Journal of Data Mining and Digital Humanities

EPI sciences overlay journals

Ambiguity in Crisis: A Multimodal and Synthetic Data Approach to Classification

Sumiko Teng[1,2]

[1]Waseda University, Japan
[2]National University of Singapore, Singapore

Abstract

Social media platforms, such as Twitter (now X), play a crucial role during crises by enabling real-time information sharing. However, the multimodal data can be ambiguous with misalignment of labels cross-modality. Being able to classify informative and not informative tweets can help in crisis response, yet they can be ambiguous and unbalanced in datasets, impairing model performance. This study explores the effectiveness of multimodal learning approaches for classifying crisis-related tweets regardless of ambiguity and addressing class imbalance through synthetic data augmentation using generative artificial intelligence (AI). Experimental results demonstrate that multimodal models consistently outperform unimodal ones, particularly on ambiguous tweets where label misalignment between modalities is prevalent. Furthermore, the addition of synthetic data significantly boosts macro F1 scores, indicating improved performance on the minority class.

Keywords

multimodal learning; crisis informatics; digital humanities; social media analysis; synthetic generation; tweet classification

I INTRODUCTION

Social media platforms, like X, have become useful platforms to crowdsource real-time information during crisis stemming from natural disasters, including wildfires. 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 like 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].

However, due to the free nature of social media, where users are free to post content freely, there comes the problem of noisy information hindering the effectiveness of social media to provide timely and relevant crisis updates to first responders and humanitarian aid teams to inform them of what is happening on the ground. Effective classification of tweets in X can help harness the potential of social media to gather real-time information while reducing information overload and noise.

A key problem plagued by the use of social media content for analysis is the fact that it contains

a lot of noise. Social media posts are typically unverified and, therefore, may be a less consistent source of data. For example, to determine whether a tweet contains crucial information about a crisis, it is necessary to inspect both the text and accompanying images. It is possible that these two modes of information may misalign, making it ambiguous even for human observers to access. Previous works on using multimodal classification on crisis datasets have often focused mainly on preprocessed and cleaned data, where there is no ambiguity among the labels for both text and image modalities.

This study aims to build on my previous work on multimodal classification of crisis tweets regardless of ambiguity [Teng and Öhman, 2025]. While the earlier work focused on baseline models using weighted evaluation metrics without synthetic augmentation, this thesis introduces a new layer of analysis by incorporating the exploration of the synthetic data generation and more comprehensive evaluation metrics and analysis. This work also explores the use of more advanced multimodal models like CLIP to understand the effectiveness of large pretrained models on a task as specific as ours.

II BACKGROUND

2.1 Crisis Tweet Classification

Due to information overload when looking at social media sources [Hiltz and Plotnick, 2013], information reduction and filtering are crucial in being able to effectively gather real-time information for humanitarian response. There are many studies done on information identification for crisis social media data to alleviate the problem of information overload using the CrisisMMD dataset, where text and image labels align. These include text-only unimodal models leveraging deep learning and traditional techniques, which are able to 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 the classification [Mandal et al., 2024]. Some studies also used more advanced architectures, such as the Multimodal Cycle-GAN (MMC-GAN) by employing mixed fusion strategies and robust feature extraction techniques to achieve state-of-the-art performance for classification [Zhou et al., 2023].

2.2 Handling Class Imbalance in Social Media Datasets

There are several challenges associated with using social media datasets, one of the most prominent being class imbalance, a common characteristic of real-world data. Most classification algorithms are naturally evaluated on balanced data with distribution of data from its respective classes [Ali et al., 2013]. However, in reality, much data is inherently unbalanced, for example, in fraud detection or disease detection [Johnson and Khoshgoftaar, 2019]. Class imbalances make it complex for models to learn effectively from both classes, as there is an inherent bias that favours the majority class. Within the social media context, there is no exception to this problem. Liu et al. [2014]'s research on sarcasm detection, Liu et al. [2017]'s work on spam detection and cyberbullying detection by Agrawal and Awekar [2018] all face problems of an imbalanced dataset. Common strategies around this problem involve Random Over-Sampling (ROS), Random Under-Sampling (RUS), and Synthetic Minority Over-Sampling Technique

(SMOTE). On the algorithmic front, strategies like adjusting class weights to learn more heavily from minority classes are also used [Leevy et al., 2018].

2.3 Synthetic Multimodal Data Generation

With recent advancements in generative AI, synthetic data has emerged as a viable solution for augmenting datasets and enhancing diversity, particularly in underrepresented classes. Due to the heavy cost involved in data collection, synthetic data generation is of great interest with the capabilities of generative AI algorithms. Li and Li [2025] found that synthesizing training images for use in training vision-language models like CLIP can boost compositional understanding. Deeva et al. [2021] also devised a multimodal data generation pipeline in generating tabular and image data related to personal information, which achieves highly plausible results. While this study used Generative Adversarial Networks (GANs) [Goodfellow et al., 2014] for image generation, Borji [2023] found that stable diffusion [Rombach et al., 2022] can generate images consisting of faces superior to other models.

III DATA

This study leverages the CrisisMMD dataset, a multimodal Twitter corpus containing thousands of manually annotated tweets and images from seven major global natural disasters in 2017, including earthquakes, hurricanes, wildfires, and floods [Alam et al., 2018]. The dataset includes three annotation layers: informativeness, humanitarian category, and damage severity, making it a valuable resource for analyzing crisis-related content on social media.

For this study, the analysis is narrowed to tweets specifically related to the 2017 California wildfires. A notable limitation of the dataset is that annotations for text and images were conducted independently, leading to potential misalignment between modalities. To address this issue, only tweet-image pairs with matching labels were retained, reducing ambiguity and enhancing the reliability of the dataset for training and evaluating multimodal classification models.

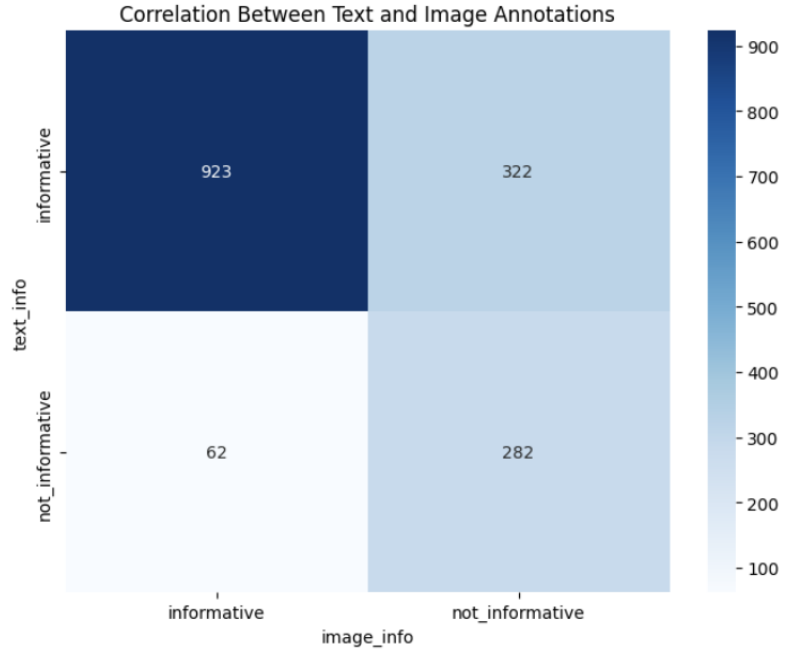


Figure 1: Correlation between text and image labels in dataset

The correlation graph in Figure 1 reveals strong alignment between text and image annotations labeled as ”informative,” with 923 instances of agreement. Nonetheless, notable discrepancies exist: 322 cases where the text is labeled ”informative” but the image is ”not informative,” and 62 cases with the opposite pattern. These mismatches underscore the inherent complexity of social media content, where text and image modalities may convey differing levels of informativeness.

Of the 1,589 tweet-image pairs analyzed, 384 were flagged as ambiguous due to misaligned modality labels. To address this, a manual re-annotation process was carried out to resolve inconsistencies. Tweets were labeled “informative” if they contained any relevant information about the California wildfires, ensuring a more consistent and meaningful multimodal dataset for subsequent analysis.

The raw multimodal label is derived from checking the alignment of unimodal labels, if they align, the respective label will be assigned. If the unimodal labels conflict, then the label will be ”ambiguous”. The second type of label is the human-annotated label assigned to ambiguous data after manual annotation. Figure 2 presents the distribution of these combined labels.

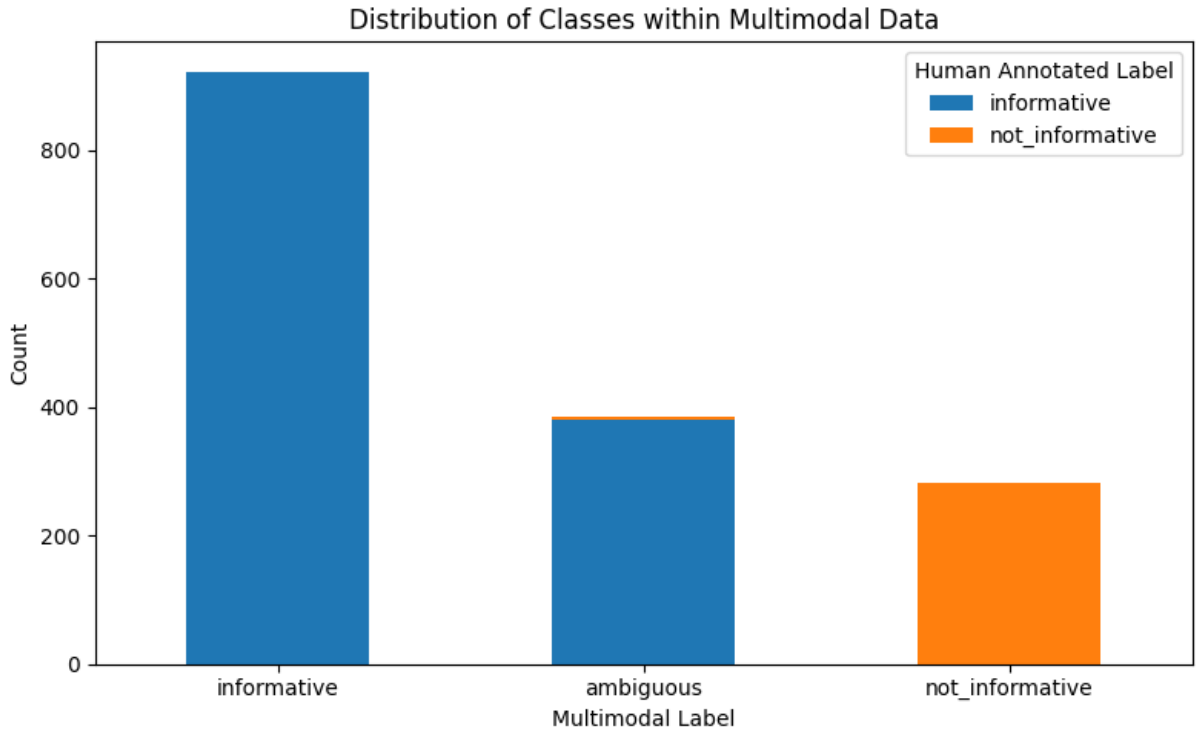


Figure 2: Distribution of class labels for the multimodal dataset showing severe data imbalance among ambiguous tweets only

As seen in the figure, a non-trivial portion of the dataset is labeled as ambiguous, comprising 384 rows, which significantly exceeds the number of rows in the not informative class. Upon closer inspection of the human annotations for these ambiguous tweets, the issue of class imbalance becomes even more pronounced: only 4 ambiguous tweets were labeled as not informative, while the remaining 380 were deemed informative. These ambiguous cases present a core challenge in the task of informativeness classification and serve as a key motivation for adopting a multimodal approach.

However, the challenge lies not only in the ambiguity itself but also in the highly imbalanced

distribution of labels of the ambiguous dataset, which may negatively impact model performance. This imbalance highlights the importance of synthetic data augmentation to better represent underrepresented classes and ensure more robust and generalizable model learning during training.

IV METHODOLOGY

This study aims to create a multimodal classification model to classify tweets, whether ambiguous or not, into two categories: "informative" and "not informative." To tackle the problem of data imbalance, a synthetic data generation process was also done to augment the dataset.

4.1 Synthetic Data Generation

Severe under-representation of minority data was a key limitation in my previous study, where class imbalance significantly hindered the model's ability to learn from the minority class [Teng and Öhman, 2025]. Building on these findings, the present study shifts the focus from overall performance to enhancing the classification of severely underrepresented classes. Specifically, it aims to improve the model's ability to accurately identify ambiguous and not informative tweets, which are typically very limited in real-world datasets. Given the scarcity of such examples, this study explores the use of generative AI techniques to synthetically augment the training data, thereby supporting more balanced and effective learning in multimodal settings.

To introduce more severe minority class data, 100 synthetic tweets, specifically ambiguous and not informative tweets, were generated using a combination of a large language model (ChatGPT) and a text-to-image model. The number of samples was deliberately set to 100 to increase the proportion of not informative tweets among ambiguous cases to at least 20%. This aimed to mitigate the skew observed in the original dataset, where the vast majority of ambiguous tweets were labeled as informative.

As the goal was to create ambiguous tweets, it was essential to ensure a conflict in informativeness between the text and image modalities. There are 2 combinations of misalignment in unimodal labels:

1. Image is informative and tweet text is not informative
2. Tweet text is informative and image is not informative

Rather than generating text-image pairs holistically, the text and image components were generated independently to minimize the coherence between them. This reduced the likelihood that the combined modalities would implicitly convey meaningful information. Thus, the final multimodal sample generated is likely to be ambiguous, and due to the disjointed or vague nature of the content, it can be deemed as not informative.

Figure 3 presents examples of synthetic tweets generated for the minority class labeled as both ambiguous and not informative. Although the tweets reference the broader context of wildfires, neither the text nor the accompanying image provides specific or actionable information. Notably, the generated images often suffer from quality limitations. In particular, human features, such as faces or limbs, are frequently rendered unrealistically or distorted. Some images may also lack interpretability even to humans, containing abstract or unrecognizable elements that do not make sense. Despite these limitations, manual inspection suggests that the generated data is generally coherent and not overly unrealistic.

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]. This BERT model was fine-tuned to classify tweets based on the text, building upon the strong semantic understanding capabilities of transformer-based architectures.

Image-Only Model. A VGG-16 model that has a 16-layer deep convolutional neural network and is pretrained on ImageNet was fine-tuned to classify tweets based on their image content [Simonyan and Zisserman, 2015]. Having a strong ability to extract relevant visual features, the VGG-16 model was fine-tuned for the classification task.

Multimodal Cross Attention Model. To leverage the complementary information from both text and image modalities, a multimodal model employing a hybrid fusion architecture that integrates the pre-trained VGG-16 model for image processing and the pre-trained BERT model for textual embeddings was used. Cross-attention mechanism was used to fuse the two information modes together effectively by aligning text and image embeddings [Khattar and Quadri, 2022]. To classify the data, the outputs of the text and cross-attention layers are designed to produce class probabilities. This cross-attention fusion design aims to allow the model to effectively capture complementary features across modes and achieve strong classification performance.

CLIP Model with Logistic Regression Classifier. CLIP (Contrastive Language–Image Pre-training) is one of the more popular multimodal vision and text model. While it is a multimodal model, CLIP was designed to be used for zero-shot classification of images. Hence, to evaluate the informativeness of crisis tweets, CLIP could be used to assess tweet images and classify them into their respective classes using text prompts. CLIP is a vision-language model developed by OpenAI that learns to align text and image embeddings within a shared latent space [Radford et al., 2021]. To use both modalities of data while using CLIP, a supervised classification pipeline was constructed based on the joint embeddings produced by the CLIP model. This method fine-tunes a classifier on top of the CLIP-generated embeddings to use both text and image for our classification.

V RESULTS

Model	Syn. Data	Macro F1	Weighted F1
BERT	No	0.63	0.80
BERT	Yes	0.72	0.82
VGG16	No	0.64	0.80
VGG16	Yes	0.66	0.78
Cross-Attention	No	0.64	0.81
Cross-Attention	Yes	0.63	0.78
CLIP w. Classifier	No	0.73	0.86
CLIP w. Classifier	Yes	0.80	0.87

Table 1: Effect of Synthetic Data on Model Performance (Entire dataset).

The results of the four classification models on the entire dataset with and without the use of synthetic data for training are summarized in Table 1. All four models perform reasonably well for the classification tasks, achieving a weighted F1-score of around 0.80 across all datasets. The CLIP and supervised classifier model consistently outperformed the other models across both datasets in both macro and weighted F1 scores.

Full dataset without synthetic data. The multimodal CLIP and supervised classifier model achieved the highest weighted F1 score of 86%, compared to 80% for both the text-only and image-only models. Both multimodal models performed better than unimodal models, indicating the strength of using both modalities for classification.

Full dataset with synthetic data. When trained with synthetic data, all models except the cross-attention model showed a boost in macro F1 score. The primary motivation for using synthetic data is to increase dataset diversity, particularly by increasing the representation of minority classes. The observed improvements in macro F1 suggest that the use of synthetic data contributes to better model performance for minority classes, addressing a key challenge faced by earlier models. The best improvement in performance through the addition of synthetic data for training is the CLIP model with a supervised classifier.

Model	Syn. Data	Macro F1	Weighted F1
BERT	No	0.45	0.86
BERT	Yes	0.80	0.86
VGG16	No	0.50	0.85
VGG16	Yes	0.64	0.76
CLIP Zero Shot	-	0.38	0.73
Cross-Attention	No	0.47	0.91
Cross-Attention	Yes	0.83	0.89
CLIP Supervised	No	0.47	0.93
CLIP Supervised	Yes	0.80	0.86

Table 2: Effect of Synthetic Data on Model Performance (Ambiguous subset only).

Ambiguous dataset. The evaluation on the ambiguous subset alone initially reveals a drop in macro F1 score across all models when trained only on the original data. This indicates that the models struggle with minority class classification in ambiguous contexts, which can be easily identified using the macro F1 score, which equally weights all classes. However, introducing synthetic data for augmentation significantly improves macro F1 scores across the board (Table 2). For instance, BERT's macro F1 improves from 0.45 to 0.80, and the multimodal cross-attention model's increases from 0.47 to 0.83, showing a notable improvement in the models' ability to generalize and correctly classify minority tweets. This signals the utility of synthetic data to improve the model's sensitivity to ambiguous and minority cases.

VI DISCUSSION

This study provides several key findings from experiments on classifying crisis-related tweets, regardless of ambiguity, using unimodal and multimodal approaches. First, harnessing the use of both text and image data consistently improved overall performance compared to single-modality models. Specifically, the CLIP model with supervised fine-tuning achieved the highest performance on the full dataset, while the multimodal cross-attention model proved most effective on the ambiguous subset.

Second, the introduction of synthetic tweets for training, specifically those of the non-informative and ambiguous class, has been shown to be useful in improving model sensitivity to the minority class. Significantly improved macro F1 scores indicate better balance across class predictions.

Third, the drop in macro F1 scores when evaluating on ambiguous tweets alone compared to the full dataset reveals the limitations of models in correctly classifying uncertain or noisy tweets.

However, similarly to the full dataset case, the inclusion of synthetic data very significantly improves the macro F1 score, showing the robustness of the augmentation method.

Superior performance of Multimodal Learning. Multimodal learning, specifically models that align text and image data through mechanisms like cross-attention, proved very effective in handling ambiguous tweets. As tweets have information scattered across both modalities, text-only or image-only models may struggle when one modality's signals are weaker than the other. Therefore, multimodal approaches are better able to resolve ambiguity by fusing complementary features or representing both modalities into one shared latent space.

Both multimodal models investigated in this study outperformed their unimodal counterparts, particularly on the subset of ambiguous tweets. This suggests that accurate classification of ambiguous tweets relies heavily on the ability to leverage both textual and visual modalities. In many ambiguous cases, modalities may contradict each other, or one modality may carry more relevant or less noisy information than the other, making multimodal integration essential for robust classification performance.

On the ambiguous subset, the cross-attention fusion model achieved higher macro and weighted F1 scores than the CLIP model with a supervised classifier, especially when trained with synthetic data. However, when evaluating on the full dataset, the CLIP supervised model outperformed the cross-attention model. This indicates that different model architectures may possess distinct strengths: cross-attention fusion appears better suited to resolving ambiguity, whereas CLIP demonstrates stronger general performance across all tweet types.

This distinction is further supported by a principal component analysis (PCA) of the learned embeddings. For the cross-attention fusion model, PCA plots reveal a tighter, more compact embedding space with a clear separation between informative and uninformative tweets. Ambiguous examples, while sometimes misclassified, are generally situated near their corresponding clusters. However, the narrowness of this space may suggest that the model has limited capacity to generalize to more varied inputs, even if it performs well on ambiguous cases. In contrast, the CLIP model's embedding space is broader and shows more dispersion, particularly for ambiguous tweets, which often lie between the two main clusters. This reflects the model's difficulty in decisively classifying ambiguous examples. However, this wider space may be indicative of better generalization, which aligns with its stronger performance on the full dataset that includes non-ambiguous tweets.

Effectiveness of Synthetic Data in Addressing Class Imbalance. Augmenting the training dataset with minority synthetic tweets led to consistent gains in macro F1 scores across most models, highlighting its value in addressing class imbalance and delivering gains for minority class classification. Macro F1 is particularly sensitive to minority class performance, and its significant increase suggests that the model becomes better at correctly classifying less represented or more ambiguous instances. Notably, models like BERT and CLIP model with a supervised classifier saw large improvements when trained on synthetic-augmented data. These findings support the idea that carefully generated synthetic data can improve generalization for underrepresented classes, especially in real-world datasets that are skewed by crisis reporting patterns.

Evaluation Metric Trade-offs and Real-World Implications. Previous work done on evaluating model performance was more naive, only focusing on weighted F1 scores as it provides a sense of how the model performs given the unbalanced nature of the data [Teng and Öhman, 2025]. However, one caveat to simply relying on the weighted F1 score may offer an incomplete perspective, as the performance for the minority class cannot be observed immediately. Hence,

macro F1, which gives equal weight to each class, provides a better reflection of model robustness across categories and is more crucial for a crisis-based dataset. A model with high accuracy but low macro F1 may only be performing well on the dominant class, failing to detect critical minority class signals, such as early warnings or localized incidents. In high-stakes applications like emergency response or misinformation filtering, optimizing for balanced performance is vital. Hence, by considering both macro and weighted F1 scores, this study demonstrates how metric choice can directly shape how models are evaluated and what will be inherently prioritized depending on the choice of metrics.

6.1 Limitations.

Annotation. The data used in this study were entirely derived from the CrisisMMD dataset. While this dataset provides multimodal labels, they were obtained through independent annotation of each modality, text, and image, rather than as an integrated whole. For tweets exhibiting ambiguous labels due to modality misalignment, additional manual annotation was conducted. However, since the annotator was not involved in the original annotation process of the CrisisMMD dataset, the newly added labels may reflect a different interpretation of what constitutes an ”informative” or ”not informative” tweet. This inconsistency may introduce subjectivity and affect the generalisability of the study.

Context Specificity. This study focused exclusively on tweets related to the 2017 California wildfires. As such, the findings may not generalize well to contemporary social media content. Language on platforms like X (formerly Twitter) evolves rapidly, with new slang, content formats, and platform norms emerging over time. Additionally, shifts in platform regulations and user behavior mean that the nature of crisis-related tweets today may differ substantially from those in the dataset, potentially limiting the applicability of the trained models to current events.

Risks of Synthetic Data. The generation of synthetic multimodal data relied on a relatively simple pipeline: textual content was created using ChatGPT and paired with images generated via the Stable Diffusion model. These components were aligned using heuristic rules designed to resemble real-world tweet-image pairs. Despite efforts to ensure realism, some generated images suffered from issues such as distorted human features, which could affect downstream model performance. While a quick visual check was done to verify the trustworthiness of the augmentation, synthetic data quality could be a limitation of this study. Furthermore, the process may have inadvertently introduced hidden biases or noise into the training data stemming from the model type used and its training data, potentially influencing our model predictions in unintended ways.

6.2 Practical Applications

Crisis Response. This study demonstrates the potential of data science in harnessing social media for real-time crisis response, even in the presence of noisy and ambiguous information. The strong performance of multimodal classification models in identifying informative crisis-related tweets suggests practical utility in emergency response systems. Such models can be integrated into automated information filtering pipelines, reducing the operational load faced by crisis response teams. By accurately detecting both the presence and severity of crisis through crowd-sourced social media content, authorities and humanitarian organizations can more effectively prioritize and allocate resources to areas in greatest need.

Use of Generative AI to Improve Models. The findings on the effectiveness of synthetic data generated using generative AI highlight a promising direction for improving model robustness and performance. In the context of social media, where data imbalance is a persistent issue,

especially for minority or underrepresented cases, synthetic data generation serves as a scalable solution. By augmenting datasets with synthetically created examples, particularly for rare or ambiguous classes, models can be trained to better recognize edge cases and avoid overfitting to majority patterns. This approach has broad implications, including enhancing misinformation detection, improving content moderation in low-resource languages, and addressing systemic biases in classification systems. Moreover, in high-stakes scenarios like natural disasters, where data collection is often difficult, generative AI can play a crucial role in filling gaps and simulating critical training data.

6.3 Future work

Other Crisis Domains. While this study focuses on California wildfires in 2017 only, future work should explore the transferability of the findings to other crisis types, such as natural disasters, pandemics, or political uprisings and time periods. Evaluating whether synthetic data generation and multimodal learning approaches generalise well across diverse scenarios will help establish their broader applications across different contexts.

Model Tuning and Prompt Engineering. Further improvement in model performance, particularly for CLIP and Cross-Attention models, could be achieved through prompt engineering and more rigorous hyperparameter tuning. This includes refining candidate label phrasing for zero-shot settings and optimising learning rates, attention parameters, or batch sizes in supervised training.

Fine-Tuning for Synthetic Data Generation. The current approach uses a pre-trained stable diffusion model to generate images using image prompts generated by large language models. Future research could involve fine-tuning generative models specifically on crisis-related data to produce more domain-relevant synthetic tweets. This could lead to higher-quality samples that better reflect the nuances and reality of how people post during a crisis.

Evaluation of Synthetic Data Quality. The use of synthetic data is crucial for our tasks, and further exploration on robust evaluation metrics to assess the realism, diversity, and utility of generated data is important. Future work can explore both automatic and human-in-the-loop methods to validate synthetic tweets before use in model training, ensuring they do not introduce noise or bias.

References

Sweta Agrawal and Amit Awekar. Deep learning for detecting cyberbullying across multiple social media platforms. In *European conference on information retrieval*, pages 141–153. Springer, 2018.

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

Aida Ali, Siti Mariyam Shamsuddin, and Anca L Ralescu. Classification with class imbalance problem. *Int. J. Advance Soft Compu. Appl*, 5(3):176–204, 2013.

Ali Borji. Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2, 2023. URL https://arxiv.org/abs/2210.00586.

Irina Deeva, Andrey Mossyayev, and Anna V Kalyuzhnaya. A multimodal approach to synthetic personal data generation with mixed modelling: Bayesian networks, gan's and classification models. In *International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services*, pages 847–859. Springer, 2021.

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

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

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

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

Tarun Jain, Dinesh Gopalani, and Yogesh Kumar Meena. 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, 2024a.

Tarun Jain, Dinesh Gopalani, and Yogesh Kumar Meena. 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, 2024b.

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

Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of big data*, 6(1):1–54, 2019.

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

Joffrey L Leevy, Taghi M Khoshgoftaar, Richard A Bauder, and Naeem Seliya. A survey on addressing high-class imbalance in big data. *Journal of Big Data*, 5(1):1–30, 2018.

Haoxin Li and Boyang Li. Enhancing vision-language compositional understanding with multimodal synthetic data. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 24849–24861, June 2025.

Peng Liu, Wei Chen, Gaoyan Ou, Tengjiao Wang, Dongqing Yang, and Kai Lei. Sarcasm detection in social media based on imbalanced classification. In *Web-Age Information Management: 15th International Conference, WAIM 2014, Macau, China, June 16-18, 2014. Proceedings 15*, pages 459–471. Springer, 2014.

Shigang Liu, Yu Wang, Jun Zhang, Chao Chen, and Yang Xiang. Addressing the class imbalance problem in twitter spam detection using ensemble learning. *Computers & Security*, 69:35–49, 2017.

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

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

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

Brett PEARY, Rajib Shaw, and Yukiko TAKEUCHI. Utilization of social media in the east japan earthquake and tsunami and its effectiveness. *Journal of Natural Disaster Science*, 34:3–18, 01 2012. doi: 10.2328/jnds.34.3.

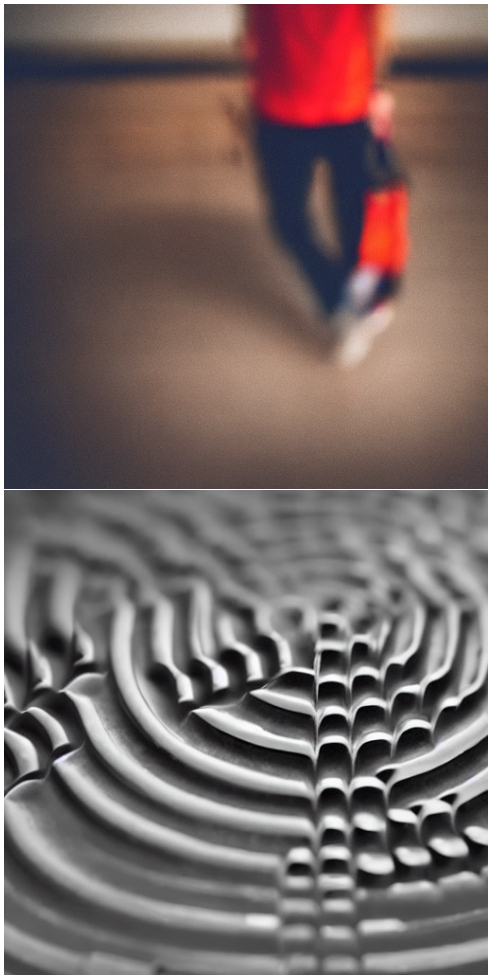
Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. URL https://arxiv.org/abs/1409.1556.

Sumiko Teng and Emily Öhman. Using multimodal models for informative classification of ambiguous tweets in crisis response. In Mika Hämäläinen, Emily Öhman, Yuri Bizzoni, So Miyagawa, and Khalid Alnajjar, editors, *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 265–271, Albuquerque, USA, May 2025. Association for Computational Linguistics. ISBN 979-8-89176-234-3. doi: 10.18653/v1/2025.nlp4dh-1.23. URL https://aclanthology.org/2025.nlp4dh-1.23/.

Jinyan Zhou, Xingang Wang, Jiandong Lv, Ning Liu, Hong Zhang, Rui Cao, Xiaoyu Liu, and Xiaomin Li. Public crisis events tweet classification based on multimodal cycle-gan. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2251–2257. IEEE, 2023.



Tweet:
We just got the evacuation alert. Staying safe and packing up.

Text Label: informative
Image Label: not informative
Final Label: not informative

Tweet:
It smells like smoke everywhere in Fresno.

Text Label: informative
Image Label: not informative
Final Label: not informative

Figure 3: Examples of synthetic ambiguous tweets paired with generated images. Each tweet contains conflicting text and image labels, resulting in an ambiguous but not informative classification.

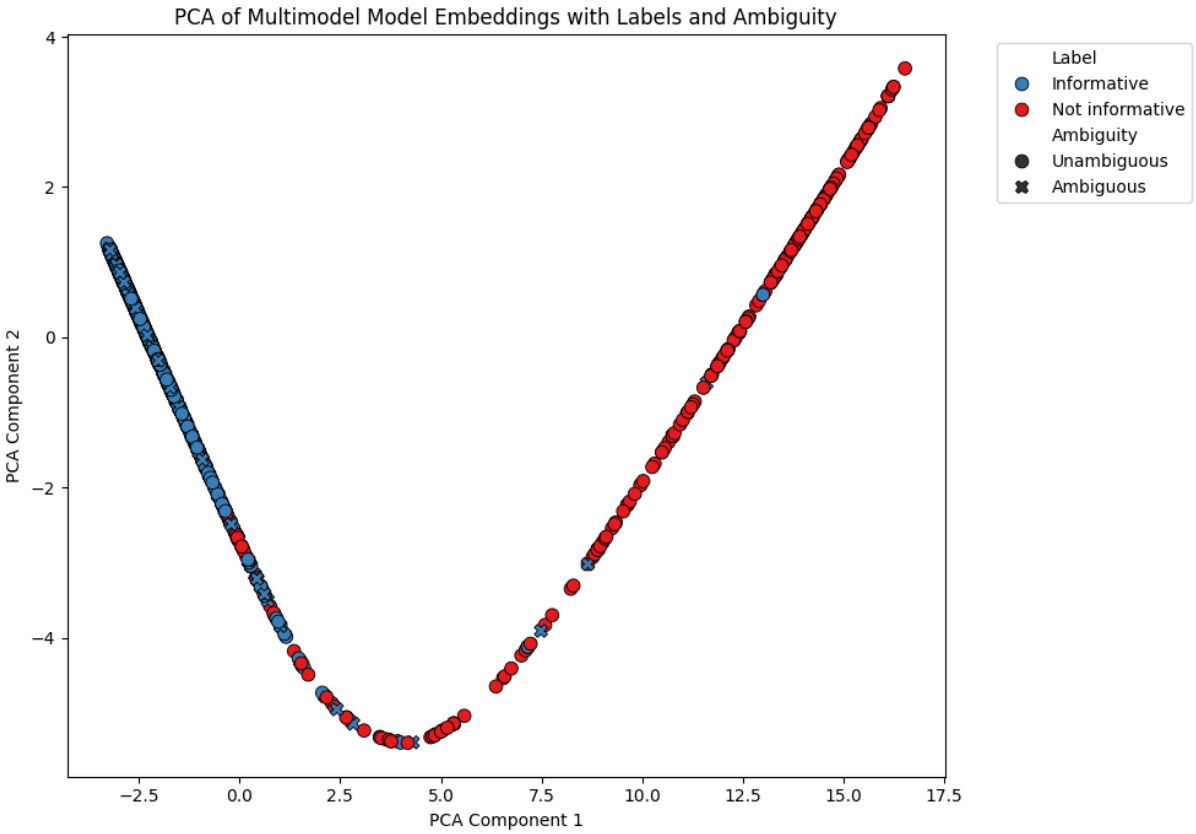


Figure 4: PCA of embeddings from Cross-Attention Multimodal model.

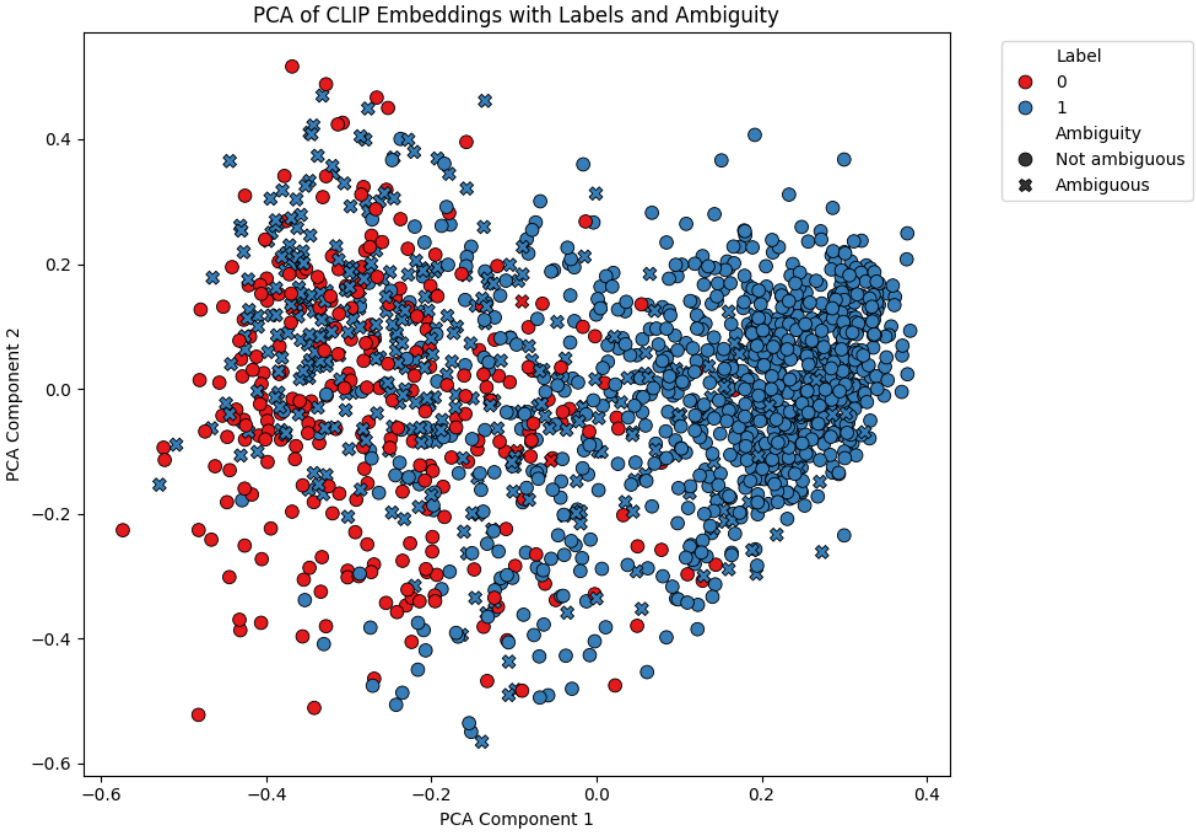


Figure 5: PCA of embeddings from CLIP model.