An Exclusion-Based Interface making Social Sensing reliable in Disasters

Yutaka Morino* [1], Mitsunori Matsushita [2], Hiroyuki Fujishiro[3],
[1][2]Kansai University, [3]Hosei University

Abstract

This research addresses the critical challenge of efficiently collecting damage information from social media immediately after a disaster happens. Social media serves as an essential "social sensor," but its data stream is heavily contaminated with noise such as mis/disinformation and extraneous content, severely hindering rapid situational awareness. Given the indispensable need for human verification to ensure information reliability, this paper proposes an interface based on a novel "subtractive strategy." This strategy systematically eliminates unnecessary information, preserving a concentrated, high-relevance information set for human judgment. The core feature is a "stepwise exclusion" mechanism that allows users to iteratively filter out irrelevant data clusters, thereby dramatically improving the visibility of critical damage posts. To verify the system's efficacy, we conducted a user study using a dataset of tweets from the July 2020 heavy rainfall event in Japan. The results demonstrated that while a simple baseline system (text and image presentation) had a slight initial lead in discovered posts, the proposed system's overall performance surpassed the baseline after approximately 20 minutes. These findings collectively indicate that our exclusion-based interface substantially improves the efficiency of extracting critical disaster intelligence from social networking services.

1 Introduction

In large-scale natural disasters such as earthquakes and heavy rain, rapid information collection and information sharing among stakeholders are critical issues to saving lives and accurately assessing damage. While information disseminated by public authorities, administrations, and mass media plays an important role in disaster response, it has also become common for victims themselves to post real-time updates on social media such as X (formerly Twitter)[17][20]. In particular, X can quickly

[1]k790414@kansai-u.ac.jp
[2]m_mat@kansai-u.ac.jp
[3]author3@email.com

disseminate highly timely information along with short texts and photos, thereby providing a certain level of credibility and enabling rapid and widespread transmission of the situation in a wide range of areas[21]. Social media now functions as a large-scale social sensing infrastructure, where innumerable users act as distributed sensors reporting real time their conditions. Municipalities have used social media to support actual rescue operations, demonstrating its usefulness[4].However, the volume of information posted to SNS is enormous, and much of it is noise such as entertainment-related contents [12], good intentions but misleading rumor spreading, and everyday posts from locations far from the disaster area, making information triage difficult [4][18]. During the Noto Peninsula Earthquake on January 1 in 2024, posts requesting rescue were indeed contributed by users who needed help, but "impression farming," mass duplication, and propagation of false information were also observed[5]. According to the Ministry of Internal Affairs and Communications in Japan, approximately 21,000 rescue requests were posted, and many were simple copies of the originals[6]Addressing such situations is increasingly important. Even in cases where local governments were able to utilize SNS effectively, they assigned many dedicated staff to sift information, handle rumors, and assess trustworthiness during such information overloaded of posts. Given that missing critical information must be avoided to save lives, and that human verification is indispensable for reliability, we need information processing systems that support collaboration between human and computer, combining machine processing with human judgment to help users access the information what/when they need it. In the context of social sensing, reliability does not solely depend on algorithmic accuracy but on how effectively humans can discern reliable information amid noisy crowded data. Thus, designing interfaces that support human judging reliability enhancement is essential.

Accordingly, this study aims to realize information triage[3] that collaboratively collects well-reliable disaster information by aggregating and sharing posts likely made by victims and cooperatively determining the veracity and priority of posts. To achieve this, it is important to create a mechanism whereby people on the ground who know the local realities, remote contributors, and computers can work together to sift and complement information[19]. As a preliminary step toward this goal, we clustered disaster-related information on social media and developed an interface to help extract information likely to have been posted by actual victims based on the clustering results.

2 Related research

During major disasters, in addition to routine posts, burst phenomena caused by the spread of disaster information have long been reported [9] and remain prevalent today[7].Similar situations have been reported on

[4]`https://jichitai.works/articles/330` (accessed 2025/10/27).
[5]`https://news.web.nhk/newsweb/na/na-k10014383261000` (accessed 2025/10/27).
[6]`https://www.soumu.go.jp/main\_content/000938666.pdf` (accessed 2025/10/27).
[7]`https://mainichi.jp/articles/20160519/k00/00m/040/059000c` (accessed 2025/10/27).

Weibo, China's largest social media platform. He et al. analyzed help-seeking posts on Weibo during flooding in China and noted that rescue requests are overwhelmed by irrelevant information [6]. To address such disaster situations, many studies have examined the potential of social media during disasters. Japan's National Institute of Information and Communications Technology (NICT) built and provided interfaces called DISAANA(now discontinued) to collect and analyze Twitter posts in real time [10] damage not only for large-scale disasters but also for daily accidents. The system displayed posts related to natural disasters, goods distribution, traffic accidents, and fires, and simultaneously showed posts that potentially contradicted them to support human judgment of veracity.

Song and Fujishiro attempted to extract features of rescue request posts found by NHK, which is one of the biggest Japanese broadcasting corporations, from the cases which actually led to rescues; however, they identified another challenge in that many posts included links to chatbots and news sites [16]. Jain et al. used the CrisisMMD dataset of tweets related to multiple disasters and showed that classifying tweets with OpenAI embeddings improved accuracy compared with conventional methods, underscoring the importance of pretrained GPT models [7]. Rezk et al. also used CrisisMMD and proposed a Multimodal Channel Attention model that integrates both text and images by assigning importance to each modality; they showed improved performance compared with text-only or image-only classification [15].

A major challenge in using social media effectively during disasters is handling rumors, disinformation and fake information. Research has shown that people are more likely to believe or share erroneous information under crisis conditions due to psychological anxiety and emotional factors[5]. Plotnick et al. surveyed how people judge misinformation during disasters; from 341 responses, they found that people tend to rely more on grammar and the credibility of the source than on the content itself[13]. However, in a disaster, posters may not have the time to write grammatically correct sentences, nor is there always enough data to assess the sender's credibility, suggesting limitations to relying on these cues alone. Kawamura and Sasaki demonstrated the potential for automatically extracting misinformation spread on SNS during major disasters[8]; interviews with local governments also identified manpower and time shortages in disaster-management divisions as obstacles to information collection and dissemination. Fujishiro argued that countermeasures against disinformation/misinformation and fake news require appropriately distinguishing among which kind of information circulating online (e.g., news, content, advertising), mentioning a large role for traditional media[2].

As discussed above, handling social media during disasters requires addressing information overload as well as rumors and misinformation. Many studies have attempted to automatically extract useful information or fake/misleading information by computational means. In contrast, considering that missing information from victims in crisis can be life-threatening, and that final reliability assessments require human judgment, our research does not directly extract "useful information." Instead, we adopt an approach that excludes irrelevant information, thereby sup-

porting final human judgment. In other words, by repeatedly removing information that is certainly noise, we aim to realize an information triage system that collaboratively collects high-quality disaster information.

3 Disaster Dataset and photo Clustering

This section describes the target disaster data and clustering. We targeted the July 2020 Heavy Rain Disaster, which occurred from July 3 to July 31, and analyzed images and videos attached to X platform, since posting with images/videos is recommended when reporting their own damage status. We first manually classified the collected tweet images/videos reflecting real-world disaster conditions. We then used CLIP, which enables multimodal processing of natural language and images, to convert images to text and performed k-means clustering. These procedures follow our prior work by Morino et al.[11].

3.1 Target Disaster and X Data

We collected X (formerly Twitter) data using the Twitter API[8] with the keywords "rescue" and "evacuation" as queries. The collection period was set to July 1–15, 2020, to include July 4–7—the period with the heaviest rainfall nationwide. We obtained a total of 476,827 tweets related to the July 2020 Heavy Rain Disaster: 110,261 tweets matched "rescue," 370,531 matched "evacuation," and 3,965 matched both, i.e., duplicates. Among these, the number of tweets that included at least one image or video was 18,197 for "rescue," 34,777 for "evacuation," and 5,540 for both, for a total of 47,434 tweets. Extracting all images and videos from these tweets granted 94,111 media files, each linked to its tweet ID.

3.2 Manual Classification of Heavy Rain Images

To classify the large volume of images, four undergraduate students (hereafter, classifiers) performed manual classification; each image was assigned to a single classifier. To ensure classification validity, the authors subsequently reviewed the results and attempted to remove misclassified data. First, the authors provided classifiers with examples frequently observed via random sampling, such as video game images, images including the "inverse L-shaped" graphic commonly shown during breaking news which are difficult to handle with simple image-processing techniques, photos of TV news screens, and news-related images regardless of media type.

Classifiers then visually inspected each image linked to a tweet (up to four images per tweet) and assigned categories at their discretion. If an image fit multiple classes, the image was duplicated into all relevant classes. As a result, the total number of images/videos after classification was 292,466. We found class variants and overlapping classes; after consulting the classifiers, we merged classes judged to be conceptually identical, we refer to the resulting image classes as "classification classes"

[8] `https://docs.x.com/x-api/introduction` (accessed 2025/10/27).

for readability. Next, among images classified "damage"—which prominently indicate disaster damage—we further subdivided classes, referring to tweet text as necessary. We confirmed that the "damage" set mixed noise such as images likely taken from TV news, online news, and live cameras in the disaster area. In this study, we treat as "damage" only images judged to have been actually taken by users.

Finally, to eliminate misclassifications, 12 other undergraduates (separate from the classifiers) verified whether each image/video was assigned to an appropriate class. This granted a dataset of 33,651 correctly classified images/videos, which we used for implementation in this study.

3.3 Language Conversion and Classification with CLIP

We used CLIP[14](ver. 1.0), which can handle multiple modalities (text and images) and enables image classification using natural language features, to perform image-to-text processing on the classified images. For videos, we used moviepy (ver. 1.0.3) to save a thumbnail image at 1 second from the start and applied the same image-to-text process. Given an input image, CLIP outputs likely words describing what appears in the image. We constrained the output vocabulary to the New General Service List (NGSL)[9] and extracted the top 10 words. We then normalized CLIP probabilities to 1 and treated them as binary vectors to eliminate variance across probability values.

Next, we performed clustering with k-means. We first used the principal content analysis (PCA) method to reduce dimensionality so that the cumulative contribution ratio exceeded 0.9, and used the silhouette score to determine the optimal number of clusters. PCA yielded 659 dimensions; the optimal number of clusters was 17 (highest silhouette score). Table1 lists the number of images/videos assigned to each cluster and "classification class,"and Figure1 shows a scatter plot of the clustering results after dimensionality reduction via PCA. Clusters 1 and 8 are spread along the first principal component (x-axis), suggesting that PCA Component 1 is a major factor in variance, while clusters 5 and 13 spread vertically (in the range of Component 2), capturing within cluster dispersion.

4 Design Guidelines for the Disaster Information Triage Interface

The goal of our interface is to present the clustering results discussed above and extract information that is directly related to the disaster damage. We modeled municipal staff as a use case workers who use SNS under severe human and time constraints to collect disaster damage information and quickly share it with relevant departments and determine the necessity of action. As stated, we assume that human judgment decides necessity and veracity to increase the visibility of posts linked to rescue or damage assessment and to avoid missing candidates for such posts, our

[9] `https://www.newgeneralservicelist.com/` (accessed 2025/10/27).

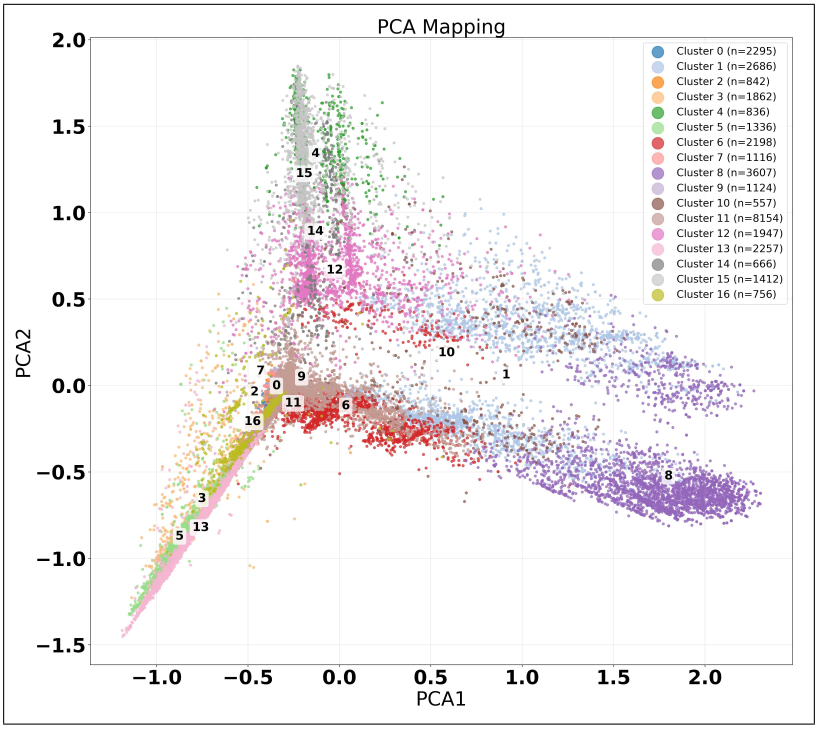

Figure 1: Scatter plot of clustering results

approach is not to extract posts directly but to go through exclusion steps for unrelated or unnecessary information. This section describes the system requirements and functions.

4.1 System Requirements and Implemented Functions

We defined seven requirements and implemented ten functions to satisfy them:

(1) Reproducibility of output data (Fig. 2-①)

Fc(a): Read posts and analysis results stored in a relational database.

(2) Support for judging necessary vs. unnecessary information (Fig 2-②)

Fc(b): 3D mapping of PCA results.

(3) Ability to present posts according to information needs (Fig 2-③)

Fc(c): Search function for posts.

Table 1: Clustering result by CLIP

	Cl:0	Cl:1	Cl:2	Cl:3	Cl:4	Cl:5	Cl:6	Cl:7	Cl:8	Cl:9	Cl:10	Cl:11	Cl:12	Cl:13	Cl:14	Cl:15	Cl:16	Total
game	4	2659	0	3	14	2	2174	0	3606	0	412	1390	418	0	0	41	9	10732
animals	2279	0	35	24	1	3	4	3	0	44	0	1617	44	26	2	2	0	4084
damage	0	0	4	747	1	20	4	0	0	0	0	310	5	1740	1	0	4	2836
landscape	5	3	59	42	4	89	4	106	0	4	1	2103	61	318	3	0	15	2817
plants	2	0	716	0	0	1	0	1007	0	7	0	293	7	25	1	1	0	2060
emergency notification	0	4	0	2	569	2	0	0	0	0	34	32	91	0	9	1188	5	1936
food	0	0	28	0	0	0	0	0	0	1064	0	602	64	0	0	0	0	1758
weather map	0	0	0	1	0	1064	0	0	0	0	1	2	0	0	0	0	71	1139
emergency supplies	0	0	0	2	10	0	0	0	0	1	17	718	126	0	10	10	0	894
Web news	2	0	0	135	42	41	3	0	0	1	20	146	158	40	110	71	34	803
rescue	0	0	0	483	0	0	2	0	0	0	0	179	19	7	0	1	2	693
Twitter	0	8	0	6	74	7	0	0	0	0	15	52	455	8	18	28	4	675
map	0	0	0	2	0	5	1	0	0	0	1	40	12	1	0	0	524	586
Inverse-L	0	2	0	217	0	44	1	0	1	0	17	104	64	38	8	5	42	543
TV news	0	4	0	166	2	53	1	0	0	0	11	120	89	34	3	13	37	533
News paper	0	0	0	12	6	0	0	0	0	0	4	11	3	0	454	0	1	491
TV show	0	5	0	7	1	3	4	0	0	1	10	327	89	5	4	1	2	459
Books	1	0	0	2	43	0	0	0	0	0	6	88	145	0	43	3	2	333
SNS	2	1	0	11	69	2	0	0	0	2	8	20	97	15	0	48	4	279
Total	2295	2686	842	1862	836	1336	2198	1116	3607	1124	557	8154	1947	2257	666	1412	756	33651

(4) Ability to secure posts where damage can be confirmed and review them later (Fig. 2-④)

Fc(d): Post display

Fc(e): Post pickup

Fc(f): Save/export picked posts

Fc(g): Export operation logs

(5) Ability to exclude information in stages (Fig. 2-⑤)

Fc(h): Exclude posts belonging to selected clusters

(6) Ability for users to record priorities (Fig. 2-⑥)

Fc(i): Memo for clusters to prioritize.

(7) Avoid information overload in the display (Fig. 2-⑦)

Fc(j): Switch the number of posts displayed (Max 500).

Fc(k): Set max sampling size for posts (Max 500).

Below, following this section, we describe the interface design (see Figure 2) and the information retrieval procedures and operation methods.

4.2 Database of Analysis Results

To satisfy requirements (1)—(3) as we described in 4.1, we created a database to store analysis results using sqlite3 (ver. 3.43.2), with two main tables: "tweets" and "images." The "tweets" table holds raw tweet text with primary key tweet_id (TEXT) and text (TEXT). As preprocessing, we excluded 4,308 cases where tweet text was missing for the tweet_id, and presented 29,343 images/videos in the interface. The "images" table stores analysis results for each media unit (image or thumbnail extracted from video) with primary key id (INTEGER, AUTOINCREMENT) and tweet_id (TEXT) as a foreign key to join with "tweets." It also stores image_file (TEXT), PCA coordinates pca1, pca2 and pca3 (REAL), cluster (INTEGER), and manual_class (TEXT) which is class information clustered in Section 3.2.

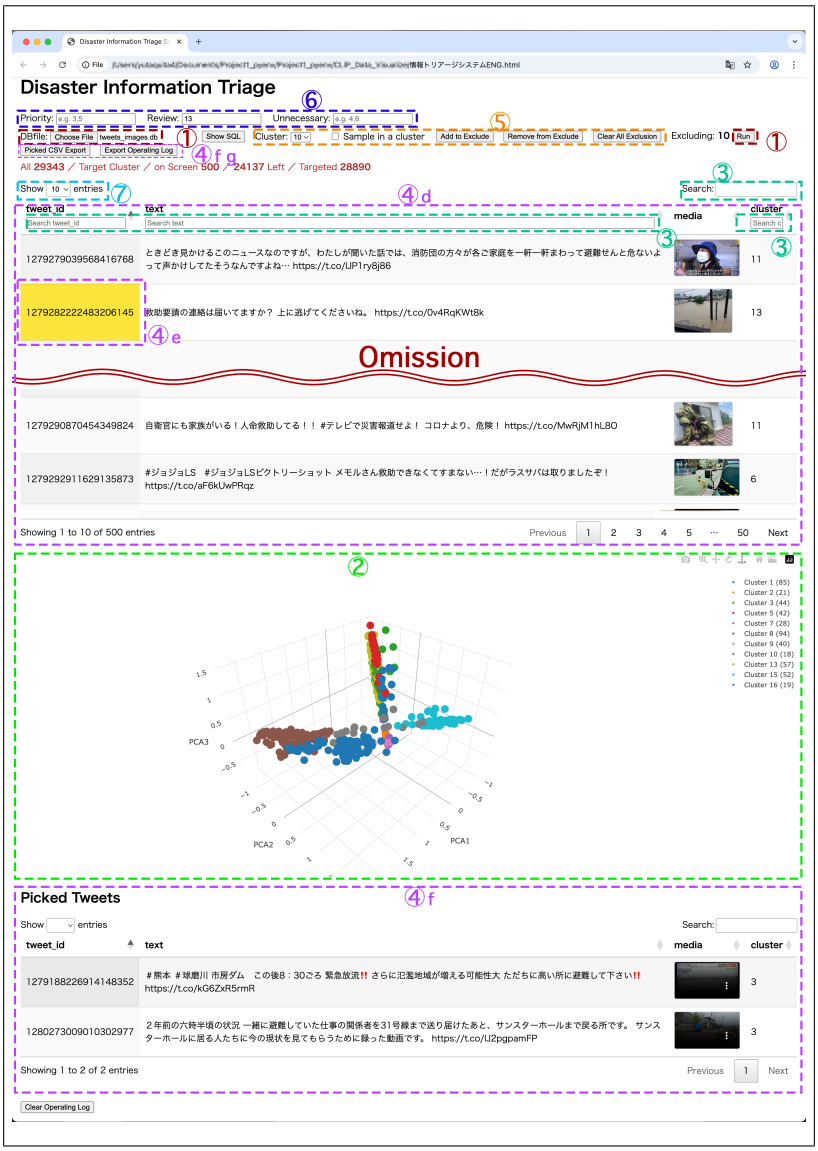

Figure 2: Snapshot of the proposed interface (after database loaded)

This structure enables: (1) consistent output by mapping media to original context via the images–tweets; (2) improved search/aggregation performance by indexing tweet_id, text, or cluster as needed; (3) direct visualization such as scatter plots using PCA values.

4.3 Interface Design and Search Procedure

After creating the relational database, we implemented the following:

(a) Database loading (Req. 1). Users click "Select DB file" (upper left) to load the DB and then "Run" (upper right) to read data (Fig. 2-①).

(b) PCA 3D mapping (Req. 2). Clusters are color-coded, and each cluster can be toggled visible/invisible. Users can zoom in/out via scroll and see post details via mouseover (Fig. 2-②).

(c) Post search (Req. 3). Users can search by entering tweet_id, text, or cluster in the central boxes, or via the "Search" box at the upper right (2-③).

(d–g) Securing and reviewing damage-confirmable posts (Req. 4). The system displays one field per image/video. Because a tweet can attach up to four media files, tweet_id may appear up to four times; duplicates within one sampling (500 items) are displayed consecutively. Clicking a tweet_id picks the post; clicking again unpicks it. Duplicate tweet_ids are automatically marked picked. Picked posts are listed at the bottom and can be saved as comma-separated value (CSV) format via the "Picked CSV Export" button. Operation logs (which clusters were excluded and which posts were picked) can be saved as JSON via the "Export Operation Log" button, facilitating multi-user information sifting. (Fig. 2-④).

(h) Stepwise exclusion by cluster (Req. 5). Users select clusters in the "Cluster:" tab, click "Add to Exclude" to list them under "Excluding:", and then click "Run" to filter out those clusters. Multiple clusters can be excluded at once; users can also "Remove from Exclude" for the selected cluster or "Clear All Exclusions" (2-⑤).

(i) Priority memos (Req. 6). Always-visible memo boxes labeled Highest Priority, To Review, and Unnecessary allow free-form notes (Fig. 2-⑥).

(j, k) Display and sampling limits (Req. 7). Users can switch max on-screen posts to 10/50/500 (all) and set the DB sampling size to a max of 500 per fetch. These limits also apply to mapping and search, reducing the user's verification cost (Fig. 2-⑦).

5 Evaluation Experiment

5.1 Purpose of experiment

We compared the proposed system (with all functions) and a simplified baseline system with restricted functionality (Fig. 3) to evaluate how many posts containing images/videos that confirm disaster damage—assumed to be posted by actual victims—participants could collect. This baseline system was intended to provide a benchmark for evaluating the pick-up efficiency when using typical social media feed. Accordingly it was designed without clustering information, the 3D map, or exclusion functions.

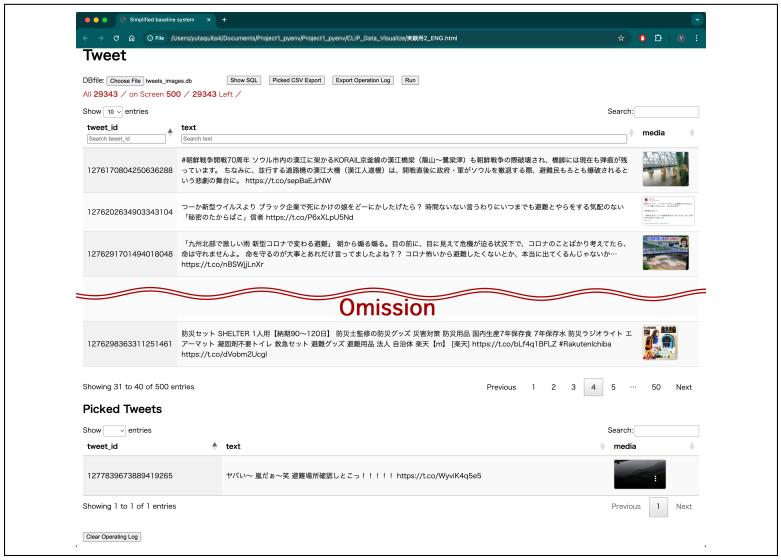

Figure 3: The simplified baseline system used in the experiment

5.2 Overview

Participants were 26 students/alumni from a university who studied informatics: 13 used the proposed system and 13 used the simplified system. Two participants in the proposed group were excluded from the analysis because observation during the experiment indicated they had not sufficiently understood the assigned task (e.g., engaging in a completely different task, such as watching only video game contents instead of searching for information). Participants were shown an experimental story assuming real disaster use and tasked with collecting SNS posts that confirm disaster damage. The story consisted of: (1) A large-scale natural disaster (heavy rain) is occurring now; (2) The participant is a municipal disaster management officer who must quickly and comprehensively find posts such as actual damage has occurred or residents are seeking rescue; (3) Posts that are difficult to judge should be picked for later discussion with other users.

5.3 Experiment Procedure

The steps were: (1) outline, (2) consent form, (3) pre-questionnaire (SNS usage, experience during disasters, PC proficiency), (4) task explanation, (5) story explanation, (6) system operation explanation, (7) practice (2 min), (8) main task (30 minutes from when posts were first displayed after DB load), and (9) post-questionnaire/interview (system improvements, whether cluster display helped, what information was prioritized, and what information allowed judging damage). For the proposed system group, we also explained clusters, how to interpret and operate the PCA

Table 2: Test results for differences in the number of damage-confirming images/videos collected by users

Period	Proposed_mean	Proposed_SD	Baseline_mean	Baseline_SD	u_stat	u_p	Cliff's d
0-5	3.6364	4.3422	8.0769	5.2195	37.5000	0.0508	-0.3986
5-10	16.7273	17.7994	22.000	12.9808	46.5000	0.1553	-0.3427
10-15	37.9091	26.2277	35.5385	17.8868	73.5000	0.9307	-0.0209
15-20	60.7273	35.4432	47.0000	25.1992	88.5000	0.3388	0.2308
20-25	97.4545	43.5692	60.3846	28.2682	108.5000	0.0342	0.5455
25-30	128.1818	63.8229	74.4615	36.7641	108.0000	0.0369	0.5105

Table 3: Accuracy of damage-confirming images/videos collected by users

Period	Proposed_mean	Baseline_mean
0-5	0.4089	0.5088
5-10	0.6017	0.6733
10-15	0.6822	0.6905
15-20	0.7983	0.6860
20-25	0.8375	0.7016
25-30	0.8434	0.7096

3D map, and the insights obtainable from it.

5.4 Experiment Result

We gathered posts with images/videos confirming damage collected within 30 minutes and compared results for the proposed group vs. the baseline group. Considering the proposed system's complexity, we aggregated every 5min period. Since the Shapiro-Wilk test rejected the assumption of normality for 4 out of the 12 items ($p<0.05$), non-parametric tests the Mann-Whitney U test were adopted for the subsequent analyses. We then U-tests to examine differences in counts (see Table 2). Fig. 4 and Fig. 5 show the cumulative number of picked posts over time and the mean for each group. The accuracy that user picked-posts fell into the "damage"(from manual classification) was calculated , and its mean is shown in Table 3.

6 Discussion

This study contributes to research on the use of social media during disasters by demonstrating —in a time-constrained experimental task— the effectiveness of a reverse design principle that prioritizes the reliable exclusion of irrelevant information over the mainstream approach of directly extracting useful information. We show that an information exclusion-based interface centered on image clustering and stepwise exclusion improves the efficiency of picking posts that confirm damage under time constraints. After the 20 minutes, the proposed group significantly outperformed the baseline group in the 20–25 min period($U = 108.000$, $p = 0.0342$, $d = 0.5455$, with an average gain of about 30.07 posts),and 25–30 min period($U = 108.000$, $p = 0.0369$, $d = 0.5105$, with an average gain

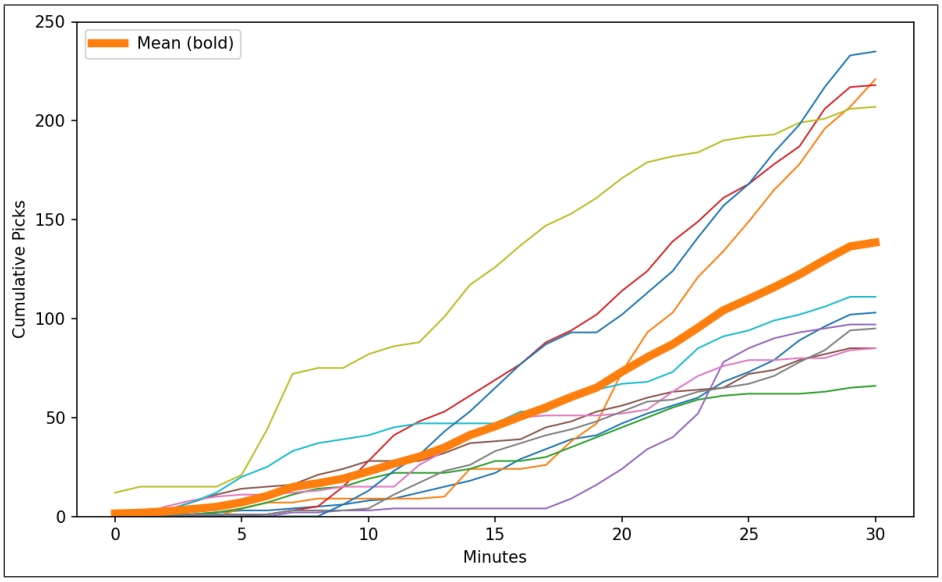

Figure 4: Cumulative number of picked posts for users of the proposed system

of about 53.99 posts). In contrast, during 0-20 min the baseline group outperformed the proposed group; performance then tended to reverse at 10-20 min period, and the gap widened toward the end. The longest time until the first pick of a damage confirming post was 554.937 sec in the proposed group versus 224.447 sec in the baseline group; the number of users who took $\geq$200 sec to make their first pick was 4 in the proposed group and 1 in the baseline group. These results suggest a learning curve: once users understand higher-level functions that carry learning costs (cluster exclusion, PCA 3D map, and combined search), their benefits compound over time.Therefore, for the future version of the system, we plan to provide additional explanation or help page to help users understand about what the cluster or 3D map indicate which was most difficult to understand for users.

In both the proposed and baseline systems, users achieved a reasonable level of accuracy. The key objective of this study is to examine how many posts users judge as damage-related and worthy of further human verification—that is, posts whose authenticity or details should be checked. Accordingly, our goal is not to maximize accuracy to 1.0. Since there was no significant difference in accuracy between the two systems, we conclude that users could identify relevant information with either system.

From the experiment result, we infer two interacting reasons why did the interface yield a significant late-stage gain: (i) stepwise exploration via exclusion that exploits the clustered concentration of noise, thereby reducing the variety of visible information; and (ii) the interaction among cluster display, 3D visualization, search, and exclusion functions that sur-

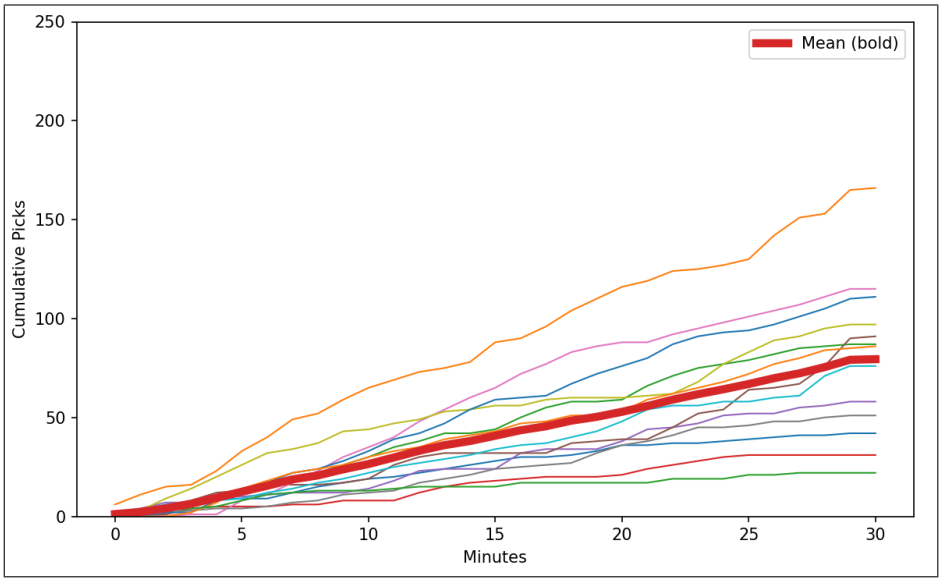

Figure 5: Cumulative number of picked posts for users of the baseline system

faces candidate images of "damage."

For (i), the cluster-wise distribution in Table 1 indicates that categories manually labeled as noise (e.g., (game: Cl 1,Cl 8), (animal: Cl 0), (TV show: Cl 11), (weather maps: Cl 5), (food: Cl 9) are concentrated in specific clusters. This observation confirms that the resulting clusters maintain semantic consistency, possibly separating semantically distinct noise categories. Excluding those clusters makes much of the non-disaster noise invisible, leaving a remainder in which clusters rich in "damage" images (e.g., Cl:3, Cl:13) become relatively more visible (Table2). For (ii), the proposed system is deliberately designed so that the PCA 3D map (function b) visualizes spatial cluster layouts; users can iteratively lower noise density via exclusion (function h) while securing candidates through search (function c). This cycle of visualization, exclusion, re-visualization helps relevant clusters "float to the interface." The PCA 3D map enables users to grasp neighboring clusters visually and intuitively; finding one noise-heavy cluster often led to chain discoveries of others, and the same mechanism helped surface clusters containing damage images. In our observations, 6 of 11 participants in the proposed group frequently consulted the PCA 3D map compared with those who did not, they tended to pick earlier and exclude more clusters. Overall, these patterns suggest that prioritizing the early exclusion of noise clusters —our stepwise exclusion strategy— amplifies the visibility of clusters likely to contain reliable "damage" images through the synergy of visualization and interaction.

2 Although these clusters included "damage," they also contained many non-disaster images (e.g., game, animals, landscape); some partici-

pants therefore excluded these clusters, raising a concern about missed information. Potential remedies include re-clustering only the broad, mixed clusters and/or adopting clustering tuned to pursue near 100% recall. Second, our dataset covers a single disaster; verification is essential to determine whether similar phenomena hold for other disasters and outside Japan. As future work, we will investigate whether applying same clustering method to CrisisMMD datasets[1] which is SNS open-data across different parts of the World.

Interface issues observed in post-experiment questionnaires and user observations include: accidental full clearing of excluded clusters; unpicking of already picked tweets; repeated pressing of the Run button during database loading that inadvertently altered the loaded sample; and, when many posts were shown at once (e.g., 50 items), moving to the next page with the scroll position retained, forcing users to scroll back to the top to resume review. Moreover, the current system lacks multi-user sharing functions. Future work should therefore (i) harden the interaction design to prevent such misoperations. (e.g., the pop-up window to confirm for such the complicated operation, (ii) support collaborative and sifting-based workflows reflective of real disaster operations, and (iii) extend the pipeline to share disaster-damage relevant information with stakeholders.

7 Conclusion

This paper presents an exclusion-based interface that improves the reliability of social sensing in disasters by the stepwise exclusion of irrelevant clusters, rather than the direct extraction of useful posts. The system combines image clustering with a controllable exclusion mechanism so that operators can progressively exclude noise and concentrate their attention on posts where disaster damage can be verified. In our study using tweets from the July 2020 heavy rainfall event in Japan, the proposed interface significantly outperformed a simplified baseline system in the final period of the task 20–25 min period($U = 108.000$, $p = 0.0342$, $d=0.5455$, with an average gain of about 30.07 posts),and 25–30 min period($U = 108.000$, $p = 0.0369$, $d=0.5105$, with an average gain of about 53.99 posts), despite a small initial lag attributable to learning the advanced functions.

The final period of task advantage is consistent with two interacting mechanisms observed in the logs and user behavior: (i) excluding clusters that concentrate noise unrelated with disaster categories (e.g., games, animals, TV show, weather maps, web news) reduces the visible variety of noise; and (ii) the synergy of cluster display, 3-D visualization, search, and exclusion surfaces clusters that contain damage images. From this point of view, doing reliable exclusion first, then collect important information offers a practical path to more trustworthy social sensing under time pressure.

The current implementation has limitations, including simplified handling of CLIP outputs (top words normalized as binary vector), a single-disaster dataset, and several interaction issues (e.g., accidental clearing of exclusions, paging/scroll quirks). Future work will need (a) refine clustering with soft probability features and re-clustering of broad mixed clusters,

(b) extend evaluation across different disaster types and regions, and (c) strengthen collaboration and sharing workflows to better match real operations.

Acknowledgment

This work was supported by JST RISTEX, Grant Number JPMJRS23L2.

References

[1] 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.

[2] Hiroyuki Fujishiro. Addressing directionality: Dis/misinformation and fake news in the digital sphere. In *IPSJ SIG Technical Report*, volume 2024-EIP-104, pages 1–6, 2024. Japanese edition.

[3] Hiroyuki Fujishiro, Mitsunori Matsushita, and Morihiro Ogasawara. Effective use of social media in large-scale disasters: The applicability of information triage. *Socio-Informatics*, 6(2):49–63, 2018.

[4] Saptarsi Goswami, Sanjay Chakraborty, Sanhita Ghosh, Amlan Chakrabarti, and Basabi Chakraborty. A review on application of data mining techniques to combat natural disasters. *Ain Shams Engineering Journal*, 9(3):365–378, 2018.

[5] Aditi Gupta, Hemank Lamba, and Ponnurangam Kumaraguru. $1.00 per rt #bostonmarathon #prayforboston: Analyzing fake content on twitter. *2013 APWG eCrime Researchers Summit*, pages 1–12, 2013.

[6] Changyang He, Yue Deng, Wenjie Yang, and Bo Li. "help! can you hear me?": Understanding how help-seeking posts are overwhelmed on social media during a natural disaster. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2), November 2022.

[7] 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.

[8] Takeshi Kawamura and Yuji Sasaki. Spreading misinformation on sns in the event of a great disaster and problems with action by local government. Technical Report 2, Hokkaido Research Organization Building Research Department Northern Regional Building Research Institute, nov 2022.

[9] Jon Kleinberg. Bursty and hierarchical structure in streams. *Data mining and knowledge discovery*, 7(4):373–397, 2003.

[10] Junta Mizuno, Masahiro Tanaka, Kiyonori Ohtake, Jong-Hoon Oh, Julien Kloetzer, Chikara Hashimoto, and Kentaro Torisawa. Wisdom x, disaana and d-summ: Large-scale nlp systems for analyzing textual

big data. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 263–267, 2016.

[11] Yutaka Morino, Hiroyuki Fujishiro, and Mitsunori Matsushita. Investigation on separation of noise information for information gathering from social media in disaster situations. volume 27, pages 133–140, 2025. Japanese edition.

[12] Yutaka Morino, Mitsunori Matsushita, and Hiroyuki Fujishiro. Vocabulary cross-contamination between entertainment content and disaster-related social media posts. In *2024 International Conference on Information and Communication Technologies for Disaster Management*, pages 1–6, 2024.

[13] Linda Plotnick, Starr Roxanne Hiltz, Sukeshini A. Grandhi, and Julie Dugdale. Real or fake? user behavior and attitudes related to determining the veracity of social media posts. *CoRR*, abs/1904.03989, 2019.

[14] 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. In *International conference on machine learning*, pages 8748–8763, 2021.

[15] Mariham Rezk, Noureldin Elmadany, Radwa K Hamad, and Ehab F Badran. Categorizing crises from social media feeds via multimodal channel attention. *IEEE Access*, 11, 2023.

[16] Chenjie Song and Hiroyuki Fujishiro. An examination of the features of t rescue-request tweets—with the case of torrential rainfall in july 2020 —. *IEICE Technical Report*, 120(166):18–23, 2020. Japanese edition.

[17] Bruno Takahashi, Edson C. Tandoc, and Christine Carmichael. Communicating on twitter during a disaster: An analysis of tweets during typhoon haiyan in the philippines. *Computers in Human Behavior*, 50:392–398, 2015.

[18] Sudha Verma, Sarah Vieweg, William Corvey, Leysia Palen, James Martin, Martha Palmer, Aaron Schram, and Kenneth Anderson. Natural language processing to the rescue? extracting" situational awareness" tweets during mass emergency. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 385–392, 2011.

[19] Megumi Yasuo, Yutaka Morino, and Mitsunori Matsushita. Media characteristics of social network services in collecting disaster information. In *Proc. 30th JSAI SIG-AM*, pages 47–54, 2023.

[20] Lei Zou, Danqing Liao, Nina S.N. Lam, Michelle A. Meyer, Nasir G. Gharaibeh, Heng Cai, Bing Zhou, and Dongying Li. Social media for emergency rescue: An analysis of rescue requests on twitter during hurricane harvey. *International Journal of Disaster Risk Reduction*, 85:103513, 2023.

[21] Arkaitz Zubiaga and Heng Ji. Tweet, but verify: epistemic study of information verification on twitter. *Social Network Analysis and Mining*, 4(1):1–12, Jan 2014.