

# HFFD: Hybrid Fusion Based Multimodal Flood Relevance Detection

Yi Shao<sup>1</sup>, Yang Zhang<sup>1</sup>, Ye Jiang<sup>2</sup>, Wenbo Wan<sup>1</sup>, Jing Li<sup>1,\*</sup> and Jiande Sun<sup>1,\*</sup>

<sup>1</sup>Shandong Normal University, China

<sup>2</sup>Qingdao University of Science and Technology, China

## Abstract

Social media, such as Twitter, has increasingly affected information dissemination and consumption, demonstrating its potential to alarm the upcoming natural disaster beforehand. This paper describes the design of a novel natural disaster event detection method that used hybrid fusion to utilize multimodal information in tweets, called **HFFD** (Hybrid Fusion based Flood Detection). The goal of this work is to discover flood-related event when related information spread at the early stage. The proposed model achieves 91.03 in F1-score, demonstrating its capacity of detecting natural disaster with satisfactory performance.

## 1. Introduction

With the development of online social media techniques, social media allows people to seek and share information more effectively and overcome the barriers of traditional communication such as time lag or geographical constraints. Such characteristic of social media shows its capacity of detecting natural disaster at early stage when the disaster related information starts to spread on social media platform, such as Twitter [2].

This paper discusses the RCTP subtask of MediaEval2022's DisasterMM task [3], which aims to detect flood-related content in tweets based on multi-modal information data on Twitter.

The proposed model adopts the hybrid fusion [4] in the multimodal fusion method, i.e., the model comprehensively adopts the early fusion [5] and the late fusion. Since the early fusion captures the low-level interactions of different modalities, and the late feature integrates a large amount of complex modal information, this method can better deal with the lack of some modalities when the flood-related information first spreads on the network. This way, the model can also make use of the existing modal information to a greater extent, so as to realize the early detection of flood information. In addition, we are also actively exploring the relationship between more modalities and task goals, such as mentions (@), hashtags (#), urls, tweet creation time, posting location, etc. contained in tweets, and strive to find more inspiration. These are described in detail in Section 3 and Section 4.

## 2. Related Work

Early fusion is more of an early exploration of multimodal research. Early fusion refers to the feature-level fusion of the features of different modalities before the decision task, which can better capture the low-level interaction between different modal information. However, due to

---


*MediaEval'22: Multimedia Evaluation Workshop, January 13–15, 2023, Bergen, Norway and Online*

\*Corresponding author.

✉ 2021020981@stu.sdnu.edu.cn (Y. Shao); 2021317099@stu.sdnu.edu.cn (Y. Zhang); ye.jiang@qust.edu.cn (Y. Jiang); wanwenbo@sdnu.edu.cn (W. Wan); lijingjdsun@hotmail.com (J. Li); jiandesun@hotmail.com (J. Sun)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

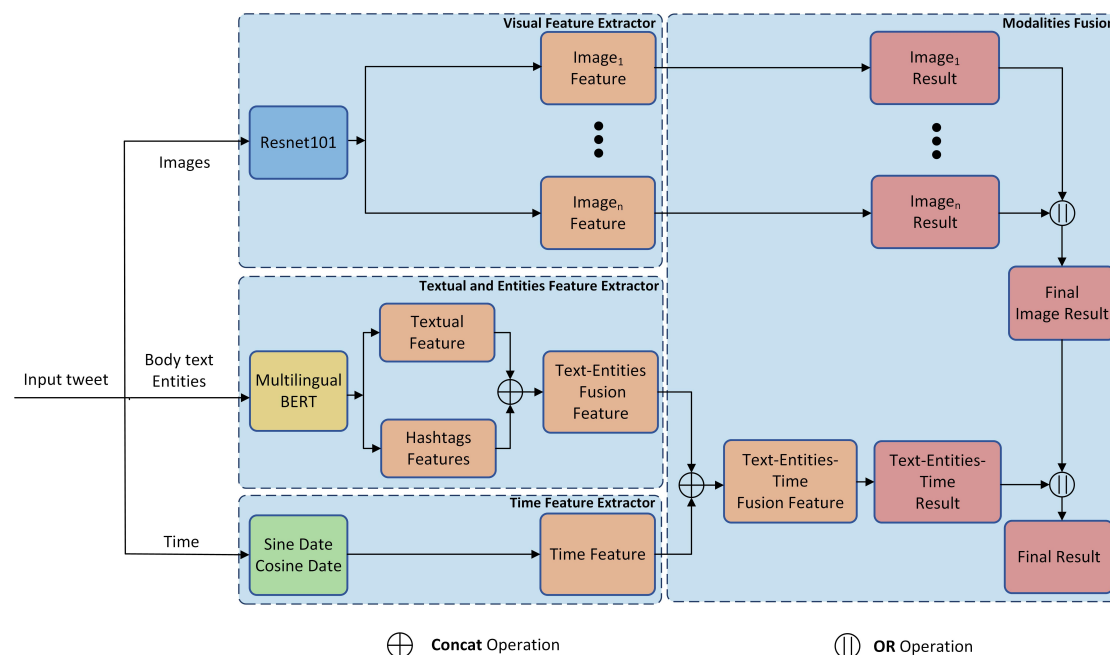
the existence of the modal gap, it is difficult to find a model that can perform transfer learning between more than two modalities, that is, early fusion cannot fully achieve cross-modal feature fusion.

In contrast, late fusion fuses at the final decision level, that is, uses features from different modalities to perform decision-making tasks separately, and finally rationally combining different results according to a cleverly designed mechanism, such as averaging [6], voting schemes [7], weighting on noise [8] or variance [9]. Due to the decision-level fusion of late fusion, it does not need to directly fuse features of different modalities, so that the overall structure of the model has great flexibility compared to early fusion. The ability of late fusion to adapt to large amounts of complex modal data gives it an advantage in flood early tweet data where modalities are often missing. But pure decision-level fusion also ignores low-level interactions between different modalities.

Synthetically, the hybrid fusion approach combines the ability of early fusion to capture feature-level interactions and the ability of late fusion to flexibly cope with complex modal situations, respectively. Hybrid fusion has been successfully applied in multimodal event detection (MED) tasks [10], and the proposed model is inspired by it.

### 3. Approach

The overall flow chart of the proposed model is shown in Figure 1. The model comprehensively utilizes the **body text**, **images**, **entities** (#, urls), and **time** features in the tweet data to detect whether the tweet is related to the specified disaster theme.



**Figure 1:** The overview of proposed model.

Each tweet contains body text and creation time, but not necessarily images and entities, i.e., hashtags (#) or urls. When a tweet sample contains both images and entities, When tweet samples contain both images and entities, the model extracts image features, text-entity features, and temporal features, respectively. The model will perform an **OR** operation based on the

prediction result of each image as the final image result. Then, the model extracts textual feature, hashtags features and time feature respectively, and performs a **Concat** operation to obtain text-entities-time fusion feature. The model performs an **OR** operation on the fusion feature prediction result and final image result again to obtain the final result. If the tweet sample does not contain images or entities, the model will simply ignore the processing of this modality and get the final result.

### 3.1. Handling of Different Modalities

The image feature extractor uses ResNet101 trained on ImageNet and fine-tuned on the task dataset. Each tweet sample contains varying numbers of images, and we input them into ResNet respectively to obtain corresponding image features  $F_i^I$ , where  $i = 1, 2, \dots, n$ .

The Italian dataset is a novel idea, for which we have tried variants of BERT models such as RoBERTa and multilingual BERT as textual feature extractors. In the end, multilingual BERT outperformed others. Entities refer to hashtags (#) and urls contained in the tweet text. It is common to mention related users or organizations in tweets, or use hashtags for topic labeling. The text in the url attached to the tweet is related to the original text of the tweet, and both can be regarded as the text content of the tweet [11]. After concatenating the first sentence of each paragraph of the text in urls with the text of the tweet, we input the multilingual BERT to get the embedding vector as textual feature  $F^{Text}$ . At the same time, each hashtag is input into the multilingual BERT **separately** to ensure that the feature vector  $F_j^H$  of hashtag $_j$  does not contain contextual information, where  $j = 1, 2, \dots, m$ .

Since flood-related tweets increase with the time of the rainy season each year, time feature is also an important modal feature for detecting flood topics. In order to extract the periodicity of time feature in long periods, the year and date of each tweet's creation time are extracted separately in the proposed model, and the time feature are encoded in the form of sine feature  $F_{sin}^{Time}$  and cosine feature  $F_{cos}^{Time}$  respectively. This, we can mine the periodic characteristics in time information.

### 3.2. Multimodal Fusion

Since both the hashtag feature  $F_j^H$  and the text feature  $F^{Text}$  are the same source text information features extracted by BERT, they are directly concatenated to obtain the text-entities fusion feature  $F^{Text-Entities} = F^{Text} \oplus F_1^H \oplus \dots \oplus F_m^H$ . After that, we further concatenate the time feature into  $F^{Text-E}$  to obtain the text-entities-time fusion feature  $F^{Text-Entities-Time} = F^{Text-Entities} \oplus F_{sin}^{Time} \oplus F_{cos}^{Time}$ . Since the number of images contained in each sample is different, We **OR** the prediction results for each image to get the final image result  $R^I = R_1^I \text{ OR } R_2^I \text{ OR } \dots \text{ OR } R_n^I$ . Finally, we consider that one of the image and text-entities-time is related to the flood, so we can conclude that the entire sample is related to the flood, so we **OR** the final result of the image and the text-entities-time result again to get the final result  $R = R^{Text-Entities-Time} \text{ OR } R^I$ .

## 4. Results and Analysis

### 4.1. Textual Feature Extractor Performance Comparison

We tested several different textual feature extractors on the Italian dataset, among which RoBERTa [12] and multilingual BERT are the models officially recommended by huggingface to

deal with Italian text problems. We give them to the plain text data in the development set to classify, and the performance results are shown in Table 1.

**Table 1**

Textual Feature Extractor performance comparison

Textual Feature Extractor	F1-score
LSTM	0.8791
RoBERTa	0.9077
<b>Multilingual BERT</b>	<b>0.9103</b>

## 4.2. Ablation Experiment

As shown in Table 2, we conduct ablation experiments with different modality feature extractors. After introducing entities features, the model performs slightly better than relying only on uncleaned text or only cleaned body text. We also found that the model relying only on image feature performs poor. This is because most sample images do not contain obvious flood-related elements, which makes the image feature extractor undertrained - in fact, the proposed multimodal model has the highest precision and recall on the development set exceeding 0.93, but the precision on the official final test set is down to 0.6741. This is because in the proposed model, there is a "path" that only passes through the image classifier, that is, when the image classifier detects a "flood-related" image, the whole model will directly output the final result. At this time, the performance of the whole model will be affected by the image classifier and become unstable. It's to say, we still need to explore a better late fusion method.

**Table 2**

Modalities Ablation Experiments

Modalities Ablation Models	Precision	Recall	F1-score
uncleaned text	0.9146	0.9161	0.9153
body text	0.9158	0.9168	0.9163
body text & entities	0.9240	0.9236	0.9238
body text & entities & time	0.9226	0.9228	0.9227
images	0.7249	0.7111	0.7113
(body text & entities & time) $\oplus$ images	0.9270	0.9272	0.9271
<b>HFFD</b>	<b>0.9368</b>	<b>0.9347</b>	<b>0.9356</b>

## 5. Discussion and Outlook

Because the development set data of the RCTP subtask was collected in a short *time span* (May 25, 2020 to June 12, 2020), time features has no significant impact on the model performance. But the dataset of the LETT subtask has a long *time span*, in which we derive the periodicity of the number of flood-related tweet creations over time relative to the dates of the rainy season. In addition, some disasters caused by special weather are also related to specific hours. For example, some areas encounter squall line weather, and heavy precipitation will occur in the afternoon and evening. But we did not find hour-level temporal characteristics in the given datasets.

## References

- [1] M. Imran, C. Castillo, F. Diaz, S. Vieweg, Processing social media messages in mass emergency: A survey, *ACM Computing Surveys (CSUR)* 47 (2015) 1–38.
- [2] R. M. Merchant, S. Elmer, N. Lurie, Integrating social media into emergency-preparedness efforts, *New England journal of medicine* 365 (2011) 289–291.
- [3] S. Andreadis, A. Bozas, I. Gialampoukidis, A. Moutzidou, R. Fiorin, F. Lombardo, T. Mavropoulos, D. Norbiato, S. Vrochidis, M. Ferri, I. Kompatsiaris, DisasterMM: Multimedia Analysis of Disaster-Related Social Media Data Task at MediaEval 2022, in: *Proceedings of the MediaEval 2022 Workshop*, Bergen, Norway and Online, 2023.
- [4] P. K. Atrey, M. A. Hossain, A. El Saddik, M. S. Kankanhalli, Multimodal fusion for multimedia analysis: a survey, *Multimedia systems* 16 (2010) 345–379.
- [5] S. K. D’mello, J. Kory, A review and meta-analysis of multimodal affect detection systems, *ACM computing surveys (CSUR)* 47 (2015) 1–36.
- [6] E. Shutova, D. Kiela, J. Maillard, Black holes and white rabbits: Metaphor identification with visual features, in: *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, 2016, pp. 160–170.
- [7] E. Morvant, A. Habrard, S. Ayache, Majority vote of diverse classifiers for late fusion, in: *Joint IAPR international workshops on statistical techniques in pattern recognition (SPR) and structural and syntactic pattern recognition (SSPR)*, Springer, 2014, pp. 153–162.
- [8] G. Potamianos, C. Neti, G. Gravier, A. Garg, A. W. Senior, Recent advances in the automatic recognition of audiovisual speech, *Proceedings of the IEEE* 91 (2003) 1306–1326.
- [9] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, Y. Avrithis, Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention, *IEEE Transactions on Multimedia* 15 (2013) 1553–1568.
- [10] Z.-z. Lan, L. Bao, S.-I. Yu, W. Liu, A. G. Hauptmann, Multimedia classification and event detection using double fusion, *Multimedia tools and applications* 71 (2014) 333–347.
- [11] A. Moutzidou, S. Andreadis, I. Gialampoukidis, A. Karakostas, S. Vrochidis, I. Kompatsiaris, Flood relevance estimation from visual and textual content in social media streams, in: *Companion Proceedings of the The Web Conference 2018*, 2018, pp. 1621–1627.
- [12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, *arXiv preprint arXiv:1907.11692* (2019).