Proceedings of CONF-CDS 2026 Symposium: Machine Learning and Neural Network Applications in Engineering
DOI: 10.54254/2755-2721/2026.AST36057

Research on Narrative Credibility Evaluation Methods for Corporate Crisis Communication Based on Large Language Models

Yiwen Zhong

School of Arts and Cultures, Newcastle University, Newcastle, UK
wellington589125@gmail.com

Abstract. As corporate crisis communication increasingly relies on public videos, social media statements, and documentary-style visual responses, public judgments of responsibility, sincerity, and remediation capacity are increasingly shaped by narrative credibility. Traditional evaluation of crisis communication mainly focuses on response strategies, responsibility attribution, and reputation repair effects, while giving limited attention to the integrated relationship among video text, visual evidence, and public feedback. To address narrative credibility evaluation in corporate crisis communication, a computational method is developed by combining film narrative analysis with large language models. Public corporate crisis videos, statement texts, and comment data from 2019 to 2025 are used as the corpus. Five evaluation dimensions are established, including factual consistency, responsibility clarity, verifiable action, emotional appropriateness, and narrative coherence. The experiment is completed through human annotation, feature extraction, LLM-based dimensional scoring, and error testing. The results show that LLM-based dimensional scoring outperforms the sentiment lexicon method, TF-IDF plus SVM, and the BERT classifier in Accuracy, Macro-F1, MAE, and Spearman correlation. Its Accuracy reaches 0.781, Macro-F1 reaches 0.742, MAE is 0.46, and Spearman correlation is 0.724. Responsibility clarity and verifiable action show relatively stable performance, while emotional appropriateness still presents higher error. These findings indicate that large language models can provide a computable and explainable technical path for evaluating narrative credibility in corporate crisis videos. They also offer methodological reference for computational film studies, crisis communication assessment, and corporate reputation governance.

Keywords: large language models, narrative credibility, corporate crisis communication, audiovisual discourse, computational film studies

1. Introduction

Corporate crisis communication has gradually shifted from announcements and press releases to public videos, brand documentary clips, executive statements, and social media responses. Public judgments of responsibility, sincerity, and corrective capacity increasingly depend on how crisis

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.AST36057

narratives are presented. Existing crisis communication research has developed mature approaches around responsibility attribution, response strategies, and reputation repair, emphasizing the alignment between corporate response, crisis type, perceived public harm, and organizational responsibility [1]. Narrative credibility analysis in film studies explains how shots, plot organization, spokesperson presence, and emotional rhythm shape viewers' judgments of corporate image, while computational methods make it possible to process subtitles, visual descriptions, and comment texts at scale. The proposed evaluation framework focuses on factual consistency, responsibility clarity, verifiable action, emotional appropriateness, and narrative coherence in corporate crisis videos, with human annotation and model explanation used to improve reproducibility and interpretability [2].

2. Literature review

2.1. Corporate crisis communication and trust repair

The core concern of corporate crisis communication literature lies in how crisis responsibility is understood by the public and how corporate responses influence trust repair. Situational crisis communication theory emphasizes the alignment among crisis type, responsibility level, and response strategy, making apology, explanation, compensation, and corrective action part of a reputation management mechanism rather than isolated statements [3]. Image repair research further places corporate responses within a framework of discursive action, examining how denial, evasion of responsibility, reduction of offensiveness, corrective action, and mortification reorganize organizational image [4]. This line of research provides two basic dimensions for evaluating narrative credibility, namely responsibility expression and repair commitment, while audiovisual crisis communication introduces visual evidence, on-screen presence, and emotional arrangement that extend credibility judgment beyond traditional text-based response analysis.

2.2. Audiovisual narrative and credibility construction

Audiovisual narrative research focuses on the relationship among story structure, narrative perspective, viewer identification, and emotional guidance. Corporate crisis videos often construct credibility through executive appearances, on-site footage, victim-related accounts, and remediation procedures. Narrative communication research shows that storytelling can reshape audiences' understanding of organizational motives and responsibility through character identification and situational immersion [5]. In corporate social responsibility and brand communication contexts, narrative richness influences perceived message credibility, organizational attitude, and behavioral intention [6]. Credibility in crisis videos is not equivalent to emotional intensity; it is jointly formed by factual evidence, accountable actors, temporal order, corrective measures, and restrained expression. Film-based narrative analysis therefore provides interpretable dimensions that can be transformed into computational features.

2.3. Large language models for narrative evaluation

Large language models are well suited to processing corporate crisis video subtitles, statement texts, and public comments because of their strengths in semantic understanding, discourse evaluation, stance recognition, and explanation generation. Research on automated text evaluation indicates that LLM-based evaluation has potential for better alignment with human judgment, although its stability depends on task definition, prompt structure, and scoring dimensions [7]. Hallucination detection

DOI: 10.54254/2755-2721/2026.AST36057

research further suggests that model outputs should be supported by consistency checks, evidence tracing, and external text verification to reduce unreliable judgments [8]. Semantic similarity metrics and explainability methods can help compare model outputs with human annotations and reveal how responsibility terms, evidential expressions, and corrective commitments affect credibility scoring [9]. LLM-based evaluation is therefore more appropriate when embedded in a process of human annotation, dimension-based scoring, and explanation validation rather than used as a single autonomous judge.

3. Experimental methods

3.1. Data collection and corpus construction

Data collection focuses on public video narratives in corporate crisis communication. The sources are limited to public crisis videos on corporate official websites, official Weibo accounts, WeChat Channels, Bilibili, YouTube, public corporate statements, press conference recordings, crisis response videos reposted by mainstream media, and their corresponding written statements. The time range is set from 2019 to 2025, covering five types of crisis scenarios, namely food safety, product quality, data leakage, labor disputes, and environmental compliance. For each case, the company name, industry category, crisis type, release time, video duration, subtitle text, written statement, public video comments, and key visual descriptions are retained. Video subtitles are generated through automatic transcription and then manually checked. Comment texts are cleaned by removing advertisements, repeated content, and emojis without semantic value. Key visual frames are structurally recorded by researchers according to scene setting, character appearance, evidence presentation, and remediation process. The final corpus is stored through a four layer structure of case, text, image description, and comment, which provides unified input for later annotation, feature extraction, and model evaluation.

3.2. Annotation scheme and feature extraction

The experimental process consists of five stages, including corpus cleaning, manual annotation, feature extraction, model scoring, and result validation. First, subtitles, statements, and comments are segmented, denoised, and aligned with the crisis timeline. Video texts, visual descriptions, and comment fragments from the same crisis case are then combined into one evaluation unit. Five evaluation dimensions are set, including factual consistency, responsibility clarity, verifiable action, emotional appropriateness, and narrative coherence. Each dimension is scored from 1 to 5, with a higher score indicating stronger narrative credibility [10]. Two annotators complete independent scoring. Samples with score differences greater than 2 points are reviewed by a third annotator. In the feature extraction stage, responsibility terms, evidence terms, commitment terms, time expressions, sentiment polarity, visual evidence type, and public questioning intensity are retained. During model evaluation, dimension based prompts guide the large language model to produce sub scores and brief explanations, which are then compared with manual annotation results.

3.3. LLM-based evaluation model and validation

The scoring function is designed to integrate three types of inputs generated from the previous procedure: subtitle and statement text, structured visual descriptions, and public comment features. For each crisis sample, the textual feature vector T_i is extracted from responsibility terms, evidence

DOI: 10.54254/2755-2721/2026.AST36057

expressions, commitment clauses, and temporal markers. The visual description vector V_i is constructed from spokesperson presence, evidence display, scene type, and remediation process. The comment vector U_i represents public questioning intensity and audience trust tendency. These three vectors are mapped into the five credibility dimensions through the large language model and a scoring layer [11]. The final narrative credibility score is calculated as shown in Formula (1), where C_i denotes the credibility score of sample i, α, β, and γ denote the weights of textual, visual, and comment features, and f_{LLM} represents the dimension based scoring function.

$$C_i=f_{LLM}(\alpha T_i+\beta V_i+\gamma U_i),\quad \alpha+\beta+\gamma=1 \tag{1}$$

To reduce the gap between model scoring and human annotation, the optimization process uses a weighted loss function. Since factual consistency and verifiable action are more decisive for crisis credibility than general emotional tone, these dimensions receive higher weights during error calculation. The model error is calculated as shown in Formula (2), where y_{ik} represents the human score of sample i on dimension k, $\hat{y}_{ik}$ represents the model score, w_k represents the dimension weight, n represents the number of samples, and K represents the five credibility dimensions.

$$L=\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}w_k\left(y_{ik}-\hat{y}_{ik}\right)^2 \tag{2}$$

4. Results

4.1. Credibility scoring performance

The experiment uses 96 public crisis communication samples as the corpus, including 67 samples in the training set and 29 samples in the test set. The five crisis scenarios are kept relatively balanced in sample size. The Cohen's kappa value of manual annotation is 0.75, indicating that the five dimensional annotation has acceptable stability. As shown in Table 1, the large language model outperforms TF IDF plus SVM, the BERT classifier, and the sentiment lexicon method in Accuracy, Macro F1, and Spearman correlation. Its scores are especially closer to human ratings in the dimensions of responsibility clarity and verifiable action. The sentiment lexicon method has the highest MAE, which shows that emotional polarity alone is insufficient for judging whether a corporate crisis narrative is credible. The BERT classifier performs steadily, but it has limited ability to use visual descriptions and cross sentence responsibility relations. The MAE of the large language model is 0.46, which falls within an acceptable error range for a five point scoring system.

Table 1. Narrative credibility evaluation results of different methods

Method	Accuracy	Macro F1	MAE	Spearman ρ
Sentiment lexicon method	0.552	0.518	0.91	0.421
TF IDF plus SVM	0.621	0.594	0.74	0.536
BERT classifier	0.704	0.681	0.58	0.642
LLM dimensional scoring	0.781	0.742	0.46	0.724

DOI: 10.54254/2755-2721/2026.AST36057

4.2. Narrative feature contribution and case interpretation

The human annotation agreement, model correlation, and error concentration of the five dimensions are further compared. The results are shown in Figure 1. Responsibility clarity has the highest human annotation agreement, with a kappa value of 0.81. Its correlation between model scores and human scores reaches 0.77, indicating that whether a company directly acknowledges responsibility and whether it explains the boundary of responsibility can be jointly recognized by annotators and the model. The kappa value of emotional appropriateness is 0.69, and its correlation coefficient is 0.63. Its error concentration reaches 29 percent. This is mainly because some videos use a restrained tone but lack factual evidence, which can lead the model to misjudge low emotional intensity as highly credible expression. The error concentration of verifiable action is 21 percent. This situation often appears in samples where companies propose commitments such as comprehensive rectification or continuous optimization without providing a timetable, responsible department, or progress evidence. Factual consistency and narrative coherence perform more stably, indicating that information alignment among subtitles, statements, and visual descriptions can effectively support model judgment.

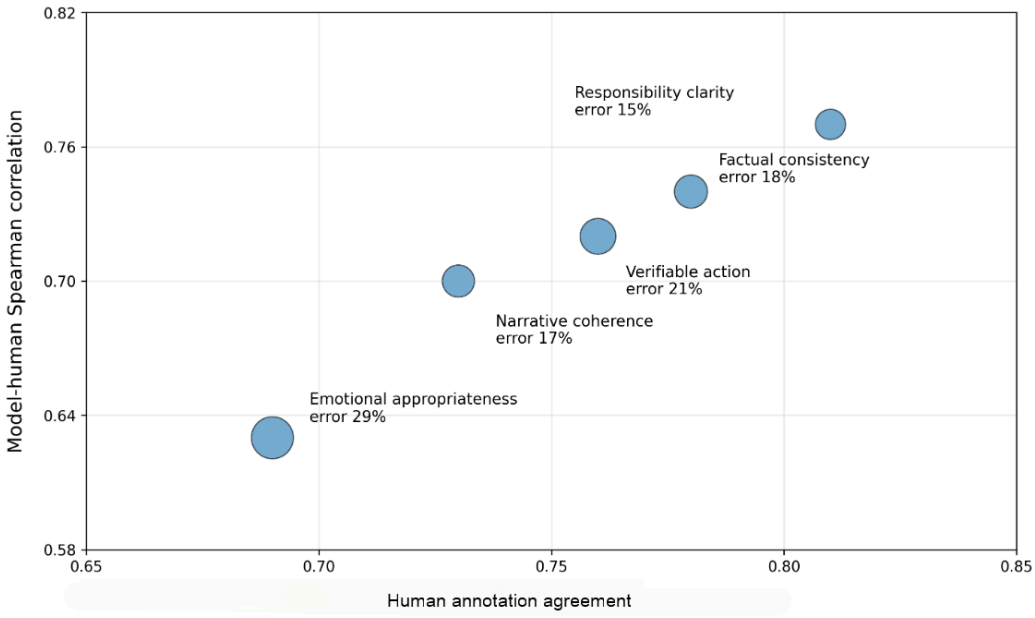

Figure 1. Bubble statistical chart of human annotation agreement and model scoring correlation

5. Discussion

The experimental results show that large language models have strong semantic integration capacity in evaluating narrative credibility in corporate crisis communication. They are especially effective in identifying the logical relationship among responsible actors, factual evidence, and corrective commitments. Compared with traditional sentiment analysis, LLM-based dimensional scoring does not simply treat positive tone as high credibility. Instead, it makes a combined judgment based on factual consistency, verifiable action, and narrative coherence. This result suggests that narrative structure, viewer identification, and visual evidence in film studies can be transformed into computational features, thereby expanding the evaluation scale of crisis communication research. At the same time, the high error in emotional appropriateness indicates that the model is still unstable when processing restrained expression, strategic apology, and ironic public comments. Future

DOI: 10.54254/2755-2721/2026.AST36057

research should introduce multimodal visual models, cross-platform contextual information, and larger human-annotated corpora to improve the model's accuracy in complex crisis video narratives.

6. Conclusion

A large language model-based experimental framework is constructed to evaluate narrative credibility in corporate crisis communication. Public crisis videos, statement texts, visual descriptions, and audience comments are integrated into a unified analytical process. Through a five-dimensional annotation scheme, an LLM scoring function, and weighted error optimization, the feasibility of using computational models to identify crisis narrative credibility is verified. The results show that LLM-based dimensional scoring can fit human judgment relatively well. It performs strongly in responsibility clarity, factual consistency, and verifiable action, which helps detect vague responsibility, insufficient evidence, and empty commitments in corporate responses. This method provides a computational path for audiovisual narrative analysis in film studies. It also offers a reproducible tool for evaluating the effects of corporate crisis communication.

References

[1] Riddell, H. (2024). Investigating social media users' preferences of content and sourcing during a crisis. *Communication and the Public, 9*(3), 277–294.
[2] Bielecka, H., Han, F., & Strzelecki, A. (2025). Antecedents and management strategies of social media crisis effects in the digital age. *European Journal of Management and Business Economics*.
[3] Bowen, S. A. (2024). "If it can be done, it will be done: " AI ethical standards and a dual role for public relations. *Public Relations Review, 50*(5), 102513.
[4] Watkins, B. A., & Woods, C. L. (2024). Confronting controversial content: Examining online narratives and frames in #CancelSpotify and #NetflixWalkout. *Public Relations Review, 50*(5), 102494.
[5] Kiambi, D., Kiousis, S., & Arceneaux, P. (2024). Analyzing media valence shifts: The association between a US PR firm's engagement and Kenya's portrayal in US media. *Public Relations Review, 50*(4), 102485.
[6] Ndone, J., & Kyriakopoulos, V. (2024). The effects of crisis type and CSR fit on organizational outcomes: The moderating role of authentic leadership in shaping organizational reputation, word-of-mouth, and purchase intentions. *Public Relations Review, 50*(5), 102517.
[7] Yue, C. A., et al. (2024). Artificial intelligence for internal communication: Strategies, challenges, and implications. *Public Relations Review, 50*(5), 102515.
[8] Liu, W., et al. (2024). Streaming disasters on TikTok: Examining social mediated crisis communication, public engagement, and emotional responses during the 2023 Maui wildfire. *Public Relations Review, 50*(5), 102512.
[9] Ki, E.-J., Ertem-Eray, T., & Hayden, G. (2024). The evolution of digital public relations research. *Public Relations Review, 50*(5), 102505.
[10] Jenner, S., et al. (2025). Using large language models for narrative analysis: A novel application of generative AI. *Methods in Psychology, 12*, 100183.
[11] Ziems, C., et al. (2024). Can large language models transform computational social science? *Computational Linguistics, 50*(1), 237–291.