LEARNING ROBUST MULTIMODAL REPRESENTATION FOR EVENT DETECTION FROM SOCIAL MEDIA DATA

ZEHANG LIN

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University
Department of Computing

Learning Robust Multimodal Representation for Event Detection from Social Media Data

Zehang Lin

A thesis submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy
June 2024

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgment has been made in the text.

Signature: _____

Name of Student: Zehang Lin

Abstract

The widespread use of social media has generated a vast amount of data, which presents unique challenges and opportunities for information processing. This massive data, characterized by its scale and complexity, demands advanced analytics to fully utilize its potential. Within this scenario, social event detection emerges as a critical analytics task, which aims at identifying and categorizing significant events from the streams of data available on social media platforms. However, the social media data used for event detection exhibit characteristics of multimodality, information fragmentation, cross-platform, and dynamic nature. The performance of current social event detection methods is hindered by two major problems.

The first problem is the limited detection accuracy. Despite the rapid advancement of deep learning methods, they face various challenges in handling modality heterogeneity inherent in multimodal social media event data and the out-of-distribution (OOD) problem caused by information fragmentation. Existing methods, although starting to leverage multimodal data for event detection, often struggle to identify the correct events when faced with fragmented information.

The second problem is the insufficient generalization capability. Current supervised event detection methods have limited generalization capability when dealing with different data sources and newly emerging events. Due to the cross-platform and dynamic nature of social event data, the lack of consideration for these aspects affects the generalizability of event detection models.

To address these problems above, our focus in this thesis is on the following objectives. Firstly, we aim to design a deep learning model to address modality heterogeneity and the OOD problem, thereby improving the accuracy of event detection. Secondly, we aim to develop an innovative manner to adapt models to implement cross-platform social event detection. Thirdly, we aim to extend existing supervised event detection methods to discover new social events in social media.

To achieve the first objective, we introduce a Multimodal Fusion with External Knowledge (MFEK) model. This method incorporates a text enrichment module that leverages image semantics to enhance textual content, along with a knowledge-aware feature fusion mechanism that effectively integrates external knowledge and multimodal data to mitigate modality heterogeneity and the OOD problem caused by the fragmentation of social event data. We find that such a method can bring a significant improvement to the performance after incorporating external knowledge, even in scenarios with fragmentation information.

To accomplish the second objective, we develop a Self-Supervised Modality Complementation (SSMC) method to enhance the model's adaptability and performance across different social media platforms. By introducing a Missing Data Complementation (MDC) module and a Multimodal Self-Learning (MSL) module, SSMC effectively addresses incomplete modalities and platform heterogeneity in the scenario of cross-platform event detection. We find that such a strategy ensures robust cross-platform event detection even in the presence of varied and incomplete data. In addition, we validate the role of cross-platform event detection in improving the quality of single-platform event data.

For the third objective, we propose a new task, generalized social event detection, which requires accurately identifying predefined events and detecting emerging new events. Specifically, we propose a Dynamic Augmentation and Entropy Optimization (DAEO) model, which utilizes adversarial learning for learning robust multimodal representation and introduces an adaptive entropy optimization technique with a

self-distillation method that promotes model adaptability to newly emerging events. We demonstrate that this combination allows for the effective identification of both known and new events, thereby enhancing the model's generalization capabilities.

To summarize, in this thesis, we propose a MFEK model by introducing external knowledge to improve the accuracy of social event detection. Furthermore, we develop a SSMC method to enhance cross-platform adaptability and a DAEO model to tackle generalized social event detection, thereby addressing key challenges in multimodal social event detection and improving overall model performance and generalization. Extensive experiments conducted on publicly available and our collected real-world datasets demonstrate their significance in the context of social event detection, outperforming the state-of-the-art baseline approaches.

Publications Arising from the Thesis

1. Zehang Lin, Jiayuan Xie, and Qing Li. "Multi-modal news event detection with external knowledge", in *Information Processing & Management (IPM)*, 61(3):103697, 2024. (corresponding to Chapter 3)

2. Zehang Lin, Zhenguo Yang, and Qing Li. "Robust Cross-platform News Event Detection via Self-supervised Modality Complementation", manuscript submitted to *IEEE Transactions on Knowledge and Data Engineering (TKDE)*. (corresponding to Chapter 4)

3. Zehang Lin, Jiayuan Xie, Zhenguo Yang, Yu Yi, and Qing Li. "Generalized News Event Discovery via Dynamic Augmentation and Entropy Optimization", in *The 32st ACM International Conference on Multimedia (ACM MM)*, pp. 10018-10026. 2024. (corresponding to Chapter 5)

Acknowledgments

During these three years of study and life as a PhD researcher, I have encountered a lot of difficulties and setbacks, but have successfully passed through them with the encouragement and help of many people. As my thesis is about to be completed, I would like to express my heartfelt thanks to all those who have given me support and help.

Firstly, I am extremely grateful to my PhD supervisor, Prof. Qing Li. He has consistently supported and guided my research work, which is a great encouragement and help to me. Prof. Li not only provided me with valuable academic advice but also offered necessary support when I faced difficulties, ensuring that I could focus on achieving my academic goals. I learned from him the importance of reflecting and slowing down even in the midst of a busy research schedule, which helped me better understand my research direction and knowledge. Moreover, Prof. Li often reminded me to maintain caution and patience in research exploration, which is key to achieving good research results. It is particularly worth mentioning that Prof. Li could always quickly and accurately point out the main problems in my work and guide me on how to resolve them, a skill I need to learn and cultivate. I am very grateful for Prof. Li's guidance and help on my academic path.

Additionally, I am also very thankful to Dr. Zhenguo Yang. Dr. Yang is rigorous in scientific research and patient with student work. During my PhD, we had numerous discussions on paper revision suggestions, which gradually taught me to

think independently and settle down to improve and refine my work. At the same time, members of my research group have also had a significant impact on me. In particular, I would like to thank Dr. Changmeng Zheng, Dr. Da Ren, Dr. Yujuan Ding, Mr. Yaowei Wang, Mr. Xingming Chen, Mr. Jun Li, Mr. Jiahao Wu from Prof. Li's group for their company and support. They encouraged and helped me in life and studies, and we ate and studied together, leaving me with unforgettable memories. Additionally, I would like to thank my friends Dr. Jiayuan Xie and Dr. Peipei Kang, who have discussed various research problems with me. I wish them smooth progress in their studies, success in their work, and happiness in life.

Last but not the least, I want to thank my family for always caring about me, supporting me, and encouraging me. No matter how far I go, I always receive their care and concern, which has given me the greatest motivation to persevere in my research. Thank you for their unconditional love and encouragement.

Finally, I would like to thank once again all those who encouraged and helped me, as well as the experts and professors who took time out of their busy schedules to review the thesis and participate in the defense.

Table of Contents

Abstract **i**

Publications Arising from the Thesis **iv**

Acknowledgments **v**

List of Figures **xii**

List of Tables **xv**

1 Introduction **1**

1.1 Background 1

1.2 Research Objectives 7

1.3 Overview of Proposed Solutions 9

1.4 Thesis Contributions 10

1.5 Thesis Outline 11

2 Related Work **14**

2.1 Social Event Detection 14

2.1.1 Single-modal Event Detection 15
2.1.2 Multimodal Event Detection 16
2.1.3 New Event Detection 17
2.2 Benchmark Datasets for Social Event Detection 18
2.3 Multimodal Data Fusion 19
2.4 Missing-modality for Multimodal Learning 21
2.5 Domain Adaptation 21
2.6 Generalized Category Discovery 24

3 Multimodal Social Event Detection with External Knowledge 25

3.1 Introduction 25
3.2 Problem Statement 30
3.3 Methodology 31
3.3.1 Text Enrichment Module 32
3.3.2 Knowledge Extraction Module 33
3.3.3 Knowledge-aware Feature Fusion Module 36
3.4 Social Event Detection (SED) Dataset 40
3.4.1 Data Collection 40
3.4.2 Statistics of SED Dataset 44
3.4.3 Comparisons with Existing Datasets 45
3.5 Experiment 46
3.5.1 Datasets and Data Partitioning 46

3.5.2 Implementation Details 47
3.5.3 Evaluation Metrics 47
3.5.4 Benchmarks 48
3.5.5 Results and Analysis 49
3.5.6 Model Ablation 51
3.5.7 Parameter Analysis 53
3.5.8 Case Study 54
3.6 Conclusion 58

4 Robust Cross-platform Social Event Detection via Self-supervised Modality Complementation 59

4.1 Introduction 59
4.2 Preliminaries and Problem Statement 63
4.3 Methodology 64
4.3.1 Overview of the Framework 64
4.3.2 Data Preprocessing 65
4.3.3 Missing Data Complementation (MDC) 65
4.3.4 Multimodal Self-learning (MSL) 67
4.3.5 Overall Objective 71
4.4 Experiment 71
4.4.1 Cross-platform Social Event Dataset (CSED) 72
4.4.2 Implementation Details 73
4.4.3 Baselines 76

4.4.4 Comparison with the State of the Arts 78
4.4.5 Ablation Study 81
4.4.6 Impact of Parameters α and β 82
4.4.7 Evaluating the Effectiveness of MSL 83
4.4.8 Visualization 84
4.4.9 Case Study 86
4.5 Conclusion 88

5 Generalized Social Event Detection via Dynamic Augmentation and Entropy Optimization 89

5.1 Introduction 89
5.2 Preliminaries and Problem Statement 93
5.3 Methodology 94
5.3.1 Overview of the Framework 94
5.3.2 Multimodal Event Feature Extraction 95
5.3.3 Multimodal Augmentation 96
5.3.4 Adaptive Entropy Optimization 98
5.3.5 Overall Formulation and Optimization 100
5.4 Experiment 102
5.4.1 Multimodal Social Event Detection (MSED) Dataset 102
5.4.2 Evaluation Metric 107
5.4.3 Implementation Details 107
5.4.4 Baselines 108

5.4.5 Comparison with the State of the Arts 108
5.4.6 Ablation Study 110
5.4.7 Parameter Analysis 114
5.4.8 Data Visualization 115
5.4.9 Case Study 116
5.4.10 Experiments on the Public Dataset 117
5.5 Conclusion 120

6 Conclusion and Future Work **121**

6.1 Conclusion 121
6.2 Future Work 122

References **125**

List of Figures

1.1 Multimodal information of the "2008 Summer Olympics Opening Ceremony" event. 3

1.2 Fragmented information of the "2013 Typhoon Haiyan" event on social media. 4

1.3 Cross-platform coverage of the "2018 California Wildfires" event. . . . 5

1.4 Dynamic emergence of new events on social media. 6

1.5 A summary of the thesis outline and the connection between thesis chapter and the research objectives. 12

3.1 Samples from CrisisMMD dataset [5] and our proposed SED dataset about the social event of "Hurricane Irma". 27

3.2 Distribution of event-related keywords in (a) Twitter posts related to 40 social events and (b) the CrisisMMD dataset. 28

3.3 The framework of MFEK. 31

3.4 Co-attention Transformer. 37

3.5 The distribution of SED dataset. 42

3.6 Examples of data samples in SED dataset. 44

3.7 F1 score on validation dataset for MFEK model under different number of attention heads (higher is better). 55

3.8 The impact of different external knowledge extraction methods. . . . 55

3.9 Success and failure examples predicted by OWSEC and MFEK. (**T**: text, **GT**: ground truth, **C**: caption, and **O**: OCR) 56

4.1 As a social event develops, different platforms provide information from different perspectives for it. 60

4.2 An overview of the proposed SSMC method when the text of the target domain is missing. The upper flow represents the source platform with labeled data (exemplified by Twitter posts with both images and accompanying text); and the lower flow depicts the target platform with unlabeled data (Flickr for example, where images are prevalent without text). Best viewed in color. 64

4.3 Missing Data Complementation Module. 66

4.4 Multimodal Self-learning Module. 67

4.5 The distribution of CSED dataset. 74

4.6 Feature visualization of the CSED dataset for different platforms. . . 75

4.7 Sensitivity of α and β ($T \rightarrow F$, missing rate: 40%). 82

4.8 Accuracy of pseudo label during training ($T \rightarrow F$, missing rate: 20%) 83

4.9 Length of consistent and inconsistent samples during training ($T \rightarrow F$, missing rate: 20%, batch size: 40) 84

4.10 Visualization of common features and modality-specific features from the target domain under $T \rightarrow F$ and $T \rightarrow O$ (missing rate: 20%). . . 85

4.11 Visualization of multimodal features from source and target domains under $T \rightarrow O$ (missing rate: 20%). 85

4.12 Success and failure examples induced by SSMC on the CSED dataset from two different scenarios, i.e., $T \rightarrow F$ and $T \rightarrow O$ (missing rate: 20%). 86

4.13 Samples detected by using SSMC for the "Typhoon Haiyan" event with Twitter as the Source platform and Flickr and Online News as the Target platforms. 87

5.1 Different settings for social event detection. Events 4 and 5 are new events that do not occur in the training set. 90

5.2 The framework of the proposed Dynamic Augmentation and Entropy Optimization (DAEO) model. 95

5.3 Multimodal augmentation module. 96

5.4 Distribution of the MSED dataset over time. 104

5.5 Distribution of the MSED dataset for different types of events over time. 104

5.6 Parameter sensitivity on the MSED dataset with the proportion of 50%. 113

5.7 TSNE visualization of multimodal features from selected social events with the proportion of 50%. 115

5.8 Failure examples of DAEO on the MSED dataset with the proportion of 50%. 116

List of Tables

3.1 The statistics of SED dataset. ("#" represents the number of samples.) 44

3.2 Comparison of existing datasets. ("#" represents the number of samples. "N.A." for Twevent indicates Not Available as the total number of events was not reported in the original dataset.) 45

3.3 Experiment results on the SED dataset. 49

3.4 Experiment results on the CrisisMMD dataset. 50

3.5 Classification performance on the test set for different variants of the MFEK model. 52

4.1 The statistics of CSED dataset. 72

4.2 Accuracy (Acc) on CSED dataset for cross-platform multimodal social event detection. 79

4.3 F1 Score on CSED dataset for cross-platform multimodal social event detection. 80

4.4 Ablation Study. M and G indicate M2M-100 model [22] and Google API respectively ($T \rightarrow F$, missing rate: 20%). 81

5.1 Statistic of the MSED dataset. 104

5.2 Comparison of existing datasets. ("#" represents the number of samples.) 106

5.3 The division of the MSED dataset in the experiments. '#New' refers to the number of new events. 106

5.4 Results on the MSED dataset. 109

5.5 Ablation study on the different components of our approach with the proportion of 50%. 111

5.6 Ablation experiment for Backbone with the proportion of 50%. 112

5.7 Ablation experiment for multilingual processing with the proportion of 50%. 112

5.8 Ablation experiment for condition of L_{Adapt} with the proportion of 50%. 114

5.9 Results of our approach for different event types with the proportion of 50%. 114

5.10 The division of the CrisisMMD dataset. '#New' refers to the number of new events. 117

5.11 Results on the CrisisMMD dataset. 118

5.12 Experimental results on the CrisisMMD dataset with closed setting (without new events). 119

Chapter 1

Introduction

1.1 Background

In January 2024, DataReportal, Meltwater and We Are Social released the "Digital 2024 Global Overview Report", which provided statistical data on global internet usage. The report highlighted that the number of active social media users has now surpassed the 5 billion mark, accounting for 62.3% of the global population, an increase of 5.6% over the same period last year. It can be observed that with the popularity of mobile devices and the mobile internet, data on social media has shown explosive growth. Platforms such as Facebook, Flickr, and Twitter have amassed large user bases. Through these media platforms, users can easily post comments, share experiences, and access news. Consequently, when a social event occurs, everyone on social networks becomes a broadcaster and commentator of the event, leading to rapid viral discussions among a vast number of online users, and generating extensive multimedia data. Particularly, social events refer to occurrences that happen in the physical world and have significant impacts on public life, which are ubiquitous and dynamic. For example, it can be natural disaster events (earthquakes, typhoons, etc.), sports (Olympic Games, FIFA World Cup Games, etc.), political events (election

activities, protests, etc.), and so on.

The detection of social events from social media aims to help people quickly and accurately understand the social events they are concerned in the huge social media data. By acquiring all kinds of news events, people can grasp the focus of society, thus providing the necessary reference for better work, study and life. The government can correctly guide social opinion by constantly detecting various emergency events, so as to maintain social stability. However, the social media data used for event detection comes from various social media platforms on the internet, and thus exhibits characteristics such as **multimodality**, **information fragmentation**, **cross-platform nature**, and **dynamic nature**. These characteristics impose higher requirements and challenges for the detection of social events, particularly in the following ways:

1. Multimodality of Social Event Data

In the early stages of the internet, text was the most common form of data presentation. With the rise of social media and mobile devices, various events are often accompanied by a large amount of multimodal data, including images, text, and videos. For example, during the Olympic events, users not only post and share a lot of text information about the opening ceremony but also upload numerous images and videos. As shown in Figure 1.1, images and videos allow users to visually understand the information about an event, while text provides more detailed analysis of the event's specifics. For the same event, although the text content posted by different users on social media may vary, the visual information is likely to be similar. Therefore, although different modalities have different expressive capabilities, these multimodal data can complement each other, helping users to understand the event comprehensively and in-depth. However, due to the heterogeneity between different modalities, how to learn the feature representation from different modalities and how to fuse these features from different modalities are current challenges that need to be addressed.

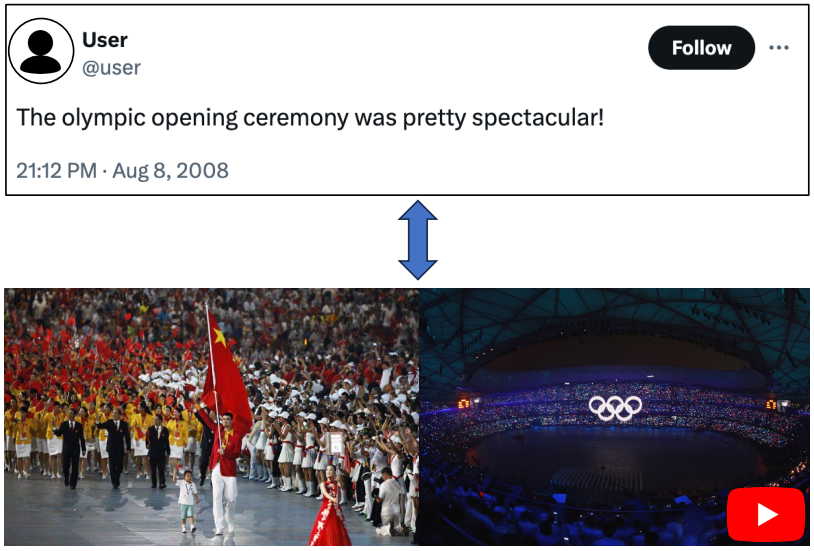


Figure 1.1: Multimodal information of the "2008 Summer Olympics Opening Ceremony" event.

2. Information Fragmentation of Social Event Data

Social media allows users to instantly share details related to events, but such sharing is often spontaneous and unstructured. Due to each user's unique perspective, knowledge background, focus, and geographical location, the content they publish varies in comprehensiveness and viewpoint, leading to information fragmentation. For instance, during a sudden public incident, different witnesses might upload various images and videos from their perspectives, accompanied by brief personal descriptions of the event. These descriptions might focus on different aspects of the event; some users might describe the causes, while others might discuss the consequences or impacts. Moreover, due to the interactive nature of social media, information can become distorted or misunderstood as it spreads among users. As shown in Figure 1.2, multiple posts are different fragments of the same event, with each post containing partial information. The fragmentation of information not only makes it more difficult to extract accurate and comprehensive reports from social media data but also requires the integration of additional background knowledge to form a complete

Figure 1.2: Fragmented information of the "2013 Typhoon Haiyan" event on social media.

understanding of the event. How to accurately understand and detect events from fragmented information has become an urgent problem to address.

3. Cross-platform Nature of Social Event Data

The cross-platform characteristic of social media data, also known as data multi-source, refers to the distribution of multimedia data related to the same event across various social media platforms. Comprehensive event detection requires gathering data from these different platforms. Currently, there are numerous social media platforms, and while they may provide similar information about events, they differ in format. For the same event, since different platforms present data from various perspectives, analyzing data from a specific platform in isolation can make it difficult to comprehensively understand the event. For example, Twitter, as a popular social media platform, primarily features users commenting on and sharing news highlights, usually in brief; Flickr, by contrast, focuses on sharing images, allowing users to convey themes through photographs. Thus, data from different platforms empha-

Figure 1.3: Cross-platform coverage of the "2018 California Wildfires" event.

size different aspects, each with unique characteristics. Utilizing the complementary perspectives of cross-platform data can provide a more comprehensive understanding of an event. As shown in Figure 1.3, reports and opinions about the "2018 California Wildfires" event quickly appear on Twitter, while Flickr and YouTube provide supplementary information through images and videos. However, the challenge of cross-platform event detection arises due to the varying data structures and focus of descriptions across platforms. Designing a universal method for cross-platform event detection has become crucial.

4. Dynamic Nature of Social Event Data

The occurrence and development of social events are ongoing and dynamic, especially on social media where the coverage of events is continuous and each new event can introduce fresh information and focal points for discussion. For instance, discussions about natural disasters, major accidents, or other sudden public events can suddenly emerge on social media, as shown in Figure 1.4. These new events often appear without any warning and may be completely different from previous incidents. Furthermore, the coverage of new events usually comes with a vast amount of user-generated content, which includes various types of data (text, images, videos, etc.), making it informative and complex. The reactions and discussions about new events by users are diverse, involving different viewpoints and emotional expressions, which further complicates the accurate detection and understanding of new events from these data sources. Given this dynamic nature, the challenge for event detection technology is how to effectively identify and respond to new events from large-scale social media

Figure 1.4: Dynamic emergence of new events on social media.

data.

In summary, due to social event data coming from different social media platforms and possessing characteristics such as multimodality, information fragmentation, cross-platform nature, and dynamic nature, research on social event detection from social media data is highly challenging. Considering the aforementioned characteristics of social media data, studying social event detection methods, and designing effective detection models are the main research focuses of this thesis.

In the following sections of this chapter, we first introduce the research objectives we aim to achieve in this thesis in Section 1.2. Then, in Section 1.3, we present our proposed solutions to these objectives. Subsequently, we summarize the main contributions of this thesis in Section 1.4. Finally, the organization of the thesis will be outlined in Section 1.5.

1.2 Research Objectives

Before introducing our research objectives, it is important to clarify that social event detection in this thesis is approached as a classification problem, where we aim to categorize social media posts into their corresponding event classes by analyzing their multimodal content. While technically implementing a classification framework, we use the term "detection" as it better describes our goal of discovering and identifying real-world events from social media streams. This classification-based detection approach is particularly suitable for our task as it enables learning discriminative features to distinguish between different types of events, provides a natural framework for fusing multimodal inputs, and allows extension to more complex scenarios like cross-platform detection and new event discovery while maintaining a consistent methodological foundation. With this framework in mind, this thesis aims to address the challenges of multimodal social event detection by focusing on two key issues: **limited detection accuracy** and **insufficient generalization capability**. For the first issue, we focus on addressing the challenges of modality heterogeneity inherent in multimodal social media event data and the out-of-distribution (OOD) problem caused by information fragmentation (Objective 1), which can improve the accuracy of event detection. Regarding the model's generalization capability, we focus on two aspects: enhancing the model's adaptability and flexibility through cross-platform capabilities (Objective 2), and improving the model's ability to detect new events to strengthen its performance on unknown events (Objective 3). The specific objectives can be summarized as follows:

1. To design a deep learning model to address modality heterogeneity and the OOD problem, thereby improving the accuracy of event detection. This objective centers on the multimodality and fragmented nature of information in social media, which means that the model must effectively integrate diverse data types such as text and images. Additionally, it must handle the inconsistencies

and gaps in information that can lead to out-of-distribution scenarios, ensuring robust performance even when the available data is incomplete or scattered. Existing studies [65, 103, 106, 113], although beginning to use basic multimodal fusion techniques to integrate images and text for improved detection accuracy, still largely rely on event keywords present in the text. When faced with the OOD problem caused by information fragmentation, their performance often deteriorates.

2. To develop an innovative manner to adapt models to implement cross-platform social event detection. Cross-platform social event detection can improve the quality of single-platform event data because information from other platforms can supplement and verify the event data on a single platform. This objective explores effective event detection across multiple social platforms, particularly how to overcome the differences in data distribution and modal incompleteness between platforms. Existing works [31, 83, 111] typically design event detection models for a specific platform, such as Twitter or Flickr, because their data is relatively easy to access. However, these models often perform poorly when applied to other platforms due to the domain gap that exists between different platforms.

3. To extend existing supervised event detection methods to discover new social events in social media. This objective focuses on the generalization capability of supervised event detection models, especially on how the model can discover and classify new types of events while maintaining accuracy in recognizing predefined events. Existing studies for new event detection include methods such as unsupervised clustering techniques [3, 8, 10] and graph-based neural networks (GNN) [13, 20, 27, 39]. Unsupervised clustering methods attempt to group similar events based on the intrinsic structure of the data without explicit labels, while graph-based methods leverage relationships between events (such as similarity or temporal connections) to enhance the model's ability to

recognize new types of events. However, these methods have limitations. Unsupervised clustering methods often rely on the intrinsic structure of the data, which might make it difficult to accurately distinguish subtle differences between events or respond quickly to new events. Graph-based methods, while able to utilize relationships between events, may not be robust enough in the absence of sufficient training data, especially in the highly dynamic social media environment. Furthermore, due to the lack of event labels for training, the learned event representations are often unstable.

1.3 Overview of Proposed Solutions

For the three objectives mentioned above, we propose corresponding solutions as follows. To accomplish the first objective, we use an attention model to integrate text, images, and their corresponding external knowledge, thereby fusing multimodal data and supplementing the incomplete event information on social media. We propose a **Multi-modal Fusion with External Knowledge (MFEK)** model, which incorporates attention mechanisms and external knowledge from Wikipedia and large language models (LLMs) to enhance the original data. This approach aims to supplement the fragmented information in social media data with external knowledge and integrate this knowledge into multimodal data, thereby improving the accuracy and completeness of event detection.

To achieve the second objective, we employs domain adaptation techniques to enhance the model's cross-platform capabilities. We propose a **Self-Supervised Modality Complementation (SSMC)** method, which not only addresses the platform heterogeneity but also considers the common issue of incomplete modalities across platforms. By leveraging self-supervised learning, this approach enables the model to adapt to different data distributions and modalities, enhancing its flexibility and effectiveness in diverse social media platforms.

For the third objective, we propose a generalize social event detection task, which requires the model to identify predefined events and distinguish various new emerging events. We design a **Dynamic Augmentation and Entropy Optimization (DAEO)** model, which utilizes data from known events to learn robust multimodal event features with a large amount of labelled data and explores different features of new events through self-distillation learning. This enables the model to quick recognition of known events and the discovery of new ones. This approach facilitates the continuous adaptation of the model to new data, enabling it to respond to the dynamic nature of social media content.

1.4 Thesis Contributions

Our main contributions in this thesis can be summarized as follows:

- We propose a MFEK model to integrate external knowledge to improve the model's accuracy in social event detection. The MFEK model features a text enrichment module to enhance the textual content, a knowledge extraction module to complement the incomplete event information, and a knowledge-aware feature fusion module to integrate external knowledge, text, and images, while filtering out irrelevant information. The proposed model has achieved better performance than the compared baseline models.

- We propose a SSMC method to tackle the challenges of incomplete modalities and platform heterogeneity presented in the cross-platform social event detection. The SSMC model consists of a Missing Data Complementation (MDC) module to complement missing modalities with modality-shared features and a Multimodal Self-Learning (MSL) module to tackle platform heterogeneity by self-learning. The proposed approach has outperformed all the baselines and achieved the new state-of-the-art performance for the cross-platform social event

detection task.

- We propose a new task, generalized social event detection, to identify both predefined events and various new emerging events, and design a DAEO model to handle the proposed task. The DAEO model includes a multimodal augmentation module to enhance the multimodal representation capability and an adaptive entropy optimization strategy to improve the model's ability to discriminate new events. The generality and effectiveness of the proposed method are validated through comprehensive experimental studies.

- We collect three large-scale datasets for social event detection tasks in different scenarios, which are annotated with real-world social events verified by Wikipedia. And we conduct extensive experiments on these datasets as well as publicly available ones, which manifest the effectiveness of our proposed models in terms of accuracy and generalization.

1.5 Thesis Outline

The main content and structure for this thesis are shown in Figure 1.5. In this thesis, we are dedicated to enhancing the accuracy and generalizability of multimodal social event detection. For research objective 1, aimed at addressing modality heterogeneity and the OOD problem, we propose the MFEK model that incorporates attention mechanisms and external knowledge to tackle these issues, which is detailed in Chapter 2. To enhance the model's generalizability, we introduce tasks for cross-platform social event detection and generalized social event detection. For cross-platform social event detection (research objective 2), we propose the SSMC model to address the challenges of modal absence and cross-platform distribution heterogeneity. For the detection of new events (research objective 3), we propose the generalized social event detection task to expand the detection scope of the original predefined events,

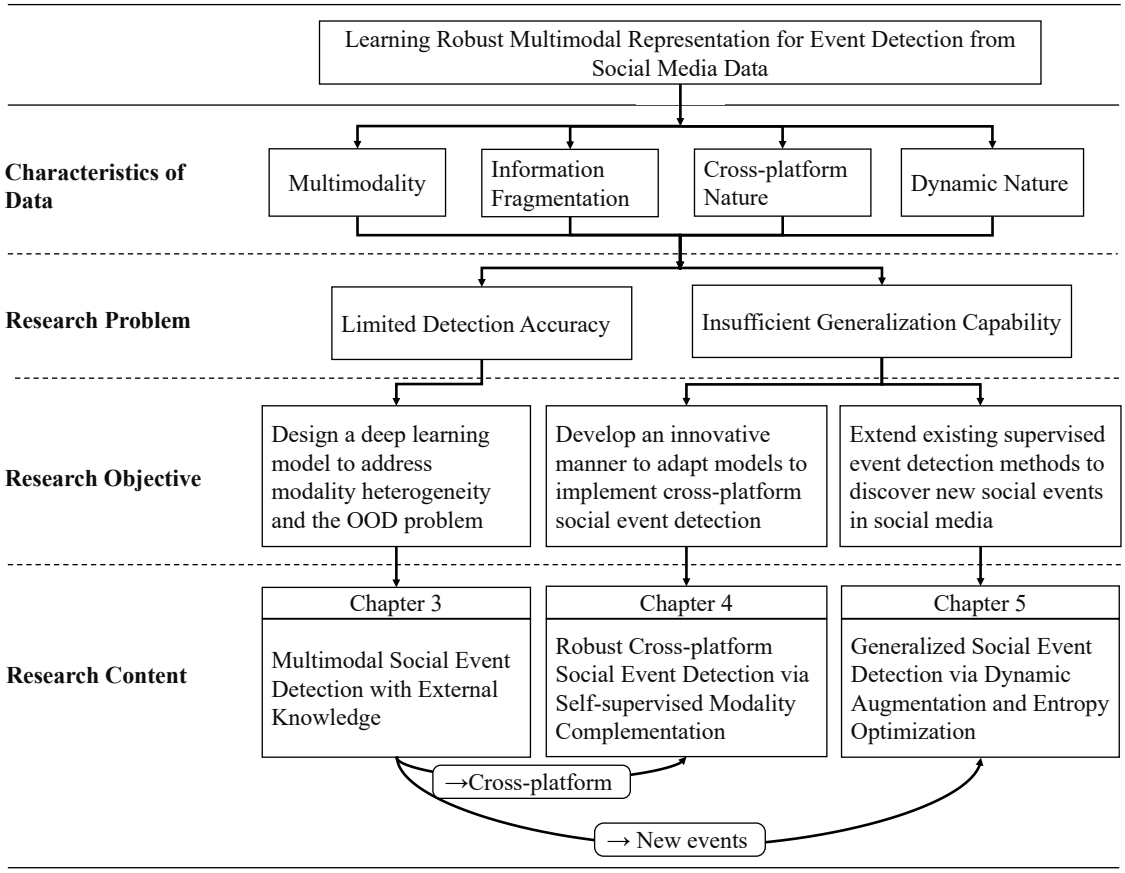


Figure 1.5: A summary of the thesis outline and the connection between thesis chapter and the research objectives.

and introduce the DAEO model to simultaneously identify both predefined and new events. We will discuss these three models in the following chapters. The detailed outline is as follows:

In Chapter 2, we investigate related work on social event detection, benchmark datasets for social event detection, multimodal data fusion, and related methods for cross-platform event detection and new event detection, including missing-modality for multimodal learning, domain adaptation and generalized category discovery.

In Chapter 3, we propose a MFEK model for multimodal social event detection, and conduct extensive experiments and analyses to evaluate its performance.

In Chapter 4, we propose a SSMC model for cross-platform social event detection, and demonstrate its performance in cross-platform scenarios through experiments and analysis.

In Chapter 5, we propose a DAEO model for generalized social event detection, and conduct extensive experiments and analyses to evaluate its performance.

In Chapter 6, the conclusion and future work are presented.

Chapter 2

Related Work

In this chapter, we first review some representative works for social event detection from social media data, which will provide the background and current state of research. Next, we explore works related to benchmark datasets used for social event detection. Furthermore, we investigate the state-of-the-art methods for multimodal data fusion. As our work also targets cross-platform event detection and new event detection, we review the developments in missing-modality for multimodal learning, domain adaptation and generalized category discovery.

2.1 Social Event Detection

Initially, social event detection referred to topic detection and tracking [6], aimed at discovering real-world events from news media articles. With the development of the internet, its search scope expanded to social media. Social event detection can be divided into single-modal event detection and multimodal event detection based on the modal data used.

2.1.1 Single-modal Event Detection

Single-modal event detection [31, 34, 51, 68, 111] refers to the detection of events using data from a single modality. These approaches can be categorized into three main streams based on the type of data they process:

Text-based Methods: These methods focus on analyzing textual content through various natural language processing techniques. Lee et al. [51] employed a naive Bayes multinomial classifier with TF-IDF features for identifying distinct trending topics, achieving efficient real-time detection but struggling with semantic understanding. In contrast, Hu et al. [42] proposed an LSTM-based model that captures temporal dependencies in text sequences, demonstrating superior performance in learning shared event representations between different tasks through a hierarchical architecture.

Image-based Methods: These approaches leverage visual features for event detection. Zaharieva et al. [116] developed a visual content analysis framework that combines low-level visual features with temporal information, particularly effective for specific social events like concerts and sports. Guo et al. [31] advanced this direction by proposing a hierarchical neural model that first extracts local features using CNNs and then models temporal relationships through a hierarchical structure, significantly improving event recognition accuracy in personal photo collections.

Video-based Methods: These methods utilize temporal visual information for event detection. Zhang et al. [118] proposed a two-stage architecture where the first stage extracts object-level knowledge through a pre-trained object detection network, and the second stage integrates temporal information using RNNs, achieving state-of-the-art performance on video event classification tasks.

However, single-modal approaches face inherent limitations in capturing complete event semantics. For instance, while text-based methods excel at extracting explicit event descriptions, they miss visual context that could be crucial for event understanding. Similarly, image-based methods might capture visual elements but

lack contextual information often present in text descriptions.

2.1.2 Multimodal Event Detection

Social media consists of rich unstructured data with multiple modalities that can complement one another. They help express the complete meaning of social event analysis. To address the limitations of single-modal event detection, multimodal event detection [56, 65, 83, 103, 106, 113, 114] has emerged as a more comprehensive approach. These methods can be categorized based on their fusion strategies:

Early Fusion Methods: Yang et al. [114] pioneered the combination of video features with metadata, using a concatenation-based fusion strategy that showed significant improvements over single-modal approaches. However, this simple fusion strategy often struggles with modality alignment issues.

Late Fusion Methods: Wu et al. [106] proposed a hierarchical fusion framework that first processes each modality independently and then combines their decisions, demonstrating better robustness to modality noise but potentially missing inter-modal correlations.

Interactive Fusion Methods: More recent works focus on interactive fusion strategies. Li et al. [56] proposed AT-CVAE, which employs a transformer-based architecture to model cross-modal interactions dynamically. This approach showed superior performance in capturing complex inter-modal relationships. Building on this, Qian et al. [83] developed OWSEC, introducing a mask transformer network that explicitly models cross-modal semantic relations, achieving state-of-the-art performance on several benchmark datasets.

However, these works do not consider the OOD problem, where the performance deteriorates when the scenario of information fragmentation emerges. In addition, their methods are limited to a single platform and they assume that the training and

test events remain consistent (i.e., a closed set), and thus the performance will drop significantly when conducting cross-platform event detection or new event detection.

2.1.3 New Event Detection

When new events emerge, classification-based methods often fail due to their closed-world assumptions. To address this, clustering-based detection methods [3, 8, 10, 13, 20, 27, 39] have been proposed. These methods can be categorized into two main approaches:

Traditional Clustering Methods: These approaches focus on grouping similar event data without requiring predefined labels. Becker et al. [18] developed an incremental clustering algorithm that leverages rich contextual information from social media data, achieving good performance in detecting emerging events. Ma et al. [64] employed K-means clustering on multimedia feature vectors, demonstrating effectiveness in distinguishing event categories but struggling with complex event boundaries.

Graph-based Methods: These methods model events as graph structures to capture complex relationships. Zhao et al. [119] represented social media data as an MC graph and used transitive segmentation for event detection, showing superior performance in capturing event evolution. Chu et al. [19] proposed a graph-shift detection method that identifies local maxima as event indicators, effectively handling complex event structures.

Despite the ability of these methods to discover new events, clustering-based methods rely on the intrinsic structure and features of the data for event detection. This means that if the data's intrinsic structure is complex or noisy, the effectiveness of clustering may be compromised, making it difficult to accurately distinguish between different events, particularly for data points that are highly similar but actually belong to different events. Additionally, most clustering algorithms require a global analysis of the entire dataset, which is computationally expensive and results in relatively

poor interpretability.

2.2 Benchmark Datasets for Social Event Detection

With the widespread use of social media, social event detection has shifted from traditional data like online news to diverse social media data (such as Twitter, Flickr, etc.), providing a richer information source but also presenting more challenges. Currently, most datasets based on social media are multimodal datasets, as multimodal data can provide more event-related elements to the model. Reuter et al. [87] collected a social event detection (SED) dataset, which comprises 427,370 images from Flickr and 1,327 videos from YouTube. Xue et al. [109] compiled a multi-modality social event dataset (MMSE) from Flickr, consisting of 74,364 documents that encompass 10 types of events. Alam et al. [5] collected data on seven crisis events that occurred worldwide in 2017, which includes 18,126 image-text pairs. Zubiaga et al. [122] gathered a PHEME dataset that contains nine categories of events, including 2,089 image-text pairs and 3,713 texts. Yang et al. [114] assembled a temporal event dataset (TED), consisting of 16,589 videos and accompanying metadata from Youtube. However, these datasets are all collected based on event keyword search, which may overlook relevant posts due to the diversity of language use. This not only simplifies the task but also weakens the role of other modalities, resulting in a substantial discrepancy with real-world scenarios. In addition, these datasets are collected from single platforms, and the event labels differ between datasets, making them unsuitable for cross-platform social event detection.

2.3 Multimodal Data Fusion

Multimodal data fusion refers to the process of integrating information from multiple modalities (e.g., visual, audio, textual) to improve the performance of a task, which can help us address modal heterogeneity in social event detection. This can be achieved through various methods, including joint-embedding-based fusion, tensor-based fusion and attention-based fusion.

One of the common methods for multimodal data fusion is joint-embedding-based fusion. This method aims to learn a common embedding space to capture shared information across different modalities. For example, Pham et al. [78] proposed a method that captures joint representations via cyclic translations from source to target modalities. Hazarika et al. [35] proposed a method that factorizes modalities into modality-invariant and modality-specific features in two distinct subspaces. The advantage of joint embedding approaches lies in their ability to directly learn shared semantic spaces. However, they face several key challenges: they typically require extensive training data to learn effective embeddings, struggle with capturing complex non-linear relationships between modalities, and may lose modality-specific information during the embedding process.

Tensor-based methods represent another sophisticated approach for multimodal data fusion. These methods treat multimodal data as tensors and leverage tensor decomposition techniques to extract cross-modal patterns. For instance, Liu et al. [59] proposed a multimodal fusion approach that utilizes modality-specific low-rank factors to reduce computational complexity while preserving inter-modal relationships. Chen et al. [17] advanced this direction by introducing adaptive tensor decomposition that automatically determines the importance of each modality. The key advantage of tensor-based methods is their ability to capture higher-order correlations between modalities. However, these methods face significant computational challenges with large-scale data and often require careful preprocessing to handle missing or noisy

inputs. Additionally, the interpretation of tensor decomposition results can be less intuitive compared to other fusion approaches.

Recently, attention-based methods [1, 48, 76, 102] have emerged as a powerful paradigm for multimodal fusion. These approaches leverage self-attention mechanisms to dynamically model relationships between different modalities. Kiela et al. [48] developed a supervised multimodal transformer that learns to project image features into text token space, enabling direct cross-modal interaction. Yao et al. [96] extended this idea by modeling temporal dependencies in multimodal sequences. Abavisani et al. [1] introduced a cross-attention module specifically designed to filter out irrelevant information from weaker modalities in social event detection. Han et al. [33] further improved fusion performance through hierarchical mutual information maximization. Attention-based methods offer several advantages: they can capture complex dependencies between modalities, adapt to varying input qualities, and provide interpretable attention weights. However, these benefits come with increased computational costs and a requirement for substantial labeled training data. Moreover, attention mechanisms may struggle with very long sequences or when modalities have significantly different temporal or spatial characteristics.

Each fusion approach offers distinct advantages and faces unique challenges in the context of social event detection. Joint embedding methods excel at learning shared semantic representations but may oversimplify complex relationships. Tensor-based approaches can capture sophisticated cross-modal patterns but face scalability issues. Attention-based methods offer flexible and interpretable fusion but require significant computational resources. The choice of fusion strategy often depends on specific application requirements, such as computational constraints, data availability, and the nature of cross-modal relationships being modeled.

2.4 Missing-modality for Multimodal Learning

The missing modality problem poses significant challenges in the field of cross-platform social event detection. Research in this area has predominantly followed two approaches to mitigate the impact of missing modalities: 1) Generative methods [55, 56, 94] focuses on using generative models to predict or recreate the missing modalities based on the modalities that are available. These methods leverage the power of generative models to fill in the gaps where data is incomplete, thereby ensuring that the system can still perform its intended function despite the absence of some modalities. For instance, Suo et al. [94] introduced a new framework focused on learning patient similarity from multimodal healthcare data, even when some of those modalities are missing. Li et al. [56] utilized a conditioned variational autoencoder for generating the missing modalities, thereby facilitating event detection. 2) Joint multimodal representation learning [52, 63, 117] focuses on creating a unified representation that contains information from all available modalities, even in the absence of some. This kind of method aims to leverage the shared information across modalities to infer missing data and maintain performance. For example, Ma et al. [63] proposed using Bayesian meta-learning to estimate the latent features of data. Similarly, Lee et al. [52] explored multimodal prompting with missing modalities for visual recognition. However, the scenarios configured by these methods typically require at least three modalities, which proves challenging to apply in some bimodal scenarios, i.e., cross-platform social event detection.

2.5 Domain Adaptation

To enable the model to learn cross-platform capabilities, we introduced domain adaptation into cross-platform social event detection. Given a source domain and a target domain with different distributions, the goal of domain adaptation is to learn a model

that can generalize well on the target domain by leveraging knowledge from the source domain. Research in this area can be divided into four categories: discrepancy-based methods, adversarial discriminative models, adversarial generative models, and self-supervision-based methods.

Discrepancy-based methods aim to minimize the discrepancy between source and target domains by measuring the distance between their distributions. One popular approach is Maximum Mean Discrepancy (MMD) [60], which minimizes the distance between the means of the source and target feature representations. The effectiveness of MMD lies in its ability to match distributions in high-dimensional feature spaces through kernel tricks. Another influential approach is Correlation Alignment (CORAL) [91], which aligns the second-order statistics between domains. CORAL has demonstrated strong performance due to its computational efficiency and theoretical guarantees. Recent advances in discrepancy-based methods have focused on incorporating local structure preservation and adaptive weighting mechanisms to better handle complex domain shifts.

Adversarial discriminative models leverage the power of adversarial training to learn domain-invariant features. In these approaches, a discriminator tries to distinguish between source and target domains, while a feature extractor aims to generate features that are indistinguishable between the two domains. Domain Adversarial Neural Network (DANN) [26] pioneered this direction by introducing a domain adversarial loss to existing neural network architectures. This encourages the feature extractor to learn domain-invariant features by forcing its output to be indistinguishable between source and target domains. Building upon DANN, Joint Adaptation Network (JAN) [61] introduced a joint maximum mean discrepancy loss that simultaneously aligns the distributions of multiple network layers. This multi-layer adaptation strategy has proven particularly effective for complex domain adaptation tasks.

Adversarial generative models take a different approach by generating synthetic samples in the target domain using a generator network. These methods typically

employ a generator trained to produce samples indistinguishable from the target domain, while a discriminator attempts to distinguish between real and synthetic samples. Adversarial Variational Domain Adaptation (AVDA) [77] combines adversarial learning with variational inference to learn a unified latent space where source and target domains can be effectively aligned. A notable advancement in this direction is Cycle-Consistent Adversarial Domain Adaptation (CyCADA) [38], which introduces a cycle-consistency constraint to ensure the preservation of semantic information during domain translation. These generative approaches have shown remarkable success in producing high-quality synthetic samples and improving classifier accuracy on target domain data.

Self-supervision-based methods represent a recent trend in domain adaptation, utilizing auxiliary tasks to learn domain-invariant features. These methods exploit the inherent structural similarities between different domains through unsupervised or self-supervised learning techniques. Deep Reconstruction-Classification Network (DRCN) [28] exemplifies this approach by introducing a reconstruction-based loss that encourages the model to preserve semantic information while discarding domain-specific details. Domain-Adaptive Meta-Learning (DAML) [115] advances this concept further by employing meta-learning to adapt model parameters to new domains with limited labeled samples. These self-supervised approaches have shown particular promise in scenarios where labeled data is scarce.

Despite these advances, significant challenges remain in domain adaptation for multimodal social event detection. The aforementioned approaches often struggle when faced with substantial domain gaps, particularly in multimodal scenarios where the complexity of the adaptation task increases significantly. Moreover, these methods typically assume complete availability of all modalities, making them unsuitable for scenarios involving missing modalities - a common occurrence in real-world cross-platform settings. These limitations highlight the need for specialized adaptation approaches that can handle both modality gaps and missing data scenarios effectively.

2.6 Generalized Category Discovery

To extend the classification-based event detection model to recognize new events, we proposed generalized social event detection, which is very similar in setup to generalized category discovery. The field of generalized category discovery has recently emerged, focusing on classifying known categories while also identifying unseen, new categories. The pioneering work in this domain was conducted by Vaze et al. [99], who introduced the idea of leveraging a universal feature representation to discover new categories. Specifically, they proposed fine-tuning a pre-trained DINO ViT [15] using a combination of one supervised and one self-supervised contrastive method. This approach is further complemented by a semi-supervised clustering for label assignment. In addition, the authors extended UNO [24] and RankStats [32] for this task, which were originally designed for novel class discovery [24, 120]. However, these methods employ a two-step training process, involving feature learning and clustering, which could potentially be sub-optimal. To address this, Wen et al. [105] suggested parametric approaches that construct a trainable classifier, enabling the joint optimization of the entire network. Similar to the idea behind DINO ViT [15], their method used the generation of pseudo cluster labels to guide the learning of new categories. This work sparked a series of follow-up studies [72, 101]. For example, Wang et al. [101] proposed the use of CLIP-generated text to guide image learning for category discovery. Nevertheless, it is challenging to apply these methods directly to generalized social event detection, which involves multimodal data and higher-level event labels. This is because these methods rely on image data and pre-trained models developed primarily for similar tasks, which emphasizes the need for specialized adaptation in generalized social event detection.

Chapter 3

Multimodal Social Event Detection with External Knowledge

In this chapter, we propose a novel deep learning network, MFEK, for multimodal social event detection, which utilizes the attention mechanism and external knowledge to deal with the modality heterogeneity and OOD problem, thus improving the accuracy of multimodal social event detection. Additionally, we introduce a dataset for multimodal social event detection and conduct extensive experimental analysis.

3.1 Introduction

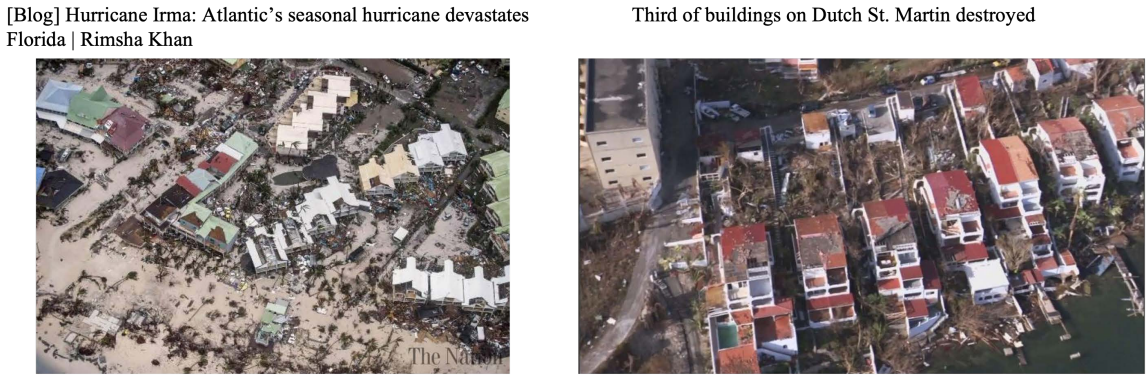
Social event detection is a critical task that involves the automated monitoring, identification, and categorization of major happenings discussed on various media, especially social media. The task primarily employs Natural Language Processing (NLP) and Machine Learning (ML) techniques to filter out the noise and identify the key events. Applications of social event detection span various domains, including crisis management [79], public sentiment monitoring [69], market analysis [89], and public safety [67]. In these contexts, accurate and prompt identification of social events

can help organizations and individuals respond more swiftly, better understand and analyze public opinion, and manage information flow more effectively.

In recent years, significant strides have been made in the field of social event detection research. Various studies [90, 25, 74, 29, 81] have been conducted to enhance the accuracy and efficiency of social event detection, leading to the development of sophisticated algorithms and models. For instance, Goyal et al. [29] proposed a novel incremental clustering algorithm to detects events and subevents within an event. However, a common limitation across many of these studies is the reliance on single-modal datasets, which may not fully capture the complexity and multifaceted nature of real-world social events.

Single-modal datasets, while useful in certain scenarios, often fail to provide a comprehensive view of social events. This is due to their inherent limitation of being able to capture only one type of data (e.g., text, images, or audio), thereby missing out on the rich information present in other modalities. Therefore, there is a growing consensus in the research community about the need to adopt multimodal datasets for a more holistic social event detection, because many events on social media are implicit, requiring the extraction of event elements from images or other domains.

There are two primary challenges for multimodal social event detection, i.e., multimodal data fusion and out-of- distribution (OOD) issues. Multimodal data fusion is a common challenge in multimodal datasets, which aims to solve the problem of multimodal data heterogeneity. Many researchers focus on this challenge in social event detection tasks [70, 1, 56], e.g., Li et al. [56] proposed an adaptive transformer network to encode the feature of images and text for social event detection. However, most research only considers directly concatenating features from different modalities, lacking interaction between modalities. The OOD problem refers to posts containing some important keywords not present in the training set, which could be a clue for social event detection. This issue has been largely overlooked, possibly because existing datasets are keyword-searched, and models can detect events based on pattern



(a) CrisisMMD (b) NED

Figure 3.1: Samples from CrisisMMD dataset [5] and our proposed SED dataset about the social event of "Hurricane Irma".

recognition of these keywords, even without additional information. However, posts lacking event keywords are far more common, as illustrated in Figure 3.1b. In such cases, addressing the OOD problem becomes critical. The reason is that "Dutch St. Martin" can serve as a significant clue to infer that the social event may be related to "Hurricane Irma" as Hurricane Irma brought tremendous destruction to Dutch St. Martin.

Another critical aspect to consider is the construction methods of current social event detection datasets. The majority of existing multimodal datasets are collected based on pre-defined event keywords, which tend to overfit to specific data characteristics and may contain the potential biases in the data collection process. In reality, numerous social media posts relevant to a social event do not necessarily mention the specific keywords associated with that event. As illustrated in Figure 3.1a, the post in the keyword-based dataset, i.e., CrisisMMD [5], inevitably contains the event keyword "Hurricane Irma". In fact, it is more common for posts not to include event-related keywords. In order to illustrate this point, we conducted an empirical analysis, as depicted in Figure 3.2a. We collected related posts for 40 social events from Twitter

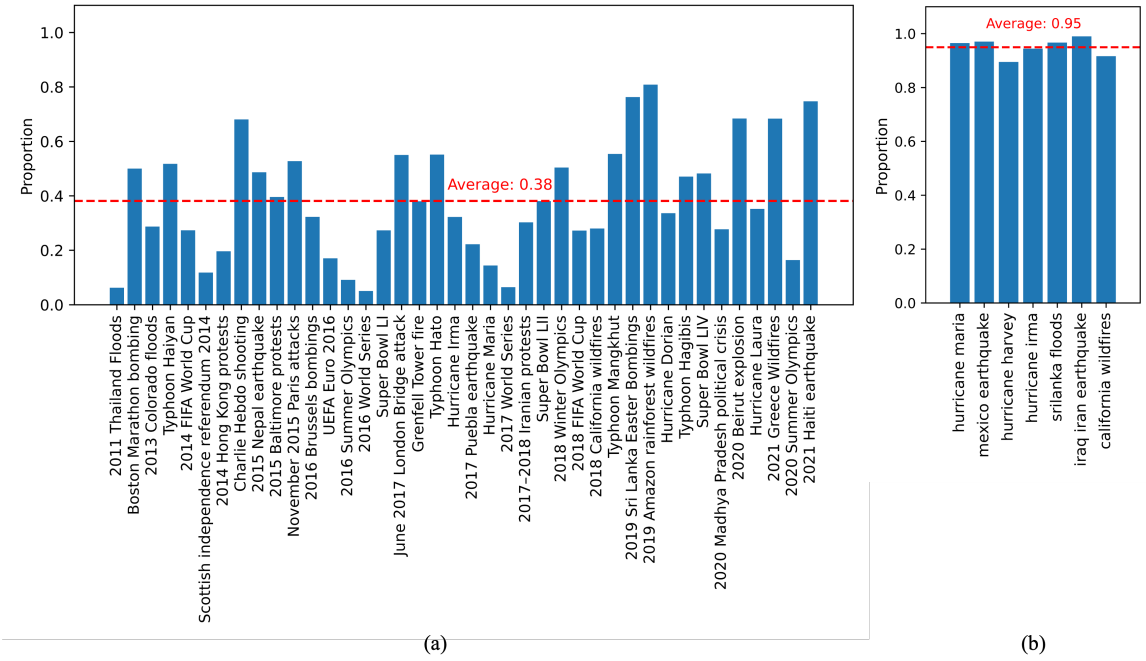


Figure 3.2: Distribution of event-related keywords in (a) Twitter posts related to 40 social events and (b) the CrisisMMD dataset.

using hashtags to simulate real-world scenarios. These posts were then analyzed using event-related keywords (i.e., event names), with a match being counted if any keyword appeared (excluding hashtags). Based on the collected data, we can find that the proportion of posts containing event-related keywords in real-world scenarios is only an average of 38%, indicating the limitations of keyword-based retrieval. At the same time, we also utilized the same method to count the existing datasets based on event-related keyword searches, i.e., CrisisMMD. As shown in Figure 3.2b, the proportion of event keywords is close to 100%. Therefore, keyword-based datasets may overlook relevant posts due to the diversity of language use and are subject to user bias, thereby simplifying event detection and compromising effectiveness in real-world scenarios.

Based on the limitations of existing work, the key objectives of this chapter

are two-fold: (1) to design an effective model that handles these key challenges in social event detection, which can serve as a benchmark for this task; (2) to develop a more realistic benchmark dataset, which can enable researchers to delve deeper into multimodal fusion and OOD challenges in multimodal social event detection task.

In this chapter, we propose a Multimodal Fusion with External Knowledge (MFEK) model, addressing the multimodal fusion and OOD issues in social event detection. MFEK integrates a text enrichment module, an external knowledge extraction module, and a knowledge-aware feature fusion module. Specifically, the text enrichment module primarily extracts image information (i.e., image captions and OCR information) to enrich text information. The external knowledge extraction module includes explicit and implicit knowledge extraction. Explicit knowledge is obtained through external sources of knowledge (e.g., Wikipedia passages [100]), while implicit knowledge is acquired using the large language model (e.g., ChatGPT [11]). The combination of these two knowledge types enables our model to handle the OOD problem effectively. The knowledge-aware feature fusion module employs multiple co-attention Transformers to integrate image, text, and knowledge data, filtering out irrelevant knowledge.

Furthermore, we collect a real-world Social Event Detection (SED) dataset comprising 17,366 posts with text-image pairs from Twitter, annotated with 40 real-world events. SED presents several advantages over existing ones. Firstly, SED leverages user hashtags for data collection, which aligns closely with real-world scenarios and reduces reliance on event-specific keywords. Secondly, SED encompasses a broad scope of social event themes, including political events (e.g., elections and referendums, political crises and protests), sports events (e.g., the Olympics and soccer), natural disasters (e.g., hurricanes and floods), etc. Lastly, SED poses more significant challenges due to the presence of numerous similar new events within each topic category. For instance, in the category of the Olympics, the dataset includes similar events like “2016 Summer Olympics” and “2018 Winter Olympics”.

The main contributions of this chapter can be summarized as follows:

1. We propose a method for social event detection that incorporates both explicit knowledge from Wikipedia and implicit knowledge from a large language model, offering a comprehensive approach to mitigate the out-of-distribution (OOD) problem.

2. We present a Multimodal Fusion with External Knowledge (MFEK) model, which employs a co-attention mechanism to effectively integrate knowledge, text, and image information, thereby enhancing the robustness and accuracy of social event detection.

3. We contribute a Social Event Detection (SED) dataset, annotated with real-world social event labels verified by Wikipedia. This dataset not only provides a more realistic benchmark for social event detection, but also enables researchers to delve deeper into multimodal fusion and OOD challenges in this domain.

4. We conduct extensive experiments on the SED dataset using various methods for social event detection. The results serve as a robust benchmark for future studies, promoting advancements in multimodal fusion and OOD solutions.

3.2 Problem Statement

Social events are defined as real-world occurrences that are reported through various media channels. These events are typically characterized by their significance, timeliness, and the impact they have on the public. For example, a social event could be a political rally that took place in Washington D.C., reported through text describing the event and images showing the crowd and key figures.

The task of multimodal social event detection aims to predict these specific social events based on the given posts. Specifically, let us define a social event detection

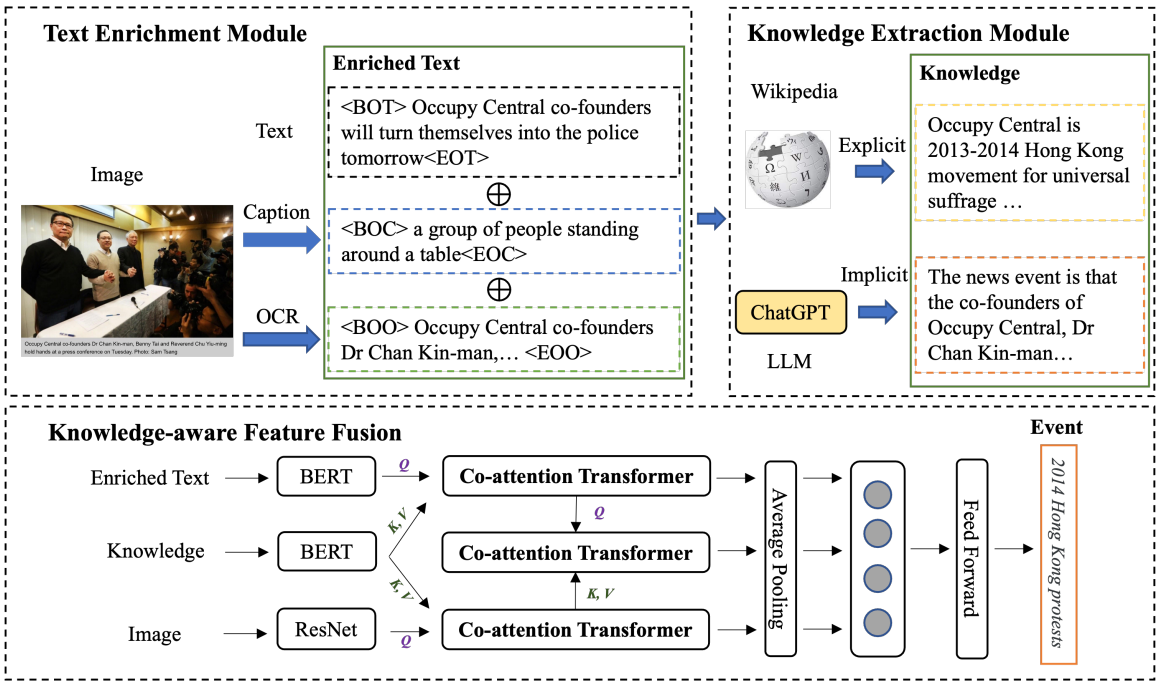


Figure 3.3: The framework of MFEK.

dataset as $D = D_{tr} + D_{te} = \{(I_i, T_i, Y_i)\}_{i=1}^{N}$, where D_{tr}, D_{te} represent the training set and the test set respectively, I_i, T_i, $Y_i \in \{1, \ldots, n_{event}\}$ represent the image, text, and social event label of the i-th input sample respectively, and N represents the total number of samples. D_{tr} and D_{te} are drawn from the same distribution with n_{event} classes. The goal of social event detection is to train a model with parameters X using D_{tr} to identify the social event E_i in D_{te} through the input I_i and T_i.

3.3 Methodology

In this section, we introduce our proposed Multimodal Fusion with External Knowledge (MFEK) method. Figure 3.3 shows a framework of our method. Specifically, MFEK consists of three main modules: a text enhancement module, a knowledge extraction module, and a knowledge-aware feature fusion module. The text enhance-

ment module enriches the original text by extracting the semantic features of the image, while the knowledge extraction model utilizes Wikipedia and large language models (LLM) to extract external knowledge from text and images. After that, the knowledge-aware feature fusion module uses the attention mechanism to merge knowledge, text, and images into a multimodal fusion to predict social events. In the following, we explain the components of the MFEK method in more detail in Sections 3.3.1, 3.3.2, and 3.3.2.

3.3.1 Text Enrichment Module

The text enrichment module aims to extract semantic information from images to enrich and supplement necessary social event elements in the text. Specifically, we consider extracting two types of semantic information from images, i.e., image-level captions and token-level optical character recognition (OCR) information. For extracting caption information from images, we utilize the state-of-the-art visual-language pre-trained (VLP) model, namely BLIP model [54]. To extract OCR information, we employ EasyOCR[1], which is a robust tool for text detection and recognition in images. Given the input images, denoted as $\{I_i\}_{i=1}^{N}$, where N is the number of samples, we obtain the image captions C_i, and OCR texts T_i, as follows:

$$C_i = \text{BLIP}(I_i), \tag{3.1}$$

$$O_i = \text{OCR}(I_i). \tag{3.2}$$

Finally, we prepend and append identifiers, such as "<BOT>" (Beginning Of Text) and "<EOT>" (End Of Text), to the caption and OCR text. This process yields the enriched text, which can be represented as

$$T_i' =< T_i, C_i, O_i >, \tag{3.3}$$

[1] https://github.com/JaidedAI/EasyOCR

where $<\cdot>$ means a merge operation.

For feature extraction, we employ the BERT model [47], which has proven effective in various tasks, including text classification [43] and question answering [84]. Given an enriched text with input length n_t, represented as $T_i' = \{w_i^1, ..., w_i^{n_t}\}$, we extract its features $F_i^{T'} = \{f_i^1, ..., f_i^{n_t}\}$ using the following method:

$$F_i^{T'} = \text{BERT}(T_i') \in \mathbb{R}^{n_t \times 768}, \tag{3.4}$$

where f_i^j represents the output feature of the j-th word.

3.3.2 Knowledge Extraction Module

The incorporation of external knowledge can effectively mitigate the out-of-distribution (OOD) issues encountered in social event detection. As illustrated in Figure 3.1 (b), "Dutch St. Martin" only appears once in the test set, and the introduction of external knowledge (i.e., Hurricane Irma brought tremendous destruction to Dutch St. Martin.) can provide a clue for the model to identify the social event "Hurricane Irma". The introduction of external knowledge is motivated by two key observations: 1) Social media posts often contain fragmented information that requires additional context to fully understand; 2) Many event-related concepts and entities have rich semantic descriptions in knowledge bases that can help bridge information gaps. Therefore, combining both explicit knowledge from Wikipedia and implicit knowledge from LLMs can provide complementary contextual information for more accurate event detection.

Explicit Knowledge Extraction

Explicit knowledge refers to the more intuitive knowledge that can be directly obtained. For instance, the description "tropical/subtropical edible staple fruit" can be

considered explicit knowledge about "banana". We utilize Wikipedia as the source of explicit knowledge, as it includes explanations and related materials of many concepts, nouns, and events. Existing methods [41] primarily consider obtaining corresponding explicit knowledge through text information. In contrast to existing methods, we consider obtaining explicit knowledge from both text and images. Specifically, for text content, we employ a text entity linkage tool, i.e., TAGME [23], which can link entities in the text with Wikipedia entries. Notably, we take the enhanced text T' as input, which can take into account information from the images. For image content, we utilize the Faster R-CNN [86] model to detect and extract important objects appearing in the images, such as characters, symbols, etc. Subsequently, we employ the BLIP model in a visual question answering (VQA) format to extract specific information from the images, e.g., inputting the object identified as a character and the question "Who is he/she?". Once this information is acquired, we utilize a pre-trained visual entity linkage model [40] to associate image entities with their corresponding Wikipedia entries. We ultimately link text and image entities to their respective Wikipedia entries, resulting in M linked entities, i.e., $e = \{e_i\}_{i=1}^{M}$.

Based on the retrieved entities e, we obtain corresponding entry descriptions from the page of English Wikipedia. Specifically, we choose brief introductions from Wikipedia as its descriptions. We encode the acquired explicit knowledge in the format: "Entity1 is Description1; Entity2 is Description2; ...". Then, we can use the BERT model to extract features. Given an explicit knowledge word sequence of input length n_{exp}, denoted as $K_i^{exp} = \{w_i^1, ..., w_i^{n_{exp}}\}$, its features $F_i^{exp} = \{f_i^1, ..., f_i^{n_{exp}}\}$ can be obtained as follows:

$$F_i^{exp} = \text{BERT}(K_i^{exp}) \in \mathbb{R}^{n_{exp} \times 768}, \tag{3.5}$$

where f_i^j represents the output feature of the j-th word.

Implicit Knowledge Extraction

Implicit knowledge refers to implicit commonsense knowledge, such as the knowledge that lemons are sour. With the advancement of large-scale language models, we can leverage these models to acquire such implicit knowledge. In this part, we utilize ChatGPT [11] to obtain event-related implicit knowledge. Specifically, it can be divided into the following five steps:

1) Obtaining an API key from OpenAI. The first step is to register an API key from OpenAI[2]. After obtaining the API key, we can call the API to ChatGPT service to get responses.

2) Formulating an appropriate prompt. To extract implicit knowledge using ChatGPT, we need to formulate a prompt that consists of a context and a question. Specifically, we take the enhanced text as input because it contains information from the image. We then design a prompt X_{gpt} to extract implicit knowledge using ChatGPT, structured as: "Context: T' (enriched text). Question: what's the social event occur? Answer:".

3) Extracting generated answers. When making the API call, we pass the designed prompt as input to the ChatGPT model. The model will generate a response, which in our case is a tentative answer A_i to the question.

4) Extracting generated explanations. In order to obtain a corresponding explanation to derive reliable implicit knowledge, we append the acquired A_i to X_{gpt}, and add "This is because" as a new prompt input to ChatGPT. By making another API call using this updated prompt, we can get the corresponding explanation EX_i from the model. This explanation provides insights into the implicit knowledge associated with the answer.

5) Utilizing the BERT model for feature extraction. The final implicit

[2] `https://platform.openai.com/docs/api-reference`

knowledge is obtained by merging the tentative answer A_i with the corresponding explanation EX_i, which can be denoted as $K_i^{imp} =< A_i, EX_i >= \{w_i^1, ..., w_i^{n_{imp}}\}$ where n_{imp} refers the word sequence of input length. Similar to explicit knowledge, we also utilize the BERT model for feature extraction. The features $F_i^{imp} = \{f_i^1, ..., f_i^{n_{imp}}\}$ can be obtained as follows:

$$F_i^{imp} = \text{BERT}(K_i^{imp}) \in \mathbb{R}^{n_{imp} \times 768}. \tag{3.6}$$

After obtaining the feature vectors derived from both explicit and implicit knowledge, we concatenate them to form a comprehensive external knowledge feature F_i^k, which can be represented as

$$F_i^k = \text{Concat}(F_i^{exp}, F_i^{imp}) \in \mathbb{R}^{(n_{exp}+n_{imp}) \times 768}, \tag{3.7}$$

where $\text{Concat}(\cdot)$ denotes the concatenation operation.

3.3.3 Knowledge-aware Feature Fusion Module

To effectively facilitate the integration of external knowledge, text, and visual content, we employ co-attention Transformers for knowledge fusion. Specifically, the knowledge-aware feature fusion module consists of three distinct co-attention Transformer encoders [97], referred to as Transformer 1, Transformer 2, and Transformer 3, as shown in Figure 3.3. The first two encoders are used to incorporate the extracted external knowledge into the text and visual content, filtering out irrelevant knowledge. The last encoder further integrates the text and visual content after knowledge fusion, filtering out content irrelevant to the task.

Within Transformer 1, the enriched text features, denoted as $F_i^{T'} \in \mathbb{R}^{n_t \times 768}$, serve as the query ($Q$) inputs, while the features of the external knowledge, denoted as $F_i^k \in \mathbb{R}^{(n_{exp}+n_{imp}) \times 768}$, are utilized as both the key (K) and value (V) inputs in the

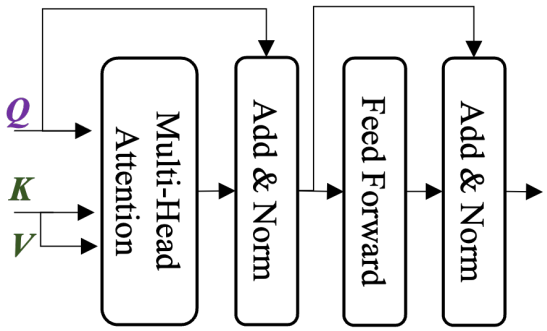


Figure 3.4: Co-attention Transformer.

attention mechanism. As shown in Figure 3.4, the co-attention Transformer employs a two-part architecture, i.e., a multi-head self-attention layer and a fully connected feed-forward network with residual connections and layer normalization.

Specifically, the multi-head self-attention layer utilizes a multi-head attention mechanism to process the inputs (Q, K and V) in parallel. It involves splitting the inputs into h heads, where each head operates on a reduced dimension $\mathbb{R}^{\frac{n_t \times 768}{h}}$. After processing by the attention mechanism, the outputs from all heads are concatenated and linearly transformed by parameter matrices W_A, resulting in an output dimension that matches the original dimension of $Q \in \mathbb{R}^{n_t \times 768}$. The output of the multi-head attention is then combined with Q with layer normalization to obtain the output $O_f \in \mathbb{R}^{n_t \times 768}$, which can be formulated as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V, \tag{3.8}$$

$$A_i = \text{Attention}(Q * W_{Qi}, K * W_{Ki}, V * W_{Vi}), \tag{3.9}$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(A_1, ..., A_h)W_A, \tag{3.10}$$

$$O_f = \text{LayerNorm}(Q + \text{MultiHead}(Q, K, V)), \tag{3.11}$$

where $\text{Attention}(Q, K, V)$ and $\text{MultiHead}(Q, K, V)$ represent the self-attention function and multi-head self-attention function. d_k is the dimension of K in Eq. 3.8, $\text{softmax}(\cdot)$ is the softmax function, $*$ denotes matrix multiplication, and $\text{LayerNorm}(\cdot)$

is the layer normalization function. Finally, we obtain the output O^{tk} through a fully connected feed-forward network with residual connections and layer normalization, which typically maintains the dimensionality of its input:

$$O^{tk} = \text{LayerNorm}(O_f + \text{FNN}(O_f)), \tag{3.12}$$

where $\text{FNN}(\cdot)$ is a feed-forward neural network, typically composed of two fully connected layers and an activation function (such as ReLU). Moreover, for the i-th sample, the variable $O_i^{tk} \in \mathbb{R}^{n_t \times 768}$ symbolizes the output derived from Transformer 1.

For Transformer 2, we take the original images as input, as the low-level features of images can sometimes help distinguish different social event scenes compared to the high-level features such as captions. To extract image features, we utilize the output from a pre-trained ResNet network [36], specifically from its penultimate pooling layer, to capture region-specific feature $R_i = \{R_i^1, \ldots, R_i^{n_r}\}$ of the image, where $n_r = 49$ represents the number of regions in the image. Since the dimension of R_i^j is 2048, we use a feed-forward neural network to map it to 768 dimensions (the same as the text) to obtain the final image feature as follows:

$$R_i = \text{FNN}(\text{ResNet}(I_i)) \in \mathbb{R}^{n_r \times 768}. \tag{3.13}$$

Similar to Transformer 1, we use the image feature $R_i \in \mathbb{R}^{n_r \times 768}$ as Q, and the feature $F_i^k \in \mathbb{R}^{(n_{exp}+n_{imp}) \times 768}$ of the extracted external knowledge as K and V, to obtain the output $O_i^{ik} \in \mathbb{R}^{n_r \times 768}$. Finally, Given the output $O_i^{tk} \in \mathbb{R}^{n_t \times 768}$ of Transformer 1 as Q, and the output $O_i^{ik} \in \mathbb{R}^{n_r \times 768}$ of Transformer 2 as K and V, we can get the output $O_i^{ti} \in \mathbb{R}^{n_t \times 768}$ of Transformer 3.

After obtaining the features fused from multiple modalities, each modality's features undergo a global average pooling process to distill the information into a unified form, which will then be concatenated to compose the final feature vector as follows:

Algorithm 1 MFEK Algorithm

Input: Text input T, image input I, and their social event label y.

Output: Learned model parameters θ.

while $t \leq$ **MaxIter**

1: Extract image caption $C = BLIP(I)$ and OCR text $O = OCR(I)$.
2: Enrich text: $T' =< T, C, O >$.
3: Extract text features $F^{T'} = BERT(T')$ according to Eq. 3.4.
4: Extract explicit knowledge K^{exp} from Wikipedia.
5: Extract implicit knowledge K^{imp} from LLM.
6: Fuse knowledge $F^k = Concat(F^{exp}, F^{imp})$ according to Eq. 3.7.
7: Extract image region features R using ResNet according to Eq. 3.13.
8: Apply co-attention transformers to obtain O_{tk}, O_{ik}, O_{ti} according to Eq. 3.11 and Eq. 3.12.
9: Generate final prediction through classifier according to Eq. 3.14 and Eq. 3.15.
10: Optimize using cross-entropy loss according to Eq. 3.16.

end while

$$O_i = \text{Concat}(\text{GAP}(O_i^{tk}), \text{GAP}(O_i^{ik}), \text{GAP}(O_i^{ti})) \in \mathbb{R}^{2304}, \tag{3.14}$$

where $\text{GAP}(\cdot)$ denotes the global average pooling operation.

Ultimately, a feed-forward neural network with a Softmax activation function projects the aggregated features into a discrete label space corresponding to different social events:

$$\hat{y}_i = \text{FNN}(O_i) \in \mathbb{R}^{n_{event}}. \tag{3.15}$$

The model is trained using a cross-entropy loss function to optimize the classification of social events, where $|D_{tr}|$ denotes the total number of samples in the training dataset. The loss function is defined as follows:

$$\mathcal{L} = -\sum_{i=1}^{|D_{tr}|} y_i \log(\hat{y}_i), \tag{3.16}$$

where y_i is the true social event category.

The detailed algorithm for the MFEK method is presented in Algorithm 1. The computational complexity of MFEK can be analyzed in several main components: 1) feature extraction from BERT and ResNet with complexity $O(n_t)$ and $O(n_r)$ respectively, where n_t is the text length and n_r is the number of image regions; 2) knowledge extraction and fusion with complexity $O(M + L)$, where M is the number of Wikipedia entities and L is the LLM processing complexity; 3) co-attention transformer operations with complexity $O(n_t^2 + n_r^2)$ due to the self-attention mechanisms. Therefore, the overall computational complexity is $O(n_t^2 + n_r^2 + M + L)$. The space complexity is $O(n_t + n_r)$ for storing the feature representations.

3.4 Social Event Detection (SED) Dataset

In this section, we present the collection and the statistics of the SED dataset.

3.4.1 Data Collection

Collection of Social Events

In this chapter, we utilize Wikipedia [100] for collecting social events, since Wikipedia operates as a crowd-sourced platform with verified social events. Specifically, we start with Wikipedia's event category page[3], which provides access to a collection of various categories of public events that have occurred or are occurring in the world. They are

[3] `https://en.wikipedia.org/wiki/Category:Lists_of_events`

organized into different subcategories, such as protests[4], disasters[5], sports events[6], and so on. Then, we can obtain a number of event-specific Wikipedia entries through the corresponding subcategories, e.g., the event "2011 Thailand floods"[7] from the floods subcategory page[8].

In our dataset, we select social event entries based on the following principles: 1) diversity of social event themes - covering as many different events as possible to mirror real-life scenarios; 2) abundant and similar sub-events under the same social event theme - increasing the challenge of social event detection and helping enhance the model's capacity to extract event elements; 3) popular social events - being able to access more multimodal data on social media platforms. Ultimately, based on these principles, we manually collected 40 social events from 2011 to 2022, each corresponding to a Wikipedia entry, as depicted in Figure 3.5.

Collection of SED Dataset

After obtaining social event entries, most existing datasets will directly perform keyword search matching based on these entries, acquiring posts corresponding to the social events as the dataset. Datasets obtained from text keyword searches not only simplify the task of social event detection but also weaken the role of non-text modalities. However, it is challenging to identify event-related posts on social media without keywords. Therefore, this chapter uses hashtags labeled by users to collect data on social media.

Specifically, in order to obtain relevant and representative hashtags for each social event, we first directly use the social event name, location, and time of occurrence for the collection of posts. Next, we count hashtag frequencies in the collected posts

[4] https://en.wikipedia.org/wiki/Category:Lists_of_protests

[5] https://en.wikipedia.org/wiki/Category:Lists_of_disasters

[6] https://en.wikipedia.org/wiki/Category:Lists_of_sports_events

[7] https://en.wikipedia.org/wiki/2011_Thailand_floods

[8] https://en.wikipedia.org/wiki/List_of_floods

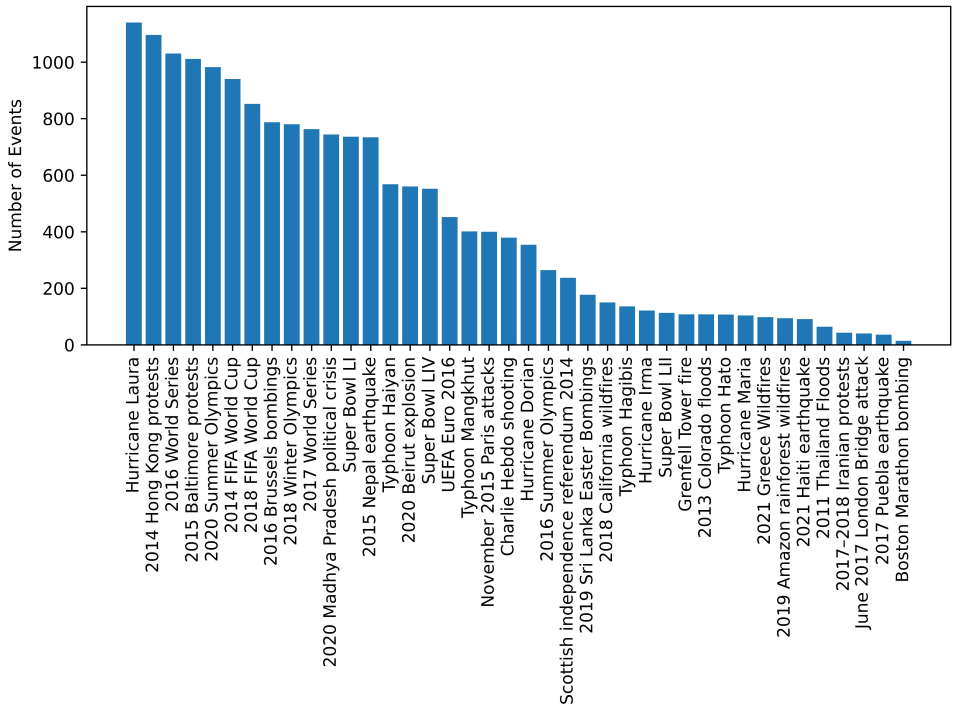


Figure 3.5: The distribution of SED dataset.

for each event. Finally, we manually screen out the high-frequency hashtags for each event based on event relevance. For example, the final high-frequency hashtags for the event "2020 Summer Olympics" are "#olympics", "#summerolympics", "#tokyoolympics", and "#olympicgames".

We utilize these selected high-frequency hashtags, along with the location and time of the event, to collect posts once again. Then, we remove these high-frequency hashtags used in the posts considering the following reasons:

- **Avoiding Bias**: Most of hashtags are user-generated and thus can be highly subjective, often reflecting the user's personal beliefs or sentiments rather than the objective facts of the event. By removing them, we can minimize the influence of these subjective elements on our dataset, thereby reducing bias.

- **Generalization**: In practical applications, not all event-related posts will nec-

essarily include specific hashtags. If the model becomes overly reliant on these hashtags for social event detection, it may fail to perform adequately when encountering posts without these hashtags. By removing them, we enable the model to generalize better and effectively detect social events even in the absence of hashtags.

- **Noise Reduction**: These hashtags often include informal language, abbreviations, or internet slang, which may introduce noise into our dataset. By removing them, we can focus on the main text of the posts, which is likely to be more informative for our news detection task.

Finally, we filtered out non-English, repetitive, and single-modal samples and manually checked according to whether the semantics had changed and whether the post matched the corresponding event, which ultimately obtained 17,366 samples as our SED dataset.

Instead of using event-specific keywords, our approach leverages user-generated hashtags to collect data. This strategy significantly increases the relevance of the collected data to the actual social events. Hashtags, as opposed to keywords, are a product of user engagement and provide a focused snapshot of how events are discussed in real time on social media. They are less likely to suffer from semantic dilution—a common issue with keyword searches where the intent and context can be lost. Consequently, the use of hashtags preserves the integrity of the original posts and captures the nuanced discourse surrounding social events. By adopting this hashtag-centric collection method, our dataset can more accurately mirror the dynamic and organic nature of how social events unfold and are talked about in real-world scenarios.

Table 3.1: The statistics of SED dataset. ("#" represents the number of samples.)

	#Event	#Text	#Image	Average Words
SED	40	17,366	21,117	16.12

(a) 2014 Hong Kong protests (b) 2017 World Series (c) 2018 California wildfires

Figure 3.6: Examples of data samples in SED dataset.

3.4.2 Statistics of SED Dataset

The statistics for SED dataset are shown in Table 3.1. Specifically, SED includes 17,366 samples from 2011 to 2022, annotated with 40 real-world social events. These events cover a broad range of topics, including political events, sports events, natural disasters, social and cultural events, and violent and terror events. To enhance the dataset's complexity and utility, each thematic category includes numerous closely related events, thereby increasing the detection task's difficulty, e.g., "Super Bowl LI" and "Super Bowl LII" in the American football events, "Hurricane Maria" and "Hurricane Dorian" in hurricane events. In addition, some visual examples are shown in Figure 3.6. The distributions of data and events are depicted in Figure 3.5.

Table 3.2: Comparison of existing datasets. ("#" represents the number of samples. "N.A." for Twevent indicates Not Available as the total number of events was not reported in the original dataset.)

Dataset	Platform	#Sample	#Event	Modality	Fine-grained	Keyword-based	Public
CE	Twitter	800	2	Single	no	yes	no
SED-14	Flickr, Youtube	427,370/1,327	21,169	Multiple	no	yes	yes
SW	Sina Weibo	4,341	2	Single	no	yes	no
ASO	Twitter	1,100	3	Single	no	yes	no
OSMNs	Twitter	3.5M	20	Single	no	yes	no
Twevent	Wikipedia,Twitter	3.2M/4.3M	N.A.	Single	no	yes	no
DHS	Twitter, Tumblr	2.1M/0.3M	600	Multiple	no	yes	no
PHEME	Twitter	2,089	9	Multiple	yes	yes	yes
CrisisMMD	Twitter	18,126	7	Multiple	yes	yes	yes
SED	Twitter	17,366	40	Multiple	yes	no	yes

3.4.3 Comparisons with Existing Datasets

Table 3.2 compares some attributes of our SED dataset and existing datasets, including CE [80], SED-14 [87], SW [58], ASO [95], OSMNs [30], Twevent [53], DHS [46], PHEME [122] and CrisisMMD [5]. From the table, we have some observations:

- Currently, most datasets are collected through event-related keywords, which simplifies the task of social event detection.

- The categories of social events in existing datasets are relatively coarse-grained (e.g., the SW dataset only distinguishes between earthquake and non-earthquake events, and the SED dataset categorizes events broadly into conferences, sports, festivals, etc.). For fine-grained event datasets (i.e., PHEME and CrisisMMD), their original intent was not for the task of social event detection; thus they only consider a limited type of social events, i.e., political and earthquake-related events.

- Compared to other datasets, the SED dataset is collected based on hashtags,

which avoids the disadvantages of keyword searches and is more aligned with real-world scenarios. Additionally, the data we have gathered is fine-grained and multimodal, which benefits researchers in their further studies. Furthermore, we have open-sourced our dataset, with the hope of fostering the development of the field of social event detection.

3.5 Experiment

In this section, we conduct comprehensive experiments to evaluate MFEK and validate its effectiveness for multimodal social event detection. We first describe the SED and CrisisMMD datasets and experimental settings. Then, we evaluate MFEK through multiple aspects: 1) comparison with state-of-the-art methods to demonstrate overall performance; 2) ablation studies to validate the effectiveness of text enrichment, knowledge extraction, and knowledge-aware feature fusion modules; 3) parameter analysis to demonstrate model robustness, particularly focusing on the impact of attention heads and knowledge extraction methods; and 4) qualitative analysis through case studies examining both successful and failed predictions.

3.5.1 Datasets and Data Partitioning

We evaluate MFEK with baselines on our proposed SED dataset and the publicly available CrisisMMD dataset [5].

SED dataset. Our proposed SED dataset is selected for evaluation. We employed a stratified sampling method to ensure that each category of events is represented proportionally in the training, validation and test sets. The data is partitioned by category randomly, allocating 55% for the training set, 10% for the validation set, and 35% for the test set.

CrisisMMD dataset. It is a multimodal crisis dataset collected using event-related keywords. This dataset is composed of seven natural disaster events from 2017, i.e., Hurricane Irma, Hurricane Maria, Hurricane Harvey, Mexico earthquake, Iraq-Iran earthquakes, Sri Lanka floods, and California wildfires. Following [83], we divide 70% of the dataset as the training set, 10% as the validation set, and 20% as the test set.

3.5.2 Implementation Details

For text data, we initialize our BERT model using the "bert-base-uncased" configuration and set a maximum sequence length of 200 word tokens. For image feature extraction, we employ a pre-trained ResNet50 model, defaulting to the first image in a multi-image post for the experiment. To derive implicit knowledge, we utilize the "GPT-3.5 Turbo" variant of ChatGPT. For parameter selection, we set the learning rate to 1×10^{-5} for stable fine-tuning of the pre-trained model, and Lion [16] is chosen as the optimizer. The model was trained for 200 epochs to ensure convergence. The batch size is 30, and the head of the co-attention Transformer h is set to 4. In addition, we randomly select different 5 random seeds in the experiment, and calculate their average as the final result.

3.5.3 Evaluation Metrics

In our experiments, we employ accuracy, macro-averaged precision, recall and F1 score to provide a more comprehensive performance evaluation, as the data from different social events is unbalanced.

3.5.4 Benchmarks

To validate the effectiveness of our MFEK model, we compare it against both single-modal and multimodal benchmark methods, including multimodal fusion methods, state-of-the-art event detection methods, and large-scale language model (LLM) methods. The specifics are as follows:

Single-modal methods:

- **ResNet50** [36] for image modality
- **BERT** [47] for text modality

Multimodal methods:

- **MMBT** [48], which utilizes Transformer modules for fusing textual and visual features to enhance classification tasks.
- **COOLANT** [102], which leverages a cross-modal contrastive learning framework to achieve more accurate image-text alignment.
- **SCBD** [1], which integrates the textual and visual features using a self-attention mechanism to detect social events.
- **AT-CVAE** [56], which exploits an adaptive Transformer-based conditioned variational autoencoder Network for incomplete social event classification.
- **OWSEC** [83], which designs a multimodal mask transformer network to capture cross-modal semantic relations and fuse fine-grained multimodal features of social events.
- **ChatGPT** [11], which is a large-scale Transformer-based language model that exhibits high performance across a variety of text-related domains. In our approach, we leverage it to extract implicit knowledge. In this part, we design

a prompt for a fair comparison, i.e., "<instructions> <in-context examples> Post:<caption C_i> <OCR O_i> <content T_i>. Q:<What are the related social events to this post:> A:". "<instructions>" introduces the task and provides all the social events, e.g., "You need to identify the social media platform Twitter post from which of the 40 social events given below? 40 social events: 2014 FIFA World Cup, 2020 Summer Olympics, ...". "<in-context examples>" randomly selects 5 examples from the training set,e.g., "Here are five examples. Please learn how to select the true label in these examples, and pay particular attention to the consistent use of the answer in these below examples.\n Post: C_1+O_1+T_1. Q: What are the related social events to this post: \n A: A_1 \n Post: C_2+O_2+T_2. Q: What are the related social events to this post: \n A: A_2 ...".

Table 3.3: Experiment results on the SED dataset.

Model	Accuracy	Precision	Recall	F1
ResNet50 [36]	0.507	0.421	0.365	0.381
BERT [47]	0.816	0.756	0.755	0.753
MMBT [48]	0.574	0.503	0.427	0.440
COOLANT [102]	0.715	0.660	0.559	0.591
SCBD [1]	0.786	0.669	0.603	0.609
AT-CVAE [56]	0.814	0.748	0.750	0.747
OWSEC [83]	0.818	0.759	0.760	0.754
ChatGPT [11]	0.640	0.570	0.621	0.547
MFEK	**0.855**	**0.824**	**0.796**	**0.809**

3.5.5 Results and Analysis

Tables 3.3 and 3.4 present a comparative analysis of our MFEK model against the aforementioned single-modal and multimodal methods, utilizing the SED and Crisis-

Table 3.4: Experiment results on the CrisisMMD dataset.

Model	Accuracy	Precision	Recall	F1
ResNet50 [36]	0.482	0.519	0.455	0.477
BERT [47]	0.961	0.960	0.968	0.964
MMBT [48]	0.658	0.728	0.648	0.675
COOLANT [102]	0.924	0.936	0.929	0.933
SCBD [1]	0.956	0.961	0.962	0.956
AT-CVAE [56]	0.961	0.961	0.968	0.964
OWSEC [83]	0.963	0.969	0.972	0.971
ChatGPT [11]	0.490	0.448	0.453	0.445
MFEK	**0.972**	**0.976**	**0.978**	**0.977**

MMD datasets for evaluation. From the table, we have the following observations:

- Our MFEK model achieves superior performance over other methods on the SED dataset, achieving the best performance . The model's enhanced performance is attributed to the incorporation of external knowledge, which effectively mitigates the OOD issues inherent in social event detection. In addition, we designed a knowledge-aware feature fusion module to fuse and filter knowledge with the input text and images, further improving the classification results.

- Our proposed method outperforms other state-of-the-art methods on the CrisisMMD dataset, even though this dataset was collected based on event-related keywords, which proves the robustness and generalization ability of our method.

- For single-modal methods, the text-based model (i.e., BERT) performs better than the image-based model (i.e., ResNet50) on both SED and CrisisMMD datasets, suggesting that text provides more effective information for social event detection compared to images alone. Specifically, we found that this gap is even larger on the CrisisMMD dataset, since this dataset is based on event-

related keyword searches, which can be achieved with good performance by text alone.

- Certain multimodal fusion approaches (e.g., MMBT and COOLANT) exhibit lower performance compared to the single-modal BERT method (i.e., BERT). One reason may be the introduction of the image modality, as the noise in the image modality in the dataset is relatively large (e.g., irrelevant images for the task). Our model utilizes the knowledge-aware feature fusion module to filter irrelevant information from images and text, which can make good use of useful image information to enrich features.
- The prompt-based LLM (i.e., ChatGPT) does not yield as high performance as other methods on both SED and CrisisMMD datasets, which means that LLMs are not yet a comprehensive substitute for the domain-specific task, i.e., social event detection. Rather than employing ChatGPT directly for classification, our method leverages it to distill valuable implicit knowledge, thereby enhancing the model's performance.

3.5.6 Model Ablation

To evaluate the contribution of each component within the MFEK model, we perform a series of ablation studies, which involve:

- w/o Text: Remove the input text.
- w/o Image: Remove the input image.
- w/o Caption: Remove the image caption generation component.
- w/o OCR: Remove the OCR-generated text from images.
- w/o Implicit Knowledge: Remove the integration of implicit knowledge.

Table 3.5: Classification performance on the test set for different variants of the MFEK model.

#	Text	Image	Caption	OCR	Implicit Knowledge	Explicit Knowledge	Co-attention	Accuracy	Precision	Recall	F1
1	✓	×	×	×	✓	✓	✓	0.833	0.787	0.770	0.776
2	×	✓	✓	✓	✓	✓	✓	0.790	0.770	0.746	0.755
3	✓	✓	×	✓	✓	✓	✓	0.842	0.809	0.778	0.791
4	✓	✓	✓	×	✓	✓	✓	0.847	0.821	0.785	0.800
5	✓	✓	✓	✓	×	✓	✓	0.846	0.805	0.784	0.793
6	✓	✓	✓	✓	✓	×	✓	0.849	0.813	0.788	0.798
7	✓	✓	✓	✓	×	×	✓	0.842	0.785	0.779	0.785
8	✓	✓	✓	✓	✓	✓	×	0.844	0.805	0.790	0.800
9	✓	✓	✓	✓	✓	✓	✓	**0.855**	**0.824**	**0.796**	**0.809**

- w/o Explicit Knowledge: Remove the integration of explicit knowledge.
- w/o Knowledge: Remove the integration of all external knowledge.
- w/o Co-attention: Remove the co-attention Transformer module and use concatenation instead.

Table 3.5 details the performance for each variant, illustrating the impact of removing specific components. From the table, we make the following key observations:

- The performance of our model decreases significantly in the absence of text or image inputs. This underscores that multimodal data has a complementary function in social event detection, which plays a crucial role in supplementing the social event elements.
- For the text enrichment module, the semantic information represented by captions or OCR text is beneficial for social event detection. This demonstrates the advantage of using visual semantic representations to enrich text.
- The integration of external knowledge, including both implicit and explicit knowledge, contributes to the performance. This highlights the benefit of leveraging external knowledge for social event detection, which can help the model mitigate the OOD problem.
- After replacing the co-attention Transformer module with a concatenation method, the performance of the model decreases. This proves that our designed knowledge-aware feature fusion module can effectively integrate the obtained knowledge with the input, filtering out some irrelevant information to the task.

3.5.7 Parameter Analysis

In our experiment, the knowledge-aware feature fusion module employs a multi-head attention mechanism, where the number of heads h can be a power of 2. As shown

in Figure 3.7, we evaluated this parameter using the validation set and found that the model achieves optimal results when the number of attention heads h is set to 4, making it our choice for the final configuration.

For extracting implicit and explicit knowledge, we utilized the "GPT-3.5 Turbo" model and brief introductions from Wikipedia entries, respectively. While any LLMs could potentially be used for implicit knowledge extraction, we specifically conducted evaluations using the "GPT-4" model. For explicit knowledge extraction, we select to evaluate using entire Wikipedia page documents as it can offer more information about the entities. As shown in Figure 3.8, we can observe:

- Using the "GPT-4" model for implicit knowledge extraction does not improve performance. This may be due to the "GPT-3.5 Turbo" model producing more diverse outputs, thereby subtly enhancing more event-related information.
- Employing entire Wikipedia page documents for explicit knowledge extraction resulted in slightly diminished performance compared to utilizing concise entity introductions. This is because, although the page documents provide richer information, they also introduce more noise.

3.5.8 Case Study

Figure 3.9 presents several success and failure examples using the OWSEC and MFEK methods on the SED dataset. We also display the visual content extracted by our model (i.e., caption and OCR text) and the external knowledge (i.e., implicit and explicit knowledge). From the figure, we have the following observations:

- Figures 3.9a and b illustrate successful predictions by both OWSEC and MFEK, attributed to the presence of event-relevant keywords within the text, such as "#HKDemocracy", "#DemiXSuperBowl", and "#LIV". In addition, we find

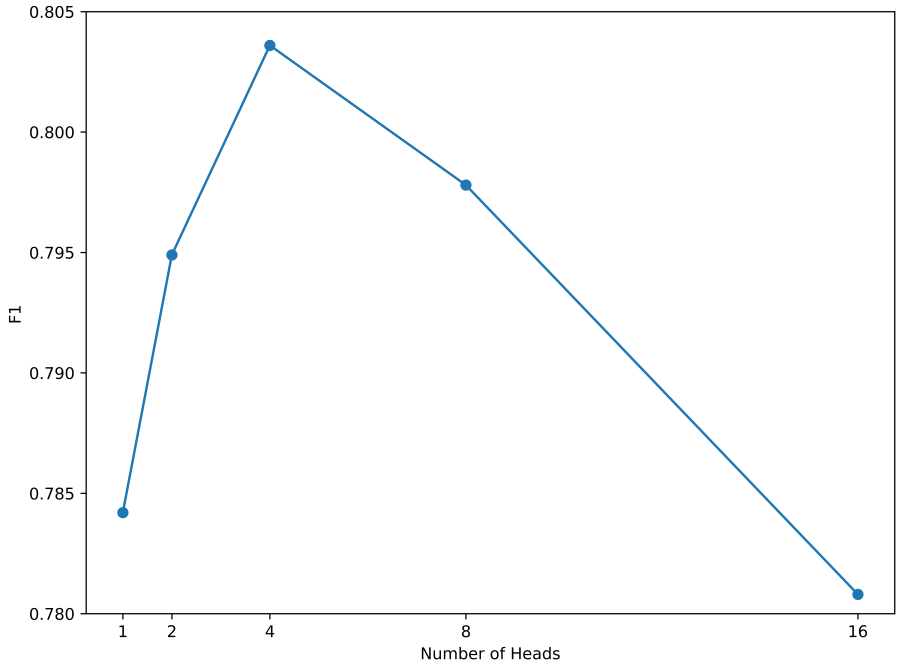


Figure 3.7: F1 score on validation dataset for MFEK model under different number of attention heads (higher is better).

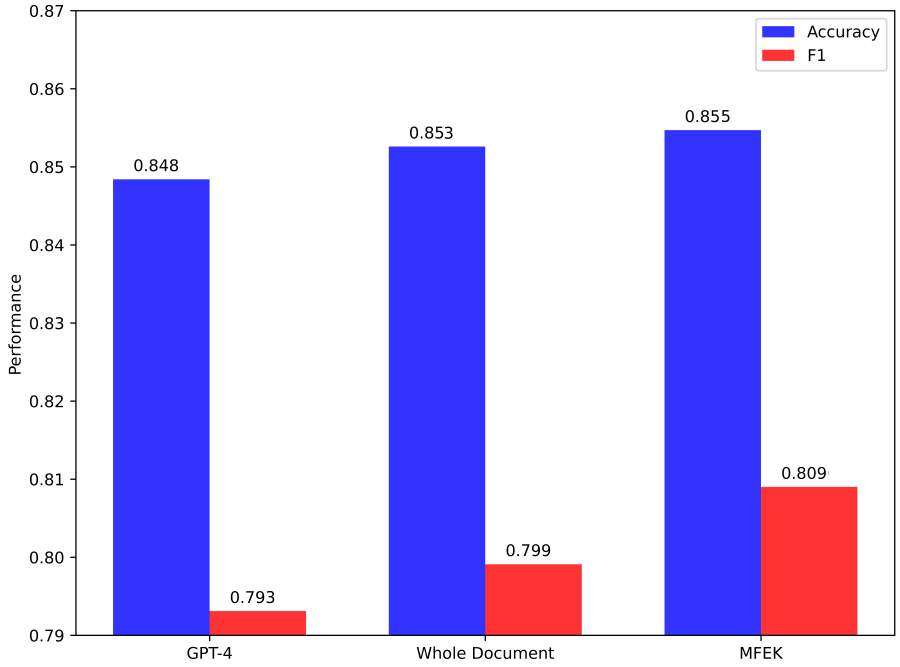


Figure 3.8: The impact of different external knowledge extraction methods.

T: Alex Chow with a yellow umbrella #HKDemocracy

OWSEC: 2014 Hong Kong protests
MFEK: 2014 Hong Kong protests
GT: 2014 Hong Kong protests

Visual Content:
C: a man holding a yellow umbrella in front of a crowd
O: NOW

Implicit Knowledge:
The news event that occurred according to the above information is the Hong Kong pro-democracy protests. This is indicated by the hashtag #HKDemocracy and the mention of Alex Chow, who was a prominent pro-democracy activist in Hong Kong. The yellow umbrella is also a symbol of the pro-democracy movement in Hong Kong.

Explicit Knowledge:
Alex Chow is Hong Kong student activist; yellow umbrella is Umbrella Revolution, a series of sit-in street protests.

(a)

T: Lovin @justdemi performance 🔥❤️🥰#2020 #DemiXSuperBowl #Niners #ChiefsKingdom #LIV

OWSEC: Super Bowl LIV
MFEK: Super Bowl LIV
GT: Super Bowl LIV

Visual Content:
C: a woman standing in front of a microphone
O: FOX

Implicit Knowledge:
The news event is the performance by Demi Lovato at the 2020 Super Bowl. This is evident from the hashtags #DemiXSuperBowl and #LIV, which refer to the Super Bowl LIV (54) that took place in 2020.

Explicit Knowledge:
Lovin is Romanian middle-distance runner; Demi Lovato is American singer (born 1992).

(b)

T: It could have been so much worse. This is from our partners at the @TB_Times #WTSP

OWSEC: 2016 Brussels bombings
MFEK: Hurricane Irma
GT: Hurricane Irma

Visual Content:
C: a newspaper with a picture of a truck on it
O: @ampa HURRICANE EDITION Bay ccom @imcs WERE Tampa Bay spared worst as Irma brings havoc; takes out power

Implicit Knowledge:
The news event is a hurricane, specifically Hurricane Irma. This is indicated by the mention of "Hurricane Edition" and "Irma" in the OCR text. The caption also mentions a truck, which could be related to the hurricane\'s impact or damage.

Explicit Knowledge:
WTSP is CBS affiliate in St. Petersburg, Florida; Irma is the 9th named storm of the 2017 Atlantic hurricane season.

(c)

T: RT @MeredithFrost Super Typhoon on a collision course with the Philippines.

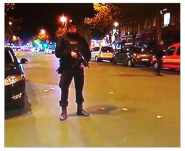
OWSEC: Hurricane Laura
MFEK: Typhoon Haiyan
GT: Typhoon Haiyan

Visual Content:
C: a view of the earth from space at night

Implicit Knowledge:
The news event is about Super Typhoon on a collision course with the Philippines. This is mentioned because it is a significant weather event that could potentially cause damage and impact the Philippines.

Explicit Knowledge:
Super Typhoon is Type of tropical cyclone that develops in the Northern Hemisphere; Typhoon is Type of tropical cyclone that develops in the Northern Hemisphere; Philippines is Archipelagic country in Southeast Asia.

(d)



T: Truppe speciali in strada #Paris #Parisattack

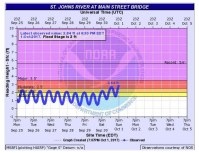
OWSEC: Charlie Hebdo shooting
MFEK: November 2015 Paris attacks
GT: November 2015 Paris attacks

Visual Content:
C: a man standing on the side of a street next to parked cars

Implicit Knowledge:
The news event is a special forces presence on the streets of Paris following an attack. This is indicated by the hashtags #Paris and #Parisattack.

Explicit Knowledge:
Truppe is Austrian alpine skier.

(e)



T: Rivers in Northeast Florida are rising to the lower end of levels last seen during water.weather.gov/ahps2/hydr…

OWSEC: Hurricane Irma
MFEK: Hurricane Irma
GT: Hurricane Maria

Visual Content:
C: the st johns river at main street bridge
O: ST. JOHNS RIVER AT MAIN STREET BRIDGE Universal Time (UTC) 232 232 232 232 232 232 232 232 232 Sep 25 Sep 26 Sep 27 Sep 28 Sep 29 Sep 30 Oct Oct Oct.

Implicit Knowledge:
The news event is that the rivers in Northeast Florida are rising to the lower end of levels last seen during a previous event. The reason for this is not mentioned in the given information.

Explicit Knowledge:
Florida is U.S. state; weather.gov is U.S. forecasting agency of the National Oceanic and Atmospheric Administration.

(f)

Figure 3.9: Success and failure examples predicted by OWSEC and MFEK. (**T**: text, **GT**: ground truth, **C**: caption, and **O**: OCR)

that the implicit and explicit knowledge we extract can effectively supplement the original semantics of the text, which contributes to the model's better prediction.

- In Figure 3.9c, we note that the text does not contain keywords directly related to the social event. Instead, the keywords are appeared in the contents of a newspaper shown in the image. In this case, the OCR text becomes essential in enriching the original text. Additionally, our proposed model, which incorporates external knowledge, successfully links "WTSP" to "Florida" where the social event took place, and associates "Irma" with hurricanes, helps the model to predict the correct event "Hurricane Irma". Compared to other state-of-the-art models (e.g., OWSEC), it is difficult to predict the correct event when it comes to this OOD problem. This demonstrates the ability of our proposed model to infer connections that are not explicitly present in the training set, thus mitigating the OOD problem to some extent.

- In Figure 3.9d, we find that the image and implicit knowledge do not provide much useful information, while the explicit knowledge offers a greater degree of supplementation. However, in Figure 3.9e, the explicit knowledge provides incorrect and irrelevant information. This underscores the importance of our proposed knowledge-aware feature fusion module in discerning and disregarding irrelevant information to ensure accurate predictions.

- In Figure 3.9f, both models fail in their predictions. The reason could be that the model can only identify "Florida", a region affected by "Hurricane Maria", as the information based on the text and image. However, the news similar event "Hurricane Irma" also impacted this area, which highlights the challenging nature of our proposed SED dataset.

3.6 Conclusion

In this chapter, we propose a multimodal fusion with external knowledge (MFEK) model for social event detection. Specifically, the text enrichment module effectively incorporates image-derived semantic information into the text, enriching the data context. The knowledge extraction module utilizes Wikipedia to extract explicit knowledge and uses large language models (LLMs) to extract implicit knowledge. Furthermore, the knowledge-aware feature fusion module fuses the acquired external knowledge with multimodal inputs and filters out task-irrelevant information. Moreover, we propose a well-labeled social event detection (SED) dataset, which includes multimodal data derived from the social media platform, i.e., Twitter. Compared to existing datasets, we utilize hashtags for data collection and annotation rather than solely relying on event-related keywords, which makes the collected data more consistent with real-world social event detection scenarios. Extensive experiments on the SED and CrisisMMD datasets demonstrate that the MFEK model exceeds the performance of current state-of-the-art methods in social event detection. With a variety of available benchmarks, the SED dataset is expected to facilitate research in social event detection.

Chapter 4

Robust Cross-platform Social Event Detection via Self-supervised Modality Complementation

In this chapter, we propose a cross-platform social event detection model, SSMC, which aims at enhancing the model's capability for cross-platform detection. Specifically, we introduce a Missing Data Complementation (MDC) module to address the issue of missing modalities in cross-platform scenarios. Moreover, a Multimodal Self-Learning (MSL) approach is proposed to mitigate the domain gap between different platforms through self-learning. Finally, we extend the SED dataset to a multi-platform social event dataset and conduct extensive experimental analysis.

4.1 Introduction

The majority of existing works for social event detection, including single-modal and multimodal methods [2, 4, 57, 83, 98, 111, 114], focus on single-platform data, which limits the scope and diversity of the detected events. Real-world social events, how-

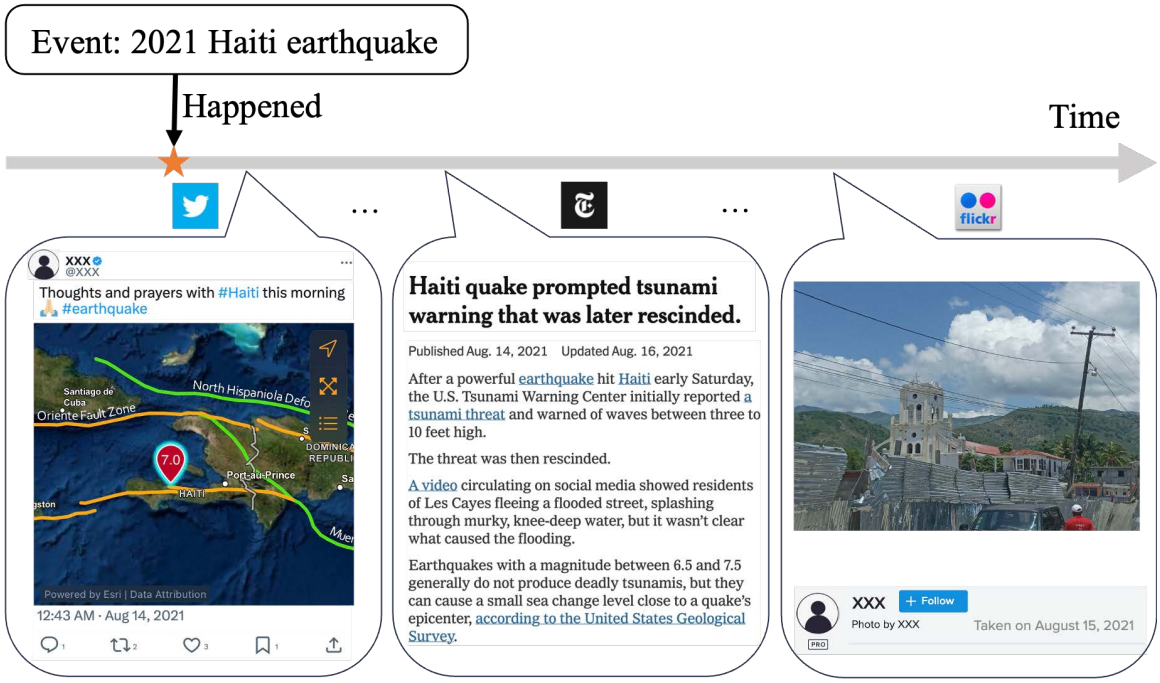


Figure 4.1: As a social event develops, different platforms provide information from different perspectives for it.

ever, manifest across multiple platforms, each offering unique perspectives and modalities of information. Consider the example illustrated in Figure 4.1, where an event "2021 Haiti earthquake" emerges. The initial wave of information often comes from social media platforms like Twitter, offering immediate firsthand accounts and public reactions. This is soon followed by the more structured and analytical coverage provided by online news media platforms, such as The New York Times, which brought in-depth reports and analyses. In parallel, other platforms like Flickr capture and share visual narratives through photographs. This progression underscores the necessity of cross-platform detection to gain a comprehensive and nuanced understanding of social events. However, current research methods tend to falter when confronted with cross-platform scenarios, demonstrating diminished effectiveness and adaptability in these more complex environments.

In this chapter, we delve into the novel task of cross-platform multimodal social event detection, aiming to enhance the precision of social event detection across different platforms. This task seeks to leverage data from source platforms with known event labels alongside unlabeled data from target platforms, training models to perform more effectively on unlabeled target platform data. However, it introduces several challenges:

- **Incomplete Modalities**: One of the challenges in cross-platform multimodal social event detection is the inherent issue of incomplete modalities. This refers to the frequent scenario where certain types of data (e.g., images, text, or videos) that may be available on one platform are absent on another. This disparity can arise due to the differing nature of platforms. As illustrated in Figure 4.1, the sample from image-sharing-oriented Flickr only contains images, while online news media platforms only provide text.

- **Platform Heterogeneity**: Another critical challenge is platform heterogeneity, which denotes the diverse characteristics and user behaviors inherent to different platforms. Each platform has its unique content presentation, user interaction mechanisms, and data formats. For instance, the way social events are reported and discussed on Twitter, with its character limit and emphasis on immediacy, differs significantly from the more detailed and narrative-driven content found on online news media. This diversity is the main reason why most existing methods underperform when applied across platforms.

- **The Scarcity of Annotated Datasets**: Lastly, the scarcity of annotated datasets poses a significant barrier to the advancement of cross-platform multimodal social event detection. Annotated datasets are important for training and evaluating machine learning models; however, the creation of such datasets is labor-intensive and requires significant domain expertise, especially when dealing with multimodal data across different platforms. The existing datasets

are predominantly single-platform [80, 5, 114], and the social event labels vary between datasets, which makes direct application challenging.

Therefore, we propose a Self-Supervised Modality Complementation (SSMC) approach designed to tackle these challenges. Specifically, our method is underpinned by two components: a Missing Data Complementation (MDC) module and a Multimodal Self-Learning (MSL) module. The MDC module employs a modality classifier to differentiate between modality-specific and modality-shared features across all modalities. When a modality is absent, it becomes possible to supplement it using the information contained within the common features of another modality. The MSL module addresses platform heterogeneity by leveraging self-learning (i.e., pseudo-labeling), which individually exploiting the relationships of semantic, image, text, and joint multimodal features in different spaces. In particular, it uses a nearest-neighbor clustering algorithm to achieve multi-views pseudo labeling and utilizes high-confidence pseudo labels for self-learning while self-penalizing low-confidence pseudo labels. In addition, we collect a cross-platform social event detection (CSED) dataset for the cross-platform multimodal social event detection task. Specifically, it contains 37,711 multimodal samples covering 40 public social events from three distinct platforms, i.e., Twitter, Flickr and online news media. Each event within our dataset is verified through Wikipedia, ensuring reliability and breadth of coverage across a wide range of topics. This dataset not only facilitates the exploration of cross-platform multimodal social event detection but also sets a new benchmark for future research in this domain. The experimental results on the CSED dataset demonstrate its effectiveness in both cross-platform and missing modality scenarios.

Our contributions are summarized as follows:

- We propose a Self-Supervised Modality Complementation (SSMC) method that effectively addresses the challenges of incomplete modalities and platform heterogeneity in cross-platform multimodal social event detection.

- We compile a comprehensive Cross-platform Social Event Detection (CSED) dataset, bridging the gap in multimodal data resources across diverse platforms.

- Through extensive experiments on the CSED dataset, we validate the effectiveness of our SSMC method, setting a new benchmark for cross-platform multimodal social event detection.

4.2 Preliminaries and Problem Statement

A social event is a significant occurrence or happening that is the subject of media coverage and public interest. In the digital era, such events are disseminated across various platforms, each with its unique presentation and content format. Cross-platform multimodal social event detection is the process of identifying and categorizing these social events across different platforms by analyzing multimodal data, such as text articles and images. A more formal definition of the problem is illustrated as follows.

Problem 1 (Cross-platform Multimodal Social Event Detection). Cross-platform multimodal social event detection aims to address the identification of social events y^T in a target platform's dataset $\mathcal{D}^T$ based on the learning from a source platform's labeled dataset $\mathcal{D}^S$ and the target platform's unlabeled dataset $\mathcal{D}^T$. Formally, the source dataset is denoted as $\mathcal{D}^S = \{(I^S, T^S, Y^S)\} = \{(i_i^S, t_i^S, y_i^S)\}_{i=1}^{n_S}$, consisting of n_S samples, where $i_i^S \in \mathbb{R}^{d_{SI}}$ and $t_i^S \in \mathbb{R}^{d_{ST}}$ are the image and text modalities of the i-th sample, respectively, and $y_i^S \in \mathcal{Y} = \{1, 2 \ldots, C\}$ is the labeled social event. The target dataset, lacking such labels, is represented as $\mathcal{D}^T = \{(I^T, T^T)\} = \{(i_i^T, t_i^T)\}_{i=1}^{n_T}$ with n_T samples. Notably, in some samples, either the image I or the text T from $\mathcal{D}^S$ and $\mathcal{D}^T$ might be missing. The goal is to predict the event labels $y^T \in \mathcal{Y}$ for $\mathcal{D}^T$, effectively bridging the gap between the multimodal data representations across platforms.

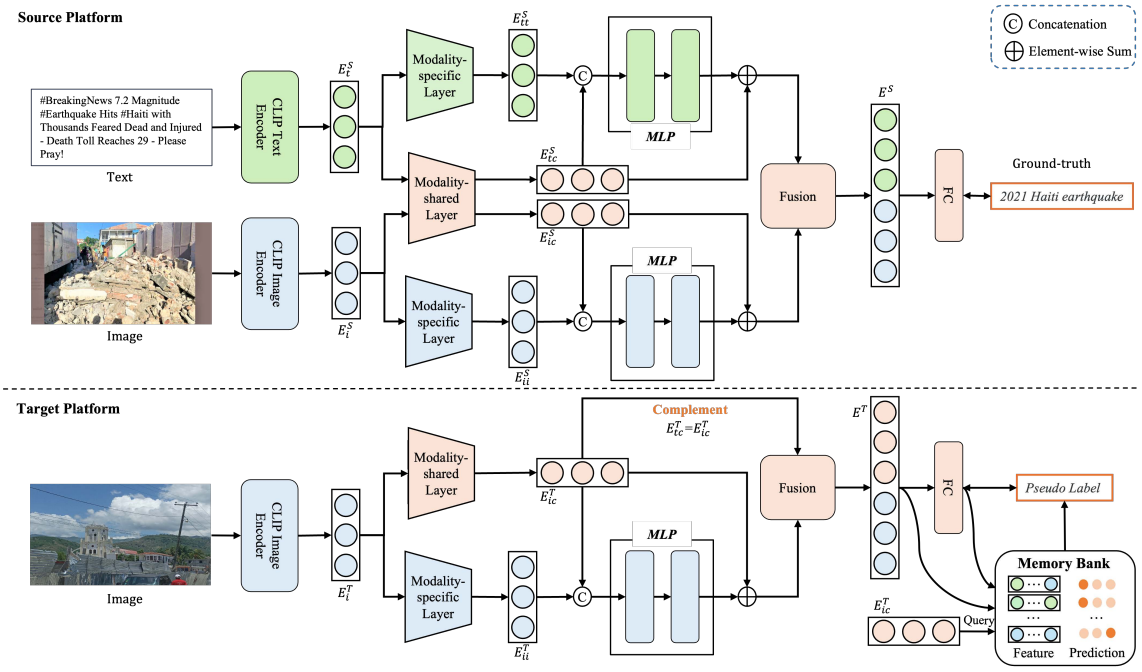


Figure 4.2: An overview of the proposed SSMC method when the text of the target domain is missing. The upper flow represents the source platform with labeled data (exemplified by Twitter posts with both images and accompanying text); and the lower flow depicts the target platform with unlabeled data (Flickr for example, where images are prevalent without text). Best viewed in color.

4.3 Methodology

In this section, we first provide an overview of the framework. Following that, we detail the processes of data preprocessing and the various submodules of SSMC.

4.3.1 Overview of the Framework

As illustrated in Figure 4.2, our SSMC model first uses CLIP [85] to extract the features of images and text separately. For each modality data, we design a modality-specific layer, i.e., F_T for text and F_I for image, to extract the specific modality information, and a modality-shared layer, i.e., F_C, to extract the common information

across different modalities. Then, we concatenate them with a residual module to fuse these two parts of features. Finally, we use a fusion module to combine two modal features and then proceed to classification. When a modality is missing, e.g., text in the target domain, the text modality-specific layer F_T becomes unavailable. We supplement it with the common features extracted from image modality, which not only plays a complementary role but also ensures end-to-end training. For the target platform, we utilize the semantic pseudo labels obtained from the output of the model and structural pseudo labels extracted from the multimodal common features (e.g., E_{ic}^T) and the multimodal fusion features (i.e., E^T) to perform high-quality pseudo label screening to achieve self-learning.

4.3.2 Data Preprocessing

Different social events can originate from various countries, thereby presenting a multilingual challenge in relevant social media. As shown in Table 4.1, there are 109 languages in the posts on Twitter about 40 different social events. We directly use Google Translation API[1] to convert multiple languages into English as our approach focuses on incomplete modalities and platform heterogeneity. Furthermore, to bring the multimodal feature distributions closer, we utilize CLIP [85] with ViT-B-32 for image and text feature extraction. The extracted features can be represented as E_i^S, E_i^T, E_t^S and E_t^T.

4.3.3 Missing Data Complementation (MDC)

Missing modality is common in multimodal learning. We consider supplementing the missing modality with information from another modality. However, there exists a modality gap between different modalities of data. To mitigate this gap, we design a separation mechanism to obtain modality-specific and modality-shared features be-

[1] `https://cloud.google.com/translate`

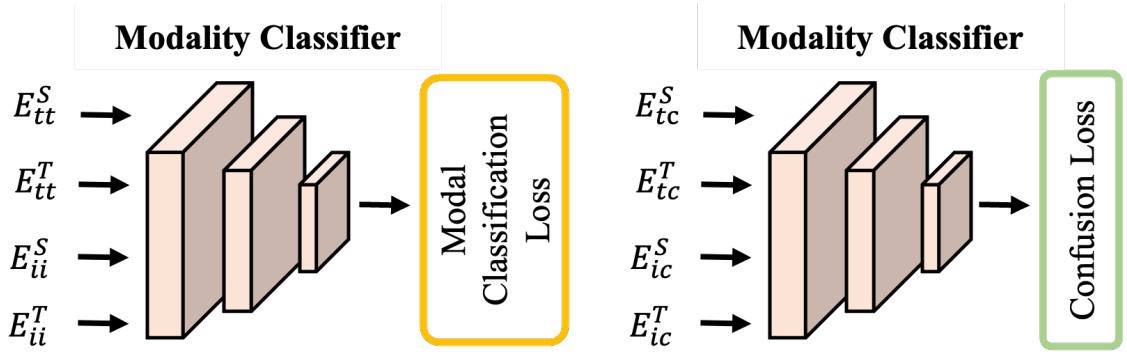


Figure 4.3: Missing Data Complementation Module.

tween modalities, and then use the common features to supplement information when a modality is missing. Specifically, we utilize two modality-specific layers to extract text-specific and image-specific features respectively, and one modality-shared layer to extract information common to both modalities.

To achieve this, we design two losses, i.e., modal classification loss ℓ_{mc} and confusion loss ℓ_d, as shown in Figure 4.3. Specifically, the modal classification loss assumes that if different modalities of data can be classified as that modality, then the modality contains modality-specific features, which can be expressed as:

$$\ell_{mc} = -\sum_{i=1}^{|\mathcal{D}_S+\mathcal{D}_T|} \sum_{j\in\{I,T\}} \left(d^{(j)}\right)^{\top} \log\left(F_{\text{mc}}\left(E_j^{(i)}\right)\right), \tag{4.1}$$

where $d^{(j)}$ represents a [1,0] or [0,1] vector when the input modalities are images (I) or text (T), respectively. F_{mc} denotes the modality classifier, which is composed of a multilayer neural network. $E_j^{(i)}$ refers to the i-th features extracted by modality-specific layers, i.e., E_{ii}^{S}, E_{ii}^{T}, E_{tt}^{S} and E_{tt}^{T}.

The confusion loss is used to achieve the goal of common feature extraction by confusing the modality classifier F_{mc}, which can be represented as:

$$\ell_d = -\sum_{i=1}^{|\mathcal{D}_S+\mathcal{D}_T|} \sum_{j\in\{I,T\}} \left(u^{(j)}\right)^{\top} \log\left(F_{\text{mc}}\left(E_j^{(i)}\right)\right), \tag{4.2}$$

where $u^{(j)}$ represents a [0.5,0.5] vector. $E_j^{(i)}$ refers to the i-th features extracted by

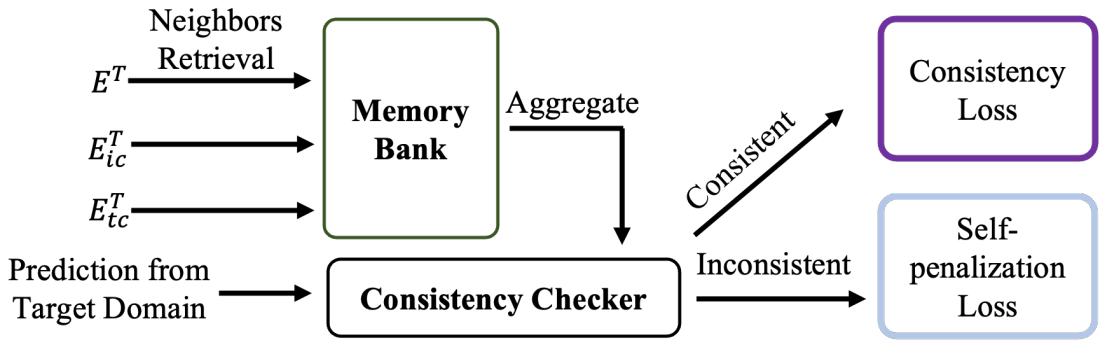


Figure 4.4: Multimodal Self-learning Module.

the modality-shared layer, i.e., E^S_{ic}, E^T_{ic}, E^S_{tc} and E^T_{tc}.

When the modality is complete, we concatenate the modality-specific features and common features for each modality, use a multilayer perceptron (MLP) for dimension reduction, and add the common features as a residual to form semantically complete modality embeddings for the text domain E_{text} and image domain E_{image}, which can be represented as:

$$E_{text} = \text{MLP}(\text{Concat}(E_{tt}, E_{tc})) + E_{tc}, \tag{4.3}$$

$$E_{image} = \text{MLP}(\text{Concat}(E_{ii}, E_{ic})) + E_{ic}, \tag{4.4}$$

where $\text{Concat}(\cdot, \cdot)$ represents the concatenation operation. For data with missing modalities, we compensate by extracting common features from another modality, e.g., $E'_{text} = E_{ic}$ if the text is missing. Finally, we obtain multimodal features (i.e., E^S and E^T) by fusing E_{text} and E_{image} for different platforms, and then use a shared classifier F_S for classification. Specifically, concatenation is utilized as the fusion method.

4.3.4 Multimodal Self-learning (MSL)

A significant domain gap exists across different platforms, especially in multimodal data. In our model, we chose a self-supervised learning strategy for cross-platform

adaptation because: 1) Traditional domain adaptation methods often fail when facing large domain gaps in multimodal data; 2) Self-supervision can leverage the intrinsic structure of unlabeled target domain data; 3) Using multiple views for consistency checking helps generate more reliable pseudo labels even with missing modalities. Specifically, we extract semantic pseudo labels and structural pseudo labels from multimodal views to obtain high-quality pseudo labels for self-learning.

For semantic pseudo labels, we directly obtain them by:

$$\hat{y}_i^{se} = \arg\max_k p_{i,k}, i = 1, 2, \cdots, n_t, \tag{4.5}$$

where $p_i = F_S\left(E_i^T\right)$ is the C-dimensional prediction and F_S denotes the shared classifier in the last layer.

However, the reliability of this pseudo label is not high as a domain gap exists. Therefore, we consider extracting structural pseudo labels, which make joint decisions based on the neighboring samples in the feature space and can achieve higher reliability [56]. In addition, multimodal data features have different views in the feature space. Compared to only using the multimodal features to generate pseudo labels, we also consider image and text features before fusion individually for extracting structural pseudo labels.

We first establish a memory bank to update the target domain features (including image common features E_{ic}^T, text common features E_{tc}^T and multimodal features E^T) and output probabilities of the model p_i. Specifically, we sharpen each probability and feature of the output via temperature scaling (i.e., $p_{i,k}^d = p_{i,k}^{\frac{1}{t}} / \sum_k p_{i,k}^{\frac{1}{t}}$ where t is the temperature scaling parameter of sharpening) and L2-normalization, respectively. During the training, we utilize these generated features and probabilities to update the memory bank by the moving average strategy. When the missing modality occurs, the memory bank only updates the features of the available modalities.

As shown in Figure 4.4, we utilize the features from the target domain to retrieve

the memory bank for each training step. Specifically, we use cosine distance to obtain the k-nearest samples for each sample from different feature spaces, i.e., image space E_{ic}^{T}, text space E_{tc}^{T} and multimodal space E^{T}. Furthermore, we find out the probabilities of the nearest neighbor samples and get their corresponding probabilities in the memory bank. Finally, we average the probabilities of the K nearest neighbor samples to obtain the final output probabilities for each space:

$$\hat{q}_i^m = \frac{1}{K} \sum_{j \in \mathcal{N}_i} p_j, \tag{4.6}$$

where $\mathcal{N}_i$ is the index of the nearest neighbors for the i-th sample. m denotes different feature spaces, i.e., image, text and multimodal spaces.

Then, we can directly use the maximum probability prediction corresponding to this probability as the structural pseudo label for each space:

$$\hat{y}_i^m = \arg\max_k \hat{q}_{i,k}^m. \tag{4.7}$$

After obtaining the semantic pseudo label $\hat{y}_i^{se}$ and structural pseudo labels $\hat{y}_i^m$, we perform a consistency check on them. We consider the pseudo label reliable when all views' pseudo labels are consistent. The consistency loss ℓ_{co} can be represented as:

$$\ell_{co} = -\frac{1}{n_{co}} \sum_{i=1}^{n_{co}} \log p_{i,\hat{y}_i}, \tag{4.8}$$

where $\hat{y}_i$ is the reliable pseudo label from voting for the i-th sample. n_{co} denotes the number of samples with a consistent pseudo label.

However, although the quality of pseudo labels obtained through consistency samples is high, even in the absence of modalities, the number of samples available for training has decreased after screening. Observations from preliminary experiments revealed a noteworthy phenomenon: after several rounds of training, the ground-truth label of an inconsistent sample is likely to be found among these pseudo labels generated from different perspectives. This insight led us to hypothesize that by penalizing

Algorithm 2 SSMC Algorithm

Input: Labeled source platform: $\mathcal{D}^S = \{i^S, t^S, y^S | i^S \in I^S, t^S \in T^S, y^S \in Y^S\}$, unlabeled target platform: $\mathcal{D}^T = \{i^T, t^T | i^T \in I^T, t^T \in T^T\}$, the feature extractors F_T, F_I and F_C, the modality classifier F_{mc} and the shared classifier F_S.

Output: Social event Y^T from target platform.

while $t \leq$ **MaxIter**

1: Compute the features E_i^S, E_i^T, E_t^S and E_t^T by the CLIP model.
2: Compute the multimodal features E^S and E^T according to Eq. 4.3 and Eq. 4.4.
3: Compute the pseudo-labels for the target domain from different views according to Eq. 4.5 and Eq. 4.7.
4: Perform a consistency check for these pseudo-labels and compute the consistency loss ℓ_{co} and self-penalization loss ℓ_{sp} according to Eq. 4.8 and Eq. 4.9, respectively.
5: Compute the cross-entropy loss ℓ_c , modal classification loss ℓ_{mc} and confusion loss ℓ_d according to Eq. 4.11, Eq. 4.1 and Eq. 4.2, respectively.
6: Optimize the overall objective in Eq. 4.10 through stochastic gradient descent.
7: Update memory bank features and output probabilities.

end while

categories not present among these pseudo labels, we could potentially increase the likelihood of the model predicting the correct category. The self-penalization loss ℓ_{sp} can be formulated as:

$$\ell_{sp} = \frac{1}{n_t - n_{co}} \sum_{i=1}^{n_t - n_{co}} \log(1 - p_{i,1_i}), \tag{4.9}$$

where 1_i is a one-hot label, with 0 at the i-th position when it is in the extracted pseudo label list and 1 elsewhere.

4.3.5 Overall Objective

The overall objective function to be minimized can be formulated as follows:

$$\ell = \ell_c + \alpha\ell_{mc} + \beta\ell_d + \lambda(\ell_{co} + \ell_{sp}), \tag{4.10}$$

$$\ell_c = -\frac{1}{n_s}\sum_{i=1}^{n_s}\log p_{i,\hat{y}_i^s}, \tag{4.11}$$

where ℓ_c denotes the loss for the source domain. α and β are hyperparameters. To reduce parameter sensitivity and ease the selection of models like [75], we adopt a gradual progressive strategy for λ, which jointly weights both l_{co} and l_{sp} as they complement each other in the self-learning process. The weighting parameter λ is set to $\frac{2}{1+exp(-10\cdot p)} - 1$, where p represents the training progress linearly changing from 0 to 1. This sigmoid-like function ensures λ starts from a small value and gradually increases, reducing the influence of potentially noisy pseudo labels in the early training stages. This gradual increase attenuates the influence of noise inherent in the pseudo labels during the preliminary iterations, consequently preventing the accumulation of errors to some degree.

The detailed algorithm for the SSMC method is presented in Algorithm 2. The computational complexity of SSMC mainly comes from: 1) CLIP feature extraction with $O(n)$ complexity where n is the number of samples; 2) modality classification and fusion with $O(d)$ complexity where d is the feature dimension; 3) nearest neighbor search for pseudo-label generation with $O(n_t^2)$ complexity where n_t is the number of target samples. Therefore, the overall complexity per iteration is $O(n + d + n_t^2)$.

4.4 Experiment

In this section, we evaluate our proposed SSMC method for cross-platform social event detection. We first introduce the CSED dataset, including its collection process and statistical analysis. Then, we conduct extensive experiments from multiple

Table 4.1: The statistics of CSED dataset.

	#Event	**#Sample**	**#Words**	**#Language**
Twitter (T)	40	24,607	14.72	109
Flickr (F)	40	9,191	46.66	51
Online News (O)	40	3,913	756.95	26

perspectives: 1) comparison with state-of-the-art domain adaptation and missing modality methods under different missing rates; 2) ablation studies to validate the effectiveness of the Missing Data Complementation (MDC) module and Multimodal Self-Learning (MSL) module; 3) parameter sensitivity analysis to investigate the impact of key parameters α and β; 4) analysis of pseudo label quality during training; and 5) visualization and case studies to demonstrate model performance across different platforms.

4.4.1 Cross-platform Social Event Dataset (CSED)

Collection and Statistics of the Dataset

To validate the performance of our proposed model in cross-platform social event detection, we extend the SED dataset from a single platform to multiple platforms, which includes 40 social events. Specifically, we choose three mainstream social media platforms for data collection, including the multimodal Twitter platform (T), the image-focused Flickr platform (F), and various online news platforms (O) that primarily feature long texts. Specifically, to avoid task simplification through direct keyword search, for Twitter, we use event-related hashtags and the time of the social event for collection; for Flickr, we utilize event-related keywords to search for related album sets, then collect related posts from them; for online news, we collect through the related links in the corresponding Wikipedia entries. Ultimately, we filter all single-modal samples and manually check whether the data semantically matches

the corresponding events, resulting in 37,711 samples. Specifically, unlike previous datasets including the SED dataset, we do not filter out non-English samples, considering that multilingual data could provide a more comprehensive perspective of social events. The statistics, distribution and feature visualization of our collected CSED dataset are shown in Table 4.1, Figure 4.5 and Figure 4.6, respectively.

Comparisons with Existing Datasets

Most of the current social event datasets are based on single-platform data, including both single-modal datasets [30, 58, 80] and multimodal datasets [5, 87, 112, 114]. Moreover, the datasets they collect are mostly in English only, which, for the task of social event detection, lacks the interpretation of different perspectives on events. Additionally, most datasets are directly collected based on keywords, which diminishes the value of multimodal data because texts retrieved solely by keywords can already perform well. These datasets are difficult to use for cross-platform social event detection directly as the social events are different. Therefore, we have collected a dataset that is multi-platform, multimodal, and multilingual, hoping to advance the development of this field.

4.4.2 Implementation Details

We evaluate all cross-platform scenarios (i.e., $T \rightarrow F$, $F \rightarrow T$, $T \rightarrow O$, $O \rightarrow T$, $F \rightarrow O$ and $O \rightarrow F$). For scenarios with missing modalities, we randomly mask the images or the text with a missing rate on both source and target domains. The missing rate is set at 0%, 20%, and 40%. Take 20% as an example (20% of the samples only contain text, 20% of the samples only contain images, and 60% of the samples contain both images and text for both source and target platforms).

For O and F, we combine the title and content as the text content. For MDC, α and β are set to 0.1 based on the observation of the performance of the labeled source

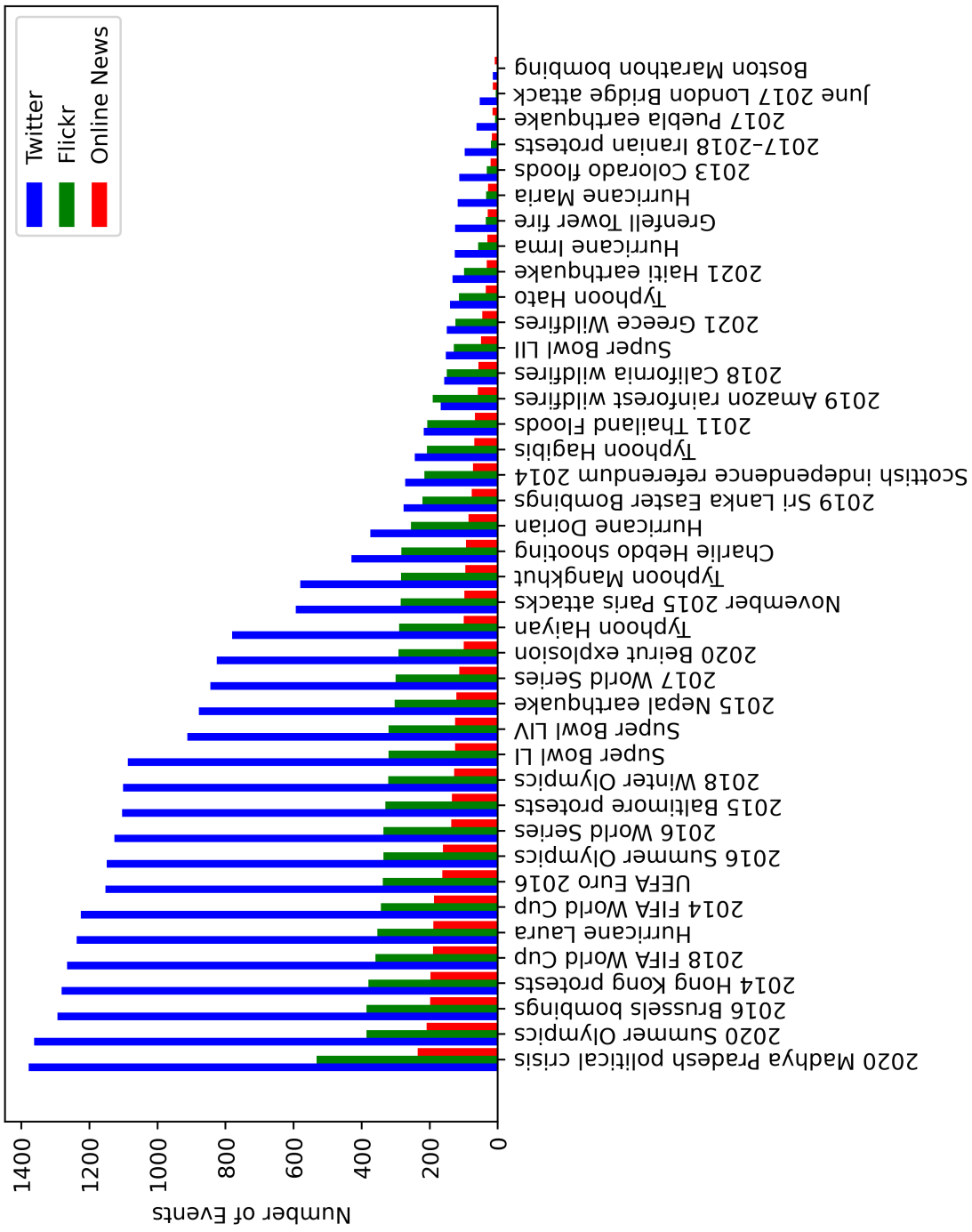


Figure 4.5: The distribution of CSED dataset.

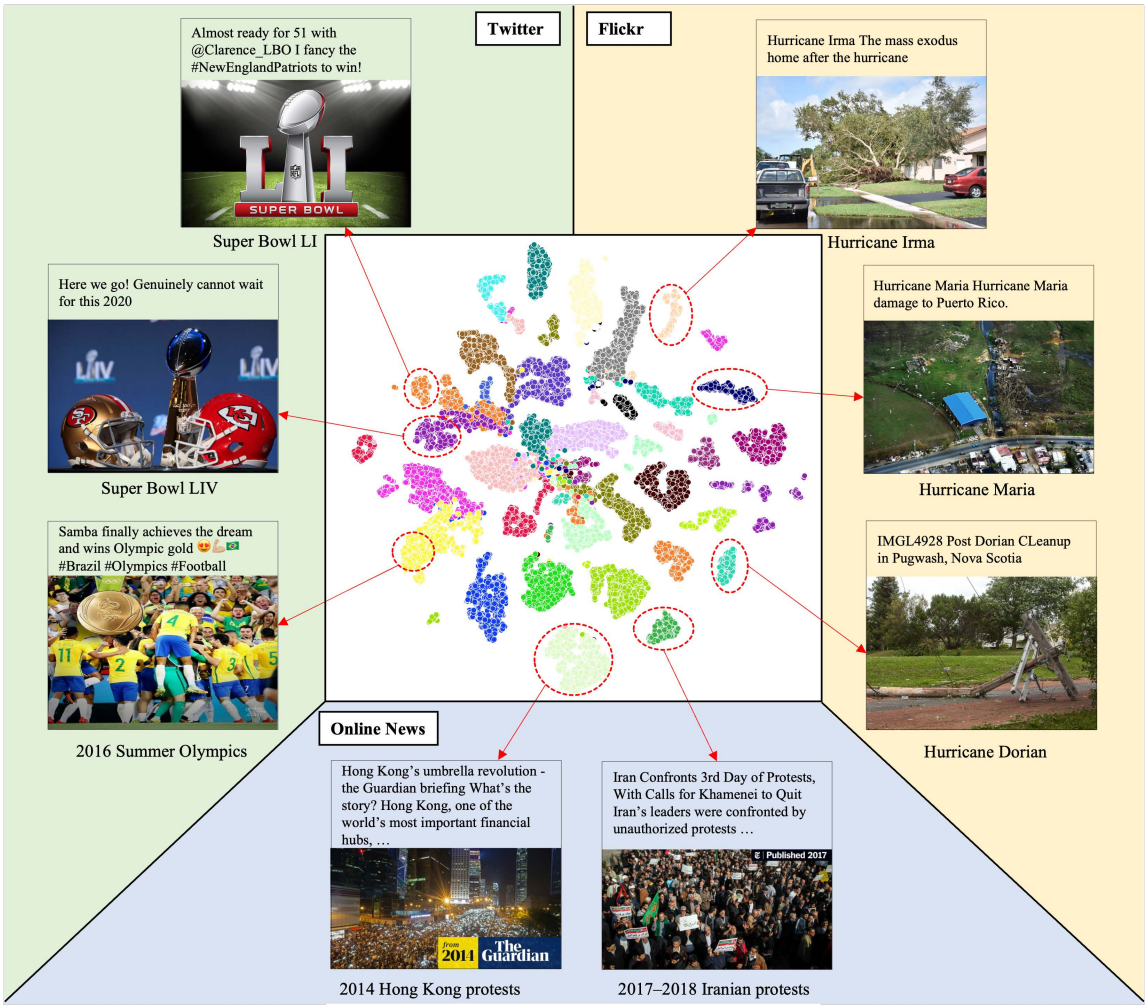


Figure 4.6: Feature visualization of the CSED dataset for different platforms.

domain as the target domain is unlabeled. In fact, they can be arbitrarily selected within a certain order of magnitude; our subsequent parameter analysis demonstrates that the results are robust to some variations of these hyperparameters. For MSL, we set the sharpening parameter t to 5 for memory bank construction, and the number of nearest neighbors N to 3 for pseudo label generation. We employ Adam as the optimizer, the learning rate as 1e-4, the batch size as 40, and the training epochs as 30. Accuracy and F1 score are used as the evaluation metric. In the experiment, we randomly choose five different seeds and compute their average to obtain the final result.

4.4.3 Baselines

We compare SSMC across two distinct scenarios: complete multimodal cross-platform social event detection (with a missing rate of 0%) and cross-platform social event detection with missing modalities (missing rates of 20% and 40%, respectively).

For the first scenario, our baselines include various domain adaptation methods as our benchmarks.

- **CORAL** [92] aligns the second-order statistics of the source and target distributions through a linear transformation.
- **DAN** [60] uses a multiple kernel variant of maximum mean discrepancies to learn transferable features in deep networks.
- **JAN** [61] aims to learn a transfer network by aligning the joint distributions of multiple domain-specific layers across domains using a joint maximum mean discrepancy criterion.
- **DANN** [26] focuses on learning domain-invariant features by adversarial learning.

- **MMDA** [82] employs multiple adversarial losses to learn common multimodal features.

- **MCC** [44] introduces a minimum class confusion constraint to achieve transfer learning.

- **ATDOC** [56] alleviates classifier bias by introducing an auxiliary classifier specifically for target data, thereby improving the quality of pseudo labels.

- **ENT** [108] integrates domain adversarial training into entropy minimization to enhance pseudo-label accuracy.

- **CDCL** [104] presents a method based on contrastive self-supervised learning aimed at feature alignment to minimize the domain gap between the training and testing datasets.

- **RCE** [21] introduces a training method that aligns with risk consistency, allowing the model to learn information from noisy pseudo-labeled data without compromising the performance.

Additionally, we report the performance of direct training on the source platform without applying domain adaptation methods, i.e., **image-only**, **text-only**, and **image+text** configurations. We also report the results of training and testing using the target domain, which serves as the **upper bound** (following an 8:2 split between the training and testing sets on the target platform).

For the second scenario, our baselines involve methods tailored for missing modalities.

- **DAL** [12] proposes incorporating available category information and adversarial training to enable the model to generate more informative domain information despite missing modalities.

- **DVAE** [45] introduces dual-aligned variational autoencoders to learn modality-invariant representations, addressing the challenge of missing modalities in data.

- **AT-CVAE** [56] proposes an adaptive transformer-based conditioned variational autoencoder network for incomplete social event classification, leveraging the capabilities of variational autoencoders and transformers to handle incomplete data.

4.4.4 Comparison with the State of the Arts

Table 4.2 and Table 4.3 summarize the performances of various methods on the CSED dataset. From the results, we have the following observations:

- Compared with other domain adaptation and missing modality methods, our method can simultaneously handle cross-platform and incomplete modalities and achieve the best performance, indicating our model's strong generalization ability for the cross-platform multimodal social event detection task.

- Many domain adaptation methods have declined results compared to those not using domain adaptation, possibly due to a large domain gap for the cross-platform multimodal social event detection task causing negative transfer.

- As the missing rate increases, the degree of decline in our method is relatively small compared to other models, which demonstrates the effectiveness of the proposed MDC.

- Our model performs relatively poorly when the target platform is Twitter compared to other platforms. This is due to Twitter having a larger number of posts and a wider variety of languages compared to other social media platforms, which inevitably contains more noise.

Table 4.2: Accuracy (Acc) on CSED dataset for cross-platform multimodal social event detection.

Methods	$T \rightarrow F$	$F \rightarrow T$	$T \rightarrow O$	$O \rightarrow T$	$F \rightarrow O$	$O \rightarrow F$	Average
Missing Rate: 0%							
Image Only	0.470	0.481	0.441	0.485	0.418	0.467	0.460
Text Only	0.669	0.593	0.784	0.653	0.828	0.725	0.709
Image+Text	0.761	0.669	0.784	0.692	0.818	0.804	0.755
CORAL [92]	0.766	0.667	0.780	0.689	0.813	0.805	0.753
DAN [60]	0.749	0.675	0.780	0.690	0.785	0.807	0.748
JAN [61]	0.724	0.647	0.694	0.637	0.723	0.788	0.702
DANN [26]	0.737	0.667	0.776	0.684	0.803	0.803	0.745
MMDA [82]	0.741	0.684	0.781	0.688	0.783	0.807	0.747
MCC [44]	0.763	0.665	0.811	0.709	0.825	0.792	0.761
ATDOC [56]	0.724	0.718	0.775	0.686	0.794	0.745	0.740
ENT [108]	0.743	0.668	0.791	0.713	0.749	0.800	0.744
CDCL [104]	0.810	0.713	0.855	0.716	0.824	0.802	0.787
RCE [21]	0.826	0.704	0.843	0.693	0.812	0.806	0.781
SSMC	**0.839**	**0.746**	**0.857**	**0.749**	**0.826**	**0.814**	**0.805**
Upper Bound	0.957	0.893	0.934	0.893	0.934	0.957	0.928
Missing Rate: 20%							
DAL [12]	0.648	0.580	0.682	0.627	0.639	0.695	0.645
DVAE [45]	0.665	0.594	0.702	0.608	0.698	0.690	0.659
AT-CVAE [56]	0.683	0.609	0.706	0.625	0.716	0.706	0.674
SSMC	**0.736**	**0.686**	**0.779**	**0.681**	**0.737**	**0.740**	**0.727**
Missing Rate: 40%							
DAL [12]	0.506	0.519	0.609	0.557	0.575	0.595	0.561
DVAE [45]	0.555	0.518	0.618	0.536	0.608	0.589	0.571
AT-CVAE [56]	0.609	0.542	0.641	0.558	0.634	0.616	0.600
SSMC	**0.644**	**0.611**	**0.692**	**0.597**	**0.653**	**0.652**	**0.641**

Table 4.3: F1 Score on CSED dataset for cross-platform multimodal social event detection.

Methods	$T \rightarrow F$	$F \rightarrow T$	$T \rightarrow O$	$O \rightarrow T$	$F \rightarrow O$	$O \rightarrow F$	Average
Missing Rate: 0%							
Image Only	0.413	0.385	0.411	0.426	0.359	0.418	0.402
Text Only	0.658	0.532	0.764	0.602	0.793	0.712	0.677
Image+Text	0.685	0.595	0.753	0.645	0.767	0.748	0.699
CORAL [92]	0.682	0.596	0.749	0.642	0.766	0.750	0.697
DAN [60]	0.663	0.582	0.749	0.630	0.735	0.738	0.683
JAN [61]	0.645	0.539	0.679	0.606	0.676	0.713	0.643
DANN [26]	0.660	0.587	0.749	0.638	0.672	0.735	0.674
MMDA [82]	0.667	0.592	0.748	0.634	0.730	0.741	0.685
MCC [44]	0.676	0.577	0.762	0.625	0.774	0.714	0.688
ATDOC [56]	0.666	0.640	0.743	0.658	0.749	0.684	0.690
ENT [108]	0.675	0.628	0.763	0.655	0.748	0.727	0.699
CDCL [104]	0.741	0.635	0.813	0.649	0.781	0.744	0.727
RCE [21]	0.742	0.620	0.795	0.630	0.766	0.727	0.713
SSMC	**0.762**	**0.670**	**0.819**	**0.680**	**0.797**	**0.769**	**0.750**
Upper Bound	0.897	0.841	0.921	0.841	0.921	0.897	0.886
Missing Rate: 20%							
DAL [12]	0.573	0.493	0.643	0.575	0.535	0.618	0.573
DVAE [45]	0.594	0.510	0.656	0.553	0.632	0.621	0.594
AT-CVAE [56]	0.612	0.525	0.676	0.577	0.660	0.655	0.617
SSMC	**0.674**	**0.601**	**0.734**	**0.607**	**0.693**	**0.686**	**0.666**
Missing Rate: 40%							
DAL [12]	0.554	0.436	0.576	0.503	0.476	0.542	0.514
DVAE [45]	0.506	0.436	0.585	0.480	0.566	0.537	0.518
AT-CVAE [56]	0.563	0.454	0.614	0.505	0.582	0.575	0.549
SSMC	**0.603**	**0.530**	**0.645**	**0.535**	**0.602**	**0.612**	**0.588**

Table 4.4: Ablation Study. M and G indicate M2M-100 model [22] and Google API respectively ($T \rightarrow F$, missing rate: 20%).

#	ℓ_{mc}	ℓ_d	ℓ_{sp}	ℓ_{co}	M	G	Accuracy	F1
1	×	✓	✓	✓	×	✓	0.717	0.636
2	✓	×	✓	✓	×	✓	0.723	0.639
3	×	×	✓	✓	×	✓	0.709	0.630
4	✓	✓	×	✓	×	✓	0.722	0.652
5	✓	✓	✓	×	×	✓	0.715	0.650
6	✓	✓	×	×	×	✓	0.708	0.649
7	✓	✓	✓	✓	×	×	0.693	0.622
8	✓	✓	✓	✓	✓	×	0.709	0.636
9	✓	✓	✓	✓	×	✓	**0.736**	**0.674**

4.4.5 Ablation Study

To investigate the effectiveness of the various modules, we conduct an ablation experiment. Specifically, we consider removing each of our proposed modules to test the performance, i.e., without (w/o) the modal classification loss ℓ_{mc}, confusion loss ℓ_d, self-penalization loss ℓ_{sp} and consistency loss ℓ_{co}. As illustrated in Table 4.4, we select one of the cross-platform scenarios with missing rate as 20%, i.e., $T \rightarrow F$. From the table, we observe that:

- The performance of our model declines after removing any module, demonstrating the effectiveness of each module we proposed.
- The significant impact is on the consistency loss, as it provides high-quality pseudo labels that directly facilitate learning in the target domain.

Furthermore, we also conducted ablation studies on the translation module. Although our method does not focus on this part, it exists in our dataset, and here we

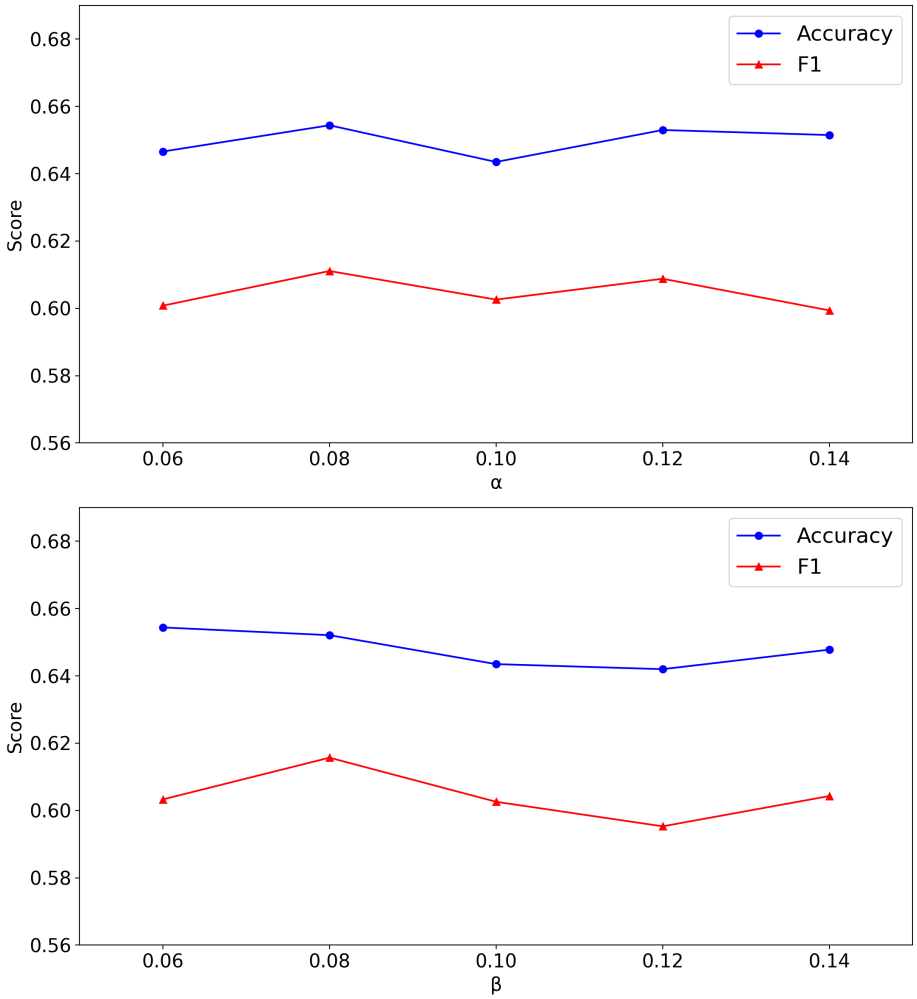


Figure 4.7: Sensitivity of α and β ($T \rightarrow F$, missing rate: 40%).

show the impact of this part on our model's results. Specifically, we choose another multilingual translation model, i.e., M2M-100 [22], for comparison. As shown in Table 4.4, we observe that the Google API performs better than the M2M-100 model, which is why we select it as our preprocessing module.

4.4.6 Impact of Parameters α and β

We analyze the impact of parameters α and β on our model. When testing the value of α, we set β to 0.1; when testing β, we set α to 0.1. The choice of 0.1 is based

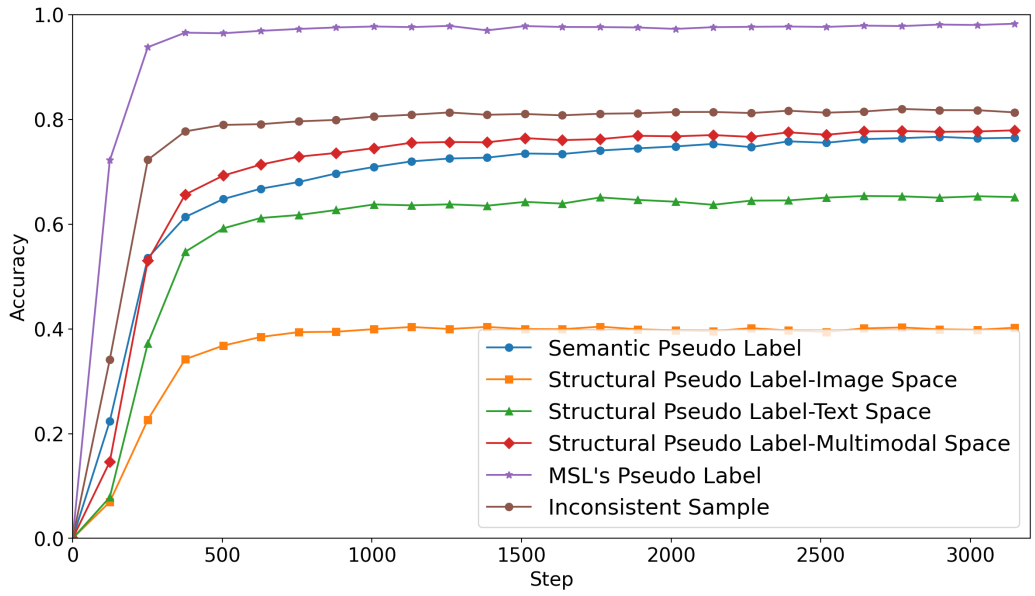


Figure 4.8: Accuracy of pseudo label during training ($T \rightarrow F$, missing rate: 20%)

on observing the performance of the source domain data, as we cannot access the target data. Here, we observe the performance of SSMC on the target platform as α and β vary. As shown in Figure 4.7, our research indicates that our method is robust within a certain range of different α or β values, with performance at α or $\beta = 0.08$ even surpassing the performance reported in the Table 4.2 and Table 4.3. Note that we did not conduct ablation studies on λ since it automatically adapts during training through a predefined function rather than being a fixed hyperparameter. The effectiveness of this adaptive strategy was instead validated through the ablation studies on l_{co} and l_{sp} that λ weights.

4.4.7 Evaluating the Effectiveness of MSL

To further explore the effectiveness of the consistency loss and self-penalization loss proposed in our MSL, we observe the accuracy of pseudo labels from different perspectives on the target domain during the training process. As shown in Figure 4.8, we find that compared to other pseudo labels, our method's consistency pseudo labels

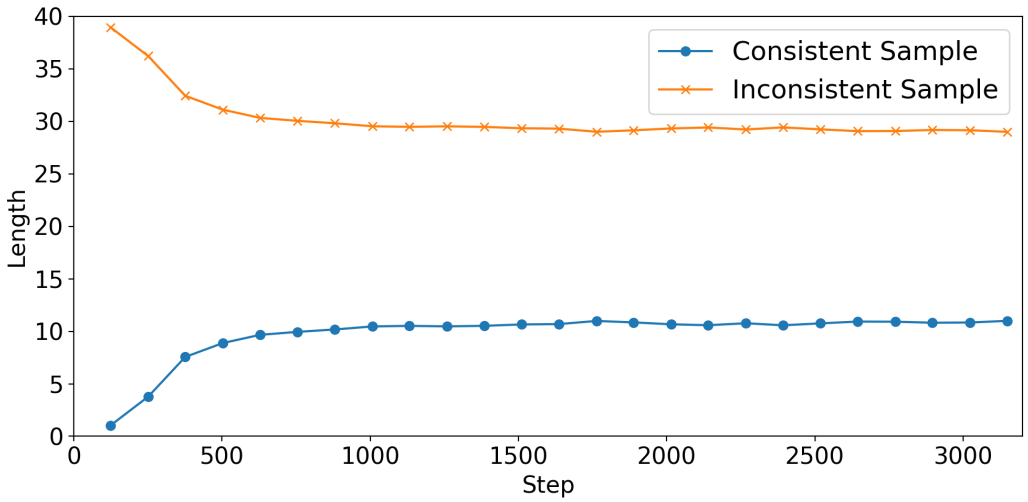
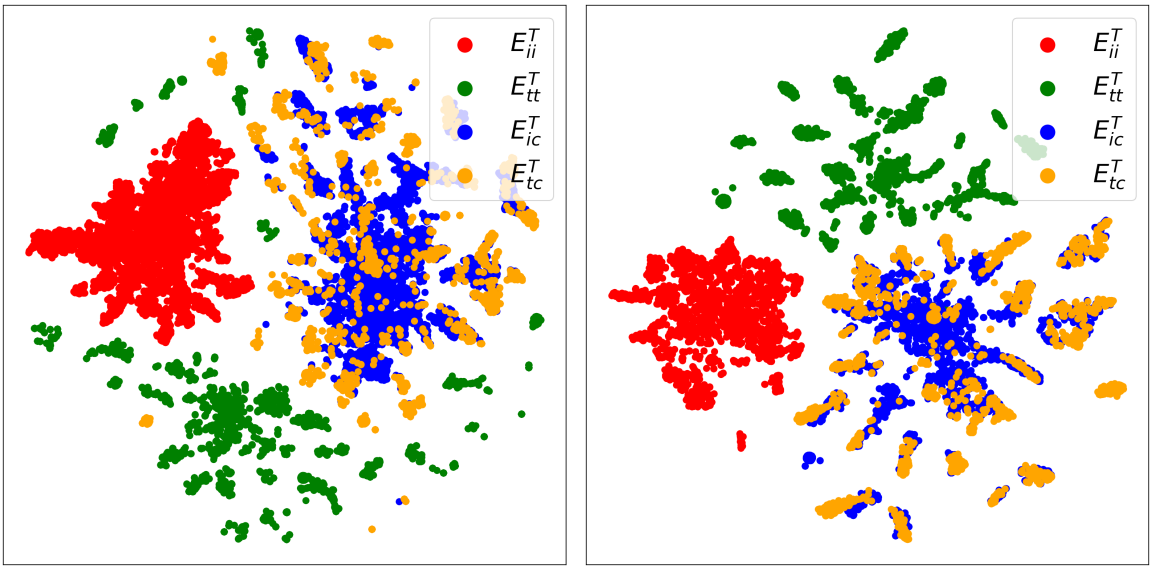


Figure 4.9: Length of consistent and inconsistent samples during training ($T \rightarrow F$, missing rate: 20%, batch size: 40)

achieve an accuracy rate close to 100% after several rounds of training, even with a missing rate of 20%, which proves the high quality of our proposed pseudo labels. Meanwhile, as shown in Figure 4.9, the proportion of these high-quality samples is only 25% as mentioned in Section 4.3.4. Therefore, we use the self-penalization loss to learn from the remaining inconsistent samples. In Figure 4.8, we also visualized the accuracy between inconsistent samples' labels and the pseudo labels list obtained from different perspectives. We can find the accuracy rate is relatively high (around 80%) compared to other pseudo labels, which verifies our previous assumption. Therefore, by combining these two strategies, our model can utilize all samples for training, which results in good performance.

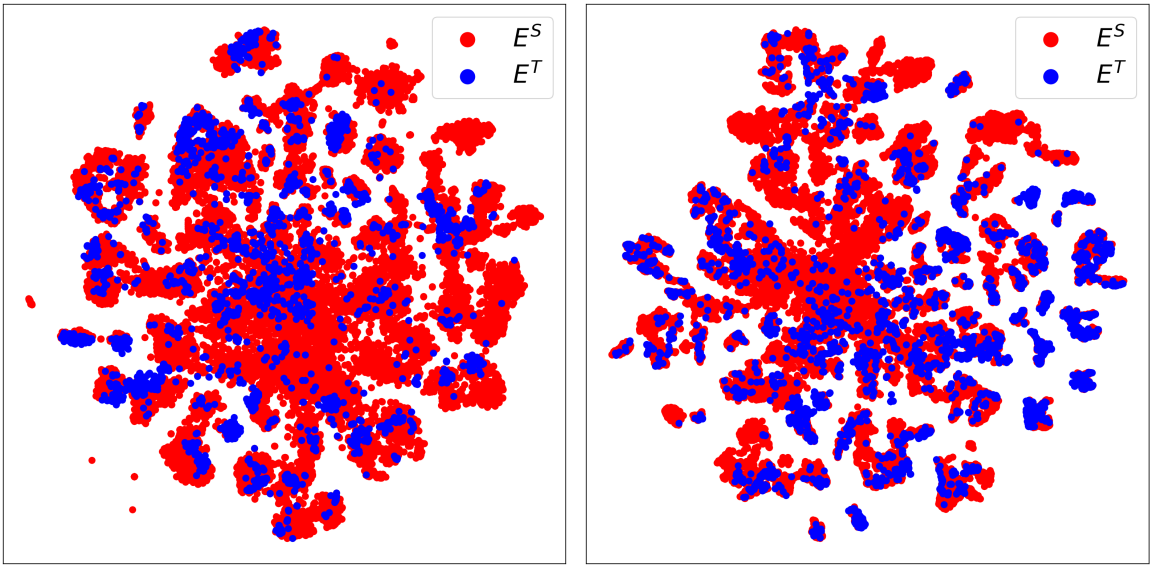
4.4.8 Visualization

To validate the effectiveness of MDC, we employ t-SNE to visualize the common features and the modality-specific features from the target domain. As shown in Figs. 4.10a and b, we can observe that the common features of images and texts are



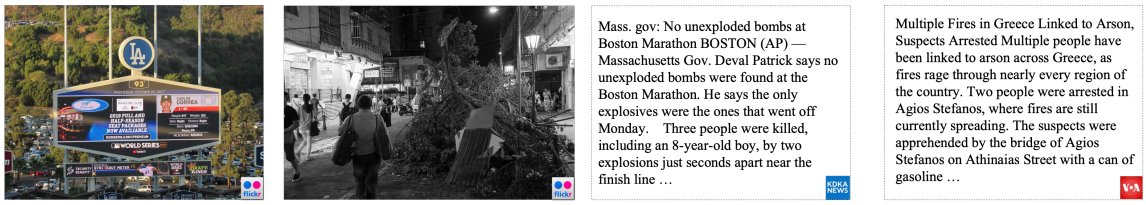
(a) $T \rightarrow F$ (b) $T \rightarrow O$

Figure 4.10: Visualization of common features and modality-specific features from the target domain under $T \rightarrow F$ and $T \rightarrow O$ (missing rate: 20%).



(a) AT-CVAE (b) SSMC

Figure 4.11: Visualization of multimodal features from source and target domains under $T \rightarrow O$ (missing rate: 20%).



	DVAE:	AT-CVAE:	SSMC:	GT:
(a)	2016 World Series	UEFA Euro 2016	2017 World Series	2017 World Series
(b)	Typhoon Mangkhut	2011 Thailand Floods	Typhoon Hagibis	Typhoon Hato
(c)	2016 Brussels bombings	2016 Brussels bombings	Boston Marathon bombing	Boston Marathon bombing
(d)	Beirut explosion	2020 Beirut explosion	2017–2018 Iranian protests	2021 Greece Wildfires

Figure 4.12: Success and failure examples induced by SSMC on the CSED dataset from two different scenarios, i.e., $T \rightarrow F$ and $T \rightarrow O$ (missing rate: 20%).

mixed together, while the modality-specific features do not overlap at all, which indicates the effectiveness of our proposed MDC module. In addition, we also visualize the multimodal features from both the source and target domains by using different methods. As illustrated in Figs. 4.11a and b, the boundaries of our SSMC's features from the source and target domains are clearer compared to AT-CVAE, which validates the effectiveness of our proposed MSL module.

4.4.9 Case Study

Figure 4.12 presents several success and failure examples in different scenarios. From the figure, we have the following observations:

- Figures 4.12a and c illustrate successful predictions by SSMC, which demonstrates that our model can make correct judgments based on another modality when one modality is missing.

- As shown in Figure 4.12b, when the target platform lacks text, if the image information does not contain more elements about the corresponding social event, our model is prone to errors (e.g., although it predicts a typhoon event based on the collapse of trees, it is still a different event).

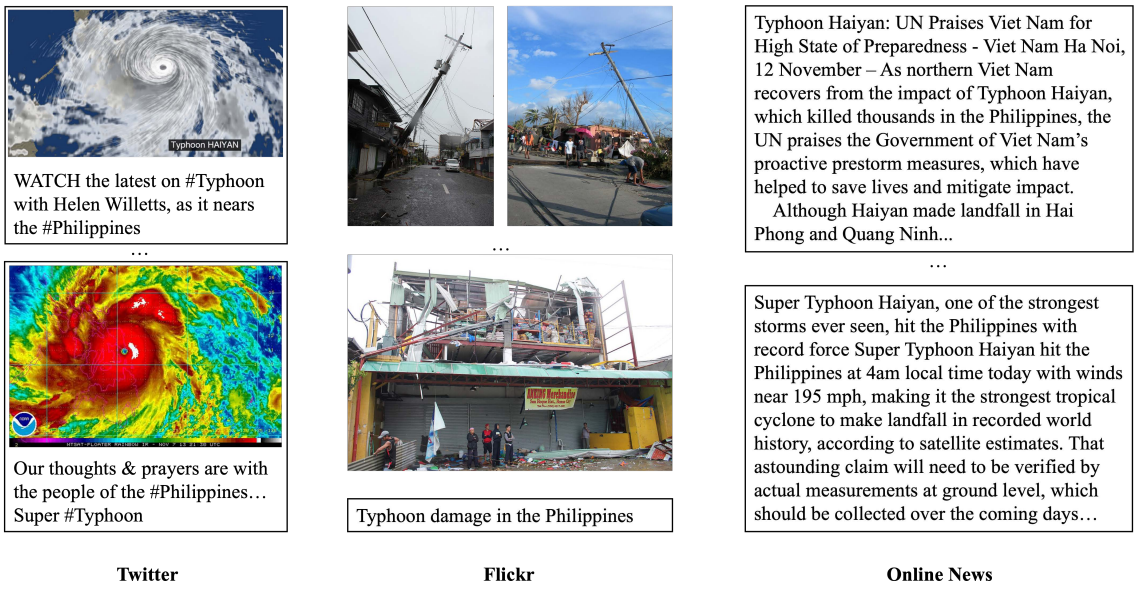


Figure 4.13: Samples detected by using SSMC for the "Typhoon Haiyan" event with Twitter as the Source platform and Flickr and Online News as the Target platforms.

- As illustrated in Figure 4.12d, When the target platform lacks images, the lengthy texts in online news may contain many descriptions unrelated to the event, which misleads our model (e.g., the example describes extensively the causes of a fire without describing wildfires).

In addition, to verify that cross-platform event detection can improve the quality of event data from a single platform, we use the SSMC model for cross-platform event detection. Specifically, we take Twitter as the source platform, and Flickr and Online News as the target platform to detect the "Typhoon Haiyan" event. As shown in Figure 4.13, we can find that Twitter, as the source platform, mainly consists of real-time updates and alert information rapidly posted by users during the natural disaster. The detected samples from target platforms like Flickr focus more on high-quality post-disaster photos showing the damage, while online news provides textual descriptions of the entire natural disaster. Therefore, by combining data from cross-platform sources, we can overcome the limitations of a single platform and obtain

more comprehensive information, which validates the importance of cross-platform event detection.

4.5 Conclusion

In this chapter, we propose a Self-Supervised Modality Complementation (SSMC) method that effectively addresses the challenges of missing modality and cross-platform scenarios in the cross-platform multimodal social event detection task. A missing data complementation module is designed to use modality-shared features to supplement missing modality scenarios, and a multimodal self-learning module generates reliable pseudo labels from multiple perspectives to achieve self-learning and self-penalization in the target domain. We have also introduced a comprehensive Cross-platform Social Event Detection (CSED) dataset, encompassing diverse platforms and a wide range of public social events. Experimental results on the CSED dataset validate the effectiveness of our proposed method and demonstrate the role of cross-platform event detection in improving the quality of event data on a single platform.

Chapter 5

Generalized Social Event Detection via Dynamic Augmentation and Entropy Optimization

In this chapter, we propose a generalized social event detection task, which aims to leverage labeled event data to learn more generalized features for detecting both known and newly occurring events. Specifically, we introduce a deep learning network, DAEO, to achieve this. On one hand, it utilizes a multimodal augmentation module to learn more robust multimodal event features. On the other hand, it combines self-distillation learning and adaptive entropy optimization to detect new events. Additionally, we expand the SED dataset by increasing the number of event types and samples, and conduct extensive experimental analysis.

5.1 Introduction

The current research [2, 56, 57, 112, 121] in social event detection primarily focuses on utilizing multimodal approaches due to the richer information provided by the

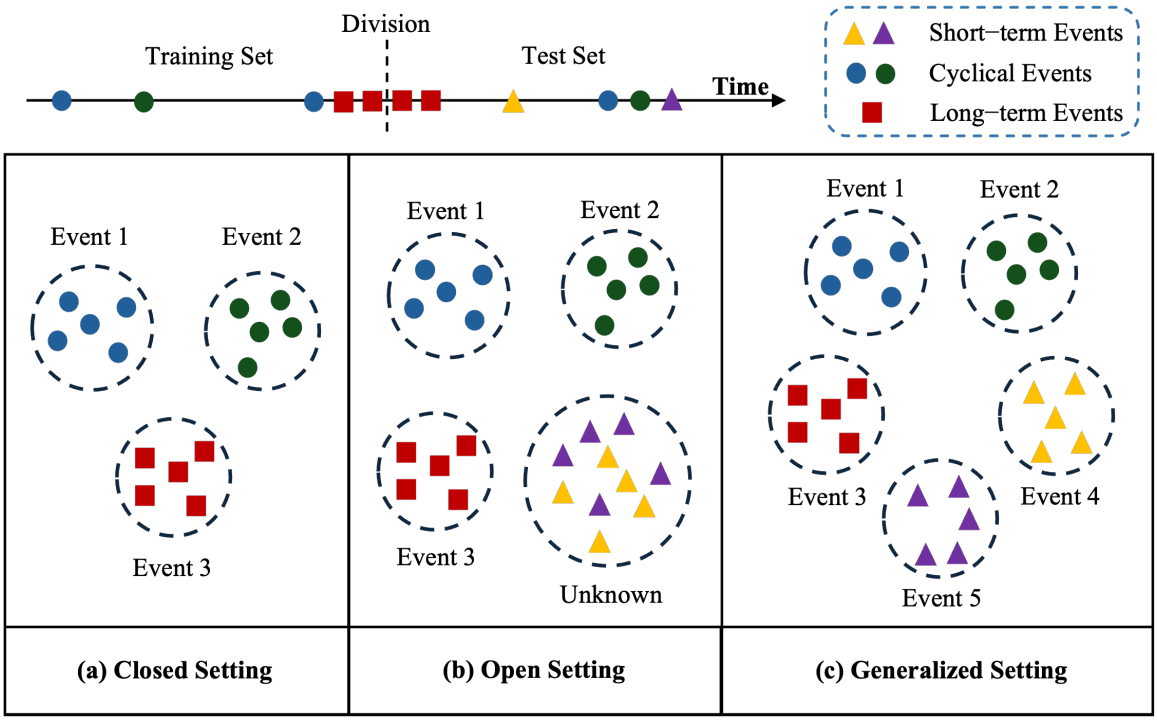


Figure 5.1: Different settings for social event detection. Events 4 and 5 are new events that do not occur in the training set.

multimodal data. However, a significant limitation of these studies is their reliance on a closed-set assumption, which greatly diminishes their applicability in practical scenarios. As illustrated in Figure 5.1a, under a closed setting, the focus is mainly on identifying social events that are already known, like cyclical or long-term events, which have happened before. This kind of approach falls short when it comes to detecting novel events that emerge over time. In response to this limitation, some researchers [83] have proposed shifting towards an open setting, aiming to identify novel events as they occur as shown in Figure 5.1b. Yet, as the volume of new events grows in real life, merely distinguishing whether an event is unknown or not is often insufficient. This has led to the exploration of the generalized social event detection problem, which seeks to extend beyond the binary classification of new events as either known or unknown. This task aims at not only recognizing previously occurred social

events but also differentiating among events that have not yet occurred, referring to this capability as general category discovery [99]. As shown in Figure 5.1c, it requires the model to both identify known events and categorize new events.

However, the task of generalized social event detection presents several challenges. The first challenge lies in dealing with multimodal features. When identifying social events, it's possible that only one modality, either text or image, provides useful information, or both modalities offer complementary insights. This variability requires the effective integration and utilization of each modality, especially for events that are closely related or similar in nature. The second challenge involves utilizing knowledge of previously occurred events to identify new events. This necessitates a model capable of distinguishing subtle differences between known and new events. Lastly, the challenge of dataset scarcity compounds the difficulty of this task. A comprehensive dataset, rich in both volume and variety of events, including temporal information, is crucial for this task. As illustrated in Figure 5.1, temporal information plays a crucial role in the division of datasets. Unfortunately, existing datasets [57, 83] lack this temporal metadata, leading to the use of random splits for training and test sets, which can not reflect the real-world scenario where events unfold over time.

To address the aforementioned challenges, we introduce a Dynamic Augmentation and Entropy Optimization (DAEO) model. For the first challenge, we design a multimodal augmentation module to learn more robust multimodal event features, which implicitly leverages the label information of events to learn the relationship between different modalities. It utilizes adversarial learning to not only encourage the generation of multimodal features that can distinguish between different similar events but also ensure the generated features are as diverse as possible. For the second challenge, we learn a unified prototypical classification head for all new and known classes with self-distillation learning. Unlike previous methods [105] that used entropy maximization for all samples, we introduce an adaptive entropy optimiza-

tion technique. Specifically, we generate various pseudo labels using a multi-view approach, including single-modal random augmentations (e.g., image augmentations) and outputs from the multimodal augmentation module. Then, when there is consistency across multiple views, the model is optimized to minimize entropy, thereby enhancing confidence in identified known events. Conversely, when views differ, entropy maximization is employed to encourage further exploration of the new events. Furthermore, we collect a multimodal social event detection (MSED) dataset for generalized social event detection from Twitter, comprising 161,350 multimodal samples annotated with 66 real-world events. Reflecting the temporal characteristics of social events, we define and collect three types of social events: short-term, cyclical, and long-term events. To ensure diversity, each event type encompasses a broad range of sub-events, including short-term events like natural and man-made disasters, terrorist attacks; cyclical events such as sporting events, political elections, international summits; and long-term events covering political conflicts, economic/social crises, and environmental/health issues. Experimental results on the MSED dataset demonstrate the effectiveness of our proposed method, validating its capability to address the challenges of generalized social event detection efficiently.

The contributions of this chapter can be summarized as follows:

- We formulate the task of generalized social event detection and introduce a Dynamic Augmentation and Entropy Optimization (DAEO) model designed to tackle this task.
- We propose a multimodal augmentation module and an adaptive entropy optimization strategy aimed at improving the representation of multimodal features and enhancing the ability to uncover new events, respectively.
- We collect a comprehensive multimodal social event detection (MSED) dataset deigned for social event detection, which encompasses a wide array of events categorized into long-term, cyclical, and short-term events, providing a rich

resource for the research community.

- Extensive experimental results on the MSED dataset demonstrate the effectiveness of our proposed model.

5.2 Preliminaries and Problem Statement

Problem 1 (Generalized Social Event Detection). Given a dataset D contains two parts: D_L containing known events and D_U including both known and new events, organized chronologically. A model is expected to be developed that can accurately categorize both known and new events in D_U.

More specifically, $D_L = \{(x_i, y_i)\}_{i=1}^{N}$ constitutes a labeled dataset containing multimodal instances x_i, each labeled with y_i from the set Y_L of known event categories. $D_U = \{(x_j)\}_{j=1}^{M}$ represents an unlabeled dataset with multimodal instances x_j, which are to be associated with labels from an expanded set Y_U. The set Y_U includes new, unseen event categories denoted by Y_{new}, and a subset of Y_L, designated as Y_{future}, which represents a subset of Y_L that will continue to happen in the future. Hence, the relationship $Y_U = Y_{new} \cup Y_{future}$, with $Y_{future} \subseteq Y_L$ as not all events from Y_L are expected to reoccur. During training, the model is concurrently trained on both D_L, to learn from the historical occurrence of events, and D_U, to anticipate and categorize future, unseen events.

In addition, in order to ensure that there are enough event types and relationships that can be used for generalized social event detection, we define three types of social events based on their temporal attributes: short-term, cyclical, and long-term events. The following are the formal definitions:

Definition 1: (Short-term Event). A short-term event is characterized by its ephemeral nature, typically unfolding and concluding within a brief time span. Examples of such events include natural disasters, sudden political upheavals, or

unexpected public incidents. These events are transient and unpredictable, hence they have a high probability of falling into both $Y_L \setminus Y_{future}$ (elements present in Y_L but absent in Y_{future}) and Y_{new} since they may not have occurred in the past or might represent entirely new scenarios.

Definition 2: (Cyclical Event). Cyclical events are those that occur at regular intervals, marked by their predictability and periodicity. An example of a cyclical event is the Olympic Games, which recur on a four-year cycle. These events are anticipated and are typically encompassed within Y_{future} due to their recurrent nature.

Definition 3: (Long-term Event). Long-term events span extended periods, often unfolding over months, years, or even decades. Wars, economic recessions, or major policy reforms are examples of long-term events. These events persist over such durations that they may be present in both Y_L and Y_U.

5.3 Methodology

In this section, we first provide an overview of the framework. Following that, we will introduce the various submodules of DAEO.

5.3.1 Overview of the Framework

As illustrated in Figure 5.2, our DAEO model begins by leveraging a pretrained CLIP model [85] to extract features from both images and texts, which are then concatenated to form multimodal event features E. Specifically, to enable self-learning from unlabeled data, we apply random data augmentation to the images of the input posts to obtain an augmented post for distillation learning. The multimodal augmentation module then employs adversarial learning to generate robust multimodal augmented features E^{Aug}, which enhances the classifier's ability to distinguish between similar

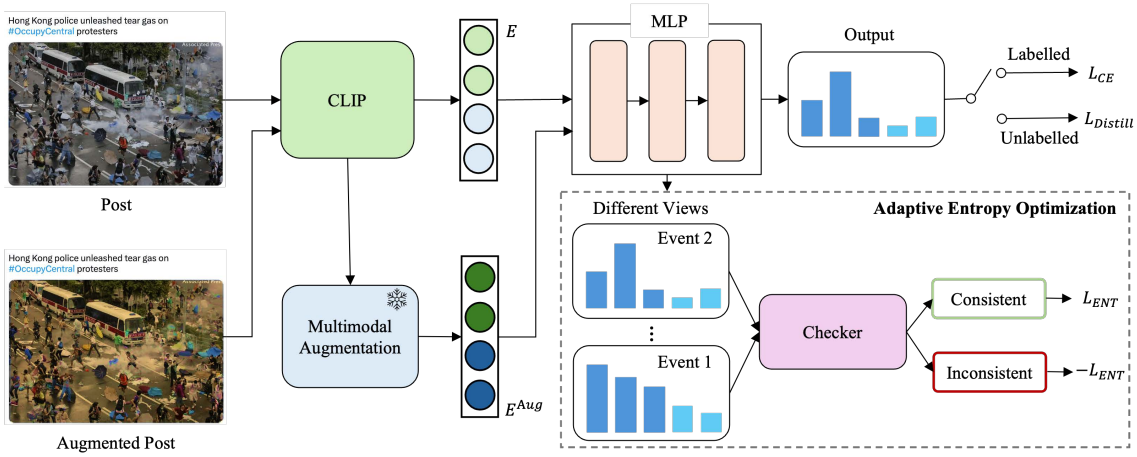


Figure 5.2: The framework of the proposed Dynamic Augmentation and Entropy Optimization (DAEO) model.

events. Then, we adopt a multilayer perceptron (MLP) as classifier f to obtain the output. For labeled data, we employ standard supervised learning techniques using the labels; for unlabeled data, we utilize distillation learning for training. Additionally, the adaptive entropy optimization module uses the generated multi-view pseudo-labels for consistency checking to selectively optimize entropy. This approach not only encourages the detection of new events but also improves the accuracy of known events.

5.3.2 Multimodal Event Feature Extraction

According to [99], it is crucial to adopt a robust pretrained model to discover new category, like DINO ViT [15]. However, most pretrained models are primarily focused on image data. Thanks to the cross-modal alignment training on very large-scale image-text pairs, CLIP [85] demonstrates strong zero-shot performance, evidencing its powerful generalization capability for multimodal joint embedding. Therefore, given an input sample x_i, we utilize the pretrained CLIP ViT-B/16 model to generate features for both the images and texts. These features are then concatenated to form

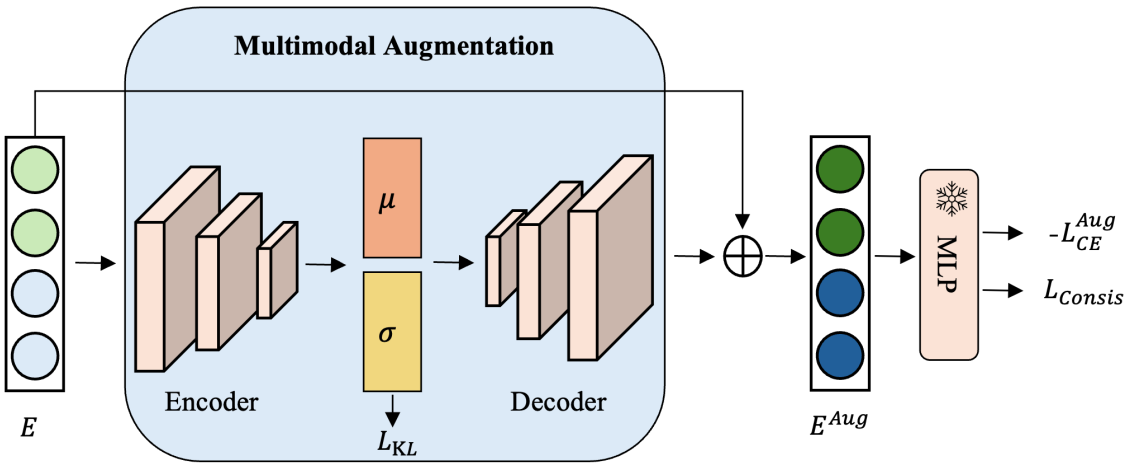


Figure 5.3: Multimodal augmentation module.

our multimodal event feature E_i, which can be represented as:

$$E_i = CLIP(x_i). \tag{5.1}$$

5.3.3 Multimodal Augmentation

In generalized social event detection, it is important for a model to distinguish similar social events finely, such as different earthquakes in disaster events. Previous methods [99] utilize the supervised contrastive learning and self-contrastive learning method to widen the decision margins between different categories. However, applying random data augmentation for contrastive learning on single modalities, such as text or images, does not seem to enhance model performance for multimodal data (see Sec. 5.4.6). A possible reason is that random augmentation, especially for text, might lead to the loss of event-related clues, causing negative optimization. For social event detection tasks, the relationship between images and text can be complementary, related, or unrelated.

In our model, we adopt a different approach to learn robust features, i.e., the multimodal augmentation module, by generating multimodal augmented features through

the adversarial method [62] at the feature level. On one hand, we aim for the generated multimodal augmented features to closely approach the decision boundary, which improves the classifier's ability to distinguish between similar events. On the other hand, we strive to ensure that the generated multimodal features retain the original event semantics, which prevents negative optimization.

As shown in Figure 5.3, we employ a Variational Autoencoder (VAE) model [49] as the generative model, denoted as G, which includes an encoder and a decoder. The VAE model has been proven effective in generating features. We utilize the KL divergence [37] to make the encoder's output as close to a standard Gaussian distribution as possible. Based on the properties of KL divergence between Gaussian distributions, this divergence is always non-negative and can be formulated in closed form as:

$$L_{KL} = -\frac{1}{2N}\sum_{i=1}^{N}\left(1+\log(\sigma_i^2)-\mu_i^2-\sigma_i^2\right), \tag{5.2}$$

where μ and σ are the mean and standard deviation parameters output by the encoder, respectively. Furthermore, we use a residual module to retain more of the original multimodal feature semantics. The augmented features E_i^{Aug} can be formulated as:

$$E_i^{Aug} = G(E_i) + E_i. \tag{5.3}$$

For learning to generate robust multimodal augmented features, we perform the adversarial training consisting of two parts. In the first part, as shown in Figure 5.2, we fix the parameters of the multimodal augmentation module G and train the CLIP and classifier model f to minimize the cross-entropy loss between the output and the true event labels, which ensures that augmented features retain their original semantics. It can be formulated as:

$$L_{CE}^{Aug} = -\frac{1}{N}\sum_{i=1}^{N}\ell_{ce}(f(E_i^{Aug}), y_L^{(i)}), \tag{5.4}$$

where $\ell_{ce}(\cdot)$ represents the cross-entropy loss function.

In the second part, as shown in Figure 5.3, we fix the parameters of the CLIP and classifier models and train G, on one hand, maximize the cross-entropy loss between the output and the true event labels as much as possible to generate more discriminative features, and on the other hand, minimize the consistency loss to align the semantics of the augmented and original multimodal feature outputs. The consistency loss can be formulated as:

$$L_{Consis} = -\frac{1}{N}\sum_{i=1}^{N} f(E_i)\log(f(E_i^{Aug})). \tag{5.5}$$

To achieve the adversarial goal, we want the optimal parameters $\hat{\theta}_{CLIP}$, $\hat{\theta}_G$ and $\hat{\theta}_f$ to jointly satisfy

$$(\hat{\theta}_{CLIP}, \hat{\theta}_f) = \arg\min_{\theta_{CLIP},\theta_f} L_{CE}^{Aug} + L_{CE}, \tag{5.6}$$

$$(\hat{\theta}_G) = \arg\max_{\theta_G} L_{CE}^{Aug} - L_{Consis} - L_{KL}, \tag{5.7}$$

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N} \ell_{ce}(f(E_i), y_L^{(i)}). \tag{5.8}$$

In this way, the generated features E_i^{Aug} will be close to the decision boundary, which further helps the classifier f to distinguish the class with some ambiguous decision boundaries.

5.3.4 Adaptive Entropy Optimization

The proposed adaptive entropy optimization strategy is designed based on several key insights: 1) When predictions from different views are consistent, it likely indicates the model has found meaningful patterns that should be reinforced; 2) Inconsistent predictions often suggest uncertainty about new categories that should be explored further; 3) Balancing between entropy minimization and maximization helps maintain accuracy on known categories while discovering new ones.

To identify new categories, we train a unified prototypical classification head for all new and known classes using a self-distillation framework. For self-distillation, we perform simple augmentations on images to obtain augmented images. Considering the potential for existing random text augmentation methods to change semantics and cause negative optimization, we compose augmented multimodal data directly from augmented images and original texts. Through the CLIP model, we obtain two different views of multimodal features, E_i and $E_i^{'}$. Then, we map these multimodal features to K-dimensional vectors as outputs using a function f, where $K = |Y_L \cup Y_U|$ is the total number of event categories. For labeled data, we optimize using a cross-entropy function in Eq. 5.8. For unlabeled data, we employ self-distillation learning. Specifically, we first randomly initialize a set of prototypes $C = \{c_1, ..., c_K\}$, each representing one category. During training, we compute the cosine similarity between the output features and prototypes to obtain soft labels p_i/q_i for each view, which can be formulated as:

$$p_i^{(k)} = \frac{\exp\left(\frac{1}{\tau}(f(E_i)/\|f(E_i)\|_2)^T(\mathbf{c}_k/\|\mathbf{c}_k\|_2)\right)}{\sum_{k'}\exp\left(\frac{1}{\tau}(f(E_i)/\|f(E_i)\|_2)^T(\mathbf{c}_{k'}/\|\mathbf{c}_{k'}\|_2)\right)}, \tag{5.9}$$

where τ is a temperature parameter for p_i and a sharper version for another view q_i. The distillation loss can be formulated as:

$$L_{Distill} = -\frac{1}{M}\sum_{i=1}^{M} q_i \log p_i. \tag{5.10}$$

We also adopt a entropy maximization regularizer [9] for the unsupervised objective, which can be formulated as:

$$L_{ENT} = -\frac{1}{M}\sum_{i=1}^{M} p_i \log p_i, \tag{5.11}$$

However, we found that maximizing entropy, while encouraging the exploration of new categories, also decreases the model's confidence in known categories, ultimately sacrificing accuracy on known categories (see Sec. 5.4.6).

To address this issue, we propose an adaptive entropy optimization strategy, aiming for the model to actively explore new categories while maintaining accuracy on known categories. Specifically, we use pseudo-label consistency across four views to decide on entropy optimization. For a sample, on one hand, we generate two pseudo-labels using p_i and its augmented view q_i; on the other hand, we generate p_i^{Aug} and q_i^{Aug} as two additional views using the multimodal augmentation module mentioned earlier, which provides a more challenging perspective as the generated feature is more discriminative. We then use a consistency checker to perform consistency checks on the pseudo-labels from these four different views for entropy optimization, which can be formulated as:

$$L_{Adapt} = \begin{cases} \alpha L_{ENT} & \text{if } n = 4 \\ -\beta L_{ENT} & \text{if } n < 4, \end{cases} \tag{5.12}$$

where n represents the number of consistency for the pseudo-labels, α and β are hyperparameters.

Through this strategy, when the model's predictions are completely consistent across different views, we increase the model's confidence in its judgment by minimizing entropy; when there is a discrepancy in the model's judgments across views, we encourage further exploration by maximizing entropy as we want the model to explore new events as much as possible when there is uncertainty, rather than blindly gravitating towards known events.

5.3.5 Overall Formulation and Optimization

In this study, we optimize a minimax problem via a straightforward back-propagation way. To summarize the previous discussions, the overall objective function of DAEO can be formulated as follows:

Algorithm 3 DAEO Algorithm

Input: Labeled data: $D_L = \{(x_i, y_i)\}_{i=1}^{N}$, unlabeled data: $D_U = \{(x_j)\}_{j=1}^{M}$, the CLIP model, the multimodal augmentation model G and the MLP classifier f.

Output: Learned model parameters $\hat{\theta}_{CLIP}$, $\hat{\theta}_f$ and $\hat{\theta}_G$.

while $t \leq$ **MaxIter**

1: Compute the multimodal features E_i according to Eq. 5.1.
2: Compute the augmented multimodal features E_i^{Aug} according to Eq. 5.3.
3: Compute the pseudo-labels from four different views.
4: Perform a consistency check for these pseudo-labels and compute the adaptive entropy loss L_{Adapt} according to Eq. 5.12.
5: Compute the cross-entropy loss L_{CE} and L_{CE}^{Aug} and distill loss $L_{Distill}$ according to Eq. 5.8, Eq. 5.4 and Eq. 5.10, respectively.
6: Optimize the objective in Eq. 5.13.
7: Recompute the multimodal features E_i and the augmented multimodal features E_i^{Aug} according to Eq. 5.1 and Eq. 5.3, respectively.
8: Recompute the cross-entropy loss L_{CE}^{Aug}, KL divergence loss L_{KL} and consistency loss L_{Consis} according to Eq. 5.4, Eq. 5.2 and Eq. 5.5, respectively.
9: Optimize the objective in Eq. 5.14.

end while

$$(\hat{\theta}_{CLIP}, \hat{\theta}_f) = \arg \min_{\theta_{CLIP}, \theta_f} L_{CE}^{Aug} + L_{CE} + L_{Distill} + L_{Adapt}, \tag{5.13}$$

$$(\hat{\theta}_G) = \arg \max_{\theta_G} L_{CE}^{Aug} - L_{Consis} - L_{KL}. \tag{5.14}$$

The detailed algorithm for the DAEO method is presented in Algorithm 3. The computational complexity of DAEO consists of: 1) multimodal feature extraction via CLIP with $O(n)$ complexity; 2) feature augmentation through the generator with $O(d)$ complexity where d is the feature dimension; 3) consistency checking across views with $O(nk)$ complexity where k is the number of views. The total computational complexity per iteration is $O(n + d + nk)$.

5.4 Experiment

In this section, we present extensive experiments to evaluate our DAEO model for generalized social event detection. We first introduce the MSED dataset, which contains various types of events across different time periods. Then, we conduct comprehensive experiments: 1) comparison with state-of-the-art methods on both known and new event detection; 2) ablation studies to validate the effectiveness of multimodal augmentation and adaptive entropy optimization; 3) parameter analysis to demonstrate model robustness to different hyperparameter settings; 4) visualization analysis to examine the learned feature representations; and 5) case studies to analyze model performance on different event types. Additionally, we validate our method's generalization capability on the public CrisisMMD dataset.

5.4.1 Multimodal Social Event Detection (MSED) Dataset

In this section, we first present the collection and statistics of the MSED dataset. Then, we detail the implementation details, baselines, and extensive experimental

analysis. Finally, we also analyze the model's performance on a public dataset.

Collection of Social Events

The collection of social events plays a crucial role in generalized social event detection, demanding a diverse array of relationships among different social events to encompass various possibilities. For instance, when using time as a divider to separate the training and test sets, the relationship between events can be identical, subset, intersecting, or entirely distinct, depending on the type and time of the event. In this chapter, we define short-term events, cyclical events, and long-term events, which are designed to cover various event relationships and align with real-world scenarios. Specifically, we collect a variety of events ranging from 2011 to 2023 for each type of event from a crowd-sourced platform Wikipedia, e.g., short-term events include natural, human-made disasters, etc.; cyclical events encompass sports competitions, political elections, etc.; long-term events involve political conflicts, social movements, etc. Ultimately, our collection comprises 66 social events, including 42 short-term events, 13 cyclical events and 11 long-term events.

Collection and Statistics of the Dataset

For data collection and statistics, we select Twitter as our primary source due to its extensive user base. We employ event-related hashtags and temporal searches to avoid oversimplification of the task. For long-term events, we sample important sub-events based on Wikipedia, e.g., 'Syrian Civil War' containing 'Ghouta Chemical Attack', 'US Troops Withdrawing from Northern Syria', and so on. Subsequently, we filter out single-modality data samples and manually verified the semantic relevance of the samples for the corresponding event, resulting in a multimodal social event detection (MSED) dataset of 161,350 samples. Data statistics and sample distribution are shown in Table 5.1, Figure 5.4 and Figure 5.5.

Table 5.1: Statistic of the MSED dataset.

#Events	#Text	#Image	Average Words	#Language
66	161,350	196,543	17.645	63

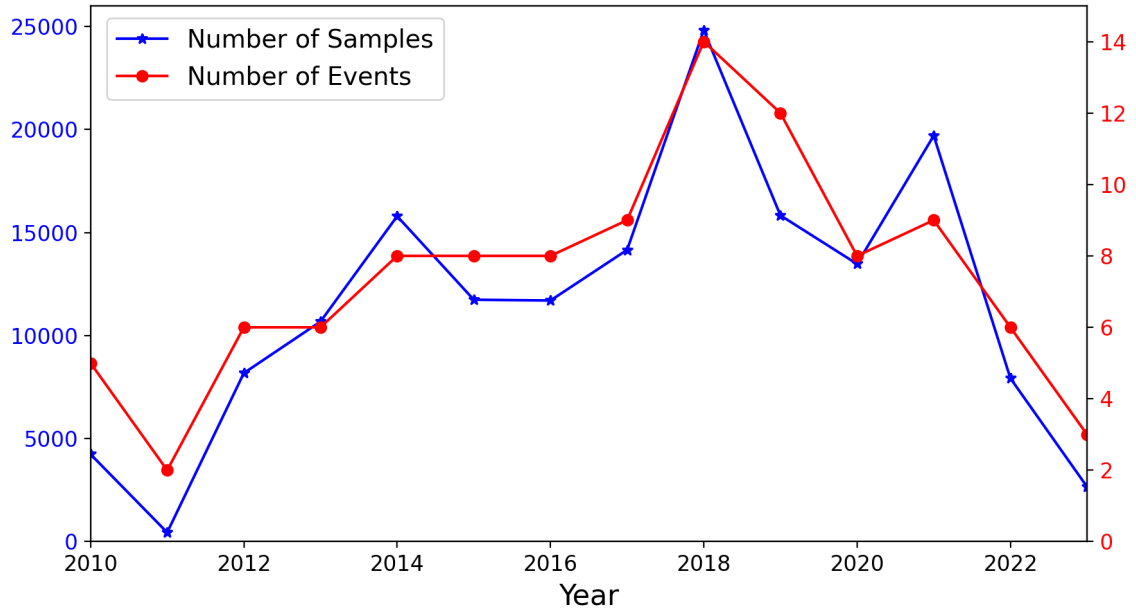


Figure 5.4: Distribution of the MSED dataset over time.

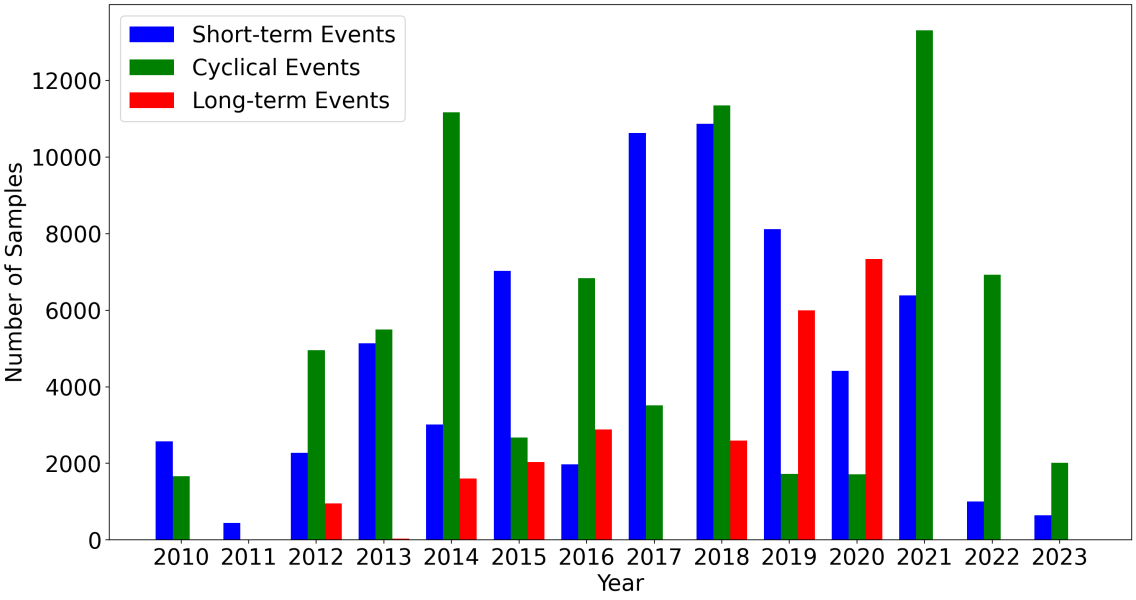


Figure 5.5: Distribution of the MSED dataset for different types of events over time.

Comparisons with Existing Datasets

Table 5.2 compares our MSED dataset with other existing datasets. From this comparison, we have the following observations:

- Most current datasets do not include temporal metadata, which is a crucial attribute for social event detection tasks. This omission is typically because these datasets are designed for the closed-setting event detection, where training and test sets are randomly split based on event categories rather than by time. In our work, we advocate for splitting training and test sets based on chronological order, which more accurately reflects the temporal nature of real-world social event detection tasks.

- Most existing datasets have relatively few samples, which is not conducive to learning social event features effectively. Events, representing complex semantic entities, require a substantial number of samples to capture their nuances fully. For the SED-14 dataset, it includes a large volume of data but is limited to coarse-grained event labels such as parties and festivals. In contrast, our dataset not only provides a large number of more fine-grained event categories but also categorizes these events into short-term, cyclical, and long-term events. This categorization is beneficial for models to learn distinctive features associated with different types of events.

- The majority of existing datasets predominantly consist of English posts. This is because the collection process intentionally filters out other languages to simplify the analysis. However, social event detection tasks inherently involve events from diverse countries, implying that multiple languages are common and that the local language of the event can offer a more authentic perspective for interpreting the event. Thus, our dataset retains posts in various languages, which, while increasing the complexity of the task, also provides multiple viewpoints that aid the model in understanding the event more comprehensively.

Table 5.2: Comparison of existing datasets. ("#" represents the number of samples.)

Dataset	Platform	#Sample	#Event	Modality	Fine-grained	Temporal Metadata	Multilanguage	Public
CE [80]	Twitter	800	2	Single	no	yes	no	no
SED-14 [87]	Flickr, Youtube	427,370/1,327	21,169	Multiple	no	yes	no	yes
ASO [95]	Twitter	1,100	3	Single	no	no	no	no
OSMNs [30]	Twitter	3.5M	20	Single	no	yes	no	no
Twevent [53]	Wikipedia,Twitter	3.2M/4.3M	N.A.	Single	no	no	no	no
DHS [46]	Twitter, Tumblr	2.1M/0.3M	600	Multiple	no	no	no	no
PHEME [122]	Twitter	2,089	9	Multiple	yes	no	no	yes
NED [57]	Twitter	17,366	40	Multiple	yes	no	no	yes
CrisisMMD [5]	Twitter	18,126	7	Multiple	yes	no	no	yes
MSED	Twitter	161,350	66	Multiple	yes	yes	yes	yes

Table 5.3: The division of the MSED dataset in the experiments. '#New' refers to the number of new events.

Proportion	Training set		Test set		
	#Sample	#Event	#Sample	#Event	#New
25%	32,270	27	121,013	55	39
50%	64,540	43	80,675	42	23
75%	96,810	56	40,338	23	10

Data Partitioning

Different from other classification tasks, the generalized social event detection task inherently involves a temporal dimension. Therefore, we organize all posts chronologically and then split the dataset into training and test sets based on sequential proportions. For this purpose, we divide the dataset into training and test sets by selecting three different time points—corresponding to 25%, 50%, and 75% of the timeline of the collected data—to determine the chronological length of the training set relative to the entire dataset. As for the validation set, we allocate 20% of the training set, chosen randomly across categories. The specifics of this division are summarized in Table 5.3.

5.4.2 Evaluation Metric

To evaluate our model's performance, we employ a clustering accuracy (ACC) followed by [99]. This metric is calculated as follows:

$$ACC = \max_{p \in P(Y_U)} \frac{1}{M} \sum_{i=1}^{M} 1\{y_i = p(\hat{y}_i)\}, \tag{5.15}$$

where P represents the set of all possible permutations that align the model's predicted labels $\hat{y}_i$ with the actual ground truth labels y_i, which utilizes the Hungarian method [50] for optimal matching. We apply this metric across three sets: the complete unlabeled set denoted as "All", a subset called "Known" which contains samples from classes already known to the model, and "New", comprising samples from classes not previously seen by the model.

5.4.3 Implementation Details

We utilize the CLIP ViT-B/16 backbone to train all methods, with fine-tuning the final block and linear projection layer of the text and visual encoders. The SGD optimizer [7] is employed with an initial learning rate of 0.001 and then decayed following a cosine schedule. The models are trained over 100 epochs with a batch size of 128. In alignment with [105], the temperature value for distillation learning is set to 0.1 and the sharper version starts at 0.07, then is gradually warmed up to 0.04 using a cosine schedule in the starting 10 epochs. The hyperparameters α and β are set to 0.03 and 2.3, respectively. For non-English text, we utilize the Google Translation API[1] to translate the content into English. For posts containing multiple images, we only use the first image. The process of tuning and testing is carried out on a separate validation set, facilitating the selection of the best hyperparameters for optimal performance. All experiments are conducted on an NVIDIA GeForce RTX A6000.

[1] `https://cloud.google.com/translate`

5.4.4 Baselines

To investigate the effectiveness of our proposed method, we compare it with three different baseline approaches to provide a comprehensive evaluation.

- **K-means method** [66]. This baseline extracts the image and text features from the pretrained CLIP model and concatenates them to form the multimodal features, followed by the K-means clustering algorithm. Previous event detection methods mostly use such unsupervised clustering approach for detecting new events.

- **Novel category discovery baselines** (i.e., UNO [24] and RankStats [32]). These are strong baselines from the field of novel category discovery. Following the setup in [99], we configure one classification head to the total number of classes in order to adapt these models to fit the task.

- **The state-of-the-art methods in generalized category discovery** (i.e., GCD [99] and SimGCD [105]). GCD utilizes semi-supervised K-means clustering based on learned features, and SimGCD employs a parametric classifier for distillation learning, which has demonstrated impressive results across various image recognition tasks.

5.4.5 Comparison with the State of the Arts

Table 5.4 shows the experimental results of our proposed DAEO method and other comparison methods on the generalized social event detection task. From these results, we observe that:

- DAEO outperforms all baselines across most scenarios with different dataset proportions, validating the effectiveness of our model for generalized social event

Table 5.4: Results on the MSED dataset.

Method	25%			50%			75%		
	Known	New	All	Known	New	All	Known	New	All
K-means [66]	0.400	**0.433**	0.419	0.370	0.446	0.404	0.286	0.320	0.297
RankStats [32]	0.275	0.050	0.147	0.441	0.165	0.320	0.461	0.446	0.457
UNO [24]	0.736	0.176	0.419	0.792	0.277	0.567	0.735	0.450	0.647
GCD [99]	0.464	0.333	0.390	0.463	0.435	0.451	0.450	0.441	0.447
SimGCD [105]	0.706	0.316	0.485	0.535	0.596	0.562	0.705	0.462	0.630
DAEO	**0.757**	0.409	**0.560**	**0.823**	**0.622**	**0.735**	**0.735**	**0.629**	**0.702**
Δ	+0.052	+0.093	+0.075	+0.287	+0.024	+0.172	+0.029	+0.167	+0.072

detection. It's noteworthy that when the training proportion is 25%, our model's performance does not surpass that of the K-means baseline. This is because the relatively small amount of training data, which hinders the model's ability to learn robust event features effectively. In fact, the likelihood of encountering such a limited amount of data is lower in real-world scenarios, especially with the continuous generation of social events.

- Employing K-means clustering directly on features extracted by the pretrained CLIP model yields impressive results, underscoring the significance of utilizing a robust pretrained model.

- Parametric learning methods (i.e., SimGCD) outperform non-parametric clustering approaches (i.e., GCD). This is attributed to the joint training of the entire model, which avoids potentially being sub-optimal.

- In the scenario of generalized social event detection, compared to the traditional unsupervised new event detection setting (performance of the K-means method under the "All" setting), our proposed DAEO model not only achieves high accuracy on known events but also performs well on new events, which demonstrates the significance of our proposed setting.

5.4.6 Ablation Study

The proposed DAEO model contains two key modules: multimodal augmentation and adaptive entropy optimization. To validate their effectiveness, we conducted ablation studies on these components. We denote Multimodal Augmentation as 'MA', the entropy minimization and maximization terms in L_{Adapt} as 'Entmin' and 'Entmax', respectively, and 'Adapt' to represent L_{Adapt}, with 'Ctr' indicating self-contrastive learning and supervised contrastive learning. From the results in Table 5.5, we have the following observations:

Table 5.5: Ablation study on the different components of our approach with the proportion of 50%.

#	MA	Entmin	Entmax	Adapt	Ctr	Known	New	All
1	×	✓	✓	✓	×	0.821	0.578	0.715
2	✓	×	✓	✓	×	0.807	0.631	0.730
3	✓	✓	×	✓	×	0.742	0.220	0.513
4	✓	×	✓	×	×	0.568	**0.659**	0.608
5	✓	×	×	×	×	0.793	0.267	0.562
6	✓	✓	✓	✓	✓	0.804	0.620	0.723
7	✓	✓	✓	✓	×	**0.823**	0.622	**0.735**

- The absence of multimodal augmentation (#1 and #7) leads to a decrease in accuracy, which underscores the contribution of the multimodal augmentation module in learning more robust features.

- By comparing #4 and #5, we note that the entropy maximization term boosts the model's performance on recognizing new events but adversely affects its ability to identify known events. The inclusion of L_{Adapt} and Entmin (#4 and #7) not only retains the model's capacity to recognize new events but also improves its accuracy on known events. This demonstrates the efficacy of the adaptive entropy optimization strategy in balancing the model's performance across known and new events.

- The addition of self-contrastive learning and supervised contrastive learning (#6 and #7) does not enhance our model's performance, which could be attributed to negative optimization introduced by random augmentations.

In addition, we also investigate the effect of the model's backbone, the handling of multilingual text in the dataset, the conditions for L_{Adapt} and the performance of our model under different event types.

Table 5.6: Ablation experiment for Backbone with the proportion of 50%.

Backbone	Known	New	All
ViT-B/32	0.803	0.570	0.701
ViT-B/16	0.823	**0.622**	0.735
ViT-L/14	**0.841**	0.611	**0.740**

Table 5.7: Ablation experiment for multilingual processing with the proportion of 50%.

Method	Known	New	All
wo Translation	0.819	0.587	0.717
Translation	**0.823**	**0.622**	**0.735**
M-CLIP [14]	0.821	0.621	0.733

- As shown in Table 5.6, employing larger pretrained models as the backbone improves performance, underscoring the importance of a robust pretrained model.
- Regarding multilingual text processing, we experiment with using the M-CLIP model [14], which is pretrained on multiple languages. According to the results in Table 5.7, utilizing such model does not outperform a straightforward approach of employing the Google Translate API for language translation, thus we select the translation method.
- For the condition of L_{Adapt}, loosening the criteria (i.e., using a consistency threshold across different views to determine entropy minimization/maximization) leads to reduced recognition rates for new events, as shown in Table 5.8. Therefore, we select the condition of consistency across all views.
- As shown in Figure 5.9, the model has a high recognition rate for cyclical events, due to their high degree of similarity and predictable recurrence. However, for long-term events, the ongoing evolution of the events makes the recognition of even known events as challenging as that of short-term events.

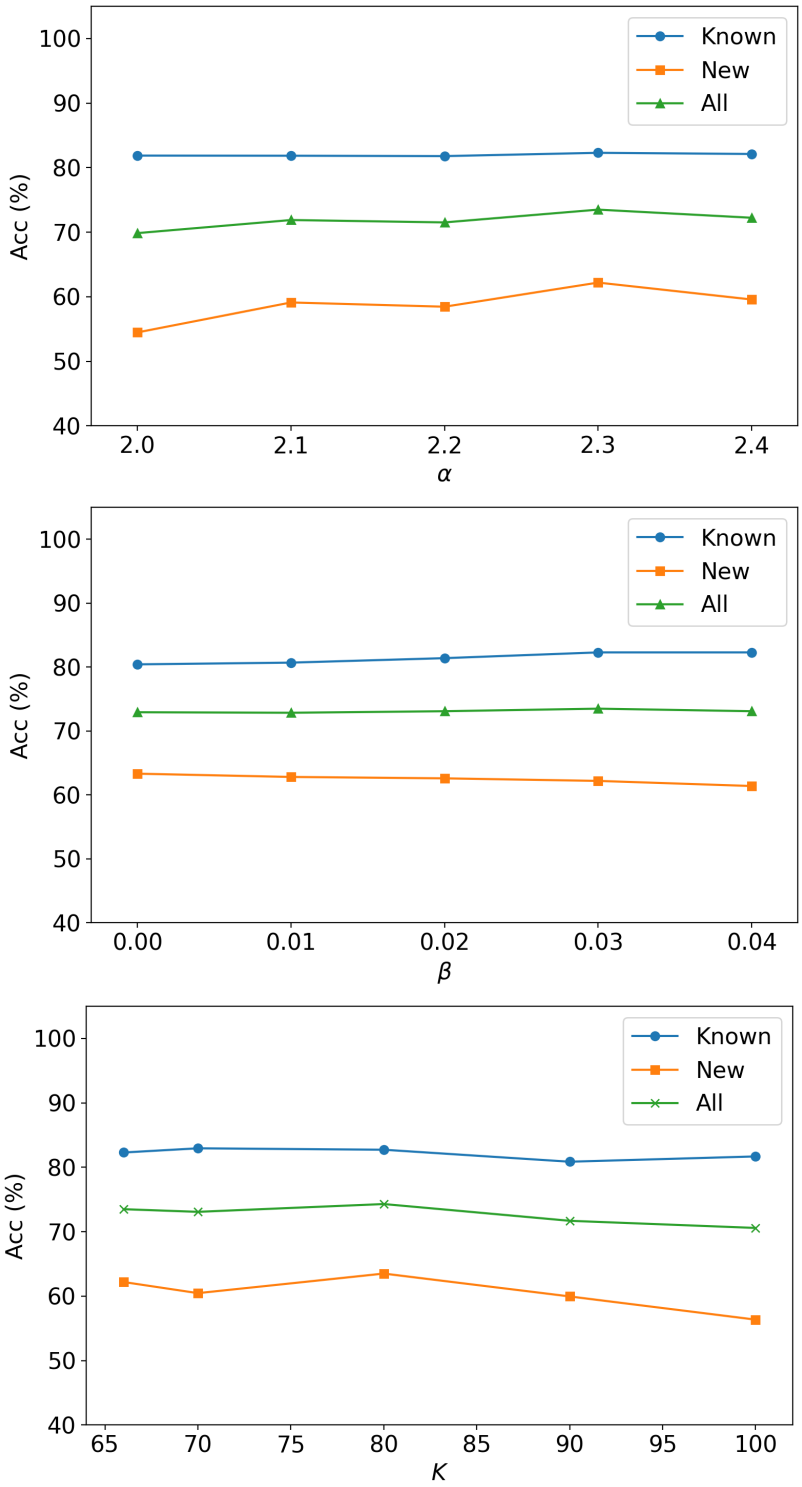


Figure 5.6: Parameter sensitivity on the MSED dataset with the proportion of 50%.

Table 5.8: Ablation experiment for condition of L_{Adapt} with the proportion of 50%.

Condition	Known	New	All
$n \geq 2$	**0.826**	0.465	0.668
$n \geq 3$	0.825	0.552	0.706
$n = 4$	0.823	**0.622**	**0.735**

Table 5.9: Results of our approach for different event types with the proportion of 50%.

Type	Known	New	All
Short-term Event	0.891	0.642	0.656
Cyclical Event	0.853	-	0.853
Long-term Event	0.662	0.527	0.600

5.4.7 Parameter Analysis

To delve into the impact of parameters α and β on the model's performance, we conducted experiments varying α from 2.0 to 2.4 and β from 0 to 0.04. As depicted in Figure 5.6, we observe that an increase in α tends to enhance the accuracy for new events at the expense of slightly reducing accuracy for known events, while β exhibits an inverse relationship. Overall, the model demonstrates moderate sensitivity to these parameters, leading to the selection of $\alpha = 2.3$ and $\beta = 0.01$ as the optimal settings.

Regarding the parameter K, which denotes the number of prototypes, we assume that the number of events is known following [105] in our model. We investigate its effect on our model using different values. As shown in Figure 5.6, although the performance on new events slightly decreases with increasing K values, the fluctuation remains minimal, which shows our model's robustness to variations in the number of prototypes.

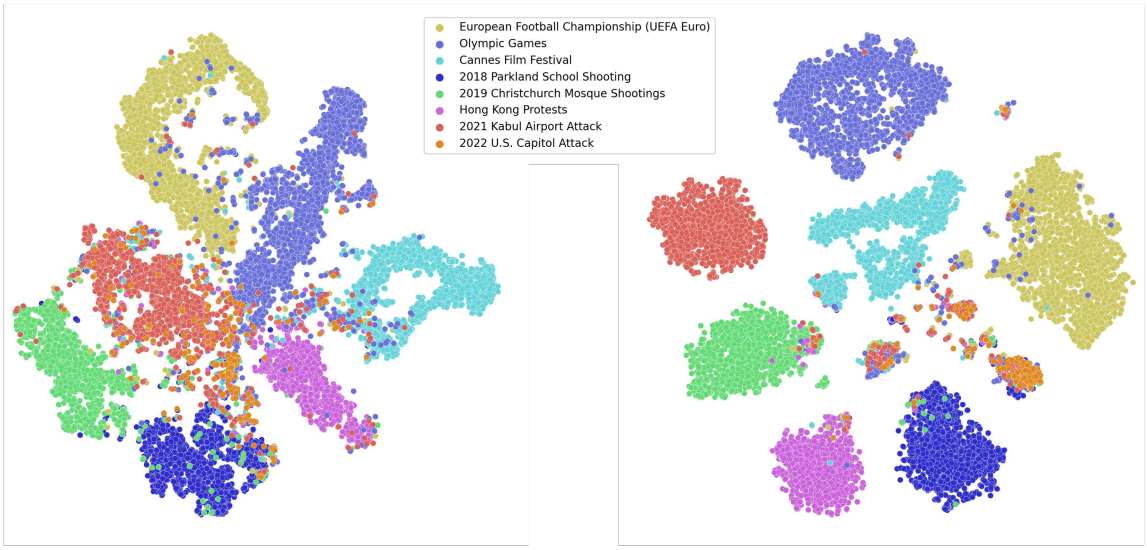


Figure 5.7: TSNE visualization of multimodal features from selected social events with the proportion of 50%.

5.4.8 Data Visualization

To further investigate the effectiveness of our proposed method, we employ t-SNE [93] visualization to illustrate the multimodal event features learned by the model. We select eight similar events, including both known and new events, for visualization. As shown in Figure 5.7, we have the following observations: 1) Compared to SimGCD, the features from our proposed method have clearer boundaries between different events, which proves the effectiveness of our approach. 2) For similar events, such as attack events, our model demonstrates a strong capability to differentiate between them, which is attributed to our multimodal augmentation module that utilizes adversarial learning to generate discriminative features.

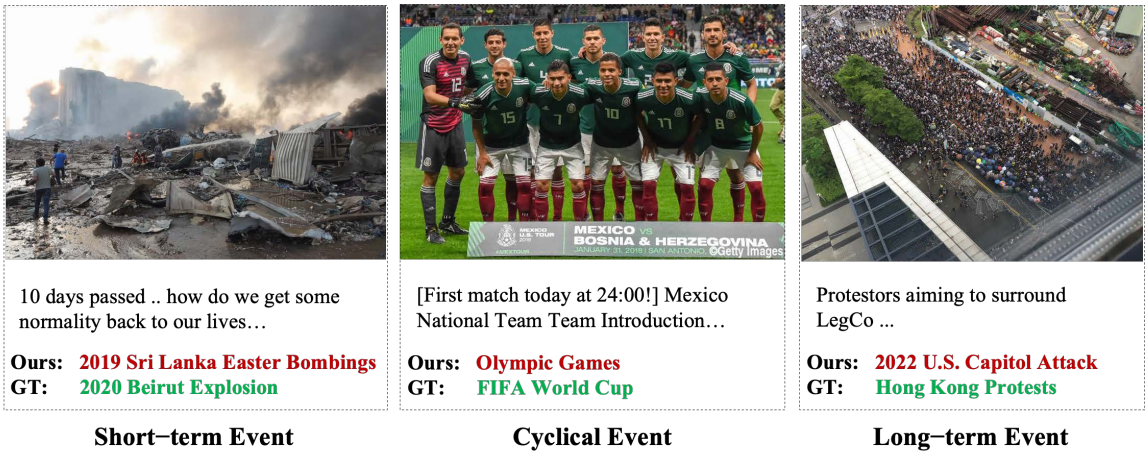


Figure 5.8: Failure examples of DAEO on the MSED dataset with the proportion of 50%.

5.4.9 Case Study

Despite the excellent performance of DAEO, Figure 5.8 shows three failure cases from different event types. We observe that their misclassification mainly stems from the lack of distinctive elements in the provided images and texts. Specifically, the first case comes from the '2020 Beirut Explosion'. Due to the absence of key information about the Beirut explosion, the event is mistakenly classified as a general explosion event. The second case, 'FIFA World Cup', included a team photo, which is common in other sports events, such as the 'Olympic Games'. The third case, 'Hong Kong Protests', featured many protesting people, leading the model to mistakenly categorize it as an attack event. These failure cases illustrate the complexity and challenges of the generalized social event detection.

Table 5.10: The division of the CrisisMMD dataset. '#New' refers to the number of new events.

Training set		Test set		
#Sample	#Event	#Sample	#Event	#New
6,047	3	10,567	7	4

5.4.10 Experiments on the Public Dataset

To validate the generalization capability of our proposed Dynamic Augmentation and Entropy Optimization (DAEO) model, we conduct experiments on a public dataset, i.e., the CrisisMMD dataset [5].

CrisisMMD Dataset

This dataset is a multimodal crisis dataset that encompasses seven natural disaster events from 2017, including Hurricane Irma, Hurricane Maria, Hurricane Harvey, the Mexico earthquake, the Iraq–Iran earthquakes, the Sri Lanka floods, and the California wildfires. Detailed statistics of the dataset are shown in Table 5.2.

Data Partitioning

Given the analysis in Section 5.4.1, the CrisisMMD dataset lacks temporal information, which is used for closed-setting event detection [56]. Therefore, we are not able to divide the training and test sets according to time for the generalized social event detection task. Following [99], we extract the first three categories of events as known events from the dataset. We set 50% of the data from these categories for the training set. The remaining 50% of data from these categories, along with all event samples from the other four categories, constitute the test set. For the validation set, we select 20% of the training set, chosen randomly across categories. The specific partitioning details are depicted in Table 5.10.

Table 5.11: Results on the CrisisMMD dataset.

Method	Known	New	All
K-means [66]	0.315	0.526	0.375
RankStats [32]	0.854	0.425	0.732
UNO [24]	0.946	0.531	0.828
GCD [99]	0.391	0.462	0.363
SimGCD [105]	0.962	0.610	0.862
DAEO	**0.968**	**0.669**	**0.883**
Δ	+0.006	+0.060	+0.021

Performance on the CrisisMMD Dataset

Table 5.11 shows the experimental results of our DAEO model on the CrisisMMD dataset. From these results, we observe the following:

- Our model achieves the best performance on this public dataset compared to other methods, which validates its strong generalization ability.

- Our model exhibits high accuracy on known classes on the CrisisMMD dataset. This is partly due to the use of random partitioning to define known and unknown events, which simplifies the task to some extent. This result also underscores the importance of partitioning training and test sets based on time to prevent potential future information leakage, which is crucial for realistic event detection tasks.

- Our model also performs well on new categories, indicating that the features generated by the proposed multimodal augmentation module are robust even for new events.

Table 5.12: Experimental results on the CrisisMMD dataset with closed setting (without new events).

Measure	f-CLSWGAN [107]	TCGAN [73]	CADA-VAE [88]	DAVAE [45]	MDL-DR [71]	Multi-RC [110]	SCBD [5]	AT-CVAE [56]	OWSEC [83]	DAEO
Accuracy	0.7582	0.8954	0.7412	0.7977	0.8677	0.8395	0.9366	0.9718	0.9672	**0.9722**
Macro F1	0.7578	0.8936	0.7406	0.7873	0.8573	0.8223	0.9510	0.9709	0.9709	**0.9758**

Performance on the CrisisMMD Dataset under Closed Setting

To validate the effectiveness of our proposed multimodal augmentation module in generating robust features, we also compare its performance in a closed-setting event detection scenario. Following [56], we divide 70% of the CrisisMMD dataset as the training set, 10% as the validation set, and 20% as the test set. The evaluation metrics used are accuracy and macro-averaged F1 score.

As shown in Table 5.12, we have the following observations:

- Our DAEO model outperforms other event detection methods, which can be attributed to our multimodal augmentation module that employs adversarial techniques. The adversarial approach in feature generation effectively enhances the variability and representational capacity of the features, which in turn improves the classifier's ability to discriminate between different event types accurately.

- Combined with Table 5.11, we observe that the accuracy for known events in our model under the generalized setting remains very close to the accuracy under a closed setting, even after the addition of new events. This is attributed to our adaptive entropy optimization strategy, which selectively optimizes for both known and new events. By maintaining accuracy for known events while encouraging exploration of new events, this strategy ensures that the model remains effective across all categories without compromising its ability to identify events it has previously learned.

5.5 Conclusion

In this chapter, we introduce a Dynamic Augmentation and Entropy Optimization (DAEO) model designed specifically to tackle the challenges of generalized social event detection. A multimodal augmentation module is designed to employ adversarial learning to generate distinctive multimodal features, which improves the model's ability to discern between similar event categories. An adaptive entropy optimization strategy with a self-distillation method leverages pseudo-labels from different views to adaptively optimize entropy, thereby enhancing the model's ability in recognizing both new and known events. Additionally, we contribute to the field by introducing the Multimodal Social Event Detection (MSED) dataset, which contains various event types and serves as a valuable resource for researchers. Extensive experiments conducted on the MSED dataset validate the effectiveness of our proposed model and demonstrate the superiority of our proposed generalized social event detection setting.

Chapter 6

Conclusion and Future Work

In this chapter, we first summarize the key contributions of this thesis and then outline potential directions for future research.

6.1 Conclusion

In this thesis, we focus on learning robust features to enhance the accuracy and generalizability of multimodal social event detection models.

First, we propose a deep learning algorithm, MFEK, which addresses the out-of-distribution (OOD) and multimodal fusion issues in event detection by incorporating external knowledge and attention mechanisms. To achieve this, we use a knowledge extraction module to extract event-related explicit and implicit knowledge from existing knowledge bases and large language models. Then, we integrate the extracted knowledge into multimodal data using attention mechanisms to improve the accuracy of the social event detection task. We find that incorporating external knowledge significantly improves the performance of the proposed MFEK model compared to other state-of-the-art methods, allowing it to accurately identify events even in scenarios with information fragmentation.

Second, we explore cross-platform social event detection to enhance the generalizability of the learned features. To this end, we propose a novel transfer learning model, SSMC, which addresses the issues of missing modalities and heterogeneous distributions in cross-platform scenarios. We first design a missing data complementation module to learn modality-shared features that supplement the missing modality information. Next, we introduce a multimodal self-learning module that adapts the model to target platform data by generating reliable pseudo labels, thereby reducing the distribution gap between different platforms. Our studies indicate that the proposed SSMC outperforms other existing state-of-the-art methods. Additionally, we verify the effectiveness of cross-platform event detection in improving the quality of single-platform event data.

Furthermore, we introduce a new task, generalized social event detection, to explore new event detection and enhance the generalizability of the learned features. To address this task, we propose a new multimodal deep learning algorithm, DAEO. This algorithm leverages adversarial learning to learn discriminative multimodal features from known event data. Additionally, it employs an adaptive entropy optimization strategy combined with a self-distillation method, which allows the model to cluster and detect unknown events while maintaining high accuracy for known events. Our findings show that the proposed method not only achieves high accuracy in detecting known events but also surpasses traditional unsupervised methods in new event detection. This success is attributed to the more robust multimodal features learned from known event data, which facilitate the identification and clustering of new events.

6.2 Future Work

In future research, we plan to improve and extend existing data representation learning models to further enhance the accuracy and generalizability of multimodal social event detection models. There are many promising directions for future research,

which we summarize as follows:

- **Increasing the diversity of multimodal data**: Social media data encompasses various types of media, including social links, geographic information, and more. This thesis mainly focuses on detecting social events using multimodal content such as images and text. Utilizing additional attributes of social media (e.g., tags, spatial, and temporal information) can further enhance the diversity of multimodal data. Therefore, the next step should consider how to leverage more available data to improve model detection accuracy.

- **Multisource and multimodal social event detection**: The cross-platform detection method in this thesis is currently limited to two platforms: source and target domains. In future work, we plan to extend this method to support multi-platform detection. Specifically, we will research multi-platform domain adaptation techniques to develop models capable of handling data from multiple platforms simultaneously, thereby further enhancing the comprehensiveness and robustness of event detection.

- **Real-time multimodal social event detection in open domains**: Detecting social events requires timely responses for rapid decision-making and emergency handling. This thesis introduces generalized social event detection, which aims to extend existing methods from detecting limited types of events to detecting all types of events without specific categories. However, the real-time nature of the algorithm leaves much room for improvement. Therefore, considering the real-time nature of the algorithm is crucial for emergency response and decision support, as it can significantly improve the efficiency and effectiveness of emergency handling.

- **Improving the quality of raw data**: The quality of raw data determines the accuracy of subsequent tasks such as detection and analysis. However, raw data

often has potential issues like incompleteness, sparsity, and imbalance. This thesis collect three datasets related to event detection, which also exhibited these issues. Although sampling algorithms were used to address the imbalance problem, how to collect and process more complete and balanced datasets remains a key focus of this research. Therefore, better addressing the issues of data imbalance and poor completeness remains one of the primary research focuses for the next stage.

- **Studying the impact of multilingualism**: Social event detection often requires data from various social media platforms, with users from around the world using different languages. Relying on data from a single language may result in missing critical event information. Additionally, many social events have clear regional and localized characteristics, with information typically published in the local language. Without studying multilingual event detection, important localized information may be missed. Although this thesis attempted to address multilingual issues by translating data into a single language, it heavily depends on the reliability of translation models. Therefore, in-depth research on multilingualism for social event detection has significant practical importance and value.

- **Enhancing the interpretability of social event detection models**: Current social event detection models are mostly black-box models that transform data information into representation vectors, lacking reasonable interpretability for the final analysis results. Although this thesis introduces external knowledge to provide some level of explanation, the models remain black-box. Therefore, future research should focus on improving the interpretability of social event analysis models.

References

[1] Mahdi Abavisani, Liwei Wu, Shengli Hu, Joel Tetreault, and Alejandro Jaimes. Multimodal categorization of crisis events in social media. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14679–14689, 2020.

[2] Imad Afyouni, Zaher Al Aghbari, and Reshma Abdul Razack. Multi-feature, multi-modal, and multi-source social event detection: A comprehensive survey. *Information Fusion*, 79:279–308, 2022.

[3] Charu C Aggarwal and Karthik Subbian. Event detection in social streams. In *Proceedings of the 2012 SIAM international conference on data mining*, pages 624–635. SIAM, 2012.

[4] Luca Maria Aiello, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, Ayse Goker, Ioannis Kompatsiaris, and Alejandro Jaimes. Sensing trending topics in twitter. *IEEE Transactions on multimedia*, 15(6):1268–1282, 2013.

[5] 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.

[6] James Allan, Ron Papka, and Victor Lavrenko. On-line new event detection and tracking. In *Proceedings of the 21st annual international ACM SIGIR*

conference on Research and development in information retrieval, pages 37–45, 1998.

[7] Shun-ichi Amari. Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(4-5):185–196, 1993.

[8] Chidubem Arachie, Manas Gaur, Sam Anzaroot, William Groves, Ke Zhang, and Alejandro Jaimes. Unsupervised detection of sub-events in large scale disasters. In *Proceedings Of The AAAI conference on artificial intelligence*, volume 34, pages 354–361, 2020.

[9] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *European Conference on Computer Vision*, pages 456–473. Springer, 2022.

[10] Bing-Kun Bao, Weiqing Min, Ke Lu, and Changsheng Xu. Social event detection with robust high-order co-clustering. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 135–142, 2013.

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[12] Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji. Deep adversarial learning for multi-modality missing data completion. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1158–1166, 2018.

[13] Yuwei Cao, Hao Peng, Zhengtao Yu, and S Yu Philip. Hierarchical and incremental structural entropy minimization for unsupervised social event detection.

In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8255–8264, 2024.

[14] Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. Cross-lingual and multilingual clip. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6848–6854, 2022.

[15] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[16] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, et al. Symbolic discovery of optimization algorithms. *Advances in Neural Information Processing Systems*, 36, 2024.

[17] Yongyong Chen, Xiaolin Xiao, and Yicong Zhou. Jointly learning kernel representation tensor and affinity matrix for multi-view clustering. *IEEE Transactions on Multimedia*, 22(8):1985–1997, 2019.

[18] Jaeyoung Choi, Eungchan Kim, Martha Larson, Gerald Friedland, and Alan Hanjalic. Evento 360: Social event discovery from web-scale multimedia collection. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 193–196, 2015.

[19] Lingyang Chu, Yanyan Zhang, Guorong Li, Shuhui Wang, Weigang Zhang, and Qingming Huang. Effective multimodality fusion framework for cross-media topic detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(3):556–569, 2014.

[20] Wanqiu Cui, Junping Du, Dawei Wang, Feifei Kou, and Zhe Xue. Mvgan: Multi-view graph attention network for social event detection. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(3):1–24, 2021.

[21] Feifei Ding, Jianjun Li, Wanyong Tian, Shanqing Zhang, and Wenqiang Yuan. Unsupervised domain adaptation via risk-consistent estimators. *IEEE Transactions on Multimedia*, 26:1179–1187, 2024.

[22] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48, 2021.

[23] Paolo Ferragina and Ugo Scaiella. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1625–1628, 2010.

[24] Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9284–9292, 2021.

[25] Claudiu S Firan, Mihai Georgescu, Wolfgang Nejdl, and Raluca Paiu. Bringing order to your photos: event-driven classification of flickr images based on social knowledge. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 189–198, 2010.

[26] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[27] Wang Gao, Yuan Fang, Lin Li, and Xiaohui Tao. Event detection in social media via graph neural network. In *Web Information Systems Engineering–WISE 2021: 22nd International Conference on Web Information Systems Engineering, WISE 2021, Melbourne, VIC, Australia, October 26–29, 2021, Proceedings, Part I 22*, pages 370–384. Springer, 2021.

[28] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 597–613. Springer, 2016.

[29] Poonam Goyal, Prerna Kaushik, Pranjal Gupta, Dev Vashisth, Shavak Agarwal, and Navneet Goyal. Multilevel event detection, storyline generation, and summarization for tweet streams. *IEEE Transactions on Computational Social Systems*, 7(1):8–23, 2019.

[30] Hansu Gu, Xing Xie, Qin Lv, Yaoping Ruan, and Li Shang. Etree: Effective and efficient event modeling for real-time online social media networks. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 300–307. IEEE, 2011.

[31] Cong Guo and Xinmei Tian. Event recognition in personal photo collections using hierarchical model and multiple features. In *2015 IEEE 17th International Workshop On Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2015.

[32] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Autonovel: Automatically discovering and learning novel visual categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6767–6781, 2021.

[33] Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9192, 2021.

[34] Mahmud Hasan, Mehmet A Orgun, and Rolf Schwitter. Real-time event detection from the twitter data stream using the twitternews+ framework. *Information Processing & Management*, 56(3):1146–1165, 2019.

[35] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131, 2020.

[36] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[37] John R Hershey and Peder A Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–317. IEEE, 2007.

[38] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018.

[39] Die Hu, Dan Feng, and Yulai Xie. Egc: A novel event-oriented graph clustering framework for social media text. *Information Processing & Management*, 59(6):103059, 2022.

[40] Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12065–12075, 2023.

[41] Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. Compare to the knowledge: Graph neural fake news detection with external knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 754–763, 2021.

[42] Linmei Hu, Shuqi Yu, Bin Wu, Chao Shao, and Xiaoli Li. A neural model for joint event detection and prediction. *Neurocomputing*, 407:376–384, 2020.

[43] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025, 2020.

[44] Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. Minimum class confusion for versatile domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 464–480. Springer, 2020.

[45] Mengmeng Jing, Jingjing Li, Lei Zhu, Ke Lu, Yang Yang, and Zi Huang. Incomplete cross-modal retrieval with dual-aligned variational autoencoders. In *Proceedings of the 28th ACM international conference on multimedia*, pages 3283–3291, 2020.

[46] Satya Katragadda, Ryan Benton, and Vijay Raghavan. Framework for real-time event detection using multiple social media sources. In *50th Hawaii International Conference on System Sciences, HICSS 2017, Hilton Waikoloa Village, Hawaii, USA, January 4-7, 2017*, pages 1–10. ScholarSpace / AIS Electronic Library (AISeL), 2017.

[47] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[48] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019.

[49] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

[50] Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 52(1):7–21, 2005.

[51] Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary. Twitter trending topic classification. In *2011 IEEE 11th international conference on data mining workshops*, pages 251–258. IEEE, 2011.

[52] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14943–14952, 2023.

[53] Chenliang Li, Aixin Sun, and Anwitaman Datta. Twevent: segment-based event detection from tweets. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 155–164, 2012.

[54] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[55] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[56] Zhangming Li, Shengsheng Qian, Jie Cao, Quan Fang, and Changsheng Xu. Adaptive transformer-based conditioned variational autoencoder for incomplete social event classification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1698–1707, 2022.

[57] Zehang Lin, Jiayuan Xie, and Qing Li. Multi-modal news event detection with external knowledge. *Information Processing & Management*, 61(3):103697, 2024.

[58] Zhihong Lin, Huidong Jin, Bella Robinson, and Xunguo Lin. Towards an accurate social media disaster event detection system based on deep learning and semantic representation. In *Proceedings of the 14th Australasian Data Mining Conference, Canberra, Australia*, pages 6–8, 2016.

[59] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2018.

[60] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

[61] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017.

[62] Daniel Lowd and Christopher Meek. Adversarial learning. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 641–647, 2005.

[63] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2302–2310, 2021.

[64] Yun Ma, Qing Li, Zhenguo Yang, Zheng Lu, Haiwei Pan, and Antoni B Chan. An svd-based multimodal clustering method for social event detection. In *2015 31st IEEE International Conference on Data Engineering Workshops*, pages 202–209. IEEE, 2015.

[65] Zhigang Ma, Yi Yang, Nicu Sebe, Kai Zheng, and Alexander G Hauptmann. Multimedia event detection using a classifier-specific intermediate representation. *IEEE Transactions on Multimedia*, 15(7):1628–1637, 2013.

[66] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[67] Vukosi Marivate and Pelonomi Moiloa. Catching crime: Detection of public safety incidents using social media. In *2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, pages 1–5. IEEE, 2016.

[68] Ida Mele, Seyed Ali Bahrainian, and Fabio Crestani. Event mining and timeliness analysis from heterogeneous news streams. *Information Processing & Management*, 56(3):969–993, 2019.

[69] Iraklis Moutidis and Hywel TP Williams. Good and bad events: combining network-based event detection with sentiment analysis. *Social Network Analysis and Mining*, 10(1):64, 2020.

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

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

[72] Rabah Ouldnoughi, Chia-Wen Kuo, and Zsolt Kira. Clip-gcd: Simple language guided generalized category discovery. *arXiv preprint arXiv:2305.10420*, 2023.

[73] Frederik Pahde, Mihai Puscas, Tassilo Klein, and Moin Nabi. Multimodal prototypical networks for few-shot learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2644–2653, 2021.

[74] Beverly Estephany Parilla-Ferrer, Proceso L Fernandez, and Jaime T Ballena. Automatic classification of disaster-related tweets. In *Proc. International Conference on innovative engineering technologies (ICIET)*, volume 62, 2014.

[75] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-adversarial domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[76] Liwen Peng, Songlei Jian, Zhigang Kan, Linbo Qiao, and Dongsheng Li. Not all fake news is semantically similar: Contextual semantic representation learning

for multimodal fake news detection. *Information Processing & Management*, 61(1):103564, 2024.

[77] Manuel Ignacio Pérez Carrasco et al. Adversarial variational domain adaptation for semi-supervised image classification. 2019.

[78] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6892–6899, 2019.

[79] Daniela Pohl, Abdelhamid Bouchachia, and Hermann Hellwagner. Automatic sub-event detection in emergency management using social media. In *Proceedings of the 21st international conference on world wide web*, pages 683–686, 2012.

[80] Ana-Maria Popescu and Marco Pennacchiotti. Detecting controversial events from twitter. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1873–1876, 2010.

[81] Beatrice Portelli, Edoardo Lenzi, Emmanuele Chersoni, Giuseppe Serra, and Enrico Santus. Bert prescriptions to avoid unwanted headaches: a comparison of transformer architectures for adverse drug event detection. In *Proceedings of the 16th conference of the European chapter of the Association for Computational Linguistics: main volume*, pages 1740–1747, 2021.

[82] Fan Qi, Xiaoshan Yang, and Changsheng Xu. A unified framework for multimodal domain adaptation. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 429–437, 2018.

[83] Shengsheng Qian, Hong Chen, Dizhan Xue, Quan Fang, and Changsheng Xu. Open-world social event classification. In *Proceedings of the ACM Web Conference 2023*, pages 1562–1571, 2023.

[84] Chen Qu, Liu Yang, Minghui Qiu, W Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. Bert with history answer embedding for conversational question answering. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 1133–1136, 2019.

[85] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[86] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

[87] Timo Reuter, Symeon Papadopoulos, Giorgos Petkos, Vasileios Mezaris, Yiannis Kompatsiaris, Philipp Cimiano, Christopher De Vries, and Shlomo Geva. Social event detection at mediaeval 2013: Challenges, datasets, and evaluation. In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop Barcelona, Spain, October 18-19, 2013*. Citeseer, 2013.

[88] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8247–8255, 2019.

[89] Mohammad Shirdel, Michele Segata, Giuseppe Di Fatta, and Antonio Liotta. Event detection in financial markets. In *2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, pages 1–8. IEEE, 2022.

[90] Irina Shklovski, Leysia Palen, and Jeannette Sutton. Finding community through information and communication technology in disaster response. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 127–136, 2008.

[91] Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation alignment for unsupervised domain adaptation. *Domain adaptation in computer vision applications*, pages 153–171, 2017.

[92] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 443–450. Springer, 2016.

[93] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification? In *Chinese computational linguistics: 18th China national conference, CCL 2019, Kunming, China, October 18–20, 2019, proceedings 18*, pages 194–206. Springer, 2019.

[94] Qiuling Suo, Weida Zhong, Fenglong Ma, Ye Yuan, Jing Gao, and Aidong Zhang. Metric learning on healthcare data with incomplete modalities. In *28th International Joint Conference on Artificial Intelligence, IJCAI 2019*, pages 3534–3540. International Joint Conferences on Artificial Intelligence, 2019.

[95] Samir Tartir and Ibrahim Abdul-Nabi. Semantic sentiment analysis in arabic social media. *Journal of King Saud University-Computer and Information Sciences*, 29(2):229–233, 2017.

[96] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference.*

Association for Computational Linguistics. Meeting, volume 2019, page 6558. NIH Public Access, 2019.

[97] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[98] Konstantinos N Vavliakis, Andreas L Symeonidis, and Pericles A Mitkas. Event identification in web social media through named entity recognition and topic modeling. *Data & Knowledge Engineering*, 88:1–24, 2013.

[99] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7492–7501, 2022.

[100] Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

[101] Enguang Wang, Zhimao Peng, Zhengyuan Xie, Xialei Liu, and Ming-Ming Cheng. Get: Unlocking the multi-modal potential of clip for generalized category discovery. *arXiv preprint arXiv:2403.09974*, 2024.

[102] Longzheng Wang, Chuang Zhang, Hongbo Xu, Yongxiu Xu, Xiaohan Xu, and Siqi Wang. Cross-modal contrastive learning for multimodal fake news detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5696–5704, 2023.

[103] Meng Wang, Hao Li, Dacheng Tao, Ke Lu, and Xindong Wu. Multimodal graph-based reranking for web image search. *IEEE transactions on image processing*, 21(11):4649–4661, 2012.

[104] Rui Wang, Zuxuan Wu, Zejia Weng, Jingjing Chen, Guo-Jun Qi, and Yu-Gang Jiang. Cross-domain contrastive learning for unsupervised domain adaptation. *IEEE Transactions on Multimedia*, 25:1665–1673, 2023.

[105] Xin Wen, Bingchen Zhao, and Xiaojuan Qi. Parametric classification for generalized category discovery: A baseline study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16590–16600, 2023.

[106] Xiao Wu, Chong-Wah Ngo, and Alexander G Hauptmann. Multimodal news story clustering with pairwise visual near-duplicate constraint. *IEEE Transactions on Multimedia*, 10(2):188–199, 2008.

[107] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5542–5551, 2018.

[108] Fangzheng Xu, Yu Bao, Bingye Li, Zhining Hou, and Lekang Wang. Entropy minimization and domain adversarial training guided by label distribution similarity for domain adaptation. *Multimedia Systems*, 29(4):2281–2292, 2023.

[109] Feng Xue, Richang Hong, Xiangnan He, Jianwei Wang, Shengsheng Qian, and Changsheng Xu. Knowledge-based topic model for multi-modal social event analysis. *IEEE Transactions on Multimedia*, 22(8):2098–2110, 2019.

[110] Li Xukun and Doina Caragea. Improving disaster-related tweet classification with a multimodal approach. In *ISCRAM 2020 conference proceedings–17th international conference on information Systems for Crisis Response and Management*, 2020.

[111] Yiming Yang, Jian Zhang, Jaime Carbonell, and Chun Jin. Topic-conditioned novelty detection. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 688–693, 2002.

[112] Zhenguo Yang, Qing Li, Haoran Xie, Qi Wang, and Wenyin Liu. Learning representation from multiple media domains for enhanced event discovery. *Pattern Recognition*, 110:107640, 2021.

[113] Zhenguo Yang, Zehang Lin, Lingni Guo, Qing Li, and Wenyin Liu. Mmed: a multi-domain and multi-modality event dataset. *Information Processing & Management*, 57(6):102315, 2020.

[114] Zhenguo Yang, Zhuopan Yang, Zhiwei Guo, Zehang Lin, Haizhong Zhu, Qing Li, and Wenyin Liu. Towards temporal event detection: A dataset, benchmarks and challenges. *IEEE Transactions on Multimedia*, pages 1–12, 2023.

[115] Tianhe Yu, Chelsea Finn, Sudeep Dasari, Annie Xie, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. One-shot imitation from observing humans via domain-adaptive meta-learning. *Robotics: Science and Systems XIV*, 2018.

[116] Maia Zaharieva, Matthias Zeppelzauer, and Christian Breiteneder. Automated social event detection in large photo collections. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 167–174, 2013.

[117] Jiandian Zeng, Jiantao Zhou, and Tianyi Liu. Robust multimodal sentiment analysis via tag encoding of uncertain missing modalities. *IEEE Transactions on Multimedia*, 25:6301–6314, 2023.

[118] Lei Zhang and Xuezhi Xiang. Video event classification based on two-stage neural network. *Multimedia Tools and Applications*, 79:21471–21486, 2020.

[119] Sicheng Zhao, Yue Gao, Guiguang Ding, and Tat-Seng Chua. Real-time multimedia social event detection in microblog. *IEEE transactions on cybernetics*, 48(11):3218–3231, 2017.

[120] Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10867–10875, 2021.

[121] Han Zhou, Hongpeng Yin, Hengyi Zheng, and Yanxia Li. A survey on multi-modal social event detection. *Knowledge-Based Systems*, 195:105695, 2020.

[122] Arkaitz Zubiaga, Maria Liakata, and Rob Procter. Exploiting context for rumour detection in social media. In *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part I 9*, pages 109–123. Springer, 2017.