2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR) | 978-1-6654-9548-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/MIPR54900.2022.00081

Automatic Evaluation of Machine Generated Feedback For Text and Image Data

Pratham Goyal*[1], Anjali Raj*[2], Puneet Kumar[1], and Kishore Babu Nampalle[1]

[1]Indian Institute of Technology, Roorkee, India, 247667
[2]Indian Institute of Technology, Kharagpur, India, 721302
pratham_m@ee.iitr.ac.in, anjali.raj@kgpian.iitkgp.ac.in, {pkumar99, kbabu89}@cs.iitr.ac.in

Abstract

In this paper, a novel system, 'AutoEvalNet,' has been developed for evaluating machine-generated feedback in response to multimodal input containing text and images. A new metric, 'Automatically Evaluated Relevance Score' (AER Score), has also been defined to automatically compute the similarity between human-generated comments and machine-generated feedback. The AutoEvalNet's architecture comprises a pre-trained feedback synthesis model and the proposed feedback evaluation model. It uses an ensemble of Bidirectional Encoder Representations from Transformers (BERT) and Global Vectors for Word Representation (GloVe) models to generate the embeddings of the ground-truth comment and machine-synthesized feedback using which the similarity score is calculated. The experiments have been performed on the MMFeed dataset. The generated feedback has been evaluated automatically using the AER score and manually by having the human users evaluate the feedback for relevance to the input and ground-truth comments. The values of the AER score and human evaluation scores are in line, affirming the AER score's applicability as an automatic evaluation measure for machine-generated text instead of human evaluation.

Index Terms *– Multimodal Feedback Analysis, Automatic Evaluation, Similarity Score, Affective Computing.*

1 Introduction

The widespread increase in the use of social media platforms by every one of us reflects a large amount of data, increasing tremendously day by day. It becomes difficult for a person to keep track of every important data with tons of information. With various advancements in technology, natural language processing (NLP) has emerged as one of the powerful tools for human-computer interaction. Notable progress has been achieved in the fields of caption generation and document summarization from large-scale data, as could be from [3]. Twitter is one of those social media platforms trending in terms of the vast amount of information it provides. Thus, it becomes imperative to develop some technology that could effectively assist us in filtering the posts according to our interests. Consequently, we need to be confident with the feed's relevance. And here comes the picture of the extent of reliability and accountability of our feedback generated. Extensive efforts have been made towards developing a model for correct reliability on the similarity of responses.

Automatic feedback evaluation has been very important and gaining popularity among the masses because it extracts the overall contextual information from large unstructured data and uses it to learn the model. However, significantly fewer developments could be found in actual evaluation based on context and similarity. Various techniques have evolved around the automatic generation of captions from visual data, particularly images [18] or summaries of large-scale documents or the creation of a new story from an existing one. Hence, it becomes important to develop a measure for the reliability of the feedback generated. Consequently, we developed a versatile form of the network called the AutoEvalNet, which uses the proposed metric, 'Automatically Evaluated Relevance Score' (AER Score).'

The AER score could efficiently depict the similarity based on human-like parameters such as the number of likes for a particular comment. Its architecture (shown in Fig. 1) contains a pre-trained feedback synthesis model and the proposed feedback evaluation model. It uses an ensemble of Bidirectional Encoder Representations from Transformers (BERT) and Global Vectors for Word Representation (GloVe) models to generate the embeddings of the ground-truth comment and machine-synthesized feedback using which the similarity score is

*Denotes Equal Contribution

2770-4319/22/$31.00 ©2022 IEEE
DOI 10.1109/MIPR54900.2022.00081

calculated. The proposed automatic evaluation metric, AER score results in an average score of 0.313 whereas the average score for the human evaluations has come out to be 0.451. The proposed system's code is available at github.com/MIntelligence-Group/AutoEvalRelScore.

Our major contributions are two-fold:

- A novel system, 'AutoEvalNet,' has been proposed to evaluate machine-generated feedback towards the multimodal data containing text and images.
- We define a new metric, 'Automatically Evaluated Relevance Score' (AER Score),' that reports the similarity between human-generated comments and machine-generated feedback as effectively as human evaluation.

2 Related Work

This Section surveys the research advances in machine-generated text and its evaluation. Among the existing research tasks, Visual Dialog, Multimodal Summarization, and Image Captioning are similar to this paper's Multimodal feedback Generation task. The *Visual Dialog* aims to infer the semantic dependencies between the underlying dialogues and generates a response to a question considering the corresponding image and previous dialog history [8, 20, 2]. In the context of *Multimodal Summarization*, Zhu et al. [21] demonstrated the summarization of multimodal data using visual attention. Still, they face the challenge of non-alignment between the text and the image pairs, leading to the difference in the contexts reported by the duo. In the context of *Image Captioning*, Lei Ke et al. [10] have proposed a model which could perceive the relative positioning of words in the sentence to come up with captions for complex cases. In another work, You et al. [19] use top-down and bottom-up approach trade-offs to form semantic attention imbibed within the feedback system. Most of the above methods generated a textual response toward multimodal data; however, they did not use actual human responses to train their models.

In the context of evaluating the machine-generated content, Badry et al. [1] proposed a content evaluator for comparing the semantic similarity between the original document and its summary. In another work, Lee et al. [12] used the vision and language BERT model, which computes the similarity score between the reference and generates texts. However, the model has been evaluated on only certain existing metrics, which does not consider the robustness of the model. For evaluating the naturalness of the ss of the speech content produced by text-to-speech systems, Jaiswal et al. [7] implemented a generative adversarial network. In another work, Len et al. [6] evaluated the tutors' feedback using the feedback instruments. Various BERT and transformer-based models have also been used for calculating semantic similarity between two sentences [17]. Studies and research from [9] reveal that automatic summarization and its evaluation have a long way to cover and need to be made more robust with the various factors and noises to adapt to various use cases. Feedback synthesis is a new problem, and its evaluation requires manual intervention. There is a need for automatic measures to evaluate the relevance of machine-generated feedback concerning the ground-truth comments. With that inspiration, we have proposed a novel feedback evaluation system, 'AutoEvalNet,' and defined the 'AER Score.'

3 Proposed Methodology

The architecture of the proposed system, 'AutoEvalNet,' has been depicted in Fig. 1. It majorly contains the Feedback Synthesis Model and the Feedback Evaluation Model.

3.1 Feedback Synthesis Model

For feedback synthesis, the pre-trained model proposed by Kumar et al. [11] has been used. It consists of the visual and textual encoder-decoder blocks, which incorporate self-attention, followed by the encoder-decoder attention blocks. The outputs of textual and visual encoders are the textual context vectors, z^* and visual context vector, g^*, using which the multimodal context vector y^* is calculated as per Eq. 1 by concatenating z^* & g^* and passing through a feed-forward layer. Further, z^* and y^* are fed as input to each decoder block using which the feedback is generated.

$$y^* = concat(z^*, g^*)^T.W \tag{1}$$

Where z^*, g^*, and y^* denote textual, visual, and multimodal context vectors; T is the transpose operation, and W denotes the weight matrix.

3.2 Automatic Evaluation Model

The proposed feedback evaluation model converts the combination of feedback and comment to a single vector representation, called the embedding vector. This embedding vector generates a similarity score between the feedback and comment. A combination of two methods is used to generate the embedding vector. One involves using BERT [5] model, and the other uses GloVe [13] representations. The feedback and comment are concatenated in the BERT path using special $[CLS]$ and $[SEP]$ tokens at the beginning of the first sentence and the end of the last sentence, respectively. They are passed through 12 encoders to generate a 768-dimensional embedding vector which is passed through a linear layer which gives a 100-dimensional vector as output. In the other path, the average of GloVe embeddings of all the comments and feedback words is used,

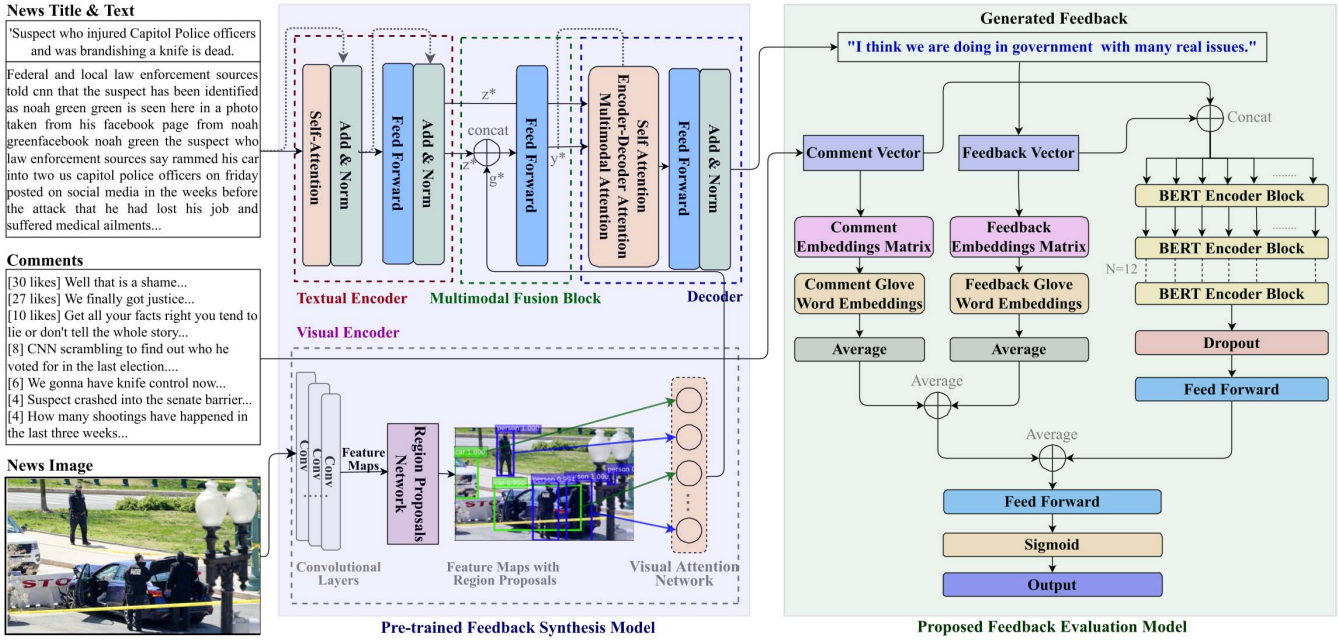

Figure 1. Architecture of the proposed system. Here, N denotes the number of encoder blocks. The feedback evaluation model's output is mapped to the similarity score between comment & feedback.

giving a single 100-dimensional vector for both comment and feedback. The average of the these two results in the combined vector representation of the comment and feedback. The vectors from both the paths are averaged and passed through a linear layer followed by sigmoid activation to generate a similarity score between 0 and 1 as depicted in Eq. 2.

$$\begin{aligned} sim(f,c) &= sigmoid(I) \\ I &= average(GE, BE)^T.W' \\ GE &= (GE\_c + GE\_f)/2 \end{aligned} \tag{2}$$

Where I is the intermediate vector, T is the transpose operation, and W' denotes the weight matrix; GE_c, GE_f are the GloVe embeddings for comment and feedback whereas BE denotes the BERT embedding combinely for the comment and feedback.

Definition 1: (AER Score)
We propose a new metric, 'Automatically Evaluated Relevance Score' (AER Score), to evaluate the feedback generated by the feedback synthesis model.

Computation of AER Score

(i) Firstly, the feedback f is generated by the feedback synthesis model based on the image and text data.

(ii) The ground-truth comments are sorted based on their number of likes.

(iii) Then, the feedback evaluation model computes the similarity score between the feedback and each comment using Eq. 3. This score indicates the extent of semantic similarity between the machine-generated feedback and the human-generated comments.

$$s_i = sim(f, c_i) \tag{3}$$

Where s_i denotes the similarity between the feedback f and i^{th} comment, c_i.

(iv) Further, as shown in Eq. 4, the AER score is computed as the weighted average of the similarity scores, with the number of likes for the corresponding comment as weights. It indicates the 'humanness' of the feedback synthesis model.

$$AER\ Score = \frac{s_1 \times l_1 + s_2 \times l_2 + ... + s_n \times l_n}{l_1 + l_2 + ... + l_n} \tag{4}$$

Where n is the number of comments, s_i is the similarity between comment c_i and feedback f_i whereas l_i denotes the number of likes for comment c_i.

Intuition behind AER Score: We assume that the most liked comment is the most relevant to the input image and text. Hence, taking weights as the number of likes is justified by this assumption as the scores generated with more relevant comments will have a greater contribution to the overall score. Moreover, this approach is not biased toward any comment as it considers all the comments' contributions.

4 Experiments & Results

4.1 Experimental Setup

The model training of 'AutoEvalNet' is carried out using an Nvidia P5000 card having 16 GB GPU memory, whereas an Intel i7 Linux OS system with 16GB RAM and 3.7GHz CPU has been used for model testing.

4.2 Datasets

The pre-trained feedback synthesis model has been trained on MMFeed dataset [11] containing 77,790 samples with images, text, comments, and the number of likes for each comment. The dataset has been compiled by crawling 9,479 Tweets corresponding to the news articles using Tweepy API[1] and NLTK[2] libraries.

For the automatic evaluation of generated feedback, we propose the feedback evaluation model has been trained on the Quora Question Pairs dataset [15]. The dataset comprises 4,00,000 pairs of questions and the similarity score, which denotes whether they have the same meaning.

4.3 Evaluation

The generated feedbacks' relevance is evaluated against the ground-truth comments using the following strategies.

4.3.1 Rank based Evaluation

The generated feedbacks are evaluated using 'Mean Reciprocal Rank' [4] and 'Recall@k' [14]. The aforementioned metrics have been defined as follows.

- **Mean Reciprocal Rank (MRR)**: If the p^{th} feedback is most similar with the k^{th} most liked comment, then the rank and reciprocal ranks of the p^{th} feedback, i.e., $rank_p$ and $rrank_p$ are calculated as per Eq. 5.

$$rank_p = k \qquad rrank_p = 1/k \tag{5}$$

  Mean Reciprocal Rank (MRR) is computed as per Eq. 6 and denotes the average of the reciprocal ranks of all feedback samples.

$$MRR = (\frac{1}{M}) \sum_{q=1}^{M} \frac{1}{rank_q} \tag{6}$$

  Where M is the number of feedback samples and $rank_q$ denotes the rank of the q^{th} feedback.

- **Recall@k**: Recall@k denotes whether a data sample matches with any of the top k relevant samples. As shown in Eq. 7, if the comment with which feedback shows maximum similarity score is in the top k comments, it will get a score of 1 for Recall@k, else 0.

$$Recall@k = 1 \text{ if } rank_r \in [1, \ldots, k] \tag{7}$$

  Where k and r denote the k^{th} comment sorted by number of likes and the r^{th} feedback.

4.3.2 Automatic Evaluation

The generated feedbacks have been evaluated automatically using the AER score proposed in Section 3.2.

4.3.3 Human Evaluation

The manual evaluation of the feedbacks has been performed by having 20 human users evaluate the feedback for relevance to the input and ground-truth comments.

4.4 Ablation Studies

Following ablation studies have been conducted to determine the feedback evaluation model's architecture.

4.4.1 Effect of using BERT Embeddings

BERT was advantageous over other embedding models because it generates word embeddings that depend on the surrounding words instead of other embedding models that generate fixed representations of words that don't depend on the context.

4.4.2 Effect of using LSTM/GRU Embeddings

BERT is known to outperform the Seq2Seq models such as the long short term memory (LSTM) and Gated Recurrent Units (GRU) in generating the contextual embeddings of the sentences [16]. In our experiments also, BERT was observed to perform better than them, where GRU performed better than LSTM.

4.4.3 Effect of using GloVe Embedding

GloVe provided static embeddings for each word irrespective of the context within which the word appears. We require the embeddings for each sentence. Hence, using GloVe embedding alone is not suitable.

4.4.4 Ensembling of BERT and GloVe

An ensemble of BERT and LSTM performed better than the embeddings mentioned above alone or in combination. Hence, the final architecture of the feedback evaluation model uses an ensemble of BERT and GloVe to generate the embedding vectors of the comments and feedbacks.

[1] https://docs.tweepy.org/en/stable/
[2] https://nltk.org/

4.5 Models

Feedback Synthesis Model: The pre-trained feedback synthesis model described in Section 3.1 has been used.

Automatic Evaluation Models: The following baselines and proposed model have been evolved based on the ablation studies performed in Section 4.4.

1. **Baseline 1** – Single BERT: Comments and feedback are concatenated and passed through a single BERT head to generate a combined vector representation of the comment and feedback. The vector is passed through a feedforward layer to generate a similarity score.
2. **Baseline 2** – Two BERTS: Embedding of comments and feedback are generated by passing them individually through two BERT heads and then averaging to generate a combined vector which is then passed through a linear layer to generate similarity score.
3. **Baseline 3** – Ensemble (GRU + BERT): Embedding of comment and feedback is generated through two paths using Seq2Seq GRU and BERT models. The average of both the paths' representations is used as final embedding from which the similarity score is calculated.
4. **Proposed Eval Model** – Ensemble (BERT + Glove): Embedding of comments and feedback is generated using GloVe and BERT models. The average of the representations from both is used as the final embedding from which the similarity score is calculated.

4.6 Results and Discussion

This Section presents the results computed as the the strategies discussed in Section 4.3. Table 1 presents the MRR and Recall@k scores.

Table 1. Rank evaluation. 'MRR' & 'R@k' denote 'Mean Reciprocal Rank' & 'Recall@k.'

Model	MRR	R@1	R@3	R@5
AutoEvalNet	**0.6182**	**37.93**	**82.76**	**93.10**

The mean MRR of 0.6182 indicates that most of the feedbacks show similarity with the top 2 comments, whereas 93.10 of the feedbacks are relevant to one of the top 5 ground-truth comments. As the comments and feedback may represent similar contexts using different words, their similarity and relevance to the inputs have been evaluated using the proposed automatic evaluation metric, AER score and manually. Table 2 shows the 'AER Score' computed automatically using the proposed system and the 'Human Score' computed through manual evaluations.

Table 2. Automatic & manual evaluation using 'AER Score' & 'Human Score', respectively.

Model	AER Score	Human Score
Baseline 1	**0.048**	**0.451**
Baseline 2	**0.307**	
Baseline 3	**0.008**	
AutoEvalNet	**0.313**	

Fig. 2 compares the 'AER Score' and 'Human Score' for randomly picked 20 samples along with showing the differences between the two scores. The sample results have been depicted in Fig. 3.

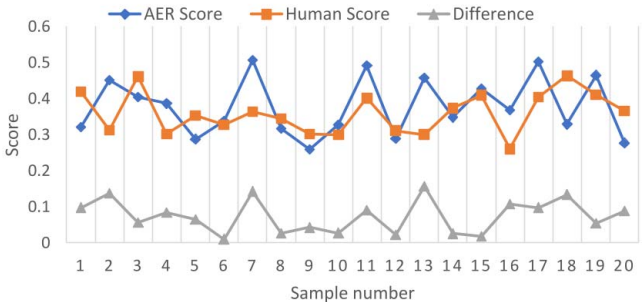

Figure 2. 'AER Score' vs. 'Human Score'.

Discussion: As depicted in Fig. 2, the AER scores and human evaluation scores are in-line, and the difference in their values is very small. The little variations among the scores are because of human subjectivity as different users have different perceptions while evaluating the feedbacks. The proposed system considers the importance of the ground-truth comments while evaluating the generated feedbacks by taking the number of likes for the comments into account while computing the AER score.

5 Conclusion

The proposed system, 'AutoEvalNet,' automatically computes the AER score denoting the similarity between ground-truth comments and machine-generated feedback. It does not require human intervention, thus saving time and manual efforts to evaluate the feedback. The experiments performed for the MMFeed dataset show that the automatically computed scores are in-line with the manually evaluated scores. In the future, we will extend the proposed system to report the relevance of the generated feedback with the input image and text along with the ground-truth comments. It is also planned to work on improving the syntax and semantics of the generated feedbacks.

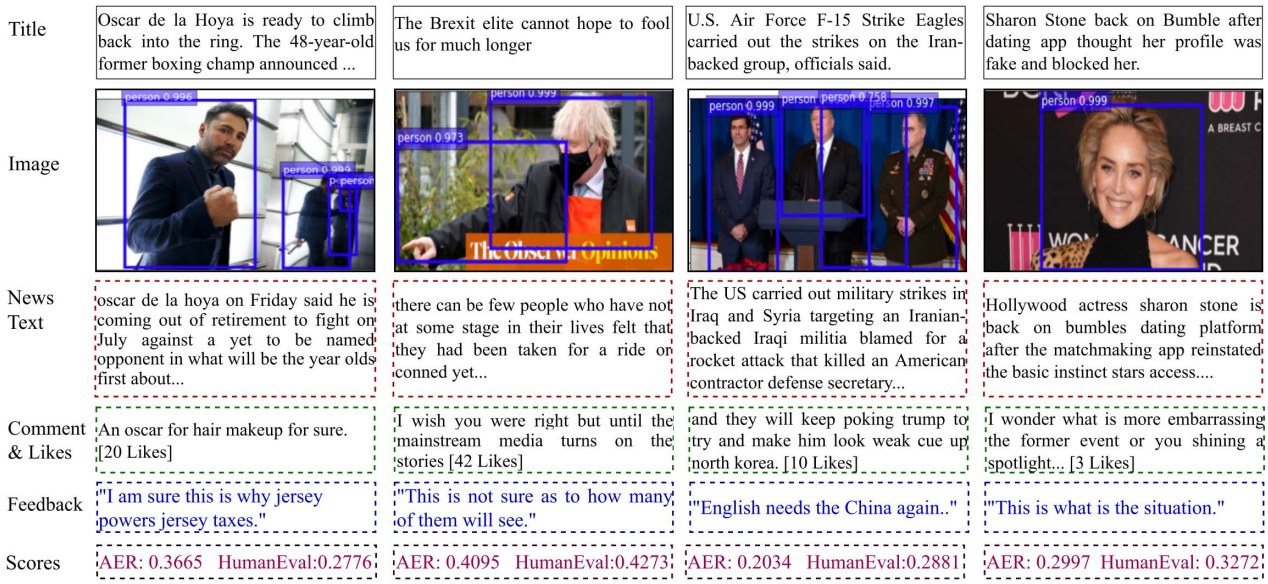

Figure 3. Sample Results. Here, 'AER Score' denotes the automatically computed similarity score using the proposed model and 'HumanEval Score' is the score computed through human evaluation.

Acknowledgement

This work has been performed at the Machine Intelligence Lab, Indian Institute of Technology Roorkee, India.

References

[1] R. M. Badry et al. Text Summarization Within the Latent Semantic Analysis Framework: Comparative Study. *International Journal of Computer Applications*, 81:40–45, 2013.

[2] F. Chen et al. DMRM: A Dual Channel Multi Hop Reasoning Model for Visual Dialog. In *The 34th AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[3] X.-W. Chen and X. Lin. Big Data Deep Learning: Challenges and Perspectives. *IEEE access*, 2:514–525, 2014.

[4] N. Craswell. Mean Reciprocal Rank. *Encyclopedia of Database Systems*, 1703, 2009.

[5] J. Devlin, M.-W. Chang, K. Lee, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv Preprint ArXiv:1810.04805*, 2018.

[6] L. Hand and M. Rowe. Evaluation of Student Feedback. *Accounting Education*, 10(2):147–160, 2001.

[7] J. Jaiswal et al. A Generative Adversarial Network based Ensemble Technique for Automatic Evaluation of Machine Synthesized Speech. In *Asian Conference on Pattern Recognition (ACPR)*, pages 580–593. Springer, 2019.

[8] G. Kang, J. Lim, and B.-T. Zhang. Dual Attention Networks for Visual Reference Resolution in Visual Dialog. In *The Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2024–2033, 2019.

[9] W. S. E. Kassas et al. Automatic Text Summarization: A Comprehensive Survey. *Expert Systems with Applications*, 165:113679, 2021.

[10] L. Ke, W. Pei, R. Li, X. Shen, et al. Reflective Decoding Network for Image Captioning. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[11] P. Kumar, G. Bhat, O. Ingle, D. Goyal, and B. Raman. Affective Feedback Synthesis Towards Multimodal Text and Image Data. *ArXiv Preprint ArXiv:2203.12692*, 2022.

[12] H. Lee et al. ViLBERTScore: Evaluating Image Caption Using Vision and Llanguage BERT. In *The First Workshop on Evaluation and Comparison of NLP Systems*, Nov. 2020.

[13] J. Pennington, R. Socher, and C. D. Manning. GloVe: Global Vectors for Word Representation. In *The conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[14] P. Runeson, M. Alexandersson, and O. Nyholm. Detection of Duplicate Defect Reports using Natural Language Processing. In *The 29th IEEE International Conference on Software Engineering (ICSE)*, pages 499–510, 2007.

[15] L. Sharma, L. Graesser, N. Nangia, and U. Evci. Natural Language Understanding with the Quora Question Pairs Dataset. *ArXiv Preprint ArXiv:1907.01041*, 2019.

[16] Y. Takahashi. NLP in the Financial Market — Sentiment Analysis. https://lethain.com/genetic-algorithms-cool-name-damn-simple/, 2020. Accessed on 2022-05-15.

[17] A. Thakur. BERT Architectures for Semantic Similarity. youtu.be/D-BlhDFXt30, 2020. Accessed on 2022-05-15.

[18] O. Vinyals et al. Show and Tell: A Neural Image Caption Generator, 2014.

[19] Q. You et al. Image Captioning with Semantic Attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[20] Z. Zheng et al. Reasoning visual dialogs with structural and partial observations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[21] J. Zhu et al. MSMO: Multimodal Summarization with Multimodal Output. In *The Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.