) are given as inputs to the Multimodal Fusion block. The Multimodal Fusion block computes the multimodal context vectors as follows. The output of the text encoder is a time series of vectors. Each of these vectors is concatenated with $g^*$ and passed through a feedforward network that outputs a time series of vectors with the same dimension as the text encoder vectors, which is referred to as Multimodal context vectors ($y^*$). The attention module incorporates textual, visual, and multimodal
attention. The textual and visual attention were described earlier, whereas multimodal attention is as follows.

**Multimodal Attention:** Multimodal context vector $y^*$ is calculated in Equation (6) by concatenating $z^*$ and $g^*$ and passing through a feedforward layer. $z^*$ and $y^*$ are fed as input to each decoder block.

$$y^* = \text{concat}(z^*, g^*)^T W,$$

where $y^*$, $z^*$, and $g^*$ denote multimodal, textual, and visual context vectors, respectively; $W$ is the weight matrix, and $T$ denotes the transpose operation. Each vector of the textual context (encoder output) is fused with the visual context vector where the positional information remains as is.

### 3.3.4 Decoder

The decoder block takes multimodal context vectors $y^*$ and textual context vectors $z^*$ as input. During training, it takes the ground-truth comments as additional input against which the model is trained. During the training phase, the decoder stack generates the output vector, which is used to calculate loss against the vector corresponding to the ground-truth comments. The comments’ embeddings are produced, and positional encoding is added for them in a similar way to how the encoder encodes the text. During testing, the decoder synthesizes the feedback using $z^*$ and $y^*$. Various layers of the decoder block are listed as follows. Each attention layer has a residual connection around it, and a normalization layer follows it.

(a) **Self-attention layer:** Where query, keys, and values matrices are initialized in the decoder step. The self-attention layer is only allowed to attend to earlier positions in the output sequence. This is achieved by masking future positions.

(b) **Encoder-decoder attention layer:** Where the queries are initialized in the decoder step and $z^*$ is assigned to keys and values matrices.

(c) **Multimodal-attention layer:** Implemented to achieve multimodal attention where queries are initialized in the decoder step and $y^*$ is assigned as keys and values matrices.

(d) **Feedforward layer:** Which finally generates the feedback vector as the decoder’s output. The decoder stack is followed by a linear layer and softmax layer that maps the feedback vector into the vocabulary vector space.

### 3.3.5 Similarity Module

The similarity module consists of teacher-student models. The teacher model implements a pretrained Sentence Bidirectional Encoder Representations from Transformer (SBERT) [48] whereas the student model implements a fine-tuned RoBERTa [35]. The teacher model converts the comments to language-agnostic vectors and produce their embeddings, $emb_c$. Likewise, the student model generates the language-agnostic embeddings for the feedbacks, $emb_f$. The language-agnostic vectors produce similar embeddings for the sentences using different words but portraying similar meaning. The teacher model $M$ maps the comments $c$ to $emb_c$. The student model $\hat{M}$ maps the comments $c$ and feedbacks $f$ to $emb_c$ and $emb_f$, respectively.

The student model has been fine-tuned as per the following strategy. Overall, 80% of the data samples were used to train the proposed system. Among the other 20% of samples, 1,000 samples were randomly selected as the validation set; the leftover samples made the test set. Human evaluation was performed on both the validation set and the test set. Among the 1,000 samples of the validation set, the similarity score (of human evaluation) was more than 0.8 (on a scale of 0 to 1) for 386 samples. We used these 386 samples to fine-tune the similarity module’s student model. The test set was used for the model testing. The student model $\hat{M}$ was such that $\hat{M}(emb_{c_i}) \approx M(emb_{c_i})$ and $\hat{M}(emb_{f_i}) \approx M(emb_{f_i})$. The similarity score ($S_{score}$) between $emb_c$ and $emb_f$ is minimized as per Equation (7). Taking MSE between them helps compute the closeness of their embeddings and, hence the similarity between them. We have fine-tuned the similarity module.
using parallel comment-feedback pairs \(((c_1, f_1), \ldots, (c_n, f_n))\).

\[
S_{score} = \frac{1}{n} \sum_{i \in n} \left( (\hat{M}(\text{emb}_{c_i}) - M(\text{emb}_{c_i}))^2 + (\hat{M}(\text{emb}_{f_i}) - M(\text{emb}_{c_i}))^2 \right), \tag{7}
\]

where \(n\) is the number of feedback-comment pairs, whereas \(\text{emb}_c\) and \(\text{emb}_f\) denote comment embedding and feedback embedding. An English SBERT model \([48]\) has been used as the teacher model \(M\), whereas RoBERTa \([35]\) has been used as the student model \(\hat{M}\). The semantic similarities of \(\text{emb}_c\) and \(\text{emb}_f\) have been found as \((\text{emb}_c)^T \text{emb}_f\) using cosine similarity where \(\text{emb}_c\) and \(\text{emb}_f\) are the embeddings for comment and feedback, respectively, and \(T\) denotes the transpose operation.

To evaluate the feedbacks against ground-truth comments, Mean Reciprocal Rank (MRR) and Recall@\(k\) (also known as Recall Rate@\(k\); defined in Section 4.2) are utilized. The MRR and Recall@\(k\) are computed to denote whether the feedback synthesized is similar to the input comments, which are sorted based on the number of likes. The Recall@\(k\) is computed for the synthesized feedback, denoting whether it is similar to top \(k\)-ranked comments. The comments for an image are ranked based on their number of likes. Then, the semantic similarity between the comments’ embeddings and the embeddings of synthesized feedback is found using cosine similarity. Finally, Recall@\(k\) is computed for the synthesized feedback, denoting whether it is similar to top \(k\)-ranked comments, where \(k\) is a user-definable integer.

4 EXPERIMENTS

4.1 Training Strategy and Parameter Tuning

The baselines and proposed system described in the following sections have been trained for 30 epochs on a GTX 1080Ti GPU with 3584 NVIDIA CUDA Cores and 11 GB GDDR5X memory. The experiments have been performed using 5-fold cross-validation. Pretrained GloVe embeddings with 6 billion tokens and 100 dimensional vectors \([46]\) have been used in the embedding layer. The hyper-parameter values are described as follows.

- General parameters – batch-size: 32, learning rate: \(5 \times 10^{-4}\) for text encoder and 0.01 for visual feature extraction model, network optimizer: Adam, loss function: cross entropy loss, activation function: ReLU.
- Parameters for the transformer model – input size: 512, output size: variable (as transformer model can generate the output with dynamic length), encoder embedding dimensions: 100, decoder embedding dimensions: 100, encoder hidden units dimensions: 128, decoder hidden units dimensions: 128, encoder dropout: 0.5, decoder dropout: 0.5, encoder number of layers and attention heads: 6 and 8, decoder number of layers and attention heads: 6 and 8, metric: accuracy.
- Parameters for the Faster R-CNN model – number of epochs: 18, metric: mAP (mean Average Precision), number of proposals: 36, number of classes for anchor-boxes: 1601, network optimizer: adaDelta.

4.2 Evaluation Metrics

The synthesized feedbacks are evaluated using qualitative and quantitative metrics for their relevance with the ground-truth comments and inputs. We have incorporated two phases of automatic evaluation using five quantitative metrics (BLEU, CIDEr, ROUGE, SPICE, and METEOR) and two qualitative metrics (Recall@\(k\) and MRR) along with the human evaluation to evaluate the synthesized feedbacks holistically. For the quantitative evaluation of the feedbacks, the following metrics have been used in the first phase of automatic evaluation. These metrics can evaluate
machine-synthesized sentences such as summaries, image descriptions, and translations against benchmark results and human references based on recall, precision, and sensitivity.

- **BLEU Score** [44]: BLEU (bilingual evaluation understudy) is a \textit{precision}-based metric that compares candidate sentence with reference sentences to judge the quality of the candidate translation. In this work, the 4-gram BLEU score was used.

- **ROUGE** [34]: ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a \textit{recall}-based metric that analyzes automatic translation and summaries against a ground-truth reference set.

- **Meteor** [30]: METEOR (Metric for Evaluation of Translation with Explicit ORdering) considers \textit{precision} and \textit{recall}'s harmonic mean to analyze automatically generated output at the sentence level.

- **CIDEr** [56]: CIDEr (Consensus-based Image Description Evaluation) automatically evaluates machine translation and image caption outputs by considering the agreement of various reference descriptions.

- **SPICE** [2]: SPICE (Semantic Propositional Image Caption Evaluation) is a metric based on the \textit{sensitivity} of the n-grams for automated evaluation of caption and sentences.

The second phase of automatic evaluation uses Recall@k [50] and Mean Reciprocal Rank [10] to judge the relevance of the synthesized feedbacks against the ground-truth comments. These metrics are defined as follows.

- **Mean Reciprocal Rank (MRR)**: If the \textit{j}th feedback is most similar with the \textit{k}th most liked comment, then the rank and reciprocal ranks of the \textit{j}th feedback, that is, \( \text{rank}_j \) and \( \text{rrank}_j \) are calculated as per Equation (8).

  \[
  \text{rank}_j = k \quad \text{rrank}_j = 1/k, \tag{8}
  \]

  where \( k \) denotes the \textit{k}th comment sorted by the number of likes whereas \( \text{rank}_j \) and \( \text{rrank}_j \) are the Rank and Reciprocal rank of the \textit{j}th feedback, respectively.

  Mean Reciprocal Rank (MRR) denotes the average of the reciprocal ranks of all feedback samples considered in the study, given by Equation (9).

  \[
  \text{MRR} = \frac{1}{Q} \sum_{j=1}^{Q} \frac{1}{\text{rank}_j}, \tag{9}
  \]

  where \( Q \) is the number of feedback samples and \( \text{rank}_j \) denotes the rank of the \textit{j}th feedback.

- **Recall@k**: In general, Recall@k denotes whether a data sample matches with any of the top \( k \) relevant samples. It has been adapted to evaluate whether the synthesized feedback matches with top \( k \) relevant ground-truth comments. We have found the similarity score of the feedback and find the ranks for all of the comments for which rank denotes the rank of the corresponding comment among all of the comments when sorted by number of likes.

  As shown in Equation (10), if the comment whose feedback shows that the maximum similarity score is in the top \( k \) comments, it will get a score of 1 for Recall@k, else 0.

  \[
  \text{Recall@k} = 1 \text{ if } \text{rank}_q \in [1, \ldots, k], \tag{10}
  \]

  where \( q \) is the \textit{q}th feedback; \( k \) is the \textit{k}th comment sorted by number of likes, and \( \text{rank}_j \) denotes the rank of the \textit{j}th feedback.
Table 3. Ablation Studies on Using Textual and Visual Modalities, Attention, and Region Proposal

| Architecture | BLEU   | CIDEr  | ROUGE  | SPICE  | METEOR |
|--------------|--------|--------|--------|--------|--------|
| T Only       | 0.2137 | 0.1365 | 0.2667 | 0.1282 | 0.1192 |
| T + V        | 0.2073 | 0.1482 | 0.2486 | 0.1344 | 0.1132 |
| T + V + A    | 0.3046 | 0.1884 | 0.3515 | 0.1783 | 0.1422 |
| T + V + A + R| 0.3023 | 0.1945 | 0.3842 | 0.1792 | 0.1638 |

Here, T, V, A, and R denote Textual and Visual features, Attention, and Region extraction.

Table 4. Ablation Studies on Using Data Samples with Various Ranges of Comments

| Comments per image | BLEU   | CIDEr  | ROUGE  | SPICE  | METEOR |
|--------------------|--------|--------|--------|--------|--------|
| Complete Data      | 0.3023 | 0.1945 | 0.3842 | 0.1792 | 0.1638 |
| Low comments (up to 5) | 0.1734 | 0.1204 | 0.2263 | 0.1456 | 0.1345 |
| Mid comments (13–50) | 0.2992 | 0.2082 | 0.3717 | 0.1768 | 0.1559 |
| High comments (30+) | 0.2656 | 0.1737 | 0.3218 | 0.1362 | 0.1235 |

Moreover, to ensure that the synthesized feedbacks represent and are relevant to the given inputs, they have been evaluated by human readers alongside the objective measures. Manual evaluation has also been carried out by having 50 human evaluators read the synthesized feedbacks and evaluate them against input image, text, and comments for their similarity.

4.3 Ablation Studies

Ablation studies have been performed to analyze the effect of using visual information, attention mechanism, and region extraction. Table 3 summarizes these studies.

4.3.1 Effect of Using Visual Modality. The feedbacks are first synthesized considering only the text input. Then, the complementary visual information has been considered. The qualitative scores improved on including the visual modality along with the textual modality.

4.3.2 Effect of Using Attention. The model is first trained on textual attention only. Then, the visual attention is fed to fine-tune the output. It has been observed that the output with both textual and visual attention was more human-like than the one with textual attention only. It should be noted that the scores reduced a bit on including visual features along with textual, though they improved significantly on including the attention mechanism.

4.3.3 Effect of Using Region Proposal. Better performance has been observed on incorporating region extraction along with using textual and visual information with attention. In Section 4.3.1, global visual features using VGG and ResNet were included, whereas this section incorporates local features using Faster R-CNN–based region proposals. The aforementioned ablation studies on using visual modality, attention, and region proposal are summarized in Table 3.

4.3.4 Effect of Using Data Samples with Varying Ranges of Comments. The experiments are performed using complete data, data with low (up to 5), mid (between 13 and 50), and high (more than 30) number of comments per image. The split was done using the following thresholds – low comments, up to 5 comments per image or tweet; mid comments, between 13 and 50 comments per image; and high comments, more than 30 comments per image. The performance for various combinations is shown in Table 4. Experimenting with the complete dataset produced similar results as samples with 13 to 50 comments; however, it was computationally expensive.
4.3.5 Computational Time Analysis. Table 5 shows the time taken by various configurations to train the model for 1 epoch. Though including region extraction is computationally expensive, it resulted in the best performance (see Table 3). Using up to 13-50 comments per image (mid comments) resulted in similar quantitative performance scores as training with complete data (see Table 4). The final implementation was carried out using the region proposal along with visual and textual attention and the data samples with 13 to 50 comments per image.

4.4 Models

The architectures of the baselines and proposed system have been determined based on the ablation studies performed in Section 4.3. As observed earlier, the feedbacks synthesized considering textual and corresponding visual features had better scores. Hence, all of the baseline models include textual and visual encoders.

- **Baseline 1**: The first baseline uses Gated Recurrent Units (GRU) for the textual encoder, whereas residual network (ResNet) is used for the visual encoder network. The choice of GRU over Long Short-Term Memory (LSTM) architecture is guided through experimental observations. LSTM corresponded to 20% more memory consumption as compared with GRU, whereas their performance was comparable.

- **Baseline 2**: The second baseline retains ResNet for visual encoding. However, it replaces the textual encoder network with Bidirectional GRU (BiGRU). The intuition behind using BiGRU was to enable a particular word’s embedding to embody its contextual meeting, which changes according to the words appearing before and after it.

- **Baseline 3**: This model further replaces the textual encoder with a text transformer model, whereas the attention mechanism is incorporated along with the ResNet-based visual encoder. The choice of using a textual transformer is governed by its off-the-shelf performance in various language analysis problems, such as summarization, translation, caption generation. [55].

- **Proposed System**: The proposed system uses a textual transformer as textual encoder, whereas a Faster R-CNN–based region proposal mechanism is incorporated along with ResNet–extracted features for visual encoding. The incorporation of local features extracted by R-CNN resulted in a better performance as compared with using only the global features extracted by ResNet.

The proposed system’s implementation code and MMFeed dataset constructed in this article can be accessed at github.com/MIntelligence-Group/MMFeed.

5 RESULTS AND DISCUSSION

The feedbacks synthesized by the proposed system have been evaluated using the quantitative and qualitative measures described in Section 4.2 and compared with the baselines.
Table 6. Quantitative Evaluation of the Synthesized Feedbacks

| Model         | BLEU  | CIDEr | ROUGE | SPICE | METEOR |
|---------------|-------|-------|-------|-------|--------|
| Baseline 1    | 0.1942| 0.1342| 0.2524| 0.1025| 0.0924 |
| Baseline 2    | 0.2124| 0.1735| 0.2745| 0.1654| 0.1393 |
| Baseline 3    | 0.3096| 0.1835| 0.3374| 0.1554| 0.1412 |
| Proposed      | 0.3023| 0.1945| 0.3842| 0.1792| 0.1638 |

Table 7. Automatic Evaluation

| Model         | MRR   | R@1   | R@3   | R@5   | R@7   |
|---------------|-------|-------|-------|-------|-------|
| Baseline 1    | 0.2412| 24.36 | 70.56 | 91.52 | 93.63 |
| Baseline 2    | 0.2643| 25.23 | 71.42 | 93.53 | 96.32 |
| Baseline 3    | 0.2923| 26.42 | 79.97 | 93.76 | 95.92 |
| Proposed      | 0.3042| 29.33 | 84.56 | 98.32 | 98.67 |

Here, MRR and R@k denote Mean Reciprocal Rank and Recall@k.

5.1 Quantitative Results

In the first phase of automatic evaluation, the synthesized feedbacks are quantitatively evaluated using BLEU, CIDEr, ROUGE, SPICE, and METEOR metrics. The results are provided in Table 6, which indicate the increased informativeness for the feedbacks synthesized by the proposed method as compared with the baseline models.

5.2 Qualitative Results

The second automatic evaluation phase evaluates the synthesized feedbacks using MRR and Recall@k metrics. As seen in Table 7, 98.67% of the feedbacks are relevant to one of the top 7 ground-truth comments, whereas they show an MRR of 0.3042, denoting that a majority of the feedbacks portrayed the most similarity with the top 3 (corresponding to MRR of 0.33) or top 4 (corresponding to MRR of 0.25) comments. The sample results are shown in Figure 4.

The ground-truth comments and synthesized feedback portray similar sentiment and context using different words. To ensure that the synthesized feedbacks represent and are relevant to the given inputs, they have also been evaluated by the human readers alongside the objective measures.

Table 8 presents a comparative analysis between the human-generated comments and the feedbacks synthesized by the proposed system. Here, the human readers have evaluated the semantic similarity of the synthesized feedbacks as compared with input text and images and the ground-truth comments. The ground-truth comments show 72.86% and 74.90% similarity with the input text and images, respectively. These numbers are 65.34% and 67.96%, respectively, for the synthesized feedbacks. At the same time, the similarity between the feedbacks and comments has been found to be 80.17%. The “Score” described in Figure 4 is for the automatic evaluations of the synthesized feedbacks while “s” described in Table 8 corresponds to their human evaluation.

5.3 Discussion

The proposed system’s architecture has evolved through progressive experiments. The model is first trained on textual attention only. Then, visual attention is also fed to fine-tune the output. The output with both textual and visual attention has been observed to be more human-like than the one with textual attention only. The incorporation of the region proposal increased the computational cost 63X because of the large size of the bounding boxes. On the other hand, using
Fig. 4. Sample results. Here, “Score” denotes the highest of the similarity scores of the predicted feedback with all of the comments and “Rank” denotes the rank of the corresponding comment among all of the comments when sorted by number of likes.

the data samples with 13 to 50 comments per image for training resulted in similar performance as training with complete data, although it took 5.5X less computational time. Finally, a trade-off was made between training the proposed system with region proposals and using the data samples with 13 to 50 comments per image. Further, the baselines and proposed system converged in
Table 8. Human Evaluation

| Model   | $s_{et}$ | $s_{ci}$ | $s_{ft}$ | $s_{fi}$ | $s_{cf}$ |
|---------|----------|----------|----------|----------|----------|
| Baseline 1 | 47.90    | 41.12    | 43.40    | 36.07    | 45.12    |
| Baseline 2 | 49.33    | 45.93    | 46.33    | 41.67    | 58.22    |
| Baseline 3 | 64.11    | 64.33    | 58.78    | 59.67    | 71.44    |
| Proposed | 72.86    | 74.90    | 65.34    | 67.96    | 80.17    |

Here, $s_{et}$, $s_{ci}$, $s_{ft}$, $s_{fi}$, and $s_{cf}$ denote semantic similarities between comments and text, comments and images, feedbacks and text, feedbacks and images, and comments and feedbacks as judged by human readers.

terms of validation loss in 18 to 23 epochs. The models have been trained for 30 epochs as a safe upper bound. An important factor that has been successfully solved by the proposed method is the relevance aspect. We are feeding semantically the most similar comments (sorted by the number of likes) for the proposed model’s training. That has enabled the generated feedbacks to be more semantically similar to the multimodality input.

The synthesized feedbacks have been evaluated for relevance with the input images and text and ground-truth comments using two automatic and one manual evaluation phases. The improvements in the evaluation scores on incorporating the attention and region proposal indicate that the synthesized feedbacks represent the corresponding text and images. The automatic and human evaluation results affirm that the synthesized feedbacks are relevant to the input text and images. The synthesized feedbacks have been observed to learn the in-context information from the training data. For example, many training data samples contained information about politics and coronavirus. The context about these subjects was reflected in some of the synthesized feedback.

The proposed system pays attention to the modality-missing as well as modality-bias problems. Multimodal summarization and VisDial systems are prone to the modality-bias problem; they tend to consider one of the input modalities more than the others. However, in the context of the proposed multimodal feedback synthesis task, subjective human evaluation has been performed along with the automatic evaluation of the synthesized feedbacks to evaluate whether the output is relevant to both input modalities. The proposed task of multimodal feedback synthesis generated textual feedback for multimodality input containing text and images. The text is the primary modality and the image is the secondary modality here. Though the proposed system has been trained to ensure that both textual and visual contexts are considered while generating feedback, it is capable of generating feedback even when the image modality is not available. The ablation studies concerning feedback synthesis using only textual modality and using visual modality (with and without attention mechanism and region proposal) along with textual modality are described in Section 4.3.

One challenge with the evaluation of affective feedback generation is that multiple feedbacks can be contextually similar and may convey the same information. Thus, given the ground truth, the evaluation is difficult as our goal is to produce contextually similar feedback. The meaning of the term *human-like* is to be taken more in the sense that the model can generate contexts similar to how a human would generate them. Though the syntax and semantics of the feedbacks are not entirely correct, the proposed system can learn and generate the words with respect to the text and image inputs. The minor errors in the model’s understanding of the human language can be attributed to the noise (special characters, sentence phrases, and multilingual symbols) present in the comments of the training data. The proposed feedback synthesis system has been trained on human-generated comments. It can generate feedbacks that would be nondifferentiable from actual human comments. The potential negative impacts, including fake feedback generation,
can be dealt with by creating and implementing fair usage policies during the distribution of the proposed system. Every automated generative system carries a small possibility of being misused. However, the research community should not allow this to prevent them from building new systems to handle novel problems. As the literature survey suggests, the researchers have created and implemented fair usage policies while distributing these systems [5, 19, 57].

6 CONCLUSIONS AND FUTURE WORK
In this work, we have introduced a novel task to generate human-like feedback for text and image data and proposed an affective feedback synthesis system. We have also constructed a large-scale multimodal feedback synthesis dataset. Automatic and human evaluations have been carried out to evaluate the synthesized feedbacks’ relevance with the input text and images and similarity with the comments. In the future, we aim to improve the syntactic and semantic correctness while generating long feedback sentences and extend the dataset to include more than one image per news article. We also plan to work on the news articles’ sentiment classification and genre classification. For the automatic evaluation of the synthesized feedbacks, existing evaluation metrics such as SPICE, CIDEr, and ROUGE have been used, which are broadly used for various types of machine-synthesized sentences such as automatic translations, summaries, and image captions. There is a need to design an automatic evaluation metric specifically for evaluating multimodal feedbacks that would aid or replace the human evaluation process. We will focus on that as well in our future work.
APPENDIX

A DATA CRAWLING AND PREPROCESSING PROCEDURE

Algorithm 1 shows the procedure to crawl and preprocess the data instances during the construction of the MMFeed dataset.

**Algorithm 1:** MMFeed Dataset Crawling and Pre-processing Procedure

```
Define user_name: Twitter handle of the news channel
Define tweet_id: A unique numeric ID of particular Tweet
Define tweet: Tweet contents
Define URL: URL link of the original news article
Define text: Text fetched from the original news article
Define image: Image fetched from the original news article
Define replies: Comments fetched from Tweet contents
Define replies_iter: Iteration, denoting the count of replies

procedure crawl(curr_url):
    article = Article(curr_url)
    article.download()
    article.parse()
    for full_tweets in tweepy.Cursor():
        tweet = full_tweets.full_text
        replies_iter = tweepy.Cursor()
        While True
            reply = replies_iter.next()
            replies.append(reply.full_text)
            favorite.append(reply._json['favorite_count'])
        end
    end
    news_text = crawl(curr_url)
    with open (str(iter) + 'twitter_data.csv', 'w', encoding='utf-8') as csv_file:
        writer = csv.DictWriter(csv_file, fieldnames=['Tweet', 'Comment', 'Likes'])
        writer.writerow('Tweet': tweet, 'Comment': ':'.join(replies), 'Likes': ':'.join(favorite))

procedure normalised_text(text):
    Parse using 'BeautifulSoup' library and remove HTML tags
    Expand contractions (text)

with open (str(iter) + 'twitter_data.csv', 'w', encoding='utf-8') as csv_file:
    writer = csv.DictWriter(csv_file, fieldnames=['Tweet', 'Comment', 'Likes'])
    writer.writerow('Tweet': tweet, 'Comment': ':'.join(replies), 'Likes': ':'.join(favorite))
```

Data pre-processing

```
Data = read_csv(twitter_data.csv)
title = data['Tweet'][0]
text = article.text
image = urllib.request.urlopen(article.top_image)
replies = data['Comment'][0]
likes = data['Likes'][0]

with open (str(iter) + 'mmfeed_data.csv', 'w', encoding='utf-8') as csv_file:
    writer = csv.DictWriter(csv_file, fieldnames=['Title', 'Text', 'Image', 'Comment', 'Likes'])
    writer.writerow('Title': title, 'Text': text, 'Image': image, 'Comment': replies, 'Likes': favorite)
```

ACM Trans. Multimedia Comput. Commun. Appl., Vol. 19, No. 6, Article 190. Publication date: May 2023.
The MMFeed dataset along with the proposed system’s implementation code can be accessed through this link: github.com/MIntelligence-Group/MMFeed.

REFERENCES

[1] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K. Marks, Chiori Hori, Peter Anderson, et al. 2019. Audio visual scene aware dialog. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’19). 7558–7567.

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic propositional image caption evaluation. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14 , Springer. 382–398.

[3] Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In Conference on Empirical Methods in Natural Language Processing (EMNLP’20). 9347–9359.

[4] Paulo Blikstein and Marcelo Worsley. 2016. Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. Journal of Learning Analytics 3, 2 (2016), 220–238.

[5] Christos Bouras, Anestis Filopoulos, Vasileios Kokkinis, Sotiris Michalopoulos, Dimitris Papadopoulos, and Georgia Tseliou. 2014. Policy recommendations for public administrators on free and open source software usage. Telematics and Informatics 31, 2 (2014), 237–252.

[6] Feilong Chen, Fandong Meng, Jiajing Xu, Peng Li, Bo Xu, and Jie Zhou. 2020. DMRM: A dual channel multi hop reasoning model for visual dialog. In 34th AAAI Conference on Artificial Intelligence (AAAI), Vol. 34. 7504–7511.

[7] Jingqiang Chen and Hai Zhuge. 2018. Abstractive text-image summarization using multimodal attention hierarchical RNN. In Conference on Empirical Methods in Natural Language Processing (EMNLP’18). 4046–4056.

[8] Kuan-Yu Chen, Shih-Hung Liu, Berlin Chen, Hsin-Min Wang, Ea-Ee Jan, Wen-Lian Hsu, and Hsin-Hsi Chen. 2015. Extractive broadcast news summarization leveraging recurrent neural network language modeling techniques. IEEE/ACM Transactions on Audio, Speech, and Language Processing 23, 8 (2015), 1322–1334.

[9] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In International Conference on Machine Learning (ICML’21). 1931–1942.

[10] Nick Craswell. 2009. Mean reciprocal rank. Encyclopedia of Database Systems 1703, Springer. https://www.bibsonomy.org/bibtex/2e8204c7513a94917d206b34004d87a54/dblp.

[11] Benjamin Dornel. 2020. New York Times Articles & Comments Dataset. Retrieved February 20, 2022 from https://www.kaggle.com/benjaminawd/new-york-times-articles-comments-2020. (2020).

[12] Wafaa S. El-Kassas, Cherif R. Salama, Ahmed A. Rafea, and Hoda K. Mohamed. 2021. Automatic text summarization: A comprehensive survey. Expert Systems with Applications 165 (2021), 113679.

[13] Fabio R. Gallo, Gerardo I. Simari, Maria Vanina Martinez, and Marcelo A. Falappa. 2020. Predicting user reactions to Twitter feed content based on personality type and social cues. Future Generation Computer Systems 110 (2020), 918–930.

[14] Shen Gao, Xiuying Chen, Zhaochun Ren, Dongyan Zhao, and Rui Yan. 2020. From standard summarization to new tasks and beyond: Summarization with manifold information. In 29th International Joint Conference on Artificial Intelligence (IJCAI’20).

[15] Ross Girshick. 2015. Fast R-CNN. In IEEE/CVF International Conference on Computer Vision (ICCV’15). 1440–1448.

[16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’14). 580–587.

[17] Xiaodong Gu, Kyunghyun Cho, Jung-Woo Ha, and Sunghun Kim. 2019. DialogWAE: Multimodal response generation with conditional Wasserstein auto encoder. In International Conference on Learning Representations (ICLR’19).

[18] Xiaodong He and Li Deng. 2017. Deep learning for image-to-text generation: A technical overview. IEEE Signal Processing Magazine 34, 6 (2017), 109–116.

[19] Matthew Hoekstra, Reshma Lal, Pradeep Pappachan, Vinay Phegade, and Juan Del Cuvillo. 2013. Using innovative instructions to create trustworthy software systems. Hardware and Architectural Support for Security and Privacy, ISCA 11, 10.1145 (2013), 2487726–2488370.

[20] Chiori Hori, Huda Alamri, Jue Wang, Gordon Wichern, Takaaki Hori, Anoop Cherian, Tim K. Marks, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, et al. 2019. End-to-end audio visual scene aware dialog using multimodal attention-based video features. In 44th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP’19). 2352–2356.

[21] Anthony Hu and Seth Flaxman. 2018. Multimodal sentiment analysis to explore the structure of emotions. In ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD’18). 350–358.
[46] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP’14)*. 1532–1543.

[47] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real time object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’16)*. 779–788.

[48] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP’19)*. 3982–3992.

[49] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28* (2015), 91–99.

[50] Per Runeson, Magnus Alexandersson, and Oskar Nyholm. 2007. Detection of duplicate defect reports using natural language processing. In *29th IEEE International Conference on Software Engineering (ICSE’07)*. 499–510.

[51] Alexander M. Rush, SEAS Harvard, Sumit Chopra, and Jason Weston. 2017. A neural attention model for sentence summarization. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1756–1765.

[52] Ian Simon, Noah Snavely, and Steven M. Seitz. 2007. Scene summarization for online image collections. In *11th IEEE/CVF International Conference on Computer Vision (ICCV’07)*. IEEE, 1–8.

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[54] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2020. VMSMO: Learning to generate multimodal summary for video-based news articles. In *Conference on Empirical Methods in Natural Language Processing (EMNLP’20)*. Association for Computational Linguistics, 3981–3991.

[55] Xinnuo Xu, Ondřej Dušek, Ioannis Konstas, and Verena Rieser. 2018. Better conversations by modeling, filtering, and optimizing for coherence and diversity. In *Conference on Empirical Methods in Natural Language Processing (EMNLP’18)*. Association for Computational Linguistics, 3981–3991.

[56] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 654–664.

[57] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *32nd AAAI Conference on Artificial Intelligence (AAAI’18)*.