This paper introduces a dynamic features-based method for image retrieval and sentiment polarity analysis within the context of digital media. The authors aim to address the challenges posed by the vast amount of digital images and the complexity of sentiment extraction. They propose an innovative multi-modal approach that integrates image captioning with visual features, allowing for more nuanced semantic and emotional analysis of the content. The method leverages deep learning techniques to extract dynamic features from both image and text, providing an effective way of improving both retrieval accuracy and emotion detection. The experimental results presented in the paper demonstrate the method's impressive performance, achieving high accuracy scores and maintaining operational efficiency. However I have few concens and you are advised to revise your paper.

1. The proposed method of combining image captioning with dynamic features is a novel approach. It would be helpful to provide more detailed comparisons with traditional methods that use only visual features or textual descriptions.
2. The accuracy values mentioned (0.951 at 1, 0.985 at 5, and 0.989 at 10) are impressive. Could the authors explain how these values compare with the current state-of-the-art in digital image retrieval and sentiment analysis?
3. The description of the image captioning model (SEIC) is well-detailed. It would be beneficial to include an example of the actual captions generated for a few images to better illustrate its effectiveness.
4. Dynamic feature extraction is mentioned as a key part of the model. Can the authors clarify how dynamic temporal changes in images are captured and whether this has an impact on image retrieval accuracy?
5. The captioning module relies heavily on the Transformer architecture. Could the authors compare its performance with other models such as BERT or GPT for better context?
6. The multi-modal feature fusion mechanism is interesting, but the paper does not provide detailed analysis of how the different weight adjustments are made dynamically. A brief discussion on this process would add value.
7. The paper would benefit from more in-depth analysis on how the method scales with larger datasets. Do the results hold when tested on datasets significantly larger than CIRCO?
8. In the results section, the authors mention using Accuracy at K for evaluation. It might be beneficial to provide additional metrics such as Precision, Recall, and F1-score to give a more complete assessment of the method’s performance.
9. In Figure 1 and 2, the architecture and formula decomposition are presented well. However, the exact role of each component (like weight matrices and their significance) is not entirely clear. A more detailed explanation could help readers understand their impact on the model.
10. The method’s real-time processing efficiency (112.4 ms) is impressive. However, how does this performance compare with similar models under high-load conditions or when applied to real-time video analysis?
11. In Section 5.3, the authors compare their method with others such as MD-SAN, DFT, and OSCARB. Could the authors include a discussion of the weaknesses of their method relative to these competitors, especially in handling highly dynamic content?
12. The paper mentions that the method could be extended to handle temporal data (video). It would be valuable to explore how video frame sequence analysis can be integrated into the current framework and how this might affect both retrieval accuracy and sentiment polarity prediction.