

Explicit and Implicit Data Augmentation for Social Event Detection

Congbo Ma^{1,2,3}, Yuxia Wang², Jia Wu¹, Jian Yang¹, Jing Du¹,
Zitai Qiu¹, Qing Li², Hu Wang², Preslav Nakov²

¹ Macquarie University, Sydney, Australia

² Mohamed bin Zayed University of Artificial Intelligence, UAE

³ New York University Abu Dhabi, UAE

Abstract

Social event detection involves identifying and categorizing important events from social media, which relies on labeled data, but annotation is costly and labor-intensive. To address this problem, we propose **Augmentation framework for Social Event Detection (SED-Aug)**, a plug-and-play dual augmentation framework, which combines explicit text-based and implicit feature-space augmentation to enhance data diversity and model robustness. The explicit augmentation utilizes large language models to enhance textual information through five diverse generation strategies. For implicit augmentation, we design five novel perturbation techniques that operate in the feature space on structural fused embeddings. These perturbations are crafted to keep the semantic and relational properties of the embeddings and make them more diverse. Specifically, SED-Aug outperforms the best baseline model by approximately 17.67% on the Twitter2012 dataset and by about 15.57% on the Twitter2018 dataset in terms of the average F1 score. The code is available at [GitHub](https://github.com/congboma/SED-Aug)¹.

1 Introduction

Social event detection (SED) identifies and classifies notable events on social media platforms (Peng et al., 2021a; Cao et al.; Ren et al., 2022b; Cao et al., 2021; Ma et al., 2025). These events, distinct from general occurrences, are characterized by their origin and propagation through user interactions, reflecting collective activity. SED is commonly defined as a classification task that analyzes social network data, including textual content (e.g., messages) and structural features (e.g., user metadata and activity logs), to detect emerging events. These insights can be important for various applications such as crisis management (Pekar et al., 2020), public opinion analysis (Peng et al., 2021b),

and financial market analysis (Nisar and Yeung, 2018). However, one key challenge in SED is its reliance on limited labeled data, which demands human annotation and hampers the generalization of models to diverse event contexts. (Qiu et al., 2024b).

Large language models (LLMs) excel in natural language understanding and generation, enabling the creation of diverse textual variations. Despite their potential, LLMs have not yet been applied to the SED task. When used for textual augmentation, LLMs generate social message variations, enriching SED training data and improving model robustness. This approach also reduces computational overhead during testing by handling intensive computations in advance. Once data are augmented, the event detection can be conducted without invoking the LLMs, thus eliminating recurring time costs associated with model inference (usually, LLMs inference will be time consuming). Moreover, Direct LLM-based predictions are expensive due to API or cloud fees. Using LLMs solely for augmentation enhances data diversity while optimizing computational and financial resources.

Focusing solely on textual data overlooks crucial structural information that captures user and event interactions in social media (Qiu et al., 2024b). Common SED methods use graph-based approaches to generate structure-fused embeddings (Cao et al., 2021; Peng et al., 2022; Ren et al., 2022a; Qiu et al., 2024b; Li et al., 2024; Ma et al., 2024). However, while LLMs excel at augmenting textual data, they are less effective at processing graph-based data (Jin et al., 2024). To address this, we extend data augmentation to structural domain, which captures both the information of social messages and the relational patterns among associated metadata. This helps incorporate the structural information and diversify the data from a different perspective to further improve SED performance.

¹<https://github.com/congboma/SED-Aug>

In this paper, we propose **Augmentation** framework for **Social Event Detection (SED-Aug)**, a dual plug-and-play data augmentation framework for SED. It includes explicit augmentation using LLMs to enhance textual diversity, and implicit augmentation, which perturbs structure-fused embeddings in the feature space. Explicit augmentation consists of one-stage and two-stage strategies. One-stage includes paraphrasing, context addition, style transfer, and paraphrasing with entity preservation. Two-stage first extracts key information using LLMs, then rewrites it into diverse messages. These strategies enable the generation of a wide range of augmented messages, increasing the variability and robustness of the training data. Implicit augmentation introduces five novel perturbation techniques that operate directly in the feature space of structure-fused message embeddings. These methods include Gaussian Perturbation (GP), Proportional Gaussian Perturbation (PGP), In-Distribution Gaussian Perturbation (IDGP), Clipped Gaussian Perturbation (CGP), and Frequency-Domain Perturbation (FDP). Each perturbation is designed to modify the embeddings while preserving their semantic and relational properties, ultimately enhancing the model to capture intricate patterns within data. This dual augmentation framework leverages both the textual and structural information inherent in SED, effectively enhancing data diversity and capturing more complex patterns within the underlying social graph. By addressing the unique characteristics of SED, it maximizes the potential of both semantic and structural embeddings, leading to more robust and accurate detection outcomes. Our main contributions can be summarized as follows:

- We propose a novel dual data augmentation framework SED-Aug tailored for SED, integrating both explicit and implicit augmentation strategies. This framework is plug-and-play and can be seamlessly integrated into a SED model to enhance its performance and robustness.
- We are the first to incorporate LLMs in addressing the SED task, developing multiple augmentation strategies for message content. These include five one-stage and two-stage techniques that increase data diversity without the need for manual labeling.
- We propose five perturbation methods for the feature space, which operate on structural fused message embeddings. This integration

of structural information into the augmentation process enhances the robustness and performance of SED models.

- The proposed augmentation framework achieves state-of-the-art results on several SED datasets, outperforming the best baseline model by approximately 17.67% on the Twitter2012 dataset and about 15.57% on the Twitter2018 dataset in terms of average F1 score, supported by comprehensive analysis validating its effectiveness.

2 Related Works

2.1 Social Event Detection

Early SED research primarily used content-based methods focusing solely on text semantics (Wurzer et al., 2015; Wang and Zhang, 2017; Yan et al., 2015), neglecting the importance of social interactions and the heterogeneous nature of social media data, such as user connections, interactions, and spatiotemporal information (Toivonen et al., 2019; Martí et al., 2019). As a result, this limitation restricts the ability to capture a comprehensive range of information, leading to missed insights from diverse data types (Ren et al., 2022b). Graph-based methods (Kipf and Welling, 2017; Velickovic et al., 2018; Hamilton et al., 2017; Wu et al., 2020; Zhang et al., 2019) address this by leveraging both textual and structural data through heterogeneous information networks (Sun and Han, 2012), modeling complex social media interactions (Cao et al., 2021; Peng et al., 2022; Ren et al., 2022a; Qiu et al., 2024b; Li et al., 2024). However, these methods do not fully leverage existing data to enhance data diversity, overlooking the potential of data augmentation techniques to improve model performance, particularly in situations with limited labeled data.

2.2 Data Augmentation

Data augmentation is crucial in NLP for enhancing textual diversity and model performance, particularly with limited labeled data (Feng et al., 2021). It can be categorized into (1) input text level and (2) feature space level augmentation. For (1) text level, various token and sentence level modifications diversify training data. Common strategies include insertion (Xie et al., 2020), deletion (Wei and Zou, 2019), and masking (Ghosh et al., 2023; Yu et al., 2023) to improve generalization. Replacement techniques (Kobayashi, 2018) include

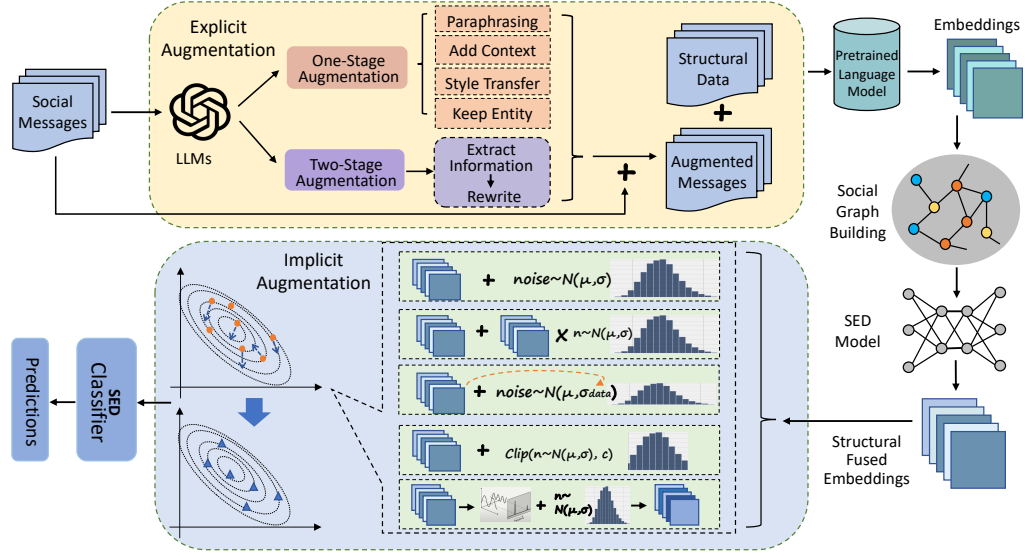


Figure 1: The framework of SED-Aug model.

synonym substitution, entity replacement (Liu and Cui, 2023), and semantic modifications (Zhuang et al., 2022). Other approaches include paraphrasing and back-translation (Sennrich et al., 2016) for generating semantically equivalent variations. For (2) feature space level, augmentation transforms feature representations (DeVries and Taylor, 2017). For example, Ang et al. (Lv et al., 2023) modeled dialogue trajectories using a Gaussian process, while Wang et al. (Wang et al., 2019) estimated class-wise covariance matrices to generate synthetic data, optimizing cross-entropy loss. However, these methods focus on textual and overlook structural information, limiting their effectiveness in social network analysis. Unlike prior work that treats textual and structural augmentation separately, we propose SED-Aug, a dual data augmentation framework for SED. By integrating explicit and implicit strategies, we enhance data diversity and model robustness in low-resource settings.

3 Methodology

In this section, we present the SED-Aug framework, as illustrated in Figure 1. We first apply explicit augmentation to social media messages using LLMs, leveraging five strategies to enhance the user-generated content. We then combine these augmented messages with the original messages and integrate them with the corresponding structural data of each message. Next, we use a pre-trained language model to extract embeddings from this enriched data. We construct a social graph that captures the connections between

the textual content and the structural information. Through a graph aggregation method, we generate structural-fused message embeddings that encapsulate both semantic and relational information. These structural-fused embeddings are then subjected to implicit augmentation to create more diverse embeddings in the feature space. Finally, we integrate these augmented embeddings with the original structural-fused embeddings for the downstream classification task, improving the robustness and adaptability of the SED model.

3.1 Explicit Data Augmentation through LLMs

We introduce five explicit augmentation techniques using LLMs to enhance SED. These methods modify the message text while preserving metadata to create diverse messages that can enhance the robustness and generalization capabilities of models for SED. Given an input message:

$$m_i^k = LLM^k(m_i) \quad (1)$$

where k represents different augmentation types. The five methods are described as follows:

Paraphrasing. Generating alternate versions of the original message while preserving its meaning, maintaining semantic integrity with variation in wording and structure.

Adding Context. Expanding the original message by adding relevant contextual information. This augmentation helps provide a broader understanding of the content, enriching the message with details that enhance its clarity and relevance.

Style Transfer. Modifying the writing style of the message without altering its core meaning. The style perturbation can include shifts in tone, formality, or other stylistic attributes to produce varied versions of the same message.

Keep Entity Unchanged. Entities are crucial for SED, representing key components like people, places, dates, or topics that provide context and relevance to the message content. Maintaining these entities preserves critical event details, helping the model distinguish and understand events. This approach modifies the message text while ensuring key entities remain unchanged, generating diverse variations while retaining essential information.

Extract Informative Information and Rewrite. This method uses an LLM to extract key components from the original message and generate a revised version that preserves essential information while modifying its presentation. We focus on (1) the selection of keywords is determined by the background knowledge of the LLM; (2) given the importance of entities in SED, we specifically guide the LLM to concentrate on entity extraction for more targeted augmentation. The LLM extracts key entities (e.g., names, locations, dates) and integrates them into a restructured message, enhancing diversity while preserving core information; (3) a knowledge graph is a structured framework that captures relationships among entities, providing a clearer understanding of the connections within the data. The LLM rewrites the message using this structured information, enriching the text while preserving key relationships.

3.2 Implicit Data Augmentation on the Latent Space

In the latent space, we aim to introduce minor variations that enhance data diversity by applying perturbations to augment the structural fused message embeddings g^i , where $g^i \in G, i = 1, 2, \dots, N$. G is the set of structural fused message embeddings. N is the number of message samples. This augmentation process improves the model’s generalization capability. During training, a probability threshold α is employed to determine whether to train on the augmented embeddings or the original embeddings, which can be expressed as follows:

$$g^i = \begin{cases} g_{\text{augmented}}^i, & \text{if } p < \alpha, \\ g^i, & \text{if } p \geq \alpha, \end{cases} \quad (2)$$

$$p \sim \text{Uniform}(0, 1). \quad (3)$$

Here, p is a real number sampled from a uniform distribution. If $p < \alpha$, we train on the augmented embeddings to introduce diversity; otherwise, we use the original embeddings to maintain stability. We have developed five different implicit data augmentation methods for enhancing the latent space in this manner.

Gaussian Perturbation (GP). For each input structural fused message embeddings g^i , we add Gaussian noise sampled from a normal distribution with a mean of zero and a standard deviation of σ . The noise is applied element-wise to each dimension, generating a perturbed version of the original feature vector g_{GP}^i . Formally, the augmented feature is computed as:

$$g_{GP}^i = g^i + n_{GP}, n_{GP} \sim \mathcal{N}(0, \sigma^2), \quad (4)$$

Proportional Gaussian Perturbation (PGP). In the proposed GP, the scale of the perturbation is independent of the original data scale. This can lead to noise that is disproportionately large or small relative to the magnitude of the input data, potentially distorting the underlying structural information. To address this, we propose proportional Gaussian perturbation (PGP), where the added noise is scaled in relation to the input data in the feature space. Specifically, the noise is generated such that its magnitude is proportional to the values of the feature vector itself. Formally, we have:

$$g_{PGP}^i = g^i + n_{PGP}, n_{PGP} \sim \mathcal{N}(0, \sigma^2) \cdot G, \quad (5)$$

In this perturbation, the noise is not only dependent on the standard deviation but also scales with the absolute value of the original feature vector. This helps that the perturbation is contextually relevant, enhancing the robustness of the feature augmentation process while preserving the intrinsic relationships within the data.

In-Distribution Gaussian Perturbation (IDGP). In the GP and PGP, the standard deviation σ should be manually predefined, rather than being dynamically adapted based on the distribution of input data. Therefore, we proposed In-Distribution Gaussian Perturbation (IDGP) that the noise is generated based on the statistical properties of the input data, specifically its standard deviation. For each feature dimension, the standard deviation σ is computed from the structural fused message embeddings g^i , and Gaussian noise with zero mean and a standard deviation proportional to this g^i is added. Formally:

$$g_{IDGP}^i = g^i + n_{IDGP}, n_{IDGP} \sim \mathcal{N}(0, \alpha * \text{std}(G)^2), \quad (6)$$

where α is the variance control parameters. By adapting the noise magnitude to the inherent variability of the data (through $\text{std}(G)$), this perturbation ensures that the noise is scaled appropriately for each feature dimension. This prevents excessive perturbation in low-variance features and ensures sufficient noise is applied in high-variance features. Furthermore, as the perturbations stay within the range of the original data’s distribution, this method maintains the data’s overall structure and prevents the creation of unrealistic or outlier data points.

Clipped Gaussian Perturbation (CGP). This method constrains Gaussian noise within a specified range to maintain balanced perturbation. Noise is sampled from a standard normal distribution and clipped within $[-c, c]$, where c is a small constant. This prevents the noise from being negligible or excessive, maintaining a balanced perturbation effect. The augmented feature is calculated as follows:

$$g_{CGP}^i = g^i + n_{CGP}, \quad \text{Clip}(n_{CGP} \sim \mathcal{N}(0, \sigma^2), c), \quad (7)$$

CGP transforms its distribution into a truncated normal distribution, characterized by finite bounds that exclude values outside the specified range. While the original noise has tails that extend infinitely, clipping focuses on the central portion of the distribution, resulting in a noise distribution with higher density near the mean and bounded at the edges.

Frequency-Domain Perturbation (FDP). To enhance message embedding features, we propose Frequency-Domain Perturbation (FDP), which applies a Fourier transform to the embeddings, adds noise in the frequency domain, and then transforms the data back to the time domain. This process introduces controlled perturbations to capture diverse embedding characteristics, improving model robustness and generalization. The process begins by converting g^i from the time domain to the frequency domain using the Fourier transform:

$$F^i = \mathcal{F}(g^i), \quad (8)$$

where $\mathcal{F}$ denotes the Fourier transform, and F^i represents the transformed embedding in the frequency domain. We selectively use the frequency components based on a specified keep ratio r . For the high mode, we retain the high-frequency components while attenuating the low-frequency components:

$$F_{filtered}^i = F^i[N - r * N : N], \quad (9)$$

where N is the number of message sample. We then add noise directly in the frequency domain to the filtered embedding:

$$F_{FDP}^i = F_{filtered}^i + n, n \sim (\mathcal{N}(0, \sigma^2) + i \cdot \mathcal{N}(0, \sigma^2)) \cdot \eta, \quad (10)$$

η is a noise level. We apply the inverse Fourier transform to convert the perturbed frequency-domain representation back into the time domain:

$$g_{FDP}^i = \mathcal{F}^{-1}(F_{FDP}^i), \quad (11)$$

By applying the Fourier transform, this method enables controlled perturbations in the frequency domain, allowing for a targeted enhancement of specific frequency components, which leads to a more nuanced and robust representation of the embedding’s features.

4 Experiments

The research questions are: **Q1:** How does the proposed dual augmentation framework compare to strong baseline models? **Q2:** What are the individual effects of explicit and implicit data augmentation on model performance? **Q3:** Which types of extracted information are most effective for rewriting in explicit augmentation? **Q4:** How do different frequency-domain perturbations (e.g., high-frequency noise, band filter, low-frequency noise) impact the model? **Q5:** How do the augmentation methods perform with limited data? **Q6:** How does implicit augmentation alter the data distribution?

4.1 Dataset, Evaluation Metrics and Baselines

We conduct experiments on three datasets: Kawarath6 (Alharbi and Lee, 2021), Twitter2012 (McMinn et al., 2013), and Twitter2018 (Mazoyer et al., 2020). For evaluation, we use Micro F1 and Macro F1. Micro F1 evaluates the overall performance in classifying instances across all classes, while Macro F1 checks each class individually before averaging these scores. The baseline models we have compared with are: TF-IDF (Aizawa, 2003), Word2Vec (Mikolov et al., 2013), FastText (Joulin et al., 2017), FinEvent (Peng et al., 2022)², BERT (Kenton and Toutanova, 2019), GraphMSE (Li et al., 2021), ETGNN (Ren et al., 2022b)³, KPGNN (Cao et al., 2021)⁴, HGT (Hu et al., 2020), and GraphHAM (Qiu et al., 2024a).

²We also tried FinEvent on Twitter2012 and Twitter2018, but it adopts GAT for message passing that is very memory-intensive and casted out-of-memory errors.

³We report the performance of ETGNN from the paper (Ren et al., 2022b), they evaluated the model on Kawarath7, which have different data scales with ours.

⁴KPGNN encounter OOM errors on the Twitter2012 and Twitter2018, and thus their results are not included.

Table 1: Overall performance comparison, with all results presented as percentages. The improvement represents the percentage increase of our method over the best baseline model.

Datasets	Kawarith6			Twitter2012			Twitter2018		
Models	Micro F1	Macro F1	Average	Micro F1	Macro F1	Average	Micro F1	Macro F1	Average
TF-IDF	92.87	92.49	92.68	67.89	34.05	50.97	42.59	20.00	31.30
Word2Vec	65.64	56.23	60.94	57.14	28.13	42.64	53.77	22.24	38.01
FastText	85.12	82.06	83.59	17.25	0.70	8.98	1.06	0.56	0.81
FinEvent	92.59	91.36	91.96	-	-	-	-	-	-
BERT	76.95	75.02	75.99	68.89	51.58	60.24	55.45	32.00	43.73
GraphMSE	94.70	94.00	94.35	80.57	67.16	73.87	76.71	66.21	71.46
ETGNN	-	-	-	84.80	75.65	80.23	-	-	-
HGT	91.93	91.18	91.56	71.11	58.41	64.76	80.46	68.92	74.69
KPGNN	78.63	76.91	77.77	-	-	-	-	-	-
GraphHAM	95.10	94.57	94.84	84.14	71.00	77.57	80.54	71.77	76.16
SED-Aug	98.41	98.29	98.35	93.03	89.53	91.28	89.61	86.43	88.02
Improvement	3.48 ↑	3.93 ↑	3.70 ↑	10.57 ↑	26.10 ↑	17.67 ↑	11.26 ↑	20.43 ↑	15.57 ↑

4.2 Overall Performance (Q1)

The experimental results presented in Table 1, highlighting the overall performance comparison of various models. the SED-Aug model demonstrated state-of-the-art performance across all datasets, showing clear improvements in both Micro F1 and Macro F1 metrics, indicating its robustness and effectiveness in handling diverse SED tasks. Specifically, on the Kawarith6 dataset, SED-Aug achieved a Micro F1 score of 98.41%, a Macro F1 score of 98.29%, and an average of 98.35%, surpassing the performance of the next best models, GraphHAM, which had averages of 94.84%. Our proposed model, SED-Aug, achieved notable improvements on both the Twitter2012 and Twitter2018 datasets. On the Twitter2012 dataset, SED-Aug showed a 10.57% increase in Micro F1 and a 26.10% increase in Macro F1 compared to the best baseline model. Similarly, on the Twitter2018 dataset, SED-Aug demonstrated an improvement of 11.26% in Micro F1 and 20.43% in Macro F1 over the best-performing baseline. These results clearly demonstrate the effectiveness of our dual data augmentation approach in enhancing the performance for SED.

Table 2: Different explicit augmentation methods, with all results presented as percentages.

Datasets	Kawarith6		Twitter2012		Twitter2018	
Methods	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1
Paraphrase	97.69	97.35	92.70	89.68	87.64	81.68
Style transfer	97.12	96.69	91.60	86.92	86.33	79.29
Add context	97.94	97.75	91.74	87.19	87.02	79.85
Keep entity	98.20	98.08	92.76	89.33	88.70	82.14
Extract rewrite	97.99	97.84	92.34	88.65	88.47	81.57

4.3 Ablation Studies (Q2)

Table 2 shows that all five explicit augmentation methods improve the performance of the SED

Table 3: Performance of combining the most and least effective explicit augmentations with all five implicit augmentations on the Twitter2012 dataset.

Twitter2012	Keep Entity		Style Transfer	
Methods	Micro F1	Macro F1	Micro F1	Macro F1
GP	91.83	87.16	92.91	89.47
PGP	91.73	86.97	92.85	89.65
IDGP	91.79	86.93	93.01	89.58
CGP	91.70	87.15	93.03	89.53
FDP	91.81	87.19	92.89	89.96

model. Among them, the "keep entity" method consistently performs best across all datasets, achieving the highest Micro F1 and Macro F1 scores on both Kawarith6 and Twitter2018, and the highest Micro F1 and second-highest Macro F1 on Twitter2012. Preserving these entities ensures that the model retains essential context, enabling it to accurately identify and detect significant events. We acknowledge that using LLMs to add context may introduce hallucinations. However, the brevity of our additions limits the risk of false information. Studies show that GPT-4 generates about 6% false claims in long-form responses (Wang et al.). To ensure quality, we sampled 50 examples and found false information in 3, indicating an acceptable factuality level for this augmentation strategy.

To explore the interactions between explicit and implicit augmentations, we further selected a subset of combinations from the 75 possible cases (5 explicit methods $\times$ 5 implicit methods $\times$ 3 datasets) based on the individual effectiveness of each augmentation. Specifically, we selected the most and least effective explicit augmentation among the five and combined each with all five implicit augmentations to analyze their interactions. The results 3 show that implicit augmentation consistently provides additional benefits when combined with explicit augmentation, with no observed cases of performance degradation.

Using "keep entity" as the baseline, we then evaluate different implicit augmentation strategies in all three datasets. Table 4 shows that all five methods improve Micro and Macro F1 scores in most cases, with variations across datasets. PGP achieved the highest F1 scores on both the Kawarith6 and Twitter2018 datasets, attributed to its method of scaling the added perturbation relative to the input data's feature values, ensuring that the perturbations remain contextually relevant to the data's magnitude. CGP attains the best Micro F1 on Twitter2012 by clipping extreme noise values, preventing outliers and preserving feature distribution.

It is worth noting that implicit augmentation methods consistently improved Macro F1 scores on the Twitter2018, increasing the baseline from 82.14% to 86.43%. To understand this, we analyzed the event distribution in both the Twitter2012 and Twitter2018 datasets, as shown in Figure 2. The analysis revealed that Twitter2018 has a more severe class imbalance compared to Twitter2012, with some event categories containing up to 12,000 messages. In datasets with higher class imbalance, feature space augmentation can better assist the model in learning the characteristics of rare classes, thereby improving its performance across different categories. In contrast, the more balanced Kawarith6 and Twitter2012 datasets saw less improvement in Macro F1 scores, indicating that implicit augmentation is particularly effective for underrepresented classes. These methods directly transform structural-fused message embeddings, capturing the semantic and structural information in the data, which strengthens the model's performance across different classes and relieves the challenges of class imbalance.

Table 4: Different implicit augmentation methods, with all results presented as percentages.

Datasets	Kawarith6		Twitter2012		Twitter2018	
Methods	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1
GP	98.30	98.21	92.91	89.47	88.86	85.47
PGP	98.41	98.29	92.85	89.65	89.61	86.43
IDGP	98.30	98.15	93.01	89.58	88.12	84.30
CGP	98.30	98.22	93.03	89.53	88.44	85.02
FDP	98.25	98.13	92.89	89.96	88.43	84.32

Table 5: Different extraction methods, with results in percentages. KG denotes knowledge graph.

Datasets	Kawarith6		Twitter2012		Twitter2018	
Information	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1
Keywords	97.99	97.84	92.34	88.65	87.52	80.30
Entities	97.58	97.33	92.15	88.21	88.47	81.57
KG	97.58	97.32	91.92	87.10	87.22	79.57

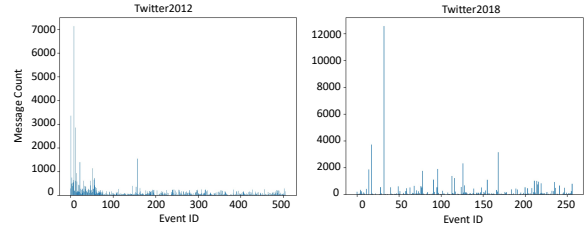


Figure 2: Data imbalanced in the SED datasets.

Table 6: Different mode used in Fourier transfer, with all results presented as percentages.

Datasets	Kawarith6		Twitter2012		Twitter2018	
Mode	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1
High	98.41	98.29	92.89	89.96	88.43	84.32
Low	95.88	95.41	92.74	89.39	87.48	84.49
Band	91.15	91.35	92.94	89.88	88.10	84.21

4.4 Effectiveness of Different Types of Information for Extraction and Rewriting (Q3)

Table 5 summarizes the performance of extract and rewrite strategies using keywords, entities, and knowledge graphs. While all are valuable, keywords are generally the most effective, followed by entities and knowledge graphs. Results indicate that using keywords extracted by the LLM yields the highest Micro F1 and Macro F1 scores across the Kawarith6 and Twitter2012 datasets. This can be attributed to the LLM's ability to identify specific terms that carry significant semantic weight in SED, helping the model focus on the most informative aspects of the message. In the Twitter2018 dataset, the entity-based approach achieved the best scores in both Micro F1 and Macro F1, emphasizing the importance of entities like names, locations, and dates in capturing key event details. This ensures the most relevant message components are preserved, which is essential for distinguishing between different events in SED tasks.

Table 7: Performance on different training ratio, with all results presented as percentages.

Ratio	Without augmentation			With augmentation			Improve
	Micro F1	Macro F1	Average	Micro F1	Macro F1	Average	
10%	76.55	61.16	68.86	82.29	69.59	75.94	10.29
20%	80.75	70.84	75.80	87.01	78.40	82.71	9.12
30%	85.65	75.44	80.55	87.83	81.97	84.90	5.41
40%	85.75	76.27	81.01	89.68	83.67	86.68	6.99
50%	85.50	78.64	82.07	91.15	86.17	88.66	8.03
60%	88.50	82.30	85.40	90.37	87.05	88.71	3.88
70%	87.72	83.69	85.71	93.03	89.53	91.28	6.50

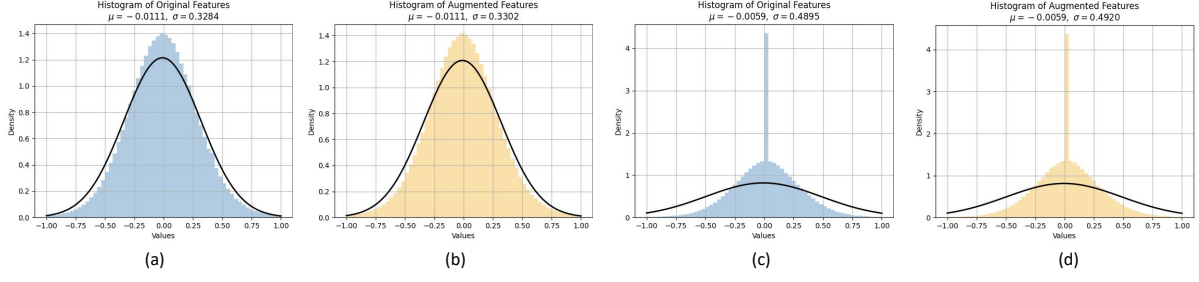


Figure 3: Data histogram distribution before and after implicit augmentation (GP). Subfigure (a) and (c) are the histograms of original features before implicit augmentation; (b) and (d) are the histograms of features after augmentation.

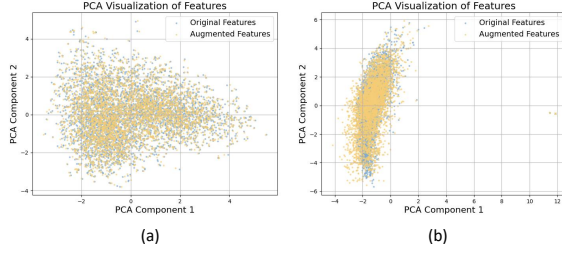


Figure 4: The PCA visualization before and after implicit augmentation (with GP noises). The blue dots in subfigures (a) and (b) represent original features; yellow dots denote augmented features.

4.5 Impact of Frequency-domain Perturbations on Different Mode (Q4)

This section investigates how different types of frequency-domain perturbations, specifically mask part of high-frequency noise, low-frequency noise, and band filtering, affect model performance across various datasets. The results, summarized in Table 6, indicate that retain most of the high-frequency noise achieves the best results on most of cases. It can be attributed to its ability to retain the low-frequency components that capture essential semantic information while effectively attenuating high-frequency noise that may obscure meaningful patterns. Low-frequency components often contain critical information about the overall structure of the messages, enabling the model to maintain a robust understanding of the content. In contrast, high-frequency noise may introduce irrelevant fluctuations that hinder the model’s ability to learn important features.

4.6 Performance of SED-Aug with Limited Data (Q5)

To evaluate the effectiveness of SED-Aug under varying data scarcity, we conducted experiments

using the Twitter2012 dataset with training sets of 10%, 20%, ..., 70% of the total data. We compared results with and without data augmentation, using both explicit and implicit methods. As shown in Table 7, dual augmentation consistently demonstrated its utility across all data volume scenarios. Notably, our dual augmentation strategy significantly improved model performance, especially when data was limited. For example, when using only 10% of the data, augmentation led to a performance increase of 10.29%. Without augmentation, performance gains diminished after 30% of the data, stagnating at 85.40% to 85.71% with 60% of the data. In contrast, dual augmentation steadily improved performance, from 88.71% to 91.28% with 60% and 70% of the data, respectively. This steady improvement can be attributed to the ability of our approach to leverage both explicit and implicit augmentation techniques, enhancing model robustness and generalization across different data settings.

4.7 Visualizing Distribution Changes in Explicit Data Augmentation (Q6)

We visualize the histograms of the structural fused embedding before and after implicit augmentation based on Twitter2012 dataset. In Figure 3, we find out the histogram distribution of original and augmented features keep most likely the same shapes, but slight differences, especially in variances. For subfigure (a) and (b), the mean μ of both Figures are -0.0111, but with a slight increment of variance σ from 0.3284 to 0.3302. Similar phenomenon happens on subfigure (c) and (d). This phenomenon meets our expectations as we do not want to change the mean of the data, but just change a little bit of the features in the embedding space by adding noises (with Gaussian Perturbation noises) sampled from another Gaussian distribution to increase the sample diversity.

We further conduct PCA visualizations on the features before and after Gaussian Perturbation implicit augmentation (in Figure 4). Resonant with the histogram visualization, the PCA visualization also shows the general overlapping between two groups of features, but with slightly differences. This is because the implicit augmentation can increase more data diversity by adding noises in the feature space from Gaussian Perturbation.

5 Conclusion

In this study, we present SED-Aug, a dual data augmentation framework for SED. By combining explicit and implicit augmentation, SED-Aug enhances model performance and robustness without manual labeling. SED-Aug significantly outperforms state-of-the-art baselines, achieving improvements of 17.67% on the Twitter2012 dataset and 15.57% on the Twitter2018 dataset in average F1 score. Both quantitative and qualitative experimental results show that the implicit augmentation is valuable for enhancing model robustness in imbalanced datasets and empowering data sample with more diversity, which further complements the explicit augmentation methods that are more effective in generating information-rich textual variations.

6 Limitations

One limitation of our work is using large language models for text augmentation lacks a clear criterion for determining the optimal amount of augmented data. While augmentation can improve model performance by enhancing data diversity, adding too much augmented data may introduce noise or redundancy. Conversely, too little augmentation may not provide sufficient variation to improve generalization. The ideal balance depends on multiple factors, including the task complexity, the quality of the original dataset, and the specific augmentation strategy used. However, there is no universally accepted guideline for how much augmentation is necessary, making it challenging to determine the best augmentation ratio.

Acknowledgement

This work was supported by the Australian Research Council Project with No. DP230100899, Macquarie University Data Horizons Research Centre and Applied Artificial Intelligence Centre. Corresponding author: Jia Wu.

References

- Akiko N. Aizawa. 2003. An information-theoretic perspective of tf-idf measures. *Information Processing and Management.*, 39(1):45–65.
- Alaa Alharbi and Mark Lee. 2021. Kawarith: an arabic twitter corpus for crisis events. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 42–52.
- Yuwei Cao, Hao Peng, Jia Wu, Yingdong Dou, Jianxin Li, and Philip S. Yu. 2021. Knowledge-preserving incremental social event detection via heterogeneous gnn. In *The Web Conference 2021*, pages 3383–3395, Virtual Event / Ljubljana, Slovenia.
- Yuwei Cao, Hao Peng, Zhengtao Yu, and Philip S. Yu. Hierarchical and incremental structural entropy minimization for unsupervised social event detection. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pages 8255–8264.
- Terrance DeVries and Graham W. Taylor. 2017. Dataset augmentation in feature space. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*.
- Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard H. Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 968–988.
- Sreyan Ghosh, Utkarsh Tyagi, Manan Suri, Sonal Kumar, Ramaneswaran S., and Dinesh Manocha. 2023. ACLM: A selective-denoising based generative data augmentation approach for low-resource complex NER. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 104–125. Association for Computational Linguistics.
- William L. Hamilton, Zhitaoying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems (NeurIPS 2017)*, pages 1024–1034, Long Beach, USA.
- Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. Heterogeneous graph transformer. In *The Web Conference 2020*, pages 2704–2710, Taipei, China.

- Bowen Jin, Gang Liu, Chi Han, Meng Jiang, Heng Ji, and Jiawei Han. 2024. Large language models on graphs: A comprehensive survey. *IEEE Transactions on Knowledge and Data Engineering*.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomáš Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, pages 427–431, Valencia, Spain.
- Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.
- Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations (ICLR 2017)*, Toulon, France.
- Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 452–457. Association for Computational Linguistics.
- Pu Li, Xiaoyan Yu, Hao Peng, Yantuan Xian, Linqin Wang, Li Sun, Jingyun Zhang, and Philip S Yu. 2024. Relational prompt-based pre-trained language models for social event detection. *ACM Transactions on Information Systems*.
- Yi Li, Yilun Jin, Guojie Song, Zihao Zhu, Chuan Shi, and Yiming Wang. 2021. Graphmse: efficient meta-path selection in semantically aligned feature space for graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2021)*, volume 35, pages 4206–4214.
- Wenzhong Liu and Xiaohui Cui. 2023. Improving named entity recognition for social media with data augmentation. *Applied Sciences*, 13(9):5360.
- Ang Lv, Jinpeng Li, Yuhan Chen, Gao Xing, Ji Zhang, and Rui Yan. 2023. Dialogps: Dialogue path sampling in continuous semantic space for data augmentation in multi-turn conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1267–1280. Association for Computational Linguistics.
- Congbo Ma, Zitai Qiu, Hu Wang, Jing Du, Shan Xue, Jia Wu, and Jian Yang. 2025. Enhanced social event detection through dynamically weighted meta-paths modeling. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1184–1188.
- Congbo Ma, Hu Wang, Zitai Qiu, Shan Xue, Jia Wu, Jian Yang, Preslav Nakov, and Quan Z Sheng. 2024. Learning to sample the meta-paths for social event detection. *arXiv preprint arXiv:2411.12588*.
- Pablo Martí, Leticia Serrano-Estrada, and Almudena Nolasco-Cirugeda. 2019. Social media data: Challenges, opportunities and limitations in urban studies. *Computers, Environment and Urban Systems*, 74:161–174.
- Béatrice Mazoyer, Julia Cagé, Nicolas Hervé, and Céline Hudelot. 2020. A french corpus for event detection on twitter. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 6220–6227, Marseille, France.
- Andrew J McMinn, Yashar Moshfeghi, and Joemon M Jose. 2013. Building a large-scale corpus for evaluating event detection on twitter. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 409–418.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.
- Tahir M Nisar and Man Yeung. 2018. Twitter as a tool for forecasting stock market movements: A short-window event study. *The journal of finance and data science*, 4(2):101–119.
- Viktor Pekar, Jane Binner, Hossein Najafi, Chris Hale, and Vincent Schmidt. 2020. Early detection of heterogeneous disaster events using social media. *Journal of the Association for Information Science and Technology*, 71(1):43–54.
- Hao Peng, Jianxin Li, Yangqiu Song, Renyu Yang, Rajiv Ranjan, Philip S. Yu, and Lifang He. 2021a. Streaming social event detection and evolution discovery in heterogeneous information networks. *ACM Trans. Knowl. Discov. Data*, 15(5):89:1–89:33.
- Hao Peng, Jianxin Li, Yangqiu Song, Renyu Yang, Rajiv Ranjan, Philip S Yu, and Lifang He. 2021b. Streaming social event detection and evolution discovery in heterogeneous information networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–33.
- Hao Peng, Ruitong Zhang, Shaoning Li, Yuwei Cao, Shirui Pan, and S Yu Philip. 2022. Reinforced, incremental and cross-lingual event detection from social messages. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):980–998.
- Zitai Qiu, Congbo Ma, Jia Wu, and Jian Yang. 2024a. An efficient automatic meta-path selection for social event detection via hyperbolic space. In *Proceedings of the ACM on Web Conference 2024 (WWW 2024)*, pages 2519–2529, Singapore.
- Zitai Qiu, Jia Wu, Jian Yang, Xing Su, and Charu C Aggarwal. 2024b. Heterogeneous social event detection via hyperbolic graph representations. *IEEE Transactions on Big Data*.

- Jiaqian Ren, Lei Jiang, Hao Peng, Yuwei Cao, Jia Wu, Philip S. Yu, and Lifang He. 2022a. From known to unknown: Quality-aware self-improving graph neural network for open set social event detection. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM 2022)*, pages 1696–1705, Atlanta, USA.
- Jiaqian Ren, Lei Jiang, Hao Peng, Zhiwei Liu, Jia Wu, and Philip S. Yu. 2022b. Evidential temporal-aware graph-based social event detection via Dempster-Shafer theory. In *IEEE International Conference on Web Services (ICWS 2022)*, pages 331–336, Barcelona, Spain.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.
- Yizhou Sun and Jiawei Han. 2012. *Mining heterogeneous information networks: principles and methodologies*. Morgan & Claypool Publishers.
- Tuuli Toivonen, Vuokko Heikinheimo, Christoph Fink, Anna Hausmann, Tuomo Hiippala, Olle Järv, Henrikki Tenkanen, and Enrico Di Minin. 2019. Social media data for conservation science: A methodological overview. *Biological Conservation*, 233:298–315.
- Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *The 6th International Conference on Learning Representations (ICLR 2018)*, Vancouver, Canada.
- Yulin Wang, Xuran Pan, Shiji Song, Hong Zhang, Gao Huang, and Cheng Wu. 2019. Implicit semantic data augmentation for deep networks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 12614–12623.
- Yuxia Wang, Minghan Wang, Hasan Iqbal, Georgi N. Georgiev, Jiahui Geng, and Preslav Nakov. Openfactcheck: A unified framework for factuality evaluation of llms.
- Zhongqing Wang and Yue Zhang. 2017. A neural model for joint event detection and summarization. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017)*, pages 4158–4164, Melbourne, Australia.
- Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 6381–6387. Association for Computational Linguistics.
- Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24.
- Dominik Wurzer, Victor Lavrenko, and Miles Osborne. 2015. Twitter-scale new event detection via k-term hashing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pages 2584–2589, Lisbon, Portugal.
- Qizhe Xie, Zihang Dai, Eduard H. Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2015. A probabilistic model for bursty topic discovery in microblogs. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI 2015)*, pages 353–359, Austin, USA.
- Jianfei Yu, Qiankun Zhao, and Rui Xia. 2023. Cross-domain data augmentation with domain-adaptive language modeling for aspect-based sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1456–1470. Association for Computational Linguistics.
- Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. 2019. Heterogeneous graph neural network. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 793–803.
- Haojie Zhuang, Wei Emma Zhang, Jian Yang, Congbo Ma, Yutong Qu, and Quan Z Sheng. 2022. Learning from the source document: Unsupervised abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4194–4205.

A Appendix

A.1 Experimental Setting

All datasets are divided into training, validation, and test sets, with the splits allocated in the ratio of 70%, 10%, and 20%, respectively. To construct the social graph, we utilize three distinct node types: messages, users, and entities extracted from those messages. For generating message embeddings, we employ the BERT model, specifically the "bert-base-uncased" variant. The text node features comprise 768-dimensional embeddings derived from

the textual content, which are subsequently combined with 2-dimensional embeddings that capture temporal information. For the user node features, we integrate 768-dimensional embeddings obtained from filtered words with 2-dimensional location embeddings, ensuring a comprehensive representation of users in the graph.

We evaluated the data quality using GPT-4o-mini, GPT-4o, and GPT-o1 on 100 examples. The results showed that GPT-4o-mini, the most cost-efficient model, produced competitively high-quality data. Consequently, we used GPT-4o-mini for all subsequent experiments, setting the maximum token limit to 1000 while keeping all other settings at their default values. In the implicit data augmentation process for the Kawarith6 dataset, we establish a probability threshold α of 0.3, a standard deviation σ of 0.01, and a clipping range c of 0.005. When applying frequency-domain perturbation, the keep ratio r is set to 0.98, and the noise level η is adjusted to 0.02. For the Twitter2012 dataset, we modify the probability threshold α to 0.6, the standard deviation σ to 0.1, and the clipping range c to 0.05. The frequency-domain perturbation settings for this dataset include a keep ratio r of 0.95 and a noise level η of 0.02. Similarly, for the Twitter2018 dataset, we set the probability threshold α at 0.6, the standard deviation σ at 0.1, and the clipping range c at 0.0006. The frequency-domain perturbation parameters for this dataset maintain a keep ratio r of 0.98 and a noise level η of 0.02.

A.2 Datasets and Evaluation Metrics

We conduct experiments on three datasets: Kawarith6 (Alharbi and Lee, 2021), Twitter2012 (McMinn et al., 2013), and Twitter2018 (Mazoyer et al., 2020). Kawarith6 contains 4,860 messages belonging to six unique event classes; Twitter2012 contains 68,841 messages from 503 unique event classes; Twitter2018 contains 64,516 messages from 257 unique event classes. For evaluation, we use Micro F1 and Macro F1. Micro F1 evaluates the overall performance in classifying instances across all classes, while Macro F1 checks each class individually before averaging these scores. The classes refer to the predefined event categories in the dataset being used. These event classes vary depending on the specific dataset. For instance, in the Twitter2012 dataset, the classes include events such as the 2012 Nobel Prize in Literature, 2012 Presidential debates, the Bolivian radio man set on

fire, etc. The exact classes depend on the annotations and categories defined within the dataset.

A.3 Removing the explicit augmentation

Table 8 presents the results of removing explicit augmentation methods on the Kawarith6 dataset to assess the contribution of explicit augmentation. "w/o Aug" denotes the performance without any explicit augmentation. As shown, all the explicit augmentation methods lead to improved performance compared to the baseline without explicit augmentation, indicating their effectiveness. Among them, FDP achieves the highest performance in both Micro F1 (95.88) and Macro F1 (95.41), suggesting it is the most beneficial augmentation technique for this dataset.

Table 8: Performance of removing the explicit augmentation methods on Kawarith6 dataset.

Augmentation	Micro F1	Macro F1
GP	95.78	95.20
PGP	95.78	95.39
IDGP	95.78	95.36
CGP	95.68	95.29
FDP	95.88	95.41
w/o Aug	95.13	94.58

Table 9: Performance of different probability threshold.

α	Kawarith6		Twitter2012		Twitter2018	
	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1
0.1	96.24	95.87	92.73	89.28	88.31	84.40
0.3	96.35	95.98	92.97	89.18	88.78	85.32
0.6	98.41	98.29	93.03	89.53	89.61	86.43

A.4 Performance of different probability threshold α

Table 9 presents the model performance under different probability thresholds α on the Kawarith6, Twitter2012, and Twitter2018 datasets. In all three data sets, increasing the value of α leads to consistent improvements in both Micro F1 and Macro F1 scores. In the Kawarith6 dataset, when α is set to 0.1, the Micro F1 score is 96.24 and the Macro F1 is 95.87. As the threshold increases to 0.6, the scores rise substantially to 98.41 for Micro F1 and 98.29 for Macro F1, indicating a clear performance gain. A similar pattern is observed in the Twitter2012 dataset, where increasing α from 0.1 to 0.6 improves Micro F1 from 92.73 to 93.03, and Macro F1 from 89.28 to 89.53. In the Twitter2018 dataset, Micro F1 increases from 88.31 at $\alpha = 0.1$ to 89.61 at $\alpha = 0.6$, while Macro F1 improves from 84.40 to 86.43.

A.5 Evaluating LLM Performance under Zero-shot and Few-shot Settings

We evaluated label predictions using an LLM under both zero-shot and few-shot settings on the Kawaristh6 dataset. The results are as follows:

Zero-shot: Micro F1 = 92.8%, Macro F1 = 92.3%;

Few-shot: Micro F1 = 98.1%, Macro F1 = 98.0%.

These results confirm that LLMs can achieve strong performance which are comparable to the proposed method, particularly in the few-shot setting. However, as we discussed, LLM-based approaches come with significant computational and financial costs, especially when scaling to large datasets. Our method, SED-aug, provides a more efficient alternative while maintaining better performance by including valuable graph information inside.

A.6 Example of explicit augmentation and prompt

Figure 5 shows the example of different explicit augmentation strategies. Figure 6 is the prompt we have used for explicit augmentation.

Original message: Geoffrey E. Hinton, awarded the 2024 #NobelPrize in Physics, was born in 1947 in London, UK.	
One-stage augmentation	Two-stage augmentation
Paraphrase: Geoffrey E. Hinton, who won the 2024 #NobelPrize in Physics, was born in London, UK, in 1947. Adding context: Geoffrey E. Hinton, a pioneering figure in artificial intelligence and machine learning, has been awarded the 2024 #NobelPrize in Physics. Born in London, UK, in 1947, Hinton's groundbreaking work continues to shape the future of technology and science. Style transfer: In a historic win, Geoffrey E. Hinton, born in London in 1947, has claimed the prestigious 2024 #NobelPrize in Physics, honoring his trailblazing contributions to the scientific community. Keep entity unchanged: Geoffrey E. Hinton, recipient of the 2024 #NobelPrize in Physics, was born in London, UK, in 1947.	Extract entities and rewrite: <i>Entities:</i> "Geoffrey E. Hinton", "2024 #NobelPrize", "Physics", "1947", "London, UK" <i>Rewrite:</i> Geoffrey E. Hinton, born in 1947 in London, UK, has won the 2024 #NobelPrize in Physics. Extract keywords and rewrite: <i>Keywords:</i> Geoffrey E. Hinton, "2024 #NobelPrize", "Physics", "born in 1947", "London, UK" <i>Rewrite:</i> Geoffrey E. Hinton, an influential figure in Physics, born in 1947 in London, UK, has been honored with the 2024 #NobelPrize. Extract knowledge graphs and rewrite : <i>Knowledge graph:</i> { "Geoffrey E. Hinton": { "born": "1947", "place_of_birth": "London, UK", "awarded": { "award": "2024 #NobelPrize", "category": "Physics" } } } <i>Rewrite:</i> Geoffrey E. Hinton, a London-born scientist from 1947, has been recognized with the 2024 #NobelPrize in Physics.

Figure 5: Example of explicit augmentation.

system_prompt = "You are a helpful assistant." augment_prompt = "" You are asked to perform data augmentation by paraphrasing, adding contenxt, and transferring style on a social media post for social event detection task. Given the post: {post}, Directly Output all augmented posts by json format like below, No Other Words: <pre>{{ "paraphrase": "paraphrase this post.", "paraphrase_unchanged_entity": "paraphrase this post and keep entities unchanged.", "add_context": "enhance this post by adding more context and providing additional information about the event.", "style_transfer": "transform this post into a different style while preserving its original meaning.", }}</pre>
extrac_augment_prompt = "" You are asked to first extract entity, important words, sentences or knowledge graph for a social media post and then perform data augmentation by rewriting or paraphrasing. Given the post: {post}, Directly output extracted entity, key words, and knowledge graph, and all augmented posts by json format like below, No Other Words: <pre>{{ "extracted_NER": "extract entities from this post.", "NER_rewrite": "generate a new post based on the extracted entities.", "extract_keywords": "extract the important words or sentences that are important for social event detection, and output as a list of strings", "keywords_rewrite": "generate a new posts based on those key words or sentences.", "extract_KG": "extract the knowledge graph that helps social event detection", "KG_rewrite": "generate new posts based on the knowledge graph." }}</pre>

Figure 6: Prompt for the explicit data augmentation.