Using Multimodal Models for Informative Classification of Ambiguous Tweets in Crisis Response

Sumiko Teng
Waseda University
National University of Singapore
sumiko@fuji.waseda.jp

Emily Ohman
Waseda University
ohman@waseda.jp

Abstract

Social media platforms like X (formerly Twitter) provide real-time information during crises but often include noisy, ambiguous data, complicating analysis. This study examines the effectiveness of multimodal models, particularly a cross-attention-based approach, in classifying tweets related to the California wildfires as "informative" or "uninformative," leveraging both text and image modalities. Models were evaluated for their ability to handle real-world noisy data with the help of a dataset containing both ambiguous and unambiguous tweets. Results show that the multimodal model outperforms unimodal counterparts, especially for ambiguous tweets, demonstrating the resilience and ability to integrate complementary modalities of multimodal approaches. These findings highlight the potential of multimodal approaches to enhance humanitarian response efforts by reducing information overload.

1 Introduction

Advancements in image and text analysis have unlocked the potential to combine these two modes of information for use in data science and analytics. One prominent application of multimodal information is in social media, where content is no longer limited to a single modality but often integrates audio, video, image, and text. This multimodal nature of social media has opened new avenues for data analysis, enabling deeper insights and richer interpretations while requiring new methodological approaches to facilitate multimodality.

This project leverages social media content, specifically tweets from X, to extract information related to crises (Palen, 2008). Since its advent, social media has served as a vital communication channel, allowing individuals on the ground to share real-time updates about ongoing events such as during the 2011 East Japan Earthquake and Tsunami (PEARY et al., 2012). This study focuses on tweets about the California wildfires in 2017, aiming to classify them as either "informative" or "not informative." Such classifications can aid humanitarian efforts by providing timely, relevant information while filtering out noise, ultimately reducing information overload and enhancing situational awareness (Imran et al., 2020).

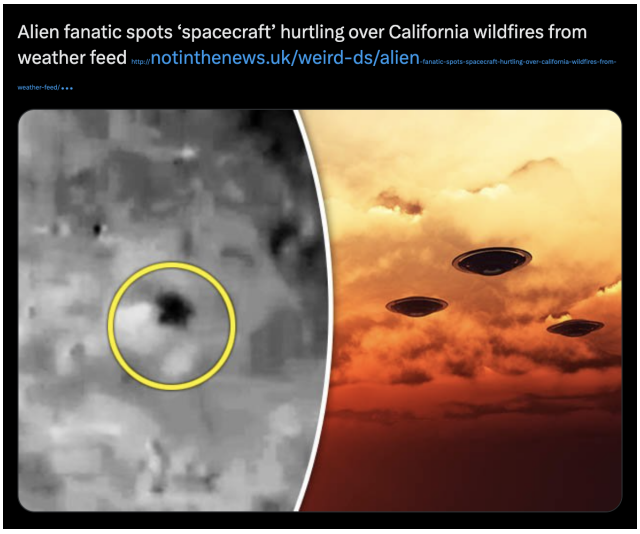

Figure 1: Example of an Ambiguous Tweet with Misaligned Image and Text Labels

However, social media content presents significant challenges as a reliable information source (e.g. Zhang and Cheng, 2024). Posts are typically unverified, and the noisy nature of multimodal data complicates analysis. For instance, a tweet's text might convey crucial information about a crisis, while its accompanying image may not align, or vice versa. An example of such misalignment is shown in Figure 1, where a tweet contains both text and image labels that do not match. Even for human observers, assessing whether a tweet is informative often requires careful consideration of both modalities, introducing ambiguity and inconsistency. Previous studies on multimodal classification for crisis datasets have often relied on pre-

Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 265–271

processed and cleaned data, where text and image labels are aligned (e.g. Imran et al., 2020).

This project, however, examines the performance of multimodal models on both ambiguous and unambiguous tweet data to address this gap. Given that multimodal models are typically based on deep neural network architectures capable of learning from complex and noisy data (Sleeman IV et al., 2022), this study hypothesizes that such models can effectively classify tweets as "informative" or "not informative," regardless of the ambiguity in the data.

2 Background

There are many aspects about multi-modal models that can be studied; how the multiple modes get fused together and how each mode of information is derived to be fused together. There are usually three key methods of designing multi-modal models: early fusion, hybrid fusion, or late fusion (Atrey et al., 2010).

2.1 Multimodal Classification for Social Media Analysis

There is no consensus on which fusion model works best so there are many different studies related to social media analysis using various fusion techniques. For example, previous work on multi-modal sentiment analysis which combines modalities of audio, text, and visual forms demonstrated superior performance in capturing sentiment cues compared to unimodal approaches (Chandrasekaran et al., 2021; Das and Singh, 2023). (Zeppelzauer and Schopfhauser, 2016) did a study using early and late fusion methodologies to classify social events using content posted on social media platforms, and found the early fusion strategy to be more superior than the late fusion strategy.

(Mouzannar et al., 2018) studied a multimodal deep learning algorithm to create a damage identification model from social media posts that was able to achieve a very high accuracy of 92.62%. It is also noted that the integration of deep learning methods significantly improving classification accuracy across various datasets. In the social media context, multimodal models are widely used to detect hateful religion memes (Hamza et al., 2024), fake news (Hangloo and Arora, 2022), medical misinformation (Wang et al., 2020) and even depression and suicide behaviour (Malhotra and Jindal, 2020), showing the potential of multimodal algorithms in addressing diverse and complex challenges involving social issues.

2.2 Social Media in Humanitarian Responses

The advent of social media has catalyzed the use of technology in humanitarian workflows and responses. Social media posts can be leveraged upon for humanitarian purposes to create alert systems (Stollberg and De Groeve, 2012), detect damages (Mouzannar et al., 2018) or even to anticipate humanitarian response during disasters (David et al., 2022). Social media platforms, like Twitter/X, can bridge communication between victims and witnesses of crises to humanitarian aid groups and authorities (Mullaney, 2012).

Social media platforms like Twitter are great at bridging communication between victims and witnesses of crises to humanitarian aid groups and authorities (Mullaney, 2012; Eriksson, 2018). Social media has become a key source of information during crises and disaster relief and (Kumar et al., 2022) presented a new application that can help relief organizations to monitor, track, and conduct analysis of tweets. These tweets can help first responders gain situational awareness immediately after a disaster or crisis to direct their response.

2.3 Annotation of Crisis Tweets

Due to information overload when looking at social media sources (Hiltz and Plotnick, 2013), information filtering is crucial for effectively gathering real-time information for humanitarian responses. Works include text-only unimodal models leveraging deep learning and traditional techniques which can capture semantic nuances within textual data (Jain et al., 2025, 2024a). Similarly, image-only models, such as those utilizing VGG-16, have been employed to extract informative visual features, achieving precise classification of images (Jain et al., 2024b).

Multimodal learning approaches that integrate traditional machine learning and deep learning techniques through early feature-level fusion are used to better address the interplay between modalities (Ofli et al., 2020). Additionally, contrastive learning models like CLIP have shown remarkable success in aligning textual and visual embeddings using contrastive loss, making them effective for classification (Mandal et al., 2024).

3 Data

This study utilizes the CrisisMMD dataset, a multimodal Twitter corpus comprising thousands of manually annotated tweets and images collected during seven major natural disasters, including earthquakes, hurricanes, wildfires, and floods that occurred globally in 2017 (Alam et al., 2018). The dataset offers three types of annotations: informative versus non-informative, humanitarian categories, and damage severity categories, providing a rich resource for analyzing crisis-related social media data.

For this study, the focus is scoped down specifically to tweets related to the Californian wildfires. While the dataset provides valuable insights, a key limitation is that the labels for text and images are collected separately. As such, a key problem with creating multimodal dataset from the collected data is that some rows have text and image labels that do not align. To address this ambiguity and ensure consistency, the multimodal data is filtered to include only instances where the text and image labels align. This filtering step mitigates potential noise and ambiguity and ensures the reliability of the dataset for training and evaluating multimodal models.

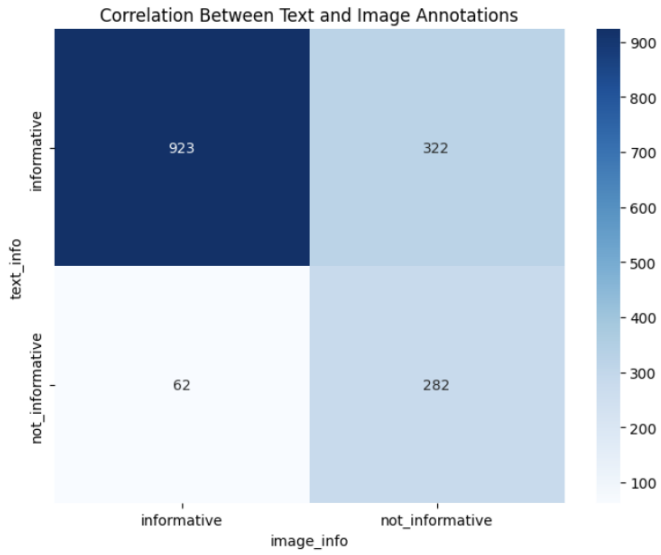

Figure 2: Correlation between text and image labels in dataset

The correlation graph in figure 2 shows substantial alignment between text and image annotations classified as "informative," with 923 instances of agreement. However, discrepancies are observed in 322 cases where the text is "informative" but the image is "not informative," and 62 cases where the image is "informative" but the text is "not informative." These misalignments highlight the complexity of social media posts, where texts and images' labels might not align, regardless of their overall informativeness.

Out of the 1,589 rows of data, 384 rows were identified as ambiguous, where the text and image labels for the tweets did not align. A manual annotation process was conducted programmatically to reclassify these rows, determining tweets as "informative" if they have any relevant information related to the California wildfires. Following this process, the dataset comprises 1,303 "informative" instances and 286 "not informative" instances, revealing a notable class imbalance that could impact the model's performance.

4 Methodology

This study aims to create a multimodal classification model to classify tweets, whether ambiguous or not, into two categories: "informative" and "not informative." The model's performance is evaluated against the entire dataset (1,589 tweets) and the subset of ambiguous data (384 tweets), testing its effectiveness in handling both aligned and misaligned text-image annotations. The data for this pilot study was limited in order to enable reliable error analysis and qualitative interpretations.

4.1 Data Preparation

To prepare the dataset for modeling, a train-test split of 80:20 was applied, ensuring that the test set retained the same proportion of ambiguous data (24%) as in the training dataset. To address the issue of class imbalance, where "informative" instances significantly outnumbered "uninformative" ones, random oversampling and class weights optimization were applied to the training data, balancing the two classes for more robust model training. The raw tweets in the dataset were pre-processed to improve text quality by converting all letters to lowercase and removing URLs, mentions, and retweet tags.

4.2 Model Experiments

Three models were experimented with to analyze the effectiveness of unimodal and multimodal approaches for classification:

Text-Only Model. A BERT base model was used to process textual data (Devlin et al., 2018).

The model was fine-tuned to classify tweets based solely on their text content, leveraging the semantic understanding capabilities of transformer-based architectures.

Image-Only Model. A VGG-16 model, a 16-layer deep convolutional neural network pre-trained on ImageNet, was utilized to classify tweets based on their image content (Simonyan and Zisserman, 2015). Renowned for its ability to extract relevant visual features, the model was fine-tuned using the image data to optimize its performance for the classification task.

Multimodal Model. A multimodal model was developed to leverage the complementary information from both text and image data. This model employs a hybrid fusion architecture that integrates the pre-trained VGG-16 model for image processing and the pre-trained BERT model for textual embeddings, using a cross-attention mechanism to fuse the two modalities efficiently. The cross-attention mechanism aligns text and image embeddings, learning from complementary features between the two modalities while filtering noise to identify relevant visual features in relation to the text (Khattar and Quadri, 2022). To classify the data, the outputs of the text and cross-attention modules are concatenated and passed through dense layers to produce class probabilities. This comprehensive fusion design allows the model to effectively capture complementary features from both modalities, enabling accurate classification of tweets as "informative" or "uninformative."

5 Results

Table 1: Weighted Average Results for Text, Image, and Multimodal Models

Dataset	Model	Acc.	Prec.	Rec.	F1
All	Text	81.0	79.0	81.0	80.0
	Image	81.0	79.0	81.0	80.0
	Multi.	**84.0**	**81.0**	**84.0**	**81.0**
Ambig.	Text	82.0	92.0	82.0	86.0
	Image	79.0	**93.0**	79.0	85.0
	Multi.	**90.0**	92.0	**90.0**	**91.0**

The results of the three classification models on both the entire dataset, "All Data", and the subset of ambiguous data are summarized in Table 1. All three models perform reasonably well for the classification tasks, achieving an F1-score of 0.80 across all datasets. The multimodal model, which combines text and image features, consistently outperformed the text-only and image-only models across both datasets in terms of accuracy, precision, recall, and F1-score.

Full dataset. The multimodal model achieved the highest accuracy of 84%, compared to 81% for both the text-only and image-only models. It also showed improved precision (0.81) and recall (0.84), resulting in a weighted F1-score of 0.81, indicating its robustness in leveraging complementary features from text and images. It should be noted that the text- and image-only models perform almost identically for the full dataset.

Ambiguous dataset. The multimodal model demonstrated a clear superiority, achieving the highest accuracy of 90% and a weighted F1-score of 0.91. It also maintained balanced precision and recall values of 0.92 and 0.90, respectively. In comparison, the text-only model showed strong performance with an accuracy score of 82% and an F1-score of 0.86, while the image-only model performed slightly lower, with an accuracy of 79% and an F1-score of 0.85.

Minority class. However, all models exhibited poor performance on the minority class of "not informative" tweets, particularly in the ambiguous dataset. None of the models achieved a recall greater than 0.33 for this class, and the multimodal model failed to classify any "not informative" tweets correctly in the ambiguous subset. This poor performance highlights the challenges posed by class imbalance and lack of minority data.

6 Discussion

The study has once again reaffirmed the effectiveness of using multimodal models for social media analysis, especially for the classification of tweets during crisis. The superior performance of multimodal models can be attributed to their ability to leverage both text and image information, similar to how humans make more informed decisions when provided with additional context.

A notable finding is the ability of multimodal models to classify ambiguous tweets more effectively than unimodal models. The incorporation of a cross-attention mechanism enables these models to focus on the most relevant features from

both modalities, reducing the impact of noise often present in ambiguous data. Compared to using unimodal models which inherently filters away one source of information and completely relying on one mode, multimodal models can make fairer and more informed decisions about whether a tweet is "informative" or "not informative".

Our results highlight the relative contributions of individual modalities in determining informativeness. For the ambiguous dataset, the text model achieved a slightly higher F1-score (0.86) compared to the image model (0.85). This suggests that, in ambiguous cases, the text modality often carries more critical information than the image modality, allowing the text-based model to perform marginally better. This finding underscores the importance of text in providing context, which can be particularly useful in determining ambiguity.

Multimodal models have proven to be effective in classifying tweets as "informative" or "uninformative" for humanitarian purposes, demonstrating their potential to enhance crisis response efforts. This study emphasizes that while noisy data, characterized by ambiguity between text and image modalities, poses challenges, it should not be dismissed. In our results, multimodal models have shown resilience in handling such ambiguity, leveraging complementary information from both modalities to make accurate predictions. The focus on the California wildfires dataset provided valuable insights into the applicability of multimodal models in real-world crises, as this dataset reflects the complexity of social media content during natural disasters. Overall, this project underscores the importance of incorporating multimodal approaches in analyzing ambiguous social media data, especially in the case of classifying tweets in order to reduce information overload and support timely humanitarian work.

6.1 Limitations.

One major limitation of this project was the difficulty all three models faced in predicting the minority class of "not informative" tweets. This challenge was particularly pronounced in the ambiguous dataset, where the test set contained only three rows for this class. To address this imbalance, weighted F1 scores were used to assess model performance, reflecting the practical reality that ambiguous "not informative" tweets are indeed rare. However, the limited representation of the minority class in the test dataset remains a significant issue, making it difficult to determine whether the models' poor performance on this class is due to inherent limitations in their predictive capabilities or simply the result of insufficient data points for evaluation. This limitation underscores the need for a more balanced dataset or alternative evaluation strategies to better assess the models' performance on minority classes.

Another limitation of this project was the manual annotation of the ambiguous dataset conducted by a single person. During this process, the tweets were labeled "informative" if they provided any information about the California wildfires, including tweets about topics like UFO sightings that may not offer significant humanitarian value. Since these manually annotated labels were integrated with the existing CrisisMMD dataset labels, discrepancies in annotation criteria between the original annotation guidelines and the ones conducted for this project could conceivably have lead to inconsistencies, potentially impacting the models' performance.

Finally, this project faced the challenge of class imbalance within the dataset, which was addressed through random oversampling to balance the classes during training. The minority class, "not informative," was duplicated to match the number of instances in the majority class. This duplication may have caused the model to overfit on these specific rows, potentially contributing to its poor performance on the minority class during test evaluation. However, employing more sophisticated data balancing techniques, such as SMOTE, is challenging for multimodal datasets due to the complexity of generating synthetic data across multiple modalities.

6.2 Future work

Some future work for this project could focus on addressing data imbalance by collecting more data, particularly for the minority class, to reduce reliance on oversampling methods to balance the dataset. To enhance label reliability especially for ambiguous cases, the annotation process could be improved to include multiple annotators and inter-annotator agreement metrics. Additionally, systematic hyperparameter tuning, which was not explored in this project, could be used to optimize model performance. Testing more advanced models, such as CLIP or other state-of-the-art multimodal architectures could further improve the classification of multimodal social media data.

References

Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16:345–379.

Ganesh Chandrasekaran, Tu N Nguyen, and Jude Hemanth D. 2021. Multimodal sentimental analysis for social media applications: A comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5):e1415.

Ringki Das and Thoudam Doren Singh. 2023. Multimodal sentiment analysis: a survey of methods, trends, and challenges. *ACM Computing Surveys*, 55(13s):1–38.

Walter David, Beatriz Garmendia-Doval, and Michelle King-Okoye. 2022. Artificial intelligence support to the paradigm shift from reactive to anticipatory action in humanitarian responses. In *International Conference on Modelling and Simulation for Autonomous Systems*, pages 145–162. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Mats Eriksson. 2018. Lessons for crisis communication on social media: A systematic review of what research tells the practice. *International Journal of Strategic Communication*, 12(5):526–551.

Ameer Hamza, Abdul Rehman Javed, Farkhund Iqbal, Amanullah Yasin, Gautam Srivastava, Dawid Połap, Thippa Reddy Gadekallu, and Zunera Jalil. 2024. Multimodal religiously hateful social media memes classification based on textual and image data. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(8):1–19.

Sakshini Hangloo and Bhavna Arora. 2022. Combating multimodal fake news on social media: methods, datasets, and future perspective. *Multimedia systems*, 28(6):2391–2422.

Starr Roxanne Hiltz and Linda Plotnick. 2013. Dealing with information overload when using social media for emergency management: Emerging solutions. In *ISCRAM*. Citeseer.

Muhammad Imran, Ferda Ofli, Doina Caragea, and Antonio Torralba. 2020. Using ai and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions.

Tarun Jain, Dinesh Gopalani, and Yogesh Kumar Meena. 2025. Informative task classification with concatenated embeddings using deep learning on crisismmd. *International Journal of Computers and Applications*, pages 1–18.

Tarun Jain, Dinesh Gopalani, and Yogesh Kumar Meena. 2024a. Classification of humanitarian crisis response through unimodal multi-class textual classification. In *2024 International Conference on Emerging Systems and Intelligent Computing (ESIC)*, pages 151–156. IEEE.

Tarun Jain, Dinesh Gopalani, and Yogesh Kumar Meena. 2024b. Image tweet classification for crisis informative task. In *2024 International Conference on Integrated Circuits, Communication, and Computing Systems (ICIC3S)*, volume 1, pages 1–6. IEEE.

Anuradha Khattar and SMK Quadri. 2022. Camm: cross-attention multimodal classification of disaster-related tweets. *IEEE Access*, 10:92889–92902.

Sameer Kumar, Chong Xu, Nidhi Ghildayal, Charu Chandra, and Muer Yang. 2022. Social media effectiveness as a humanitarian response to mitigate influenza epidemic and covid-19 pandemic. *Annals of Operations Research*, 319(1):823–851.

Anshu Malhotra and Rajni Jindal. 2020. Multimodal deep learning based framework for detecting depression and suicidal behaviour by affective analysis of social media posts. *EAI Endorsed Transactions on Pervasive Health and Technology*, 6(21).

Bishwas Mandal, Sarthak Khanal, and Doina Caragea. 2024. Contrastive learning for multimodal classification of crisis related tweets. In *Proceedings of the ACM on Web Conference 2024*, pages 4555–4564.

Hussein Mouzannar, Yara Rizk, and Mariette Awad. 2018. Damage identification in social media posts using multimodal deep learning. In *ISCRAM*. Rochester, NY, USA.

Mark J Mullaney. 2012. Optimizing social media in humanitarian crisis responses. *The Macalester Review*, 2(1):3.

Ferda Ofli, Firoj Alam, and Muhammad Imran. 2020. Analysis of social media data using multimodal deep learning for disaster response. *arXiv preprint arXiv:2004.11838*.

Leysia Palen. 2008. Online social media in crisis events. *Educause quarterly*, 31(3):76–78.

Brett PEARY, Rajib Shaw, and Yukiko TAKEUCHI. 2012. Utilization of social media in the east japan earthquake and tsunami and its effectiveness. *Journal of Natural Disaster Science*, 34:3–18.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *Preprint*, arXiv:1409.1556.

William C Sleeman IV, Rishabh Kapoor, and Preetam Ghosh. 2022. Multimodal classification: Current landscape, taxonomy and future directions. *ACM Computing Surveys*, 55(7):1–31.

Beate Stollberg and Tom De Groeve. 2012. The use of social media within the global disaster alert and coordination system (gdacs). In *Proceedings of the 21st International Conference on World Wide Web*, pages 703–706.

Zuhui Wang, Zhaozheng Yin, and Young Anna Argyris. 2020. Detecting medical misinformation on social media using multimodal deep learning. *IEEE journal of biomedical and health informatics*, 25(6):2193–2203.

Matthias Zeppelzauer and Daniel Schopfhauser. 2016. Multimodal classification of events in social media. *Image and Vision Computing*, 53:45–56.

Zeqian Zhang and Zhichao Cheng. 2024. Users' unverified information-sharing behavior on social media: the role of reasoned and social reactive pathways. *Acta Psychologica*, 245:104215.