UNIVERSIDADE
BEIRA INTERIOR

Event Detection and Tracking
Detection of Dangerous Events on Social Media

Versão final após defesa

Muhammad Luqman Jamil

Dissertação para obtenção do Grau de Mestre em
Engenharia Informática
(2º ciclo de estudos)

Orientador: Prof. Doutor Sebastião Augusto Rodrigues Figueiredo Pais
Coorientador: Prof. Doutor João Paulo da Costa Cordeiro

Agosto, 2022

Resumo Alargado

As plataformas online de redes sociais tornaram-se ferramentas essenciais para a comunicação, conexão com outros, e troca de informação nas nossas vidas. Este fenómeno tem sido intensamente estudado na última década para investigar os sentimentos dos utilizadores em diferentes cenários e para vários propósitos. Contudo, a utilização dos meios de comunicação social tornou-se mais complexa e num fenómeno mais vasto devido ao envolvimento de múltiplos intervenientes, tais como empresas, grupos e outras organizações. À medida que a tecnologia avançou e a popularidade aumentou, a utilização de termos diferentes referentes a tópicos semelhantes gerou confusão. Por outras palavras, os modelos são treinados segundo a informação de termos e âmbitos específicos. Portanto, a padronização é imperativa. O objetivo deste trabalho é unir os diferentes termos utilizados em termos mais abrangentes e padronizados. O perigo pode ser uma ameaça como violência social, desastres naturais, danos intelectuais ou comunitários, contágio, agitação social, perda económica, ou apenas a difusão de ideologias odiosas e violentas. Estudamos estes diferentes eventos e classificamos-los em tópicos para que a ténica de deteção baseada em tópicos possa ser concebida e integrada sob o termo Evento Perigosos (DE). Consequentemente, definimos o termo proposto "Eventos Perigosos" (Dangerous Events) e dividimo-lo em três categorias principais de modo a especificar as suas características. Sendo estes denominados Eventos Perigosos, Eventos Perigosos de nível superior, e Eventos Perigosos de nível inferior. O conjunto de dados MAVEN foi utilizado para a obtenção de conjuntos de dados para realizar a experiência. Estes conjuntos de dados são filtrados manualmente com base no tipo de eventos para separar eventos perigosos de eventos gerais. Os modelos de transformação BERT, RoBERTa, e XLNet foram utilizados para classificar dados de texto consoante a respetiva categoria de Eventos Perigosos. Os resultados demonstraram que o desempenho do BERT é superior a outros modelos e pode ser eficazmente utilizado para a tarefa de deteção de Eventos Perigosos. Salienta-se que a abordagem de divisão dos conjuntos de dados aumentou significativamente o desempenho dos modelos.

Existem diversos métodos propostos para a deteção de eventos. A deteção destes eventos (ED) são maioritariamente classificados na categoria de supervisonado e não supervisionados, como demonstrado nos metódos supervisionados, estão incluidos support vector machine (SVM), Conditional random field (CRF), Decision tree (DT), Naive Bayes (NB), entre outros. Enquanto a categoria de não supervisionados inclui Query-based, Statistical-based, Probabilistic-based, Clustering-based e Graph-based. Estas são as duas abordagens em uso na deteção de eventos e são denonimados de document-pivot and feature-pivot. A diferença entre estas abordagens é na sua maioria a clustering approach, a forma como os documentos são utilizados para caracterizar vetores, e a similaridade métrica utilizada para identificar se dois documentos correspondem ao mesmo evento ou não. Além da deteção de eventos, a previsão de eventos é um problema importante mas complicado que engloba diversas dimensões. Muitos destes eventos são difíceis de prever antes de se tornarem visíveis e ocorrerem. Como um exemplo, é impossível antecipar catástrofes naturais, sendo apenas detetáveis após o seu acontecimento. Existe um número limitado de recursos em ternos de conjuntos de dados de eventos. ACE 2005, MAVEN, EVIN são

alguns dos exemplos de conjuntos de dados disponíveis para a deteção de evnetos.
Os trabalhos recentes demonstraram que os Transformer-based pre-trained models (PTMs) são capazes de alcançar desempenho de última geração em várias tarefas de NLP. Estes modelos são pré-treinados em grandes quantidades de texto. Aprendem incorporações para as palavras da língua ou representações de vetores de modo a que as palavras que se relacionem se agrupen no espaço vectorial. Um total de três transformadores diferentes, nomeadamente BERT, RoBERTa, e XLNet, será utilizado para conduzir a experiência e tirar a conclusão através da comparação destes modelos.
Os modelos baseados em transformação (Transformer-based) estão em total sintonia utilizando uma divisão de 70,30 dos conjuntos de dados para fins de formação e teste/validação. A sintonização do hiperparâmetro inclui 10 epochs, 16 batch size, e o optimizador AdamW com taxa de aprendizagem 2e-5 para BERT e RoBERTa e 3e-5 para XLNet. Para eventos perigosos, o BERT fornece 60%, o RoBERTa 59 enquanto a XLNet fornece apenas 54% de precisão geral. Para as outras experiências de configuração de eventos de alto nível, o BERT e a XLNet dão 71% e 70% de desempenho com RoBERTa em relação aos outros modelos com 74% de precisão. Enquanto para o DE baseado em acções, DE baseado em cenários, e DE baseado em sentimentos, o BERT dá 62%, 85%, e 81% respetivamente; RoBERTa com 61%, 83%, e 71%; a XLNet com 52%, 81%, e 77% de precisão.
Existe a necessidade de clarificar a ambiguidade entre os diferentes trabalhos que abordam problemas similares utilizando termos diferentes. A ideia proposta de referir acontecimentos especifícos como eventos perigosos torna mais fácil a abordagem do problema em questão. No entanto, a escassez de conjunto de dados de eventos limita o desempenho dos modelos e o progresso na deteção das tarefas. A disponibilidade de uma maior quantidade de informação relacionada com eventos perigosos pode melhorar o desempenho do modelo existente. É evidente que o uso de modelos de aprendizagem profunda, tais como como BERT, RoBERTa, e XLNet, pode ajudar a detetar e classificar eventos perigosos de forma eficiente. Tem sido evidente que a utilização de modelos de aprendizagem profunda, tais como BERT, RoBERTa, e XLNet, pode ajudar a detetar e classificar eventos perigosos de forma eficiente. Em geral, o BERT tem um desempenho superior ao do RoBERTa e XLNet na detecção de eventos perigosos. É igualmente importante rastrear os eventos após a sua detecção. Por conseguinte, para trabalhos futuros, propõe-se a implementação das técnicas que lidam com o espaço e o tempo, a fim de monitorizar a sua emergência com o tempo.

Palavras-chaves

Detecção de eventos, Meios de comunicação social, Eventos perigosos, Análise de sentimentos, Extremismo

Abstract

Online social media platforms have become essential tools for communication and information exchange in our lives. It is used for connecting with people and sharing information. This phenomenon has been intensively studied in the past decade to investigate users' sentiments for different scenarios and purposes. As the technology advanced and popularity increased, it led to the use of different terms referring to similar topics which often result in confusion. We study such trends and intend to propose a uniform solution that deals with the subject clearly. We gather all these ambiguous terms under the umbrella of the most recent and popular terms to reach a concise verdict. Many events have been addressed in recent works that cover only specific types and domains of events. For the sake of keeping things simple and practical, the events that are extreme, negative, and dangerous are grouped under the name Dangerous Events (DE). These dangerous events are further divided into three main categories of action-based, scenario-based, and sentiments-based dangerous events to specify their characteristics. We then propose deep-learning-based models to detect events that are dangerous in nature. The deep-learning models that include BERT, RoBERTa, and XLNet provide valuable results that can effectively help solve the issue of detecting dangerous events using various dimensions. Even though the models perform well, the main constraint of fewer available event datasets and lower quality of certain events data affects the performance of these models can be tackled by handling the issue accordingly.

Keywords

Event Detection, Social media, Dangerous events, Sentiment Analysis, Extremism

Contents

1 Introduction **1**

1.1 Problem Statement . . . 2

1.2 Objectives . . . 4

1.2.1 Our Contribution . . . 5

2 Background and Related Work **7**

2.1 Introduction . . . 7

2.1.1 Background . . . 7

2.2 Dangerous Events . . . 13

2.2.1 Scenario-based Dangerous Events . . . 14

2.2.2 Sentiment-based Dangerous Events . . . 14

2.2.3 Action-based Dangerous Events . . . 15

2.3 Event Detection Methods . . . 15

2.3.1 Supervised Methods . . . 16

2.3.2 Unsupervised Methods . . . 17

2.3.3 Semi-Supervised Methods . . . 19

2.4 Event Detection Approaches . . . 20

2.4.1 Document-pivot Approach . . . 20

2.4.2 Feature-pivot Approach . . . 20

2.4.3 Topic Modeling Approach . . . 21

2.5 Event Tracking and Prediction . . . 22

2.6 Event Detection Datasets . . . 22

2.7 Conclusion . . . 24

3 Proposed Method **25**

3.1 Bidirectional Encoder Representations from Transformers (Bert) . . . 25

3.2 RoBERTa . . . 27

3.3 XLNet . . . 28

3.4 Conclusion . . . 29

4 Implementation and Results **31**

4.1 Experimental Setup . . . 31

4.1.1 Dataset . . . 31

4.1.2 Loading and Pre-Processing . . . 32

4.1.3 Fine-tuning and Hyperparameter tuning . . . 32

4.1.4 Performance metrics . . . 33

4.2 Results . . . 34

4.2.1 Dangerous Events . . . 34

4.2.2 Top-level Dangerous Events . . . 37

4.2.3 Sub-level Dangerous Events . . . 39

4.3 Discussion and Future Work 45

5 Conclusion **47**

Bibliography **49**

List of Figures

2.1 Dangerous Events and their categories 13
2.2 Classification of ED methods [1] 15
2.3 Topic Modeling in LDA[1] 18
2.4 Clustering-based method [1] 19
2.5 Graph-based clustering method [1] 19
2.6 Event Detection using Document-pivot approach [1] 20
2.7 Event Detection using the feature-pivot approach [1] 21
2.8 LDA - A common topic modelling technique [1] 21

3.1 Tokenized text is embedded using 12 encoders in BERT and fed into a feed-forward network and softmax function to obtain the classification probabilities. [2] 27
3.2 The architecture of XLNET model: (a)Content stream attention, which is the same as the standard self-attention. (b): Query stream attention, which does not have access to information about the content. (c): Overview of the permutation language modelling with two-stream attention. [3] 29

4.1 Total entries of dangerous events and their distribution in the original dataset 35
4.2 Top-level dangerous events and their original distribution dataset 38
4.3 Action-based dangerous events and their number of occurrences in the dataset 40
4.4 Scenario-based dangerous events and their number of occurrences in the dataset 42
4.5 Sentiment-based dangerous events and their number of occurrences in the dataset 44
4.6 Accuracy of BERT, RoBERTa and XLNet for Dangerous Events, Top-level DE, Action-based Sub-level DE, Scenario-based Sub-level DE and Sentiment-based Sub-level DE 46

List of Tables

1.1 An illustrative example of the mix-up between terms and problems in the literature . . . 3

2.1 Presumed types of dangerous events for tweets . . . 11
2.2 Dangerous Events are categorized under relevant types. . . . 12
2.3 Comparison of event detection datasets and knowledge bases . . . 24

4.1 Classification report of dangerous events using BERT . . . 35
4.2 Classification report of dangerous events using RoBERTa . . . 36
4.3 Classification report of dangerous events using XLNet . . . 37
4.4 Classification report of top-level dangerous events using the BERT model . 38
4.5 Classification report of top-level dangerous events using the RoBERTa model 38
4.6 Classification report of top-level dangerous events using XLNet model . . . 39
4.7 Classification report of action-based sub-level dangerous events using the BERT model . . . 40
4.8 Classification report of sub-level dangerous events using RoBERTa model . 41
4.9 Classification report of action-based sub-level dangerous events using XL-Net model . . . 41
4.10 Classification report of scenario-based sub-dangerous events using the BERT model . . . 42
4.11 Classification report of scenario-based sub-level dangerous events using RoBERTa model . . . 43
4.12 Classification report of scenario-based sub-level dangerous events using XL-Net model . . . 43
4.13 Classification report of sentiment-based sub-level dangerous events using the BERT model . . . 43
4.14 Classification report of sentiment-based sub-level dangerous events using RoBERTa model . . . 44
4.15 Classification report of scenario-based sub-level dangerous events using XL-Net model . . . 45

Acronyms List

DE	Dangerous Events
NLP	Natural Language Processing
ED	Event Detection
SVM	Support Vector Machine
CRF	Conditional Random Field
DT	Descion Tree
NB	Naive Bayes
KNN	K-Nearest Neighbor
LDA	Latent Dirichlet Allocation
DFT	Discrete Fourier Transformation
NER	Named Entity Relation
WT	Wavelet Transformation
CWT	Continues Wavelet Transformatio
IDF	Inverse Document Frequenc
LSH	Locality Sensitive Hashin
FSD	First Story Detection
TFIDF	Term Frequency–Inverse Document Drequency
PLSI	Probabilistic Latent Semantic Indexing

Chapter 1

Introduction

This dissertation is being developed for obtaining the degree of Master in Computer Science. It is titled "Event Detection and tracking" and is supervised by Prof. Doutor Sebastião Pais. This is the next stage of the preceding work which is titled, "Extremism and collective radicalization understanding", developed by Miguel Albardeiro.
Social media is a phenomenon that started growing with the advent of the internet. As technology advanced and become easily accessible, it evolved rapidly and become part of our daily lives. In the last decade, it has not only remained as a means of connecting with close friends and family as it was at the beginning but has become a tool for connecting with the worldwide audience. It allowed people to freely express and share their ideas, views, and emotions. Hence, it becomes very important for researchers to study such trends and their impact on people and society. This led to many new research trends, including sentiment analysis. There were many challenges these technologies brought with them, such as fake news, extremism, and lobbying, to name a few. Therefore, it was essential to develop the means to identify and detect such issues and finally track their emergence.
The extent of social media users consists of billions of people from all around the world. Initially, social media was developed with the object in mind to help connect with family and friends online. It was used to share everyday things, events, interests, and news within closed circles of family and friends. These were personal events, e.g. birthdays, weddings, vacations, graduation ceremonies, and going out. After the usability of social media was discovered, it soon caught the attention of individuals and companies that started using social media to reach more customers. Soon after, the trend became a global phenomenon where people connected worldwide based on common interests. The influence of social media on people's lives and attitudes has been widely studied and established from many different perspectives [4], [5].
Although social media is a broad term, it mainly refers to Facebook, Twitter, Reddit, Instagram, and YouTube. Some social media platforms allow users to post text, photos and videos. At the same time, many other social media applications have limited options and restrictions for sharing the type of content. YouTube allows users to post videos, while Instagram only allows users to share videos and photos. 4.66 billion active internet users worldwide, and 4.2 billion users are active on social media. As of the first quarter of 2020, Facebook has 2.6 billion monthly active users globally, making it the most extensive social media network globally. Twitter is one of the leading social media with 397 million users worldwide, becoming increasingly prominent during events and an essential tool in politics [6]. Another study [7] shows that Twitter is an effective and fast way of sharing news and developing stories. This trend has continued to grow over the last decade as the internet has become widespread.
However, the use of social media has become more complex in the last decade. It be-

came a broader phenomenon because of the involvement of multiple stakeholders such as companies, groups, and other organizations. Many lobbying and public relations firms got on board and started targeting social audiences to change people's perspectives and influence their decisions. Mostly these campaigns are related to a particular individual or a company. A similar process happens in the public sphere, where people rally against or support their target. It played a significant role in different outcomes, affecting countries, people, and the world. One example is the "Arab Spring" [8], an event that started in Tunisia and spread among other regional countries. Another example of good and bad events in the UK and US political spheres is given in the study that uses Twitter to evaluate the perceived impact on users [9].

The recent example of violence in Bangladesh can explain the link between social media with real life. On Wednesday, 15 October 2021, clashes were sparked by videos and allegations that spread across social media that a Qur'an, the Muslim holy book, had been placed on the knee of a statue of the Hindu god Hanuman. The violence continued in the following days, which resulted in the deaths of 7 people, with about 150 people injured; more than 80 special shrines set up for the Hindu festival were attacked. This case shows social media's severe and robust effect on our daily lives and ground situation [10]. This violence was termed as "worst communal violence in years" by New York Times. Similar episodes of violence are becoming a norm in India since the right of ring-wing politics. If there is the detection of events occurring on social media in advance, which alerts possible coming hazards, it can be countered in anticipation, significantly reducing the reaction immobilization of state forces while maximizing the protection of people at risk.

1.1 Problem Statement

With the advent of the latest advancements in technology, the use of social media has become a widespread and essential daily life phenomenon. It is a powerful means of communication and a source of exchanging information on a wide range of events happening in the world. The generated data are very useful for detecting real-world events and user-associated thoughts. The subject of event detection has been intensively studied in the last decade because of the popularity of social media and the invaluable information available on these platforms [1]. Using this information, many dangerous events were detected beforehand and stopped from occurring worldwide. Most of the research in this field is segregated as different terminologies referring to the same thing stir confusion [11]. This situation is given in Table.1.1 It leads to the issue of imbalanced class validation. In other words, the classifier is learned using the information from only one class. Therefore, standardization is imperative. This work aims to unite all these classes in broad and standard terms. For example, social media and social networks are used to refer the thing but using the term "social media" is more practical as it's the most frequently used and widely understood term. Similarly, hate speech, extremism, natural disasters, and violence can be grouped under the umbrella of a single term as dangerous events. The danger can be a threat like social violence, natural disasters, intellectual or community harm, con-

Paper references	Brief description	Term used
Barbara Poblete et al (2018) [12]	Natural disasters and crisis situations	Extreme Event Detection
Fatima Elsafoury. (2020) [13]	Protest repression events	Violence Event Detection
Evangelia Spiliopoulou et al. (2020) [14]	Crisis (e.g. earthquakes, floods) or attacks (e.g. bombings, shootings)	Disaster Events

Table 1.1: An illustrative example of the mix-up between terms and problems in the literature

tagion, social unrest, economic loss, or just spreading hateful and violent ideologies. We study these different events and classify them into topics so that the topic-based detection technique can be designed and integrated under the term of the dangerous event.

The term "event" implies a change, an occurrence bounded by time and space. In the context of social media, an event can be happening on the ground or online. Different mediums can broadcast events happenings on the ground while people participate in the event through social media discussion. These kinds of events can be referred to as hybrid events. An example of such a hybrid event can be a volcano eruption where people participate in the event using online discussions on social media while it is happening on the ground. While some events solely happen online, such as gaming, marketing, and learning events. Events can be communicated in text, photos and videos across social media platforms. Many events can simultaneously happen on social media platforms, providing beneficial and prosperous information. It provides information about the event itself, but it also reveals sentiments and opinions of the general public and the direction where the events are evolving shortly. This quick interaction of users and transmission of information makes it a dynamic process that sometimes proves hard to follow the latest development, making it a challenging task.

Events also have time dimensions; it is equally important to determine the time of an event after its detection. For example, an event may have occurred in the past, happened in the present, or planned to occur in the future. Based on that, further steps would be taken accordingly as per the requirements of the situation. The events occurring on social media may directly impact the personal or social life of the man/woman. Past events can tell us people's opinions and other factors; current events can be a great source of developing a story, while future events can help us prepare in advance. The study [15] reviews the existing research for the detection of disaster events and classifies them into three dimensions early warning and event detection, post-disaster, and damage assessment.

Event detection has been long addressed in the Topic of Detection and Tracking (TDT) in academia [16]. It mainly focuses on finding and following events in a stream of broadcast news stories shared by social media posts. Event Detection (ED) is further divided into two categories depending on the type of its task; New Event Detection (NED) and Retrospective Event Detection (RED) [17]. NED focuses on detecting a newly occurred event from online text streams, while RED aims to discover strange events from offline historical data. Often event detection is associated with identifying the first story on topics of interest through constant monitoring of social media and news streams. Other related

fields of research are associated with event detection, such as; event tracking, event summarization, and event prediction. Event tracking is related to the development of some events over time. Event summarization outlines an event from the given data, while the event forecasts the next event within a current event sequence. These topics are part of the Topic Detection and Tracking (TDT) field. In brief, the problem statement establishes the need to remove ambiguity between different events and filter them to be classified under the term dangerous event. For this purpose, the need for a relevant dataset is essential. Defining the problem in detail will help identify the best solution for the problem.

1.2 Objectives

Event detection is a vast research field, and various requirements and challenges exist for each task. Various terms have been used to address different events, making navigating the literature complex and sometimes confusing. We propose relevant events based on their characteristics under the umbrella term "Dangerous Events" (DE).

Different categories of these dangerous events can be based on different reasons, causes, motivations and ideologies. Identifying these motives and ideologies helps us identify these dangerous events. The impact of such events can be classified into different levels to identify them clearly according to the intensity of their nature. Classification techniques (learning-based or lexical-based) include the following categories: Document-pivot: This category comprises methods that represent documents as items to be grouped/clustered using some similarity measure. Feature-pivot: These are based on detecting abnormal patterns in the appearance of features, such as words. Once an event occurs, the expected frequency of a feature will be abnormal compared to its historical behaviour, thus indicating a potential new event. Topic modelling: This includes methods that utilize statistical/probabilistic models to identify events as latent variables in the documents.[18].

Most of the techniques for event detection cover particular topics and are limited in scope. Once an event is detected, it is crucial to classify it according to its sensitivity and act on it according to its priority. Standard techniques classify events out of which most of them can be irrelevant. Event detection from social media data must be detected efficiently and accurately. Relevant information on events of general or specific interest can be buried in mundane information (e.g., irrelevant, tainted and rumoured messages). We intend to design a strategy to classify only extreme and dangerous events to save extra work by separating the relevant events from the rest. Also, we define the different classes of dangers by defining their features. Then classify them to tackle them separately. These techniques are classified according to the type of target event specified in or detection of unspecified events. A dangerous event could be of any kind, depending on the type of event; each event has different needs to be tackled. Most of the work does the sentiment analysis of text to detect the event. We use sentiment analysis in combination with other techniques, adopting a flexible approach to find the optimal solution for our problem. In many cases, the feature-pivot approach shows more promising results [19]. Dangerous events can be other than brusty and trending and can be limited to only a specific group

and community. Therefore, it implies that only brusty and trending terms are insufficient to determine a dangerous event. They can also occur alongside the larger event, i.e. In August 2020, a right-wing opposition party tried to storm the German Parliament while hijacking the peaceful protest against corona restrictions. The same can be true for online events and can happen under the shadow of more prominent events. These events can be projected to a smaller and broader audience. First, mining the text to detect the set of informative features with strong signals that need minimal pre-processing and are positively associated with events of interest [20]. Identifying these informative features as keywords from social media can be the unsupervised approach. First, we underline the characteristics of all related terms that can be grouped under a group of clusters as dangerous events containing all classes of different categories. The semi-supervised deep learning approach will be used to classify the events at the initial stage. Then different approaches will be implemented, addressing each classified cluster based on its defined features. Secondly, we describe a model that processes documents online and detects them efficiently, finding the inter-relation between clusters and evaluating the evolution over time.[21].

1.2.1 Our Contribution

The contribution of this work can be summarized as follows:

- Defining the term “dangerous events”.
- Dividing dangerous events into top-level and sub-level categories.
- Employing deep-learning approaches to detect dangerous events and their categories.
- Comparison of the models and proposed approach.

Chapter 2

Background and Related Work

2.1 Introduction

This chapter presents the background of our problem. It shows how the same problem has been approached by using different terms. Many times the authors' goal is the same, but how it is referred to causes confusion. Few cases are discussed in the background section. The main concepts and methods are discussed in the next part of the chapter. These methods are essential for solving most of the problems related to event detection. We discuss each method briefly along with their application.

2.1.1 Background

In the last ten years, extensive research has been carried out on the importance of anomalies in various areas. There was confusion between the names and the issues in the literature. Especially when the same term is used in different disciplines but in others' meanings and vice versa. In addition, the terminology has changed over time and even in the past for the same discipline; A similar problem has been named differently at different times. This issue leads to imbalanced class validation. In other words, the classifier is learned using the information from only one class or using one specific term. Therefore, standardization is imperative.

Nourbakhsh et al. [22] address natural and artificial disasters on social media. They identified events from local news sources that may become global breaking news within 24 hours. They used Reuters News Tracer, a real-time news detection and verification engine. It uses a fixed sphere decoding (FSD) algorithm to detect breaking stories in real-time from Twitter. Each event is shown as a cluster of tweets engaging with that story. By considering different data features, they applied SGD and SVM classifier that detects breaking disasters from postings of local authorities and local news outlets.

Sakaki et al. [23] leverage Twitter to detect earthquake occurrence promptly. They propose a method to scrutinize the real-time interaction of earthquake events and, similar to detect a target event. Semantic analyses were deployed on tweets to classify them into positive and negative classes. The target for classification is two keywords; earthquake or shaking, which are also addressed as query words. Total of 597 positive samples of tweets that report earthquake occurrence is used as training data. They also implemented filtering methods to identify the location and an application called the earthquake reporting system in Japan.

Liu et al. [24] aim for crisis events. They propose a state-of-the-art attention-based deep neural networks model called CrisisBERT to embed and classify crisis events. It consists of two phases which are crisis detection and crisis recognition. In addition, another model

for embedding tweets is also introduced. The experiments are conducted on C6 and C36 datasets. According to the authors, these models surpass state-of-the-art performance for detection and recognition problems by up to 8.2% and 25.0%, respectively.

Archie et al. [25] proposed an unsupervised approach for detecting sub-events in major natural disasters. Firstly, noun-verb pairs and phrases are extracted from tweets as an important sub-event prospect. In the next stage, the semantic embedding of extracted noun-verb pairs and phrases is calculated and then ranked against a crisis-specific ontology called management of Crisis (MOAC) ontology. After filtering these obtained candidate sub-events, clusters are formed, and top-ranked clusters describe the highly important sub-events. The experiments are conducted on Hurricane Harvey and the 2015 Nepal Earthquake datasets. According to the authors, the approach outperforms the current state-of-the-art sub-event identification from social media data.

Forests fire have become a global phenomenon due to rising droughts and increasing temperatures, often attributed to global warming and climate change. The work [26] tests the usefulness of social media in supporting disaster management. However, the primary data for dealing with such incidents come from NASA satellite imagery. The authors use GPS-stamped tweets posted in 2014 from Sumatra Island, Indonesia, which experiences many haze events. As confirmed by analysing the dataset, Twitter has proven to be a valuable resource during such events. Furthermore, the authors also announced the development of a tool for disaster management.

Huang et al. [27] focus on emergency events. They consider the various type of events under the term "emergency events". It includes infectious disease, explosions, typhoons, hurricanes, earthquakes, floods], tsunamis, wildfires, and nuclear disasters. To respond in time, the model must automatically identify the attribute information 3W (What, When, and Where) of emergency events. Their proposed solution contains three phases, the classification phase, the extraction phase, and the clustering phase, and it is based on the Similarity-Based Emergency Event Detection (SBEED) framework. The experiment is done using the Weibo dataset. Different classification models such as KNN, Decision Trees, Naïve Bayes, Linear SVC (RBF), and Text-CNN are used in the classification phase. Secondly, time and location are extracted from the classification obtained. Lastly, an unsupervised dynamical text clustering algorithm is deployed to cluster events depending on the text-similarity of type, time and location information. The authors claim superiority of the proposed framework having good performance and high timeliness that can be described what emergency, and when and where it happened.

Pais et al. [28] present an unsupervised approach to detecting extreme sentiments on social networks. Online wings of radical groups use social media to study human sentiments engaging with uncensored content to recruit them. They use people who show sympathy for their cause to further promote their radical and extreme ideology. The authors developed a prototype system composed of two components, i.e., Extreme Sentiment Generator (ESG) and Extreme Sentiment Classifier (ESC). ESG is a statistical method used to generate a standard lexical resource called ExtremesentiLex, containing only extreme positive and negative terms. This lexicon is then embedded in ESC and tested on five dif-

ferent datasets. ESC finds posts with extremely negative and positive sentiments in these datasets. The result verifies that the posts previously classified as negatives or positives are, in fact, extremely negatives or positives in most cases.

The COVID-19 pandemic has forced people to change their lifestyles. Lockdown further pushed people to use social media to express their opinions and feelings. It provides a good source for studying users' topics, emotions, and attitudes discussed during the pandemic. The authors of the work [29] collected two massive COVID-19 datasets from Twitter and Instagram. They explore data with different aspects, including sentiment analysis, topic detection, emotions, and geo-temporal. Topic modelling on these datasets with distinct sentiment types (negative, neutral, positive) shows spikes in specific periods. Sentiment analysis detects spikes in specific periods and identifies what topics led to those spikes attributed to economy, politics, health, society, and tourism. Results showed that COVID-19 affected significant countries and experienced a shift in public opinion. Much of their attention was on China. This study can be very beneficial to read people's behaviour in the aftermath; Chinese people living in those countries also faced discrimination and even violence because of the Covid-19 linked with China.

Plaza-del-Arco et al. [30] investigate the link of hate speech and offensive language(HOF) with relevant concepts. Hate speech targets a person or group with a negative opinion, and it is related to sentiment analysis and emotion analysis as it causes anger and fear inside the person experiencing it. The approach consists of three phases and is based on multi-task learning (MTL). The setup is based on BERT, a transformer-based encoder pre-trained on a large English corpus. Four sequence classification heads are added to the encoder, and the model is fine-tuned for multi-class classification tasks. The sentiment classification task categorizes tweets into positive and negative categories, while emotion classification classifies tweets into different emotion categories (anger, disgust, fear, joy, sadness, surprise, enthusiasm, fun, hate, neutral, love, boredom, relief, none). The offence target is categorized as an individual, group, and unmentioned to others. Final classification detects HOF and classifies tweets into HOF and non-HOF.

Kong et al. [31] explore a method that explains how extreme views creep into online posts. Qualitative analysis is applied to make ontology using Wikibase. It proceeded from the vocabulary of annotations such as the opinions expressed in topics and labelled data collected from three online social networking platforms (Facebook, Twitter, and YouTube). In the next stage, a dataset was created using keyword search. The labelled dataset is then expanded using a looped machine learning algorithm. Two detailed case studies are outlined with observations of problematic online speech from the Australian far-right Facebook group. Using our quantitative approach, we analyzed how problematic opinions emerge. The approach exhibits how problematic opinions appear over time and how they coincide.

Demszky et al.[32] highlight four linguistic dimensions of political polarization in social media: topic choice, framing, affect and apparent force. These features are quantified with existing lexical methods. The clustering of tweet embeddings is proposed to identify important topics for analysis in such events. The method is deployed on 4.4M tweets

related to 21 mass shootings. Evidence proves the discussions on these events are highly polarized politically, driven by the framing of biased differences rather than topic choice. The measures in this study provide connecting evidence that creates a big picture of the complex ideological division penetrating public life. The method also surpasses LDA-based approaches for creating common topics.

While most typical use of social media is focused on disease outbreaks, protests, and elections, Khandpur et al. [33] explored social media to uncover ongoing cyber-attacks. The unsupervised approach detects cyber-attacks such as; breaches of private data, distributed denial of service (DDOS) attacks, and hijacking accounts while using only a limited set of event trigger as a fixed input.

Coordinated campaigns aim to manipulate and influence users on social media platforms. Pacheco et al. [34] work aim to unravel such campaigns using an unsupervised approach. The method builds a coordination network that relies on random behavioural traces between accounts. A total of five case studies are presented in the work, including U.S. elections, Hong Kong protests, the Syrian civil war, and cryptocurrency manipulation. Networks of coordinated Twitter accounts are discovered in all these cases by inspecting their identities, images, hashtag similarities, retweets, or temporal patterns. The authors propose using the presented approach for uncovering various types of coordinated information warfare scenarios.

Coordinated campaigns can also influence people towards offline violence. Xian Ng et al. [35] investigates the case of capital riots. They introduce a general methodology to discover coordinated by analyzing messages of user parleys on Parler. The method creates a user-to-user coordination network graph prompted by a user-to-text graph and a similarity graph. The text-to-text graph is built on the textual similarity of posts shared on Parler. The study of three prominent user groups in the 6 January 2020 Capitol riots detected networks of coordinated user clusters that posted similar textual content supporting different disinformation narratives connected to the U.S. 2020 elections.

Wanzheng Zhu and Suma Bhat [36] study the specific case of using euphemisms by fringe groups and organizations that is expression substituted for one considered too harsh. The work claims to address the issue of Euphemistic Phrase detection without human effort for the first time. Firstly the phrase mining is done on raw text corpus to extract standard phrases; then, word embedding similarity is implemented to select candidates of euphemistic phrases. In the final phases, those candidates are ranked using a masked language model called SpanBERT.

Yang Yang et al. [37] explore Network Structure Information (NSI) for detecting human trafficking on social media. They present a novel mathematical optimization framework that combines the network structure into content modelling to tackle the issue. The experimental results are proven effective for detecting information related to human trafficking.

The author presents a Table 2.1to clarify the intent of this work by providing an example of the collected tweets and their presumed techniques. Based on the existing methods for event detection, it gives a clear objective for using these methods for detecting dangerous events.

	Tweets	Proposed dangerous event type
1	"RT @KaitMarieox: This deranged leftist and LGBT activist named Keaton Hill assaulted and threatened to kill @FJtheDeuce, a black conservati…"	Action-based dangerous event
2	"RT @Lrihendry: When Trump is elected in 2020, I'm outta here. It's a hate-filled sewer. It is nearly impossible to watch the hateful at…."	Sentiment-based dangerous event
3	"Scientists predict a tsunami will hit Washington, DC on 1/18/2020 We Are Marching in DC… https://t.co/3af4ZhyV3J"	Scenario-based dangerous event

Table 2.1: Presumed types of dangerous events for tweets

For the sake of our work and better understanding. We approach the problem by putting all different words under the single term "dangerous events". A dangerous event can be any abnormality in a given scenario. On social media, dangerous events can be abnormal, rare, exceptional, peculiar, or outlier. For example, in security, intruders are abnormalities [38], [39]. Each scenario requires a unique approach to address them. Then comes the next task to detect them and the definition of the method to detect the event.

	Event Type	Technique	Reference	Dataset	Year
Scenario-based					
	Natural Disasters	SVM/SGD	[22]	Twitter	2017
	Earthquake	Classification(SVM)	[23]	Twitter	2010
	Crisis	CrisisBERT	[24]	Twitter (C6,C36)	2021
	Earthquake &Hurricane	Unsupervised	[25]	Twitter	2019
	Fire and Haze Disaster	Classification (hotspots)	[26]	NASA &Twitter	2017
	Emergency	Text-CNN, Linear SVC & Clustering	[27]	Weibo	2021
	...	...	...	...	...
Sentiment-based					
	Extreme Sentiments	Unsupervised learning	[28]	misc.	2020
	Covid19 Sentiments	word2vec	[29]	Twitter & Instagram	2021
	Hate speech & offensive Language	BERT	[30]	HASOC(Twitter)	2021
	Far-right Extremism	Classification	[31]	Facebook, Twitter &Youtube	2021
	Political Polarization	Clustering	[32]	Twitter	2019
	...	...	...	...	...
Action-based					
	Cyber attack	Unsupervised	[33]	Twitter	2017
	Coordinated campaigns	Unsupervised	[34]	misc.	2021
	Riots	Clustering	[35]	Parler	2021
	Drugs Trafficking	SPANBert	[36]	Text Corpus(subreddit)	2021
	Human Trafficking	Classification (NSI)	[37]	Wiebo	2018
	...	...	...	...	...

Table 2.2: Dangerous Events are categorized under relevant types.

2.2 Dangerous Events

According to Merriam-Webster [40], the word "dangerous" means involving possible injury, pain, harm, or loss characterized by danger. In that context, we define a dangerous event as the event that poses any danger to an individual, group, or society. This danger can come in many shapes and intensities. The objective is to draw a fine line between normal, harmless, unpleasant, and extreme, abnormal and harmful events. Less sensitive, unpleasant, and disliked events do not compel the person to feel threatened. While, in the case of dangerous events, the person will feel fearful, unsafe, and threatened. This provides the objective to approach the term "event" in a broader sense to address the common element of all such events. The details of dangers can always be discussed in detail, providing the necessity of the situation; for example, a natural disaster proceeds urgent hate speech. In other words, the first requires an immediate response with no time to lose, while the latter can allow some time to take action.
Dangerous events can be anomalies, novelty, outliers, and extreme. These terms can be used to refer to positive or negative meanings. However, Not all anomalies, novelties, and extremes are dangerous, but all dangerous events fulfil one or all of those conditions (extreme, anomaly, novelty). The author [28] proposed an unsupervised approach to detecting extreme sentiments on social media. Positive Extreme sentiments can be detected and differentiated from everyday positive sentiments. Therefore, it may be concluded that extreme negative sentiments will likely turn into dangerous events.
Grouping and defining dangerous events based on their characteristics is another challenging task, and it can help address the issue of approaching different types of dangerous events by narrowing it down to specific details. We will define three broad categories of dangerous events with commonality among them.

1. Scenario-based Dangerous Events
2. Sentiment-based Dangerous Events
3. Action-based Dangerous Events

The figure 2.1gives the depiction of dangerous events and their categories. In the following subsections, we will outline the definition for each type of dangerous event.

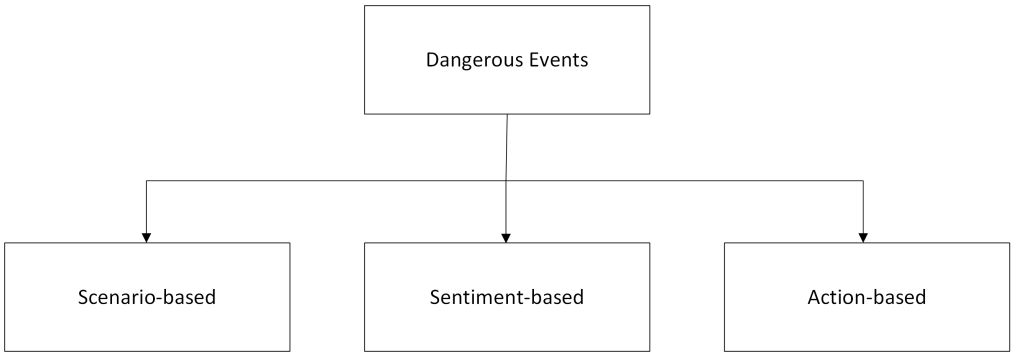

Figure 2.1: Dangerous Events and their categories

2.2.1 Scenario-based Dangerous Events

We refer to the word "scenario" as the development of events. These events are unplanned and unscripted, and most of the time, they occur naturally. Some planned events can also turn into surprising scenarios. For example, a peaceful protest can turn into a riot, like in 2020 when a peaceful protest against corona restrictions in Germany turned into an ugly situation when the rally was hijacked by right-wing extremists, which ended up storming Parliament building and exhibiting right-wing symbols and slogans [41].
Detecting and tracking natural disasters on social media have been investigated intensively, and studies [15] have proposed different methods to identify those disasters by various means. The aim of these studies has been mainly to tap into the potential of social media to get the latest updated information provided by social media users in real-time and identify the areas where assistance is required. This paper considers scenario-based dangerous events, including earthquakes, force majeure, hurricanes, floods, tornadoes, volcano eruptions, and tsunamis. Although each calamity's nature is different, the role of social media in such events provides a joint base to approach them as scenario-based dangerous events. A supposed example of scenario-based dangerous event is obtained using the crawler tool SocialNetCrawler, which can be accessed using the link[1]:
"@politicususa BREAKING: Scientists predict a tsunami will hit Washington, DC on 1/18/2020 We Are Marching in DC… https://t.co/3af4ZhyV3J"

2.2.2 Sentiment-based Dangerous Events

Sentiment Analysis (SA), also known as Opinion Mining (OM), is the process of extracting people's opinions, feelings, attitudes, and perceptions on different topics, products, and services. The sentiment analysis task can be viewed as a text classification problem as the process involves several operations that ultimately classify whether a particular text expresses positive or negative sentiment [42]. For example, A micro-blogging website like Twitter is beneficial for predicting the index of emerging epidemics. These are platforms where users can share their feelings which can be processed to generate vital information related to many areas such as healthcare, elections, reviews, illnesses, etc. Previous research suggests that understanding user behaviour, especially regarding the feelings expressed during elections, can indicate the outcome of elections [43].
Sentiments can be positive and negative, but for defining sentiment-based dangerous events, the applicable sentiments are negatives and, in some instances, negative extremes. Online radicalization can be attributed to this threat related to extreme negative sentiments towards certain people, countries, and governments. Such extreme negative sentiments can result in protests, online abuse, and social unrest. Detecting these events can help reduce their impact by allowing the concerned parties to counter beforehand. A hypothetical example of a sentiment-based dangerous event from a tweet obtained using SocialNetCrawler is given below.
"RT @Lrihendry: When Trump is elected in 2020, I'm outta here. It's a hate-filled sewer.

[1] `http://sncrawler.di.ubi.pt/`

It is nearly impossible to watch the hateful at…"

2.2.3 Action-based Dangerous Events

The action involves human indulgence in an event. Various actions happen on the ground that can be detected using social media. Actions can be of many types, but we point out actions that are causing harm, loss, or threat to any entity, which again shares the common attribute of negativity and is highly similar to previously defined types of dangerous events. Some examples of Action-based dangerous events can be prison breaks, terrorist attacks, military conflicts, shootings, etc. Several studies have been published focusing on one or more types of such action-based events. The study [44] focuses on anti-fascist accounts on Twitter to detect acts of violence, vandalism, de-platforming, and harassment of political speakers by Antifa. An assumed example of action-based dangerous event acquired using SocialNetCrawler is given below.
"RT @KaitMarieox: This deranged leftist and LGBT activist named Keaton Hill assaulted and threatened to kill @FJtheDeuce, a black conservative…."

2.3 Event Detection Methods

Many methods are proposed for the detection of events. These event detections (ED) methods are mainly classified as supervised and unsupervised, as shown in Figure 1. Supervised methods include support vector machine (SVM), Conditional random field (CRF), Decision tree (DT), Naive Bayes (NB) and others. While the unsupervised approaches include Query-based, Statistical-based, Probabilistic based, Clustering-based and Graph-based.

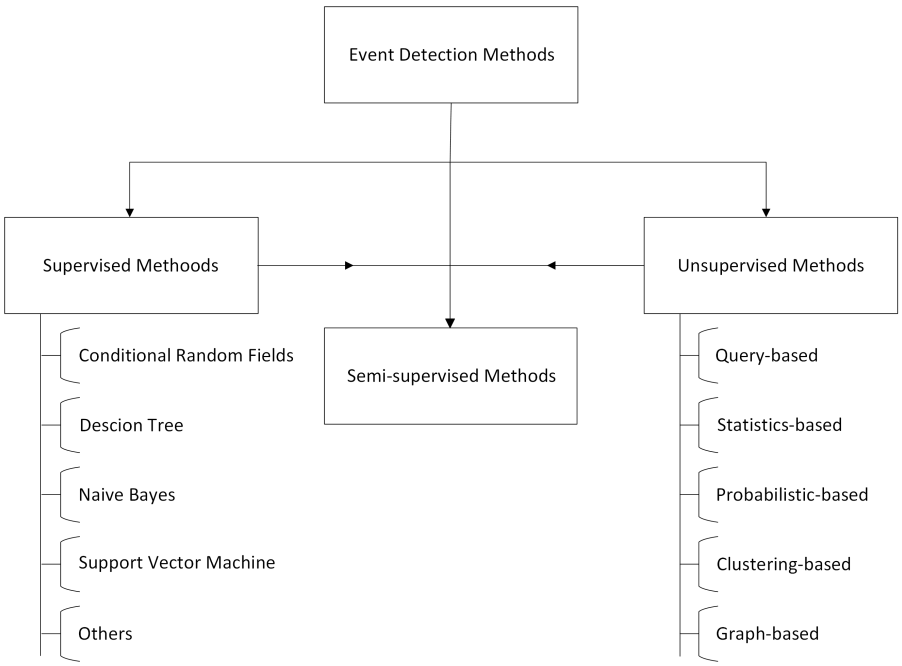

Figure 2.2: Classification of ED methods [1]

2.3.1 Supervised Methods

Supervised methods are expensive and lengthy as they require labels and training. This becomes difficult for larger data where the cost of training the model is quite higher and time-consuming. Some of the supervised methods for event detection are discussed below.

2.3.1.1 Support Vector Machines (SVM)

Support vector machines are based on computer learning theory's principle of minimizing structural risks [45]. The idea of minimizing structural risks is to find an assumption h for which we can guarantee the lowest true error. The real error in h is the probability that h will make an error in a sample test selected at random. An upper limit can be used to connect the true error of a hypothesis h with the error of h in the training set and the complexity of H (measured by VC-Dimension), the space of hypotheses that contains h [45]. The supporting vector machines find the hypothesis h, which (approximately)minimizes this limit on the true error by effectively and efficiently controlling the VC dimension of H [46].
It has been found in many studies that SVM is one of the most efficient algorithms for text classification. The accuracy was achieved in classifying the traffic or non-traffic events on Twitter. It identified valuable information regarding Twitter traffic events[47] SVM in combination with an incremental clustering technique to detect real-world and social events from photos posted on the Flicker site [48].

2.3.1.2 Conditional Random Fields (CRF)

The CRFs is an important machine learning model developed based on the Maximum Entropy Markov Model (MEMM). It was first proposed by Lafferty et al. [49] as a probabilistic model to segment and label sequence data, to inherit the previous models' advantages, increase their efficiency, overcome their defects, and solve more practical problems. A conditional Random Field (CRF) classifier was learned to extract the artist name and location of music events from a corpus of tweets [50].

2.3.1.3 Decision Tree (DT)

Decision tree learning is a supervised machine learning technique to produce a decision tree from training data. A decision tree is also referred to as a classification tree or a reduction tree. It is a predictive model that draws conclusions from observations about an item's target value. In the tree structures, leaves represent classifications (also referred to as labels), non-leaf nodes are features, and branches represent conjunctions of features that lead to the classifications. [49] A decision tree classifier called gradient boosted was used to anticipate whether tweets consist of an event that affects the target entity or not.

2.3.1.4 Naïve Bayes (NB)

Naïve Bayes is a simple learning algorithm that uses the Bayes rule and a strong assumption that the attributes are conditionally independent if the class is given. Although often

in practice, this independent assumption is violated, naïve Bayes, despite that, often provides the accuracy which is competitive. Combined with its computational efficiency and many other distinctive features, this results in naïve Bayes being extensively applied in practice.
Naïve Bayes gives a procedure for using the information in the sample data to determine the posterior probability P(y | x) of each class y, given an object x. Once we have such estimates, they can be used for classification or other decision-support applications [51].

2.3.2 Unsupervised Methods

Unsupervised methods usually do not require training or target labels. However, they can depend on certain rules based on the model and requirements. Unsupervised methods that are used for event detection are discussed in the following sections. Many unsupervised methods are developed by scientists and are grouped into different categories that are described in the following subsections.

2.3.2.1 Query Based Methods

Query-based methods are based on queries and simple rules to identify planned rules from multiple websites. e.g., YouTube, Flicker, Twitter. An event's temporal and spatial information was extracted and then used to ask other social media websites to obtain relevant information[51]. Query-based methods require predefined keywords if there is a large number of keywords to avoid unimportant events.

2.3.2.2 Statistical Based Methods

Many methods were introduced by different researchers under this category. For example, the average frequency of unigrams was calculated to find the significant unigrams (keywords) and combine those unigrams to illustrate the trending events. [52] The attempt was made to detect hot events by identifying burst features (i.e., unigram) during different time windows. Each unigram bursty feature signal was then converted into a frequency domain. using Discrete Fourier Transformation(DFT). However, DFT could not detect the time period when there is a burst, which is very important in the ED process[53].
Another technique called Wavelet Transformation (WT) was introduced to assign signals to each unigram feature. WT technique differs from DFT in tern of isolating time and frequency and provides better results[54]. A new framework was proposed that integrated different unsupervised techniques. For example, LDA, NER, and bipartite graph clustering algorithms based on relation and centrality scores to discover hidden events and extract their important information such as time, location and people involved [55].
Named Entity Relation (NER) identified increasing weights for the proper noun features. A proposed technique applied tweet segmentation to get the sentences containing one or more phrasing words instead of unigrams. Later, they computed the TFIDF of these sentences and user frequency and increased weights for the proper noun features identified by Named Entity Relation (NER). Li et al. (2012a)[56] first applied tweet classification using

K-Nearest Neighbor (KNN) to identify the events from tweets published by Singapore users[56].

Weiler et al. (2014) [57] used shifts of terms computed by Inverse Document Frequency (IDF) over a simple sliding window model to detect events and trace their evolution. Petrović et al. (2010) [58] modified and used Locality Sensitive Hashing (LSH) to perform First Story Detection (FSD) task on Twitter.

2.3.2.3 Probabilistic Based Methods

Latent Dirichlet Allocation (LDA) and Probabilistic Latent Semantic Indexing (PLSI) are topic-modelling methods used for event detection. In LDA, Each document has many topics and is supposed to have a group of topics for each document. The model is shown in Figure2.3.

LDA worked well with news articles and academic abstracts but was unsuitable for small texts. However, the LDA model has been improved by adding tweet pooling schemes and automatic labelling. Pooling schemes include basic scheme, author scheme, burst terms scheme, temporal scheme and hashtag scheme tweets published under the same hashtag. The results of the experiments showed that the hashtag scheme produced the best cluster results [59]. However, LDA defines the number of topics and terms per topic in advance, which can be inefficient when implementing it over social media.

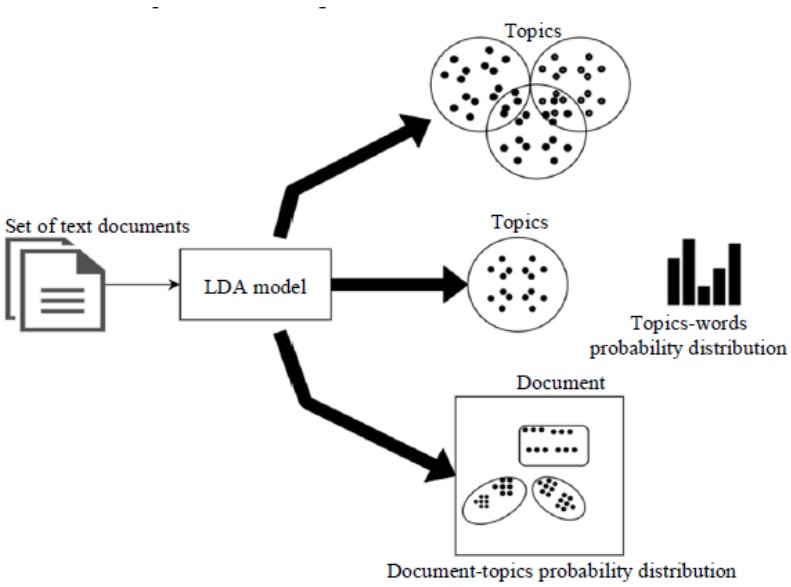

Figure 2.3: Topic Modeling in LDA[1]

2.3.2.4 Clustering Based Method

Clustering-based methods mainly rely on selecting the most informative features that contribute to event detection compared to supervised methods that require labelled data for prediction to contribute to detecting events with more precision.

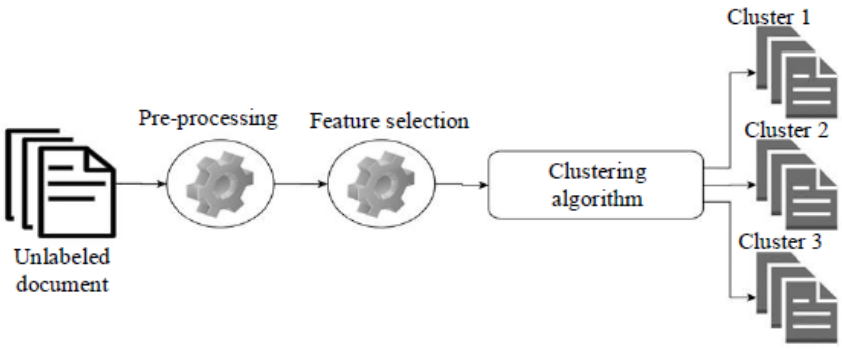

Figure 2.4: Clustering-based method [1]

Many clustering-based methods are in practice for text data. K-means is the famous clustering algorithm. A novel dual-level clustering was proposed to detect events based on news representation with time2vec[60]. Clustering-based methods have been employed in various ways along with other techniques such as NER, TFIDF and others in different tasks, but the ideal clustering technique is still yet to come.

2.3.2.5 Graph-Based Methods

Graph-based methods consist of nodes/vertices representing entities and edges representing relationships between the nodes. Valuable information can be extracted from these graphs by grouping a set of nodes based on the set of edges. Each generated group is called a cluster/graph structure, a community, cluster or module. The links between different nodes are called intra-edges; meanwhile, links that connect different communities are called inter-edges.

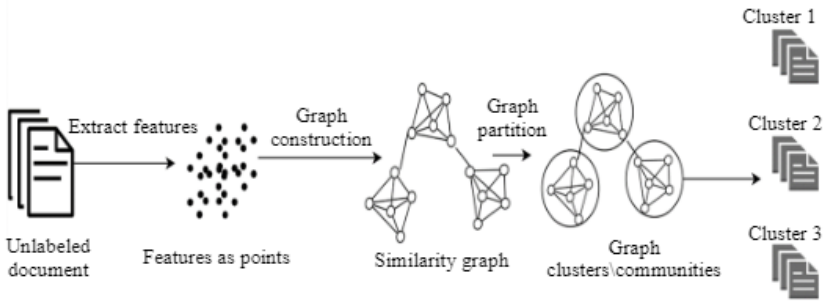

Figure 2.5: Graph-based clustering method [1]

2.3.3 Semi-Supervised Methods

Semi-supervised learning combines both supervised and unsupervised learning methods. Normally, a small number of labelled and largely unlabelled data is used for training purposes. If there is a huge number of unlabelled data combined with insufficient labelled data, it can affect the classification accuracy. It is also referred to as imbalanced training data. Similarly, if there is no labelled data for a particular class, the classification can become

inefficient and inaccurate. Some of the semi-supervised methods include self-training, generative models and graph-based methods. For such problems, a semi-supervised algorithm based on tolerance roughest and ensemble learning is recommended[61]. The missing class is extracted by approximation from the dataset and used as the labelled sample. The ensemble classifier iteratively builds the margin between positive and negative classes to further estimate negative data since negative data is mixed with the positive data. Therefore, classification is done by applying a hybrid approach without the need for training samples. It saves the cost of getting labelled data manually, especially for larger datasets.

2.4 Event Detection Approaches

There are two approaches that are being used in event detection. They are called document-pivot and feature-pivot. what differs in these approaches is mainly the clustering approach, the way documents are used to feature vectors, and the similarity metric used to identify if the two documents represent the same event or not.

2.4.1 Document-pivot Approach

Document-pivot approach detects events by clustering documents based on the document similarity. It originates from the Topic Detection and Tracking task (TDT) field and can be seen as a clustering issue. Documents are compared using cosine similarity on tf-idf representations, while a Locality Sensitive Hashing (LSH) scheme is utilized to rapidly retrieve the best match.

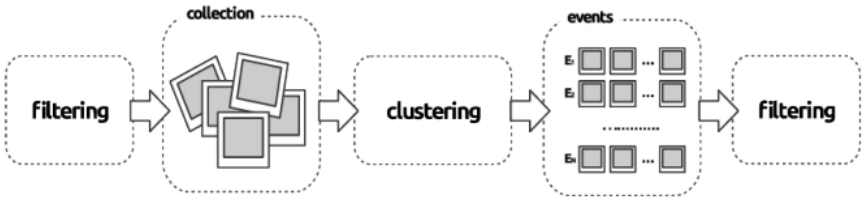

Figure 2.6: Event Detection using Document-pivot approach [1]

2.4.2 Feature-pivot Approach

Feature-pivot Approach clusters together term based on the pattern in which they occur. This technique was originally proposed for the analysis of time-stamped document streams. The bursty activity is considered an event, making some text feature more prominent. The feature can be keywords, entities and phrases. This process is shown in the figure given below.

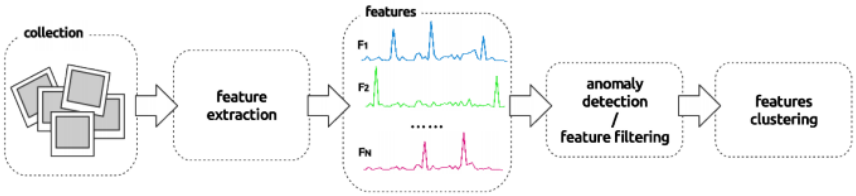

Figure 2.7: Event Detection using the feature-pivot approach [1]

2.4.3 Topic Modeling Approach

Topic modelling approaches are based on probabilistic models that detect events in social media documents, similarly to topic models that identify latent topics in text documents. In the beginning, topic models depended on word occurrences. The text corpora were given a mixture of words to model latent topics and the set of identified topics as documents. Latent Dirichlet Allocation (LDA) is the most known probabilistic topic modelling technique shown in Figure 2.2. It is a hierarchical Bayesian model where a topic distribution is supposed to have a sparse Dirichlet prior. The model is shown in the figure below, where is the parameter of the Dirichlet before the per-document topic distribution and is the word distribution for a topic. K represents the number of topics, M represents the number of documents, while the number of words in a document is given by N. If words W are the only observable variables, the learning of topics, word probabilities per topic, and the topic mixture of each document are tackled as a problem of Bayesian inference is solved by Gibbs sampling.

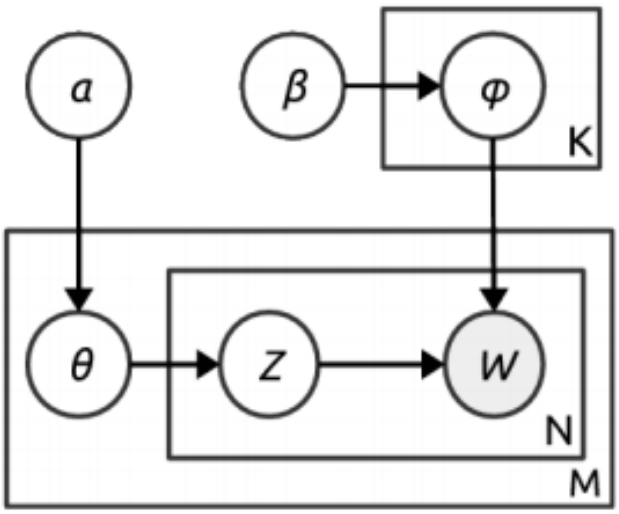

Figure 2.8: LDA - A common topic modelling technique [1]

2.5 Event Tracking and Prediction

Event prediction is a complicated problem that encompasses various dimensions. Many events are difficult to predict before they become visible and occur in real life. For instance, it is impossible to anticipate many natural disasters, which are detectable once they occur. Some events can be predicted while they are still in the development stage. Authors in [62] discover events from local news sources before they may become breaking news globally. The use case of Covid-19 can be regarded as an example where it started locally and later became a global issue. A recent dataset from the Kaggle competition suggests investigating the usability of a method for predicting disaster events mentioned in tweets. Another work [63] tests the effectiveness of BERT embeddings, an advanced contextual embedding method that constructs different vectors for the same word in various contexts. The result shows that the deep learning model surpassed other typical existing machine learning methods for disaster prediction from tweets. Zhou et al. [64] presented a novel framework called Social Media enhAnced pandemic suRveillance Technique (SMART) for predicting confirmed Covid-19 cases and fatalities. The method consists of two parts; firstly, heterogeneous knowledge graphs are constructed based on extracted events. Secondly, a module is formulated of time series prediction for confirmed cases and fatality rate at the state-level in the United States and finally discovers the risk factors for intercede COVID-19. The approach shows an improvement of 7.3% and 7.4% compared to other state-of-the-art methods. Most of the other ongoing research targets particular scenarios of event prediction with limited scope. By keeping in mind the complexity of this problem, the generalization of this problem is obscure, and only a few related works are presented and discussed.

The event tracking problem revolves around classification, Named Entity Recognition (NER), timestamp and Geo-spatial information. Each problem has unique features that can be utilized for the job. For example, the work [65] experiments with deep and non-deep learning-based models for detecting ad tracking rumours. It outlines the need for rumour tracking and how it is integral to rumour detection. Another study [66] tracks sentiments on social media by building an Arabic temporal tweets corpus labelled with positive, negative, or neutral classes. Named entities from each tweet in the corpus are extracted and used this information along with sentiment spikes. This information is then fed into the analytic system to draw useful insights. Much work has been dedicated to tracking events in different domains.

2.6 Event Detection Datasets

The internet and related technological advances have seen exponential growth in the last few decades. This growth generated an enormous amount of data that resulted in significant interest and effort in detecting the important events from the data for various purposes. However, the progress was slow in creating the benchmark event detection datasets. The possible reason for the slow growth of event detection datasets can be credited to the difficulty and expense of annotating events that require manual input from

humans. Only a few datasets are available that are related to event detection. Most of these datasets are small in size, cover very limited types of events, and focus mainly on restricted domains that cover certain features. This scenario gives birth to a problem in which deep learning usually requires balanced data. Few of the significant datasets are given in the upcoming paragraphs. Table 2.3 gives the comparison of the discussed event datasets.

Automatic Content Extraction (ACE) 2005 is a widely-used event dataset that contains approximately 1,800 files of mixed genre text in English, Arabic, and Chinese annotated for entities, relations, and events [67]. The genres include news-wire, broadcast news, broadcast conversation, weblog, discussion forums, and conversational telephone speech. It was developed by the Linguistic Data Consortium (LDC), and the data was annotated with support from the ACE Program [68] and additional assistance from LDC. It includes a total of 33 event types, 599 documents and 5, 349 instances. The ACE05 corpus is one of the sizeable corpora annotated with Coreference Resolution (CR), Entity Mention Detection (EMD) and Relation Extraction (RE), making it a plausible dataset for multi-task learning. Mention tags in ACE05 include seven types of entities Person, Organization, or Geographical Entities. Both the mentioned boundaries and the head spans are annotated for each entity. ACE05 also provides six types of relations that include Organization-Affiliation (ORG-AFF), GEN-Affiliation (GEN-AFF), and Part-Whole (PART-WHOLE).

MAVEN [69] stands for MAssive eVENt detection dataset. It provides a general domain event detection dataset manually annotated by humans. The data used for building the dataset consist of English Wikipedia and FrameNet [70] documents. The total instances of 111,611 various events and 118,732 events mentioned are given in MAVEN. The authors of MAVEN claim to be the largest human-annotated event detection dataset available on the internet. There are 164 different events, representing a much wider range of public domain events. The event types are grouped under five top-level types: action, change, scenario, sentiment, and possession.

EventWiki [71] is the event knowledge base consisting of 21,275 events containing 95 types of significant events collected from Wikipedia. EventWiki provides details: event type, event info-box, event summary, and full-text description. Authors of EventWiki claim to be the first knowledge base of significant events, whereas other knowledge bases are mostly concentrated on static entities such as people, locations, and organizations.

EVIN [72] stands for EVents In News. It describes a method to extract events from a news corpus and organize them in relevant classes. It contains 453 classes of event types and 24,348 events extracted from 300,000 different news articles. The news articles used in this work are sourced from a highly diverse set of newspapers and other online news providers (e.g., http://aljazeera.net/, http://www.independent.co.uk, etc.). These news articles were crawled from the external links mentioned on Wikipedia pages while ignoring the content of Wikipedia pages to get the articles from the original website source.

Table 2.3: Comparison of event detection datasets and knowledge bases

Name	Instances	Events	Data Source	Language	year
MAVEN	111,611	164	English Wikipedia & FrameNet	English	2020
ACE05	5,349	33	Broadcasts ,Newswire & Newspaper	Eng., Arab., Chinese	2005
EventWiki	21,275	94	English WIkipedia	English	2018
EVIN	24,348	453	News Corpus	English	2014

2.7 Conclusion

This chapter presents the background of our problem and shows some relevant examples. It has been observed that to research the methods and techniques for event detection, we need to approach literature using various means and terms. For example, the method used to tackle the problem of crisis detection [73], violence detection [13] and disaster detection [14] is relevant for our work and uses different ways to identify the problem. Hence, we grouped all these terms into one single term and refer to them as "Dangerous Event". This is supposed to help us with research and study literature in a convenient way. So that we can explore various existing methods developed for this purpose and can be used eventually to drive the solution for our problem.

Chapter 3

Proposed Method

The work of (Vikre and Wold, 2015) [74] suggests that the use of locality-sensitive hashing(LSH) along with named entity recognition(NER) achieves better performance for news detection instead of using the topic modelling approach. Moreover, sets of words are represented as topics, which may not all be bursty, therefore it may include more general topics than specific ones. Clustering techniques may lead to unimportant event yield, especially if the number of events is large. It can give huge clusters of events that might be not of our interest and can be proved difficult to filter from real dangerous events. Therefore, we propose a new strategy based on recent deep learning algorithms. Recent works demonstrate that the Transformer-based pre-trained models (PTMs) can achieve state-of-the-art performance on various tasks [75]. These algorithms mainly follow a mix of feature-pivot and document-pivot approaches where models are pre-trained on large text corpora. It learns embeddings for the words of the language or representations of vectors in a way that related words cluster together in the vector space. The embeddings then serve as input to a new classifier, which is trained on the particular task [76]. The task in our case is the detection of dangerous events in social networks. A recently published study demonstrates the efficiency of BERT for detecting extreme negative and extreme positive sentiments [2]. Even though not all extreme sentiments have the aspect of being dangerous (i.e, extreme happiness) but all dangers definitely have extreme factors rather than the normal. Therefore, a total of three different transformers-based, namely BERT, RoBERTa, and XLNet, will be used to conduct the experiment and draw the conclusion by comparing these models.

3.1 Bidirectional Encoder Representations from Transformers (Bert)

BERT is developed to pre-train deep bidirectional representations from the unlabeled text by jointly conditioning the left and right contexts. BERT accomplishes remarkably well on numerous NLP and sequence-to-sequence-based language generation-related tasks such as question answering, abstract summarization, sentence prediction, conversational response generation, polysemy and coreference (words that sound or look the same but have distinct meanings) the resolution, word sense disambiguation, natural language inference, and sentiment classification (text classification). Its pre-trained model acts as the mind, which can then master and regulate the growingly large resources of discoverable content and queries and can be fine-tuned to the user's specifications. This process is called transfer learning. The pre-trained BERT model can be fine-tuned with a single additional output layer to build state-of-the-art models for various NLP problems. It encodes text

bi-directionally and demands minimal architectural variations for a broad range of NLP problems. Using a pre-trained transformer encoder, BERT can represent any token based on its bidirectional context.

BERT is pre-trained on an extensive corpus of unlabeled text, including Wikipedia (2,500 million words) and Book pre-trained the magic that contributes to the success of BERT. As the model is trained on a large text corpus, the model begins to gain an in-depth and intimate conception of how the language works. This understanding is the backbone that is useful for almost any NLP task. The most helpful feature of BERT is fine-tuning, by which, by adding just some of the additional output layers, we can create state-of-the-art models for various NLP tasks. BERT is currently being used by Google to optimize the interpretation of search engine queries. Initially, it was limited to the English language, but by December 2019, the model had already been rolled out in more than 70 languages. The original BERT achieved state-of-the-art results in 11 NLP tasks. However, we are only interested in its classification task. [77]

BERT has two versions of different model sizes which includes BERT-base (L=12, H=768, A=12, Total Parameters=110M) and BERT-large (L=24, H=1024, A=16, Total Parameters=340M). The BERT-base model contains an encoder with 12 transformer blocks, 12 self-attention heads, and 768 units of hidden embedding parameters, a sequence of hidden states of the last layer of the model. The large model (BERT-large) uses 24 layers with 1024 hidden units and 16 self-attention heads. In particular, the former has 110 million parameters, while the latter has 340 million parameters. BERT takes an input of a sequence of up to 512 tokens and outputs the sequence representation. The sequence has one or two segments, where the first token of the sequence is always [CLS] and contains the specific classification embedding, and another special token [SEP] is used to divide the segments. BERT arranges the final hidden state h of the first token [CLS] for text classification tasks to render the complete sequence. To get the predicted probabilities from the trained model, a softmax classifier is added to the top of the BERT model. The data set must be vectorized to feed it to the classifier since it is originally in text format. BERT learns contextual embedding rather than learning context-free, such as in the case of Word2Vec. Although different models are available for text vectorization, BERT performs tokenization using the WordPiece method [78].

In addition to [CLS] and [SEP], a token called [PAD] is added to make the length of all sentences equal to the specified sequence length required or specified for the model, and an attention mask is introduced to tell the model about [PAD] tokens. These tokens are used to input the model to obtain each vector representation. Since the base model has 12 layers of encoders, tokens are fed into the first encoder, the output of the first encoder is then given as input for the second encoder, and so on until the last encoder. The last encoder, which is encoder 12 returns the embeddings for all tokens in the sentences. The representation size of each token is 768 in BERT-base model. For single-text classification applications, the BERT representation of the special classification token '[CLS]' encodes details about the whole text input. The single input text representation is fed into a small multilayer perceptron (MLP) consisting of fully connected (dense) layers to produce the

distribution of all discrete label values [79]. The Next Sentence Prediction (NSP) task allows BERT to learn relationships between sentences by predicting if the next sentence in a pair is the true next or not. For this 50%, correct pairs are supplemented with 50% random pairs and the model is trained. BERT trains both Masked language modelling (MLM) and NSP objectives simultaneously. MLM is a self-supervised pretraining task which is extensively used in natural language processing for learning text representations. MLM trains a model to predict a random sample of input tokens that have been replaced by a [MASK] placeholder in a multi-class setting over the entire vocabulary [80].
The structure of BERT is shown in Figure3.1.

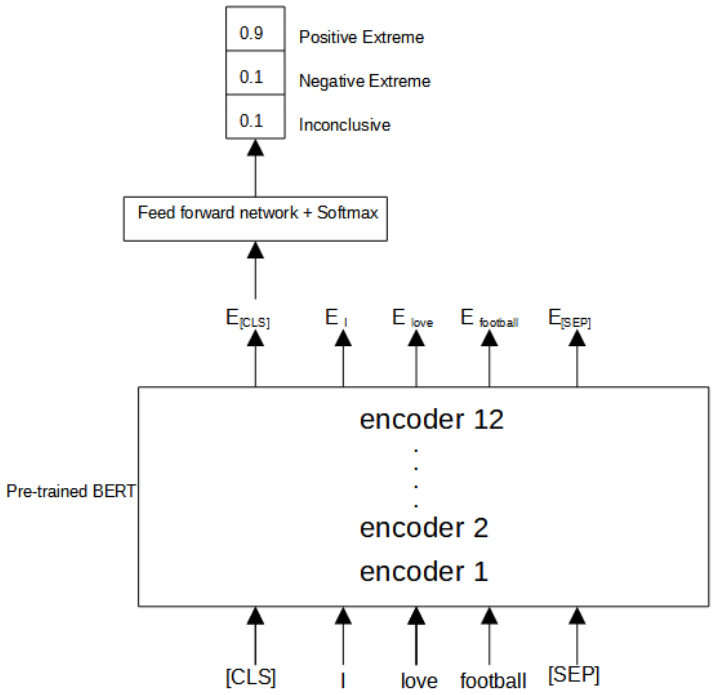

Figure 3.1: Tokenized text is embedded using 12 encoders in BERT and fed into a feed-forward network and softmax function to obtain the classification probabilities. [2]

3.2 RoBERTa

RoBERTa stands for Robustly optimized BERT approach [81] which is introduced by Facebook. It is a retraining of BERT with improved training methodology, relatively more data and computes power. The implementation of RoBERTa is the same as the Bert model with a small embedding tweak as well as a setup for Roberta pre-trained models. It has the same architecture as BERT but uses a byte-level pair encoding (BPE) as a tokenizer which is similar to GPT-2 and uses a different pretraining scheme. In particular, RoBERTa is trained with dynamic masking, FULL-SENTENCES without NSP loss, large mini-batches and a larger byte-level BPE.
To refine the training process, RoBERTa takes out the Next Sentence Prediction (NSP) task from BERT's pre-training and introduces dynamic masking so that the masked token changes during the training epochs. The experiment also showed that the larger batch-

training sizes were also found to be more useful in the training procedure. Importantly, In addition to BERT training 16GB of Books Corpus and English Wikipedia data, RoBERTa uses 160 GB of text for pre-training. The additional data includesCommonCrawl News dataset (63 million articles, 76 GB), Web text corpus (38 GB) and Stories from Common Crawl (31 GB). This combined with a massive 1024 V100 Tesla GPU's running for a day, resulted in pre-training of RoBERTa. As a consequence, RoBERTa outperforms both BERT and XLNet [3] on GLUE benchmark results.

3.3 XLNet

The XLNet [3] model is an extension of the Transformer-XL [82] model. It is pre-trained using an autoregressive method like OpenGPT [83] and bi-directional context modelling of BERT by maximizing the anticipated likelihood over all permutations of the input sequence factorization order. OpenGPT Transformer learns using left-to-right the text representation for natural language generation, while BERT uses a bidirectional transformer for natural language understanding.

XLNet is a generalized autoregressive (AR) language modelling method that uses a permutation language modelling objective to combine the advantages of AR and autoencoding (AE) methods . The XLNet neural architecture is built to work effortlessly, and harmoniously with the AR objective, including integrating Transformer-XL and the careful design of the two-stream attention mechanism. BERT is an Autoencoding (AE) based model, while XLNet is an Auto-Regressive (AR) that uses a permutation language modelling. The permutation operation during pre-training allows the context to include tokens from both left and right, making it a generalized order-aware autoregressive language model. The proposed XLNet architecture is pre-trained using nearly 10 times more data than the original BERT. It is also trained with a batch size eight times larger for half as many optimization steps, thus making it four times more sequences in pretraining compared to BERT. XLNet achieves substantial improvement over previous pretraining objectives on various tasks. It is claimed that the XLnet outperforms BERT on 20 tasks, often by a large margin.

The architecture of XLNet is shown in Figure3.2. In order to predict the word token in position 1 in a permutation 3-2-4-1, a content stream is made by joining together the positional embeddings and token embeddings of all previous words (3, 2, 4), and then a query stream is created by adding the content stream and the positional embedding of the word to be predicted (word in position 1), and in the end, the model makes the prediction based on information from the query stream. The client model is available as XLNet-Large and XLNet-Base. XLNet-Large Cased consists of 24-layer, 1024-hidden and 16-heads while

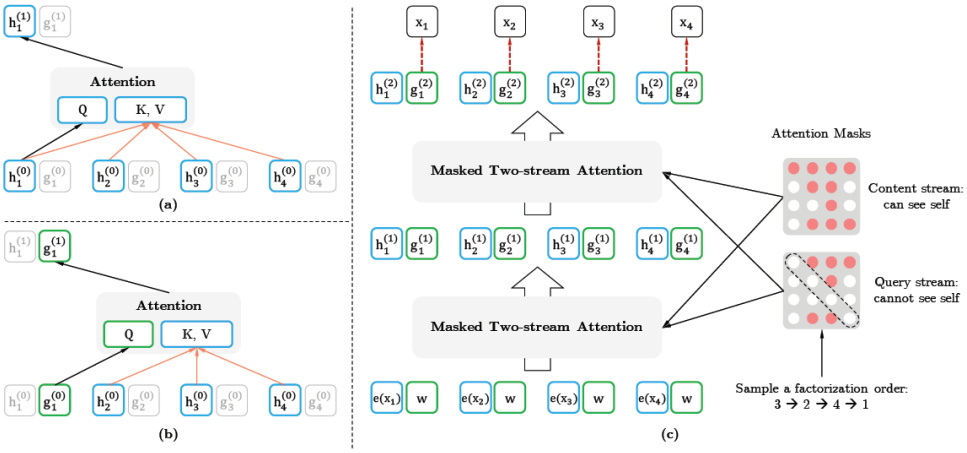

Figure 3.2: The architecture of XLNET model: (a)Content stream attention, which is the same as the standard self-attention. (b): Query stream attention, which does not have access to information about the content. (c): Overview of the permutation language modelling with two-stream attention. [3]

3.4 Conclusion

This chapter3 outlines the proposed method to address the discussed problem in the previous chapter2. It includes BERT, RoBERTa and XLNet models that will be used to carry out experiments and generate results. The models are briefly discussed by explaining their working, architecture and comparison with each other.

Chapter 4

Implementation and Results

This chapter presents important details regarding the experimental setup and briefly summarises the results. Firstly the dataset is presented along with the stages and processes involved in preparing the dataset. Secondly, fine-tuning models are given, and finally, the results are given under the sub-heading of dangerous events, top-level dangerous events, and sub-level dangerous events.

4.1 Experimental Setup

The main components of the experiment include dataset, BERT, RoBERTa, and XLNet. In the following subsections, each one is briefly described.

4.1.1 Dataset

Most of the available event datasets are either small in size or imbalanced. For instance, the most widely-used ACE 2005 English dataset [67], only contains 599 documents and 5,349 annotated instances. Furthermore, due to the deep-rooted data imbalance problem caused by the complexity of annotating events, 20 of its 33 event types only contain fewer than 100 annotated instances. As the latest deep learning-based models are usually data-hungry, these small-scale datasets are insufficient to train the model. Moreover, the covered event types in existing datasets are limited. The ACE 2005 English dataset only contains eight event types and 33 specific sub-types. The coverage of these datasets is low for general domain events, which results in the models trained on these datasets cannot being easily transferred and applied to general applications. Therefore, for this experiment, we use "MAVEN: A Massive General Domain Event Detection Dataset" [69]. MAVEN contains 4,480 documents sourced from Wikipedia that includes 118,732 instances of event mention from 168 different event types. Since this work aims to detect specific dangerous events, we manually filter out the irrelevant events data. The original dataset contains three files train, validation and test. The test file is ignored as it doesn't contain labels. The reaming files are filtered to separate dangerous events. These files will be joined together to carry out the experiment to increase the size of the dataset. The dataset contains 17309 entries of the train and 4103 entries of the validation data, totalling 21,412 entries for the whole dataset. This dangerous event dataset is then further divided into three groups for the sake of experimentation, which include the following:

- Dangerous events
- Top-level dangerous events
- Sub-level dangerous events

The dangerous event dataset includes twenty different types of dangerous events. The events include Catastrophe, Attack, Hostile encounter, Killing, Destroying, Bodily harm, Death, Damaging, Military operation, Defending, Use firearm, Dispersal, Violence, Arrest, Terrorism, Committing crime, Quarreling, Warning, Change sentiment, Protest, Bearing arms, Kidnapping, Rewards and punishments, Adducing, Imposing obligation, Prison, Revenge, Rite, Lighting, Suspicion, Incident, Risk and Emergency classes. The dataset may contain some irrelevant events, such as in the case of "change in sentiment event", it can be from happy to angry or the opposite. Since there is no automatic way of verifying each occurrence of every event, it is presumed that it still reflects the general theme of the class label and will not affect the model accuracy. The top-level dangerous events consist of 3 classes which are "Action-based dangerous events", "scenario-based dangerous events" and "sentiment-based dangerous events".

4.1.2 Loading and Pre-Processing

The pre-processing phase consists of various stages. Firstly, the dataset that is in JSONL format is unpacked into CSV format. The entries in the dataset contain a document and a list of events for each sentence in every document. While unpacking the JSONL file, the document is broken down into sentences and the event in each row. The rest of the information in columns is kept as it is. Secondly, the original dataset contains three files: train, validation and test. The text file doesn't have event information; therefore, it is dropped. The other two files are combined into one file to enlarge the dataset for training and validation. This one big file will be divided into training and validation data. Thirdly, the original dataset contains 168 different event types. These events are manually filtered out, and only events that can be regarded as dangerous are kept. The dangerous events constitute 20 total events. These dangerous events are divided into three main categories, as mentioned earlier, for the purpose of experimentation. The number of entries for each event varies greatly. Since the number of events is bigger, this can lead to underperformance of the model. To cope with the issue, and under-sampling technique has been used where all the event entries are trimmed and equal to the event with the lowest number of records. For instance, in the dangerous events classification task, the kidnapping class has 104 entries after applying the under-sampling technique to the dataset. The other 19 events also have 104 entries making the dataset balanced to feed into the model. Furthermore, the text is tokenized and converted into a tensor for training the model.

4.1.3 Fine-tuning and Hyperparameter tuning

Fine-tuning is a well-known technique for transfer learning. The target model duplicates all model layouts with their parameters from the original model except the output layer and fine-tunes these parameters based on the target dataset. On the contrary, the target model's output layer must be trained from the beginning [79]. Fine-tuning means taking the weights of a trained neural network and using it as initialization for a new model being

trained on data from the same domain (often, e.g. images). It is used to speed up the training and overcome a small dataset size.

In machine learning, hyperparameter optimization or tuning is the question of selecting a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value controls the learning process. The datasets used for fine-tuning three deep-learning models, BERT, RoBERTa and XLNet, are the same. The parameters used are similar to keep the experiment consistent. The details of each model are discussed below.

BERT base model is used for the experimentation that incorporates an encoder with 12 transformer blocks, 12 self-attention heads, and 768 units of hidden embedding parameters, a sequence of hidden states of the last layer of the model. The model is fine-tuned by dividing 70% of data for training and 30% for validation and testing purposes. The parameters selection includes a batch size of 16 with ten epochs for the training model. Hyperparameter tuning includes learning rate and optimizer. AdamW optimizer is used in the experiment, which was also used in the original BERT model for pre-training. The learning rate is chosen manually, which is 2e-5. This learning rate is required for BERT to overcome the terrible problem of forgetting. The smaller learning rates may let the model learn a more optimal or even globally optimal set of weights, while higher rates can cause failure to converge on the training set.

RoBETa utilizes a split of 70% and 30% of datasets similar to BERT. The difference in this split is the absence of a validation set, which is merged into a test set and used for validation and testing purposes. The RoBERTa base model is used for the experiment that uses manually selected parameters of 10 epochs with a 3e-5 learning rate. The hyperparameter tuning also includes an optimizer which in this case is AdamW. The batch number is 16, fed into the model with the classification layer, which is equal to the number of labels in the dataset.

XLNet follows the same dataset division for training and testing purposes. The model used for this experiment is "xlnet-base-cased". The number of training epochs is 10, with the batch size for the model is kept at 16. The model is hyperparameter tuned using the AdamW optimizer with a learning rate of 3-e5.

4.1.4 Performance metrics

This work uses common performance metrics for evaluating the trained model results. It includes accuracy, F1-score, precision, and recall. Accuracy is the fraction of correct predictions, the number of hits divided by the total number of predictions. Precision is the ability of the classifier to not predict the false label or value. The recall is the capability of the classifier to search all the positive samples. It can also be given as the fraction of the relevant labels successfully predicted. The F1-score, the balanced F-score or F-measure, is the weighted average of the precision and recall. The best value of the F1-score is 1, and the lowest is 0. The F-score is also used for calculating classification problems with more than two classes, called multi-class classification. These two classes are called micro-averaging and macro-averaging. The final score is obtained by micro-averaging, which is biased by class frequency, whereas macro-averaging takes all classes equally important. Another

type of F1-score is the weighted average. There are three types of averages, namely micro, macro and weighted. Micro-average evaluates metrics globally by calculating the total true positives, false negatives, and false positives. Macro-average computes metrics for every label and finds their unweighted mean. The imbalanced labels are not taken into account. Weighted average determines metrics for each label. It finds their average weighted by support. This changes the macro average to reckon an unbalanced label which can lead to F-score that is distinct from precision and recall. Support is the number of actual class occurrences in the specified dataset. Unbalanced support in the training data may indicate structural weaknesses in the scores of the reported classifiers and could indicate the need for stratified sampling or re-balancing. The support does not variate between the models but rather diagnoses the evaluation process.

4.2 Results

This section outlines the results obtained for each experiment. The experiments are given under each category of events for which the results from each model are provided. The categories are dangerous events, top-level dangerous events, and sub-level dangerous events.

4.2.1 Dangerous Events

Dangerous events are extracted manually from MAVEN dataset [69]. It contains 20 events, shown in Figure 4.1 along with their original distribution. As the number of events highly varies for each class, sampling is essential for balancing the data to get the true prediction of events from the model. Otherwise, the model will learn well to predict the class with higher entries and may fail to provide any result for the smaller classes. Therefore, the events with higher numbers are trimmed with respect to the lowest entry, "kidnapping" using a technique known as under-sampling. This yields an equal number of records for each class yet still provides a sizeable dataset for training the model.

BERT model provides an overall 60% of accuracy on the whole dataset of dangerous events. It runs 10 epochs for training the model using the batch size of 16. The total number of entries in the dataset is 2080, where the training set size is 1456, for validation is 312, and 312 for testing. The classification report of the experiment on dangerous events is given in Table 4.1. It shows good results for some categories such as "Death", "Warning", "Bearing_arms" and "Kidnapping " while yielding low results for some classes such as "Hostile_encounter","Attack","Killing","Military_operation","Use_firearm", and "Terrorism". The low performance of BERT for some classes can be attributed to data quality. It is expected that improving the data quality of these classes will significantly increase the performance results and increase the accuracy of the whole events dataset.

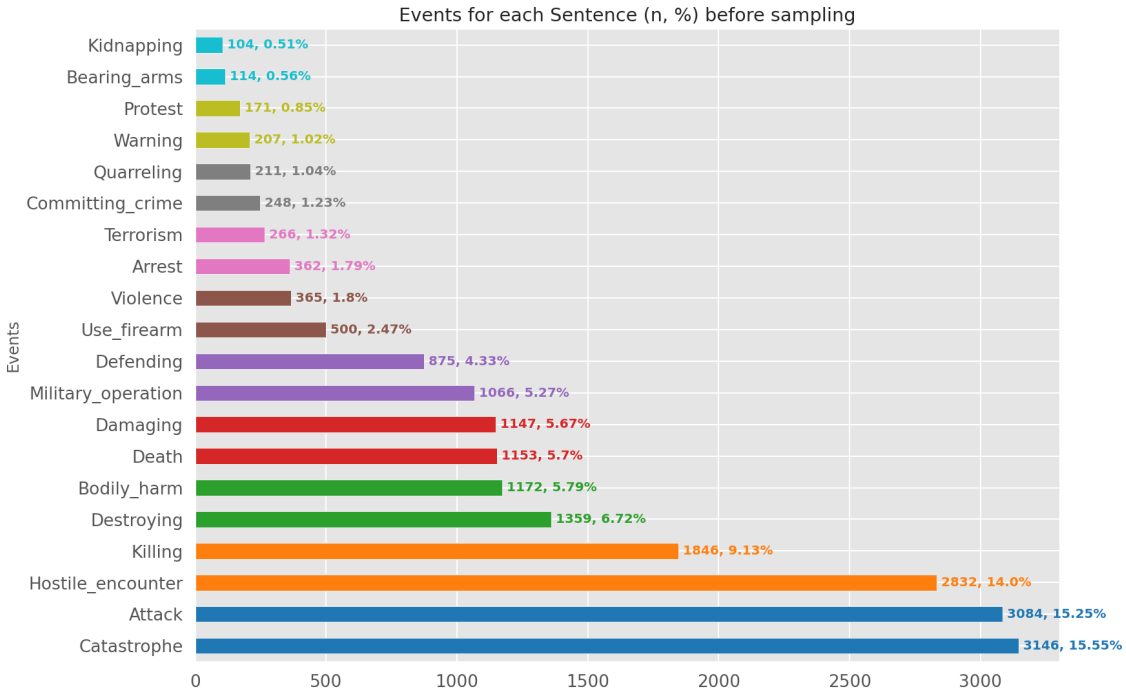

Figure 4.1: Total entries of dangerous events and their distribution in the original dataset

Table 4.1: Classification report of dangerous events using BERT

	Dangerous Events			
Events	**Precision**	**Recall**	**F-score**	**Support**
Catastrophe	0.67	0.53	0.59	15
Damaging	0.69	0.73	0.71	15
Destroying	0.50	0.53	0.52	15
Protest	0.67	0.62	0.65	16
Death	0.73	0.50	0.59	16
Warning	0.75	1.00	0.86	15
Hostile_encounter	0.35	0.38	0.36	16
Attack	0.25	0.20	0.22	15
Bodily_harm	0.62	0.53	0.57	15
Killing	0.50	0.44	0.47	16
Military_operation	0.54	0.44	0.48	16
Defending	0.56	0.62	0.59	16
Bearing_arms	0.70	0.44	0.54	16
Use_firearm	0.55	0.80	0.65	15
Committing_crime	0.62	0.62	0.62	16
Quarreling	0.67	0.62	0.65	16
Arrest	0.65	0.81	0.72	16
Violence	0.67	0.88	0.76	16
Kidnapping	0.75	0.75	0.75	16
Terrorism	0.53	0.53	0.53	15
accuracy			0.60	312
macro avg	0.60	0.60	0.59	312
weighted avg	0.60	0.60	0.59	312

RoBERTa provides similar results as BERT in terms of overall accuracy, which is 59%. It performs good for predicting certain categories, especially in the case of "Warning" and "Defending" where recall, precision and f1-score are above 70%. It performs worse

for predicting the categories of “Attack ”, “Hostile_encounter” and “Military_operation” with performance ranging between 19% to 38%. The performance of RoBERTa can be checked in Table4.2.

Table 4.2: Classification report of dangerous events using RoBERTa

	Dangerous Events			
Events	**Precision**	**Recall**	**F-score**	**Support**
Catastrophe	0.71	0.55	0.62	31
Damaging	0.66	0.61	0.63	31
Destroying	0.54	0.48	0.51	31
Protest	0.59	0.74	0.66	31
Death	0.72	0.66	0.69	32
Warning	0.82	0.90	0.86	31
Hostile_encounter	0.38	0.35	0.37	31
Attack	0.19	0.19	0.19	31
Bodily_harm	0.48	0.42	0.45	31
Killing	0.42	0.42	0.42	31
Military_operation	0.31	0.35	0.33	31
Defending	0.71	0.71	0.71	31
Bearing_arms	0.63	0.77	0.70	31
Use_firearm	0.64	0.68	0.66	31
Committing_crime	0.70	0.59	0.64	32
Quarreling	0.64	0.50	0.56	32
Arrest	0.77	0.55	0.64	31
Violence	0.56	0.74	0.64	31
Kidnapping	0.74	0.78	0.76	32
Terrorism	0.61	0.71	0.66	31
accuracy			0.59	624
macro avg	0.59	0.59	0.58	624
weighted avg	0.59	0.59	0.58	624

XLNet model performs lower than the previously mentioned models of BERT and RoBERTa. It yields only 54% overall accuracy for the experiment. It performs well for unusual categories of events which other models lack, as in the case of “Catastrophe ”, “Protest”, “Death”, “Warning” and “Terrorism” with, on average, providing results of around 70% for precision, recall and F1-score. While for the event categories of “Hostile_encounter ”,“Attack”,“Military_operation” and “Destroying”, it provides very low results. The model’s performance for each category of classified events is given in Table. 4.3.

Table 4.3: Classification report of dangerous events using XLNet

	Dangerous Events			
Events	**Precision**	**Recall**	**F-score**	**Support**
Catastrophe	0.71	0.67	0.69	15
Damaging	0.67	0.67	0.67	15
Destroying	0.39	0.47	0.42	15
Protest	0.71	0.62	0.67	16
Death	0.71	0.62	0.67	16
Warning	0.79	0.73	0.76	15
Hostile_encounter	0.27	0.25	0.26	16
Attack	0.27	0.20	0.23	15
Bodily_harm	0.40	0.27	0.32	15
Killing	0.32	0.44	0.37	16
Military_operation	0.33	0.44	0.38	16
Defending	0.46	0.38	0.41	16
Bearing_arms	0.56	0.56	0.56	16
Use_firearm	0.50	0.53	0.52	15
Committing_crime	0.78	0.44	0.56	16
Quarreling	0.53	0.50	0.52	16
Arrest	0.64	0.56	0.60	16
Violence	0.56	0.88	0.68	16
Kidnapping	0.68	0.81	0.74	16
Terrorism	0.71	0.80	0.75	15
accuracy			0.54	312
macro avg	0.55	0.54	0.54	312
weighted avg	0.55	0.54	0.54	312

Experiments reflect that the BERT provides the optimum results for the dangerous events category. There are some common patterns being repeated in the results. It includes the low performance of all models for some event types such as “Hostile_encounter”, “Attack” etc. Improving the data for these events can significantly uplift model performance, which will also increase the overall accuracy of models on the whole dataset.

4.2.2 Top-level Dangerous Events

Top-level dangerous events contain three main categories of different events. The dangerous events are manually grouped under these categories based on similarities. They share a common element in their nature, such as Action-based dangerous events reflecting the presence of an action that can be dangerous. Similarly, scenario-based events are occurrences where there is no manual factor purposefully directing them, but they turn into dangerous events such as tornadoes, tsunamis etc. Sentiment-based events are related to violent and extreme sentiments and can threaten the safety of society, groups or individuals. The low number of labels helps the model learn the features well and provide better results. The distribution of top-level events is shown in Figure 4.2.

BERT provides 71% accuracy for the top-level dangerous events. The model’s precision, recall and f1-score stay above 70% for scenario-based and sentiment-based events, while the action-based category performs slightly lower with results between 65% to 67%. The overall performance of the model can be regarded as satisfactory with regards to performance. The classification report of the BERT model is given in Table 4.4

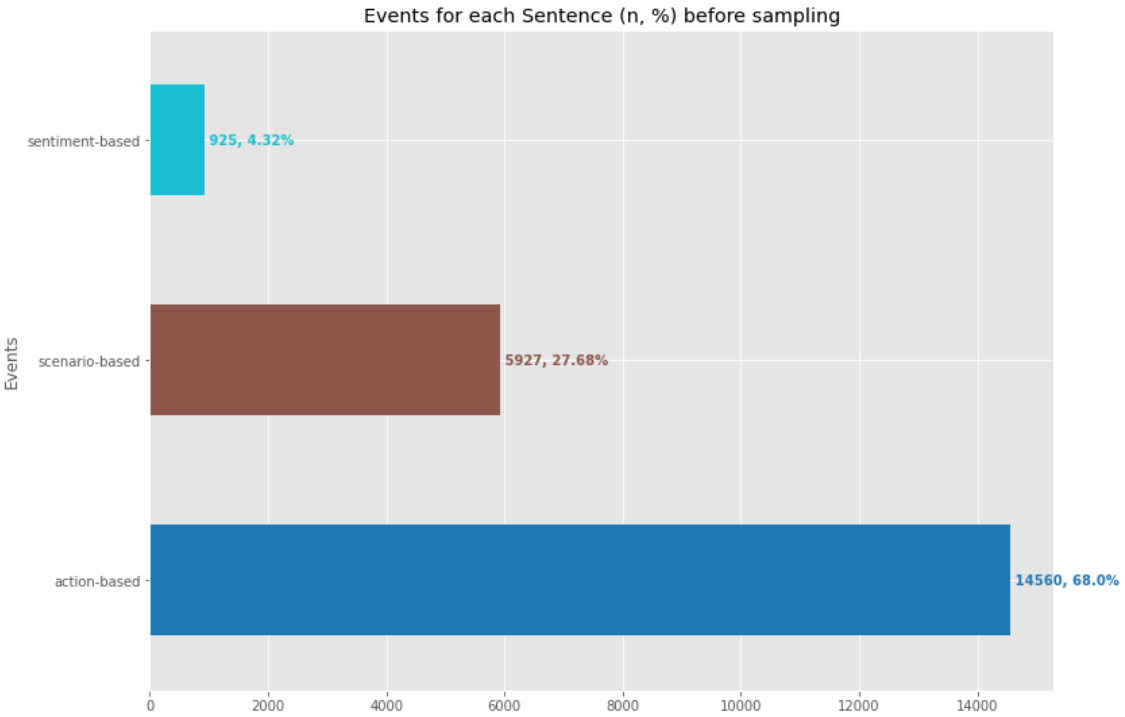

Figure 4.2: Top-level dangerous events and their original distribution dataset

Table 4.4: Classification report of top-level dangerous events using the BERT model

	Top-level Dangerous Events			
Events	**Precision**	**Recall**	**F-score**	**Support**
scenario-based	0.72	0.73	0.73	139
action-based	0.67	0.65	0.66	139
sentiment-based	0.73	0.73	0.73	139
accuracy			0.71	417
macro avg	0.70	0.71	0.70	417
weighted avg	0.70	0.71	0.70	417

RoBERTa outperforms BERT in Top-level dangerous events and achieves 74% accuracy on the test dataset. In terms of each category, it provides precision, recall and f1-score between 70% and 81%. The macro and weighted average also stand out at 74%. The classification report of the RoBERTa model is provided in Table. 4.5.

Table 4.5: Classification report of top-level dangerous events using the RoBERTa model

	Top-level Dangerous Events			
Events	**Precision**	**Recall**	**F-score**	**Support**
scenario-based	0.78	0.72	0.75	277
action-based	0.71	0.70	0.70	278
sentiment-based	0.74	0.81	0.78	278
accuracy			0.74	833
macro avg	0.74	0.74	0.74	833
weighted avg	0.74	0.74	0.74	833

XLNet yields results very similar to the BERT model. It achieves all 70% scores for accuracy, macro and weighted average. The difference between precision, recall and f1-score is very close for each category. It is between 73% to 74% for scenario-based, 64% to

65% for action-based and 71% for sentiment-based dangerous events. The classification report of the model is provided below in Table. 4.6

Table 4.6: Classification report of top-level dangerous events using XLNet model

	Top-level Dangerous Events			
Events	**Precision**	**Recall**	**F-score**	**Support**
scenario-based	0.74	0.73	0.74	139
action-based	0.64	0.65	0.65	139
sentiment-based	0.71	0.71	0.71	139
accuracy			0.70	417
macro avg	0.70	0.70	0.70	417
weighted avg	0.70	0.70	0.70	417

The above results show that grouping similar events under one category can improve the model's performance. The same dataset set for dangerous events has been used, but the results for top-level DE are higher than the previous category. It validates the approach to classifying and renaming the higher number of classes into top-level classes.

4.2.3 Sub-level Dangerous Events

Sub-level dangerous events refer to events for each category of top-level dangerous events. These are granular events of each top-level category and are separately experiments to analyze the model performance. As mentioned earlier, the top-level events have certain similarities, and this division will help verify this fact. Sub-level events are categorized under action-based, scenario-based and sentiment-based dangerous events. The results for each category using three models are given in the next sub-sections.

4.2.3.1 Action-based Dangerous Events

Action-based events include "Destroying", "Death", "Hostile_encounter", "Killing", "Attack", "Bodily_harm", "Defending", "Military_operation", Use_firearm", "Committing_crime", "Rite", "Kidnapping" and "Terrorism". The total number of events types are 11 which are given in Figure 4.3

BERT yields a 62% score for accuracy, macro and the weighted average for action-based classification. It produce better results for "Death","Defending" and "Use_firearm " while yielding lower score for "Attack" and "Hostile_encounter". The classification report of the BERT model is provided in Table 4.7

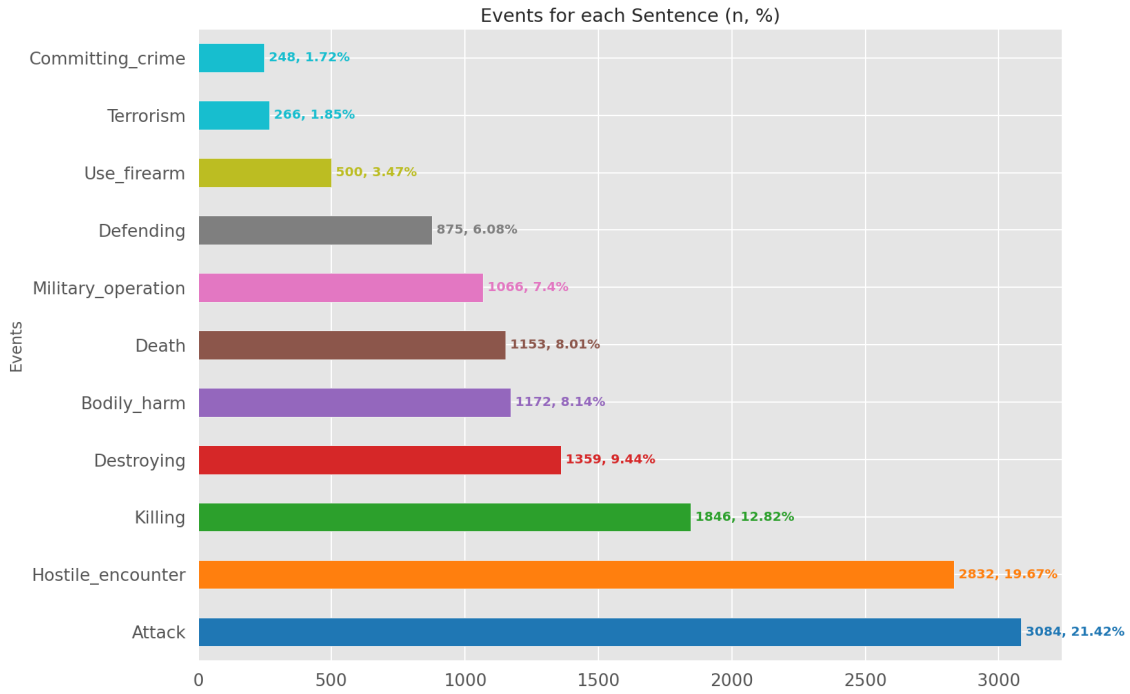

Figure 4.3: Action-based dangerous events and their number of occurrences in the dataset

Table 4.7: Classification report of action-based sub-level dangerous events using the BERT model

	Action-based Sub-level DE			
Events	**Precision**	**Recall**	**F-score**	**Support**
Destroying	0.71	0.59	0.65	37
Death	0.74	0.76	0.75	37
Hostile_encounter	0.49	0.55	0.52	38
Attack	0.43	0.35	0.39	37
Bodily_harm	0.49	0.54	0.51	37
Killing	0.51	0.50	0.51	38
Military_operation	0.58	0.49	0.53	37
Defending	0.71	0.76	0.73	38
Use_firearm	0.89	0.84	0.86	37
Committing_crime	0.63	0.65	0.64	37
Terrorism	0.64	0.78	0.71	37
accuracy			0.62	410
macro avg	0.62	0.62	0.62	410
weightedd avg	0.62	0.62	0.62	410

RoBERTa produce 62% overall accuracy for the dataset. It only performs better for "Use_firearm" while others remain lower. In terms of precision, only two events perform above 70% and in many cases, the difference between precision, recall and f1-score are big. The detailed classification report is given in Table 4.8

Table 4.8: Classification report of sub-level dangerous events using RoBERTa model

	Action-based sub-level DE			
Events	**Precision**	**Recall**	**F-score**	**Support**
Destroying	0.64	0.72	0.68	74
Death	0.66	0.81	0.73	74
Hostile_encounter	0.56	0.59	0.58	75
Attack	0.45	0.43	0.44	75
Bodily_harm	0.55	0.42	0.48	74
Killing	0.55	0.60	0.57	75
Military_operation	0.53	0.41	0.46	75
Defending	0.73	0.79	0.76	75
Use_firearm	0.86	0.68	0.76	74
Committing_crime	0.58	0.62	0.60	74
Terrorism	0.63	0.69	0.66	74
accuracy			0.61	819
macro avg	0.61	0.61	0.61	819
weightedd avg	0.61	0.61	0.61	819

XLNet yields lower results compared to the above two models. It gives an accuracy of 58% for all categories. The difference between different performance matrices is big while it predicts well for the event of "Use_firearm" with approximately 80%. The classification report of XLNet is provided in Table. 4.9

Table 4.9: Classification report of action-based sub-level dangerous events using XLNet model

	Action-based Sub-level DE			
Events	**Precision**	**Recall**	**F-score**	**Support**
Destroying	0.60	0.41	0.48	37
Death	0.64	0.76	0.69	37
Hostile_encounter	0.50	0.53	0.51	38
Attack	0.41	0.38	0.39	37
Bodily_harm	0.45	0.41	0.43	37
Killing	0.51	0.53	0.52	38
Military_operation	0.58	0.51	0.54	37
Defending	0.69	0.66	0.68	38
Use_firearm	0.79	0.81	0.80	37
Committing_crime	0.57	0.57	0.57	37
Terrorism	0.59	0.81	0.68	37
accuracy			0.58	410
macro avg	0.58	0.58	0.57	410
weighted avg	0.58	0.58	0.57	410

4.2.3.2 Scenario-based Dangerous Events

This low-level category includes total seven event types which are "Warning", "Catastrophe", "Damaging", "Bearing_arms", "Quarreling" and "Arrest". The original number of occurrences for each event is imbalanced which can be viewed in Figure 4.4. The dataset is balanced for training the model using the under-sampling technique.

BERT give overall accuracy of 85% for the scenario-based event classification task. It performs exceptionally well for the event types "Arrest" and "Damaging". The classification report of the BERT classification model is given in Table 4.10

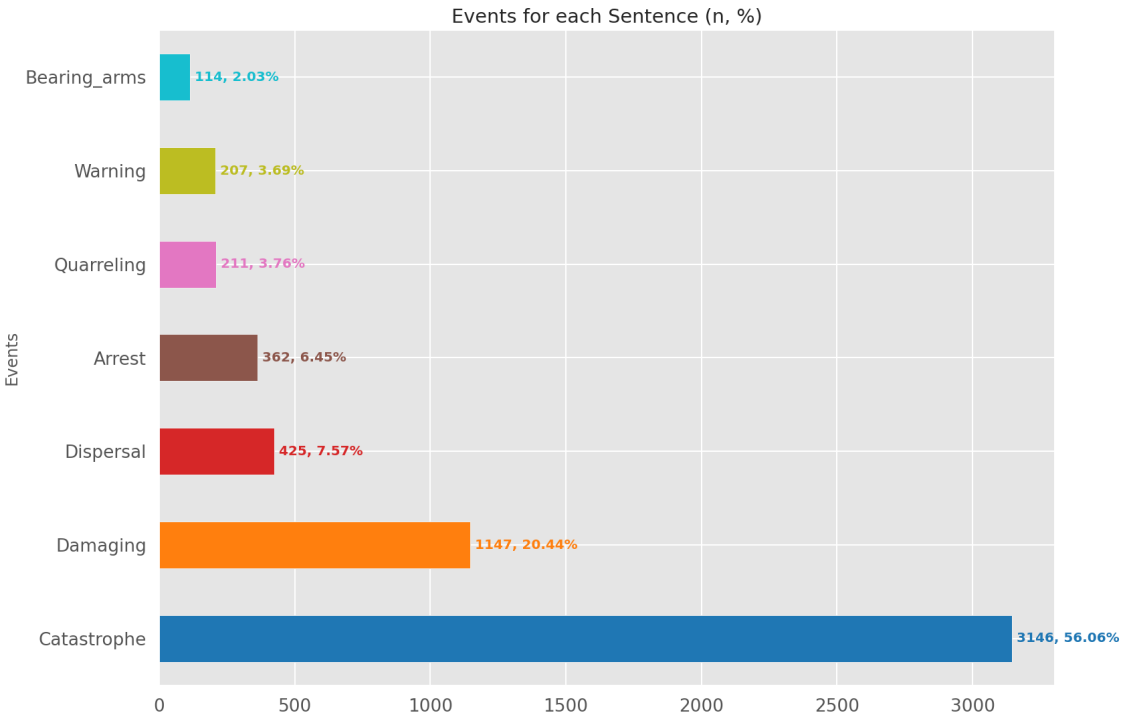

Figure 4.4: Scenario-based dangerous events and their number of occurrences in the dataset

Table 4.10: Classification report of scenario-based sub-dangerous events using the BERT model

	Scenario-based Sub-level DE			
Events	**Precision**	**Recall**	**F-score**	**Support**
Catastrophe	0.82	0.53	0.64	17
Damaging	0.72	0.76	0.74	17
Warning	0.94	1.00	0.97	17
Bearing_arms	0.86	1.00	0.92	18
Quarreling	0.84	0.94	0.89	17
Arrest	1.00	1.00	1.00	17
Dispersal	0.75	0.71	0.73	17
accuracy			0.85	120
macro avg	0.85	0.85	0.84	120
weighted avg	0.85	0.85	0.85	120

RoBERTa provides a stable score of 83% for accuracy, macro and the weighted average for scenario-based sub-level dangerous events. Similar to the BERT model, it gives the top score for the events "Arrest". The results are outlined in Table 4.11 as a classification report.

Table 4.11: Classification report of scenario-based sub-level dangerous events using RoBERTa model

	Scenario-based Sub-level DE			
Events	**Precision**	**Recall**	**F-score**	**Support**
Catastrophe	0.72	0.68	0.70	34
Damaging	0.84	0.79	0.82	34
Warning	0.89	0.94	0.91	34
Bearing_arms	0.87	0.97	0.92	35
Quarreling	0.74	0.85	0.79	34
Arrest	0.97	0.91	0.94	34
Dispersal	0.77	0.66	0.71	35
accuracy			0.83	240
macro avg	0.83	0.83	0.83	240
weighted avg	0.83	0.83	0.83	240

XLNet scores lower than BERT and RoBERTa. It gives 82% accuracy for the classification of events. The classification report of scenario-based sub-level events is given in Table 4.12

Table 4.12: Classification report of scenario-based sub-level dangerous events using XLNet model

	Scenario-based Sub-level DE			
Events	**Precision**	**Recall**	**F-score**	**Support**
Catastrophe	0.77	0.59	0.67	17
Damaging	0.74	0.82	0.78	17
Warning	0.88	0.88	0.88	17
Bearing_arms	0.77	0.94	0.85	18
Quarreling	0.88	0.88	0.88	17
Arrest	0.84	0.94	0.89	17
Dispersal	0.85	0.65	0.73	17
accuracy			0.82	120
macro avg	0.82	0.82	0.81	120
weighted avg	0.82	0.82	0.81	120

4.2.3.3 Sentiment-based Dangerous Events

Sentiment-based group contains total five typos of different events. The entries for each event in dataset is shown in Figure 4.5. The events include "Change_sentiment", "Protest", "Revenge", "Rewards_and_punishments" and "Violence".

BERT produces 81% accuracy for sentiment-based dangerous events. The details of each event and their respective precision, recall, and f-1 score is shown in Table 4.13.

Table 4.13: Classification report of sentiment-based sub-level dangerous events using the BERT model

	Sentiment-based Sub-level DE			
Events	**Precision**	**Recall**	**F-score**	**Support**
Protest	0.66	0.81	0.72	26
Change_sentiment	0.76	0.66	0.70	29
Violence	0.89	0.93	0.91	55
Rewards_and_punishments	0.92	0.75	0.83	16
Revenge	0.88	0.78	0.82	9
accuracy			0.81	135
macro avg	0.82	0.78	0.80	135
weighted avg	0.82	0.81	0.81	135

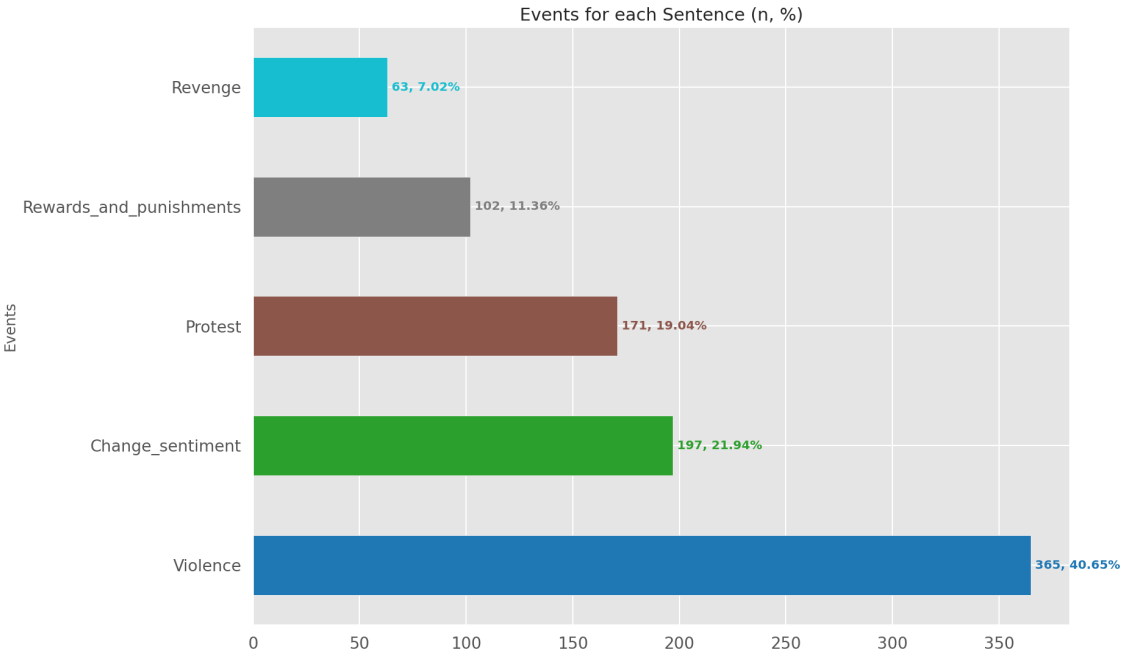

Figure 4.5: Sentiment-based dangerous events and their number of occurrences in the dataset

Roberta produce 69% of accuracy for sentiment-based classification of events. The events "Protest" and "Change_sentiment" score 58% for precision, recall and f1-score. The detailed result of the RoBERTa model is outlined in Table 4.14.

Table 4.14: Classification report of sentiment-based sub-level dangerous events using RoBERTa model

	Sentiment-based Sub-level DE			
Events	**Precision**	**Recall**	**F-score**	**Support**
Protest	0.60	0.63	0.62	19
Change_sentiment	0.61	0.58	0.59	19
Violence	0.70	1.00	0.83	19
Rewards_and_punishments	0.73	0.42	0.53	19
Revenge	0.89	0.89	0.89	19
accuracy			0.71	95
macro avg	0.71	0.71	0.69	95
weighted avg	0.71	0.71	0.69	95

XLNet slightly better results in the RoBERTa model. It provides 77% accuracy in general for the task of classifying sentiment-based events. The classification report for sentiment-based events using XLNet is shown in Table 4.15

Table 4.15: Classification report of scenario-based sub-level dangerous events using XLNet model

	Sentiment-based Sub-level DE			
Events	**Precision**	**Recall**	**F-score**	**Support**
Violence	0.86	0.60	0.71	10
Change_sentiment	0.70	0.78	0.74	9
Protest	0.67	0.60	0.63	10
Rewards_and_punishments	0.77	1.00	0.87	10
Revenge	0.89	0.89	0.89	9
accuracy			0.77	48
macro avg	0.78	0.77	0.77	48
weighted avg	0.78	0.77	0.76	48

The above approach regarding sub-level DE is based on training the model with classes of similar event types. This can help the model learn better as there are some similarities between the event types. The above results also prove this point and confirm that focusing on similar events classification can give solid results that can be useful in many ways.

4.3 Discussion and Future Work

The discussion section highlights the main findings of experimentation and takeaways. First, the important point is the need for an events dataset. This work has proven the usefulness of the events dataset for the training classification model trained for detecting events in social media texts. This work also achieved significant results using the MAVEN [69] dataset. There are a limited number of events datasets available, and the need for more events datasets is inescapable, especially in case of dangerous events. Collecting a good amount of dangerous events can be an intensive task. For that purpose, specific datasets can be complied with to build one.

BERT performs significantly well in terms of the transformer model for classifying and detecting dangerous events. The other models, RoBERTa and XLNet, provide better results for only limited and specified tasks.

As it can be visualized from Figure 4.6, BERT outperforms other models except in the case of top-level dangerous events where RoBERTa outperforms BERT. In the case of sub-level sentiment-based dangerous events, RoBERTa under-performs other models with significant differences in results. Furthermore, it can be seen that the division of events and grouping of the events under respective categories significantly improves the result. This is the case in scenario-based and sentiment sub-level dangerous events. Overall, the top-level category outperforms the dangerous events category, while the sub-level category performs better than the top-level DE in most cases. Although it seems easy to compare different categories of events, the root cause of the model performance lies with the quality of the dataset and each event provided in the dataset. Certain types of events in certain categories perform well for all models, such as the events “bearing_arms” and “Warning” in the dangerous events category, while others such as “Attack” and “Military_operation” scores very low. This can be a hypothetical case of the inferior type of data available in the dataset that the model could not train and learn properly. The insignificant data in the dataset can be improved since the “Attack” event qualifies as a good type of event to be

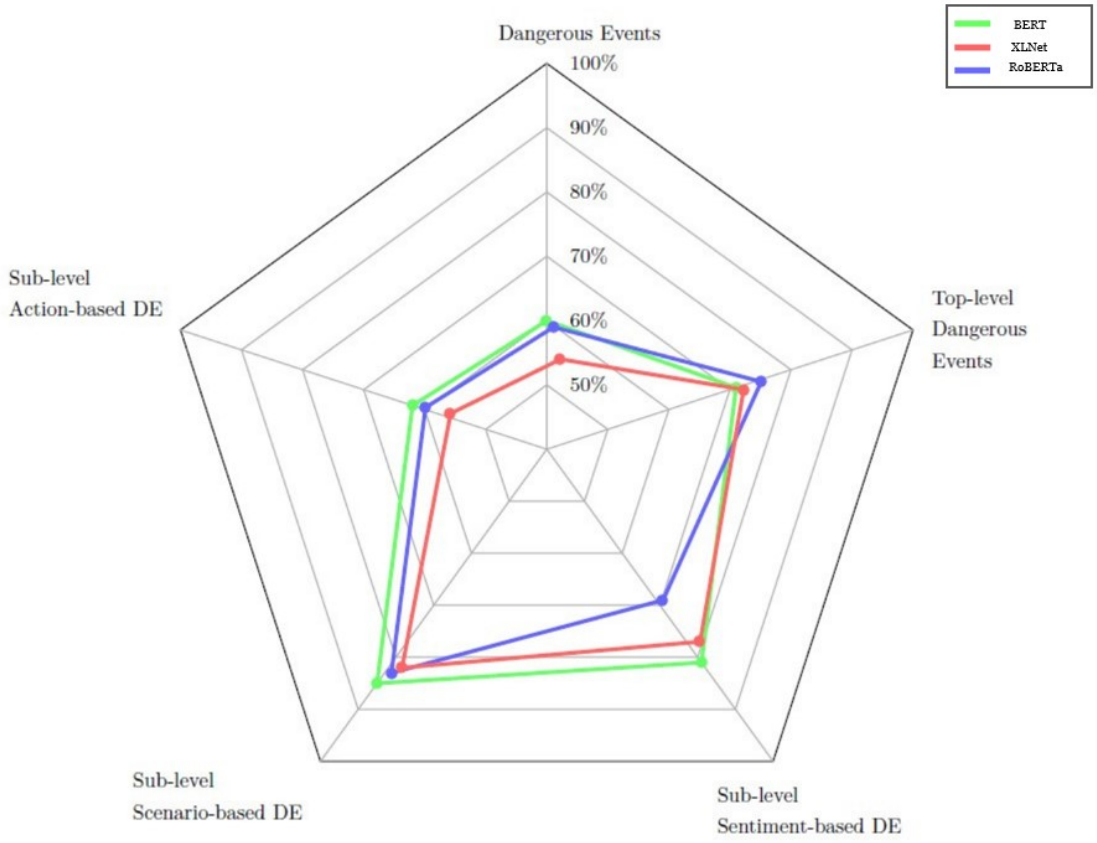

Figure 4.6: Accuracy of BERT, RoBERTa and XLNet for Dangerous Events, Top-level DE, Action-based Sub-level DE, Scenario-based Sub-level DE and Sentiment-based Sub-level DE

included in the category of action-based dangerous events. The data for each event type can significantly improve the overall performance of models and provide better results.
For future work, it is intended to build a system that detects dangerous events and maps the events over time. Event detection is an important task, but it is equally, if not less important, to track those events in real-time to counteract and apply strategies to deduce extremely important information. Event tracking is tricky because the time dimension is a very important factor. The data obtained from social networks usually contain information about time, date, user, and location along with the text. These details can help us map the time series plot for visualizing the activity of any user of interest or the evolution of any event concerning time. If a threshold is set on such data plots, they can create an alert to warn the sensitivity of any dangerous situation. For example, The time series plot can help identify the event if the event has been mentioned above the threshold and is dangerous in nature. It can alert the authorities to prepare to stop any dangerous event from taking place.

Chapter 5

Conclusion

Dangerous events are part of normal daily life that happen often. It is necessary to detect those events and prevent them from protecting the interests of any individual, group, or society. Related work shows much of the work is done for detecting different events in various domains. That work is sparse and spread around different terms referring to similar things. This way, instead of being limited to only one term, we expand our scope by approaching them under a single term, “Dangerous Events”. We then review the literature and find every work relatable under the defined term ”dangerous events”. For example, disaster event detection, violence event detection, and disaster event detection are considered dangerous events in our work. Hence, simplify the task by supposing the task as “Dangerous Event Detection”. There are few datasets available for events. The famous event datasets provide a small number of different events and are limited in scope. The dataset used in this work used is the latest and one of its kind. Yet, we came across different restrictions in this work. This implies the need for an events dataset especially related to dangerous events. Another prospect of building a dangerous events dataset can combine specific events from different datasets. It is like building a Wikipedia of dangerous events that can help refine the solution to the scarcity of events datasets.
There are many methods and approaches for event detection based on requirements and goals. These methods are supervised unsupervised and semi-supervised methods. The approaches include document-pivot, feature-pivot, and topic modeling approaches. We review these approaches and methods as the fundamental basis for our work. Based on our findings and results, we propose a deep learning classification algorithm that combines the document pivot and feature pivot approaches. The BERT, RoBERTa, and XLNet show significant potential for detecting dangerous events. Overall, the BERT model outperforms in most experiments except in the case of top-level DE, where RoBERTa has the best results. The various dimension of the experiment adds insight to understanding the efficiency of chosen models and the usability of the dataset. It can be deduced that the events that yield low performance using these models may be improved by updating these events with better data and annotation. Finally, comparing these models for the dataset also gives away the right distribution of events data as similar events may yield better results in many cases rather than having distant events that challenge the ability of models to perform beyond their capacity. For future work, it is proposed to explore state-of-the-art methods for the purpose of event tracking and prediction. A brief review has been presented in this work to provide the basis for defining the problem statement and for outlining future work.

Bibliography

[1] W. Aldyani, F. K. Ahmad, and S. Kamaruddin, "A survey on event detection models for text data streams," *Journal of Computer Science*, vol. 16, pp. 916–935, 07 2020. vii, 2, 15, 18, 19, 20, 21

[2] M. L. Jamil, S. Pais, J. Cordeiro, and G. Dias, "Detection of extreme sentiments on social networks with bert," *Social Network Analysis and Mining*, vol. 12, no. 1, pp. 1–16, 2022. vii, 25, 27

[3] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *CoRR*, vol. abs/1906.08237, 2019. [Online]. Available: http://arxiv.org/abs/1906.08237 vii, 28, 29

[4] C. Messaoudi, Z. Guessoum, and L. Romdhane, "Opinion mining in online social media: A survey," *Social Network Analysis and Mining*, vol. 12, 12 2022. 1

[5] X. Fu, M. R. Padmanabhan, R. G. Kumar, S. Basu, S. Dorius, and A. Pavan, "Measuring the impact of influence on individuals: Roadmap to quantifying attitude," *CoRR*, vol. abs/2010.13304, 2020. [Online]. Available: https://arxiv.org/abs/2010.13304 1

[6] S. R. Department, "Social media - statistics & facts," 2021. [Online]. Available: https://www.statista.com/topics/1164/social-networks/#dossierSummary___chapter1 1

[7] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" *Association for Computing Machinery*, 2010. [Online]. Available: https://doi.org/10.1145/1772690.1772751 1

[8] H. H. Khondker, "Role of the new media in the arab spring," *Globalizations*, vol. 8, no. 5, pp. 675–679, 2011. [Online]. Available: https://doi.org/10.1080/14747731.2011.621287 2

[9] I. Moutidis and H. Williams, "Good and bad events: combining network-based event detection with sentiment analysis," *Social Network Analysis and Mining*, vol. 10, 12 2020. 2

[10] Wikipedia, "2021 bangladesh communal violence," 2021. [Online]. Available: https://en.wikipedia.org/wiki/2021_Bangladesh_communal_violence 2

[11] M. Schinas, S. Papadopoulos, Y. Kompatsiaris, and P. A. Mitkas, "Event detection and retrieval on social media," *CoRR*, vol. abs/1807.03675, 2018. [Online]. Available: http://arxiv.org/abs/1807.03675 2

[12] B. Poblete, J. Guzmán, J. Maldonado, and F. Tobar, "Robust detection of extreme events using twitter: Worldwide earthquake monitoring," *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2551–2561, 2018. 3

[13] F. Elsafoury, "Teargas, water cannons and twitter: A case study on detecting protest repression events in turkey 2013," in *Text2Story@ECIR*, 2020. 3, 24

[14] S. Medina Maza, E. Spiliopoulou, E. Hovy, and A. Hauptmann, "Event-related bias removal for real-time disaster events," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 3858–3868. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.344 3, 24

[15] L. Dwarakanath, A. Kamsin, R. A. Rasheed, A. Anandhan, and L. Shuib, "Automated machine learning approaches for emergency response and coordination via social media in the aftermath of a disaster: A review," *IEEE Access*, vol. 9, pp. 68 917–68 931, 2021. 3, 14

[16] J. Allan, R. Papka, and V. Lavrenko, "On-line new event detection and tracking," *Association for Computing Machinery*, p. 37–45, 1998. [Online]. Available: https://doi.org/10.1145/290941.290954 3

[17] Z. Li, B. Wang, M. Li, and W.-Y. Ma, "A probabilistic model for retrospective news event detection," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '05. New York, NY, USA: Association for Computing Machinery, 2005, p. 106–113. [Online]. Available: https://doi.org/10.1145/1076034.1076055 3

[18] A. Carreño, I. Inza, and J. Lozano, "Analyzing rare event, anomaly, novelty and outlier detection terms under the supervised classification framework," *Artificial Intelligence Review*, vol. 53, 06 2020. 4

[19] R. Di Girolamo, C. Esposito, V. Moscato, and G. Sperlí, "Evolutionary game theoretical on-line event detection over tweet streams," *Knowledge-Based Systems*, vol. 211, p. 106563, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705120306924 4

[20] N. GabAllah and A. Rafea, "Unsupervised topic extraction from twitter: A feature-pivot approach," in *Proceedings of the 15th International Conference on Web Information Systems and Technologies*, ser. WEBIST 2019. Setubal, PRT: SCITEPRESS - Science and Technology Publications, Lda, 2019, p. 185–192. [Online]. Available: https://doi.org/10.5220/0007959001850192 5

[21] A. H. Hossny, L. Mitchell, N. Lothian, and G. Osborne, "Feature selection methods for event detection in twitter: a text mining approach," *Social Network Analysis and Mining*, vol. 10, 12 2020. 5

[22] A. Nourbakhsh, Q. Li, X. Liu, and S. Shah, ""breaking" disasters: Predicting and characterizing the global news value of natural and man-made disasters," 2017. 7, 12

[23] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," *Association for Computing Machinery*, p. 851–860, 2010. [Online]. Available: https://doi.org/10.1145/1772690.1772777 7, 12

[24] J. Liu, T. Singhal, L. T. Blessing, K. L. Wood, and K. H. Lim, "Crisisbert: A robust transformer for crisis classification and contextual crisis embedding," in *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*. New York, NY, USA: Association for Computing Machinery, 2021, p. 133–141. 7, 12

[25] C. Arachie, M. Gaur, S. Anzaroot, W. Groves, K. Zhang, and A. Jaimes, "Unsupervised detection of sub-events in large scale disasters," *CoRR*, vol. abs/1912.13332, 2019. [Online]. Available: http://arxiv.org/abs/1912.13332 8, 12

[26] M. Kibanov, G. Stumme, I. Amin, and J. G. Lee, "Mining social media to inform peatland fire and haze disaster management," *CoRR*, vol. abs/1706.05406, 2017. [Online]. Available: http://arxiv.org/abs/1706.05406 8, 12

[27] L. Huang, G. Liu, T. Chen, H. Yuan, P. Shi, and Y. Miao, "Similarity-based emergency event detection in social media," *Journal of Safety Science and Resilience*, vol. 2, no. 1, pp. 11–19, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666449620300293 8, 12

[28] S. Pais, I. K. Tanoli, M. Albardeiro, and J. Cordeiro, "Unsupervised approach to detect extreme sentiments on social networks," in *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2020, pp. 651–658. 8, 12, 13

[29] E. Abdukhamidov, F. Juraev, M. Abuhamad, and T. AbuHmed, "An exploration of geo-temporal characteristics of users' reactions on social media during the pandemic," *CoRR*, vol. abs/2103.13032, 2021. [Online]. Available: https://arxiv.org/abs/2103.13032 9, 12

[30] F. M. Plaza-del-Arco, S. Halat, S. Padó, and R. Klinger, "Multi-Task Learning with Sentiment, Emotion, and Target Detection to Recognize Hate Speech and Offensive Language," *arXiv e-prints*, p. arXiv:2109.10255, Sep. 2021. 9, 12

[31] Q. Kong, E. Booth, F. Bailo, A. Johns, and M.-A. Rizoiu, "Slipping to the extreme: A mixed method to explain how extreme opinions infiltrate online discussions," *ArXiv*, vol. abs/2109.00302, 2021. 9, 12

[32] D. Demszky, N. Garg, R. Voigt, J. Zou, M. Gentzkow, J. Shapiro, and D. Jurafsky, "Analyzing polarization in social media: Method and application to tweets on 21 mass shootings," *CoRR*, vol. abs/1904.01596, 2019. [Online]. Available: http://arxiv.org/abs/1904.01596 9, 12

[33] R. P. Khandpur, T. Ji, S. T. K. Jan, G. Wang, C. Lu, and N. Ramakrishnan, "Crowdsourcing cybersecurity: Cyber attack detection using social media," *CoRR*,

vol. abs/1702.07745, 2017. [Online]. Available: http://arxiv.org/abs/1702.07745 10, 12

[34] D. Pacheco, P. Hui, C. Torres-Lugo, B. T. Truong, A. Flammini, and F. Menczer, "Uncovering coordinated networks on social media," *CoRR*, vol. abs/2001.05658, 2020. [Online]. Available: https://arxiv.org/abs/2001.05658 10, 12

[35] L. H. X. Ng, I. J. Cruickshank, and K. M. Carley, "Coordinating narratives and the capitol riots on parler," *ArXiv*, vol. abs/2109.00945, 2021. [Online]. Available: https://arxiv.org/abs/2109.00945v1 10, 12

[36] W. Zhu and S. Bhat, "Euphemistic phrase detection by masked language model," *ArXiv*, vol. abs/2109.04666, 2021. 10, 12

[37] Y. Yang, X. Hu, H. Liu, J. Zhang, Z. Li, and P. S. Yu, "Understanding and monitoring human trafficking via social sensors: A sociological approach," *CoRR*, vol. abs/1805.10617, 2018. [Online]. Available: http://arxiv.org/abs/1805.10617 10, 12

[38] R. Ribeiro, P. Pereira, and J. Gama, "Sequential anomalies: a study in the railway industry," *Machine Learning*, vol. 105, 10 2016. 11

[39] S. Luca, D. A. Clifton, and B. Vanrumste, "One-class classification of point patterns of extremes," *Journal of Machine Learning Research*, vol. 17, no. 191, pp. 1–21, 2016. [Online]. Available: http://jmlr.org/papers/v17/16-112.html 11

[40] Merriam-Webster, "Dangerous," 2021. [Online]. Available: https://www.merriam-webster.com/dictionary/dangerous 13

[41] Euronews, "Germany slams attempt to storm reichstag after covid-19 protest," 2020. [Online]. Available: https://www.euronews.com/2020/08/29/thousands-of-anti-corona-protesters-flood-berlin 14

[42] K. O. Geddes, S. R. Czapor, and G. Labahn, *Sentiment Analysis Mining Opinions, Sentiments, and Emotions , pp. 1 - 15.* Cambridge: Cambridge University Press, 2015. 14

[43] R. Chandra and R. Saini, "Biden vs trump: Modeling us general elections using bert language model," *IEEE Access*, vol. 9, pp. 128 494–128 505, 2021. 14

[44] E. Lenihan, "A classification of antifa twitter accounts based on social network mapping and linguistic analysis," *Social Network Analysis and Mining*, vol. 12, no. 1, pp. 1–10, 2022. 15

[45] V. Vapnik, *The nature of statistical learning theory.* Springer science & business media, 1999. 16

[46] T. Joachims, "Text categorization with support vector machines," *Proc. European Conf. Machine Learning (ECML'98)*, 01 1998. 16

[47] A. Salas, P. Georgakis, and Y. Petalas, "Incident detection using data from social media," in *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, 2017, pp. 751–755. 16

[48] Y. Wang, H. Sundaram, and L. Xie, "Social event detection with interaction graph modeling," in *Proceedings of the 20th ACM International Conference on Multimedia*, ser. MM '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 865–868. [Online]. Available: https://doi.org/10.1145/2393347.2396332 16

[49] J. Lafferty, A. Mccallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, 01 2001, pp. 282–289. 16

[50] E. Benson, A. Haghighi, and R. Barzilay, "Event discovery in social media feeds," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 389–398. [Online]. Available: https://aclanthology.org/P11-1040 16

[51] G. I. Webb, *Naïve Bayes*. Boston, MA: Springer US, 2010. [Online]. Available: https://doi.org/10.1007/978-0-387-30164-8_576 17

[52] H. Becker, D. Iter, M. Naaman, and L. Gravano, "Identifying content for planned events across social media sites," in *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, ser. WSDM '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 533–542. [Online]. Available: https://doi.org/10.1145/2124295.2124360 17

[53] I. Subašić and B. Berendt, "Peddling or creating? investigating the role of twitter in news reporting," in *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ser. ECIR'11. Berlin, Heidelberg: Springer-Verlag, 2011, p. 207–213. 17

[54] J. Weng and B.-S. Lee, "Event detection in twitter," in *Proceedings of the international aaai conference on web and social media*, vol. 5, no. 1, 2011, pp. 401–408. 17

[55] K. Vavliakis, A. Symeonidis, and P. Mitkas, "Event identification in web social media through named entity recognition and topic modeling," *Data & Knowledge Engineering*, vol. 88, p. 1–24, 11 2013. 17

[56] R. Li, K. H. Lei, R. Khadiwala, and K. C.-C. Chang, "Tedas: A twitter-based event detection and analysis system," in *2012 IEEE 28th International Conference on Data Engineering*, 2012, pp. 1273–1276. 17, 18

[57] A. Weiler, M. Grossniklaus, and M. Scholl, "Event identification and tracking in social media streaming data," *CEUR Workshop Proceedings*, vol. 1133, 03 2014. 18

[58] S. Petrović, M. Osborne, and V. Lavrenko, "Streaming first story detection with application to twitter," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, ser. HLT '10. USA: Association for Computational Linguistics, 2010, p. 181–189. 18

[59] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, "Improving lda topic models for microblogs via tweet pooling and automatic labeling," in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 889–892. [Online]. Available: https://doi.org/10.1145/2484028.2484166 18

[60] S. Yu and B. Wu, "Exploiting structured news information to improve event detection via dual-level clustering," in *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*, 2018, pp. 873–880. 19

[61] L. Shi, X. Ma, L. Xi, Q. Duan, and J. Zhao, "Rough set and ensemble learning based semi-supervised algorithm for text classification," *Expert Systems with Applications*, vol. 38, no. 5, pp. 6300–6306, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417410013072 20

[62] A. Nourbakhsh, Q. Li, X. Liu, and S. Shah, ""breaking" disasters: Predicting and characterizing the global news value of natural and man-made disasters," 2017. 22

[63] A. K. Chanda, "Efficacy of BERT embeddings on predicting disaster from twitter data," *CoRR*, vol. abs/2108.10698, 2021. [Online]. Available: https://arxiv.org/abs/2108.10698 22

[64] Y. Zhou, J. Jiang, X. Chen, and W. Wang, "#stayhome or #marathon? social media enhanced pandemic surveillance on spatial-temporal dynamic graphs," *CoRR*, vol. abs/2108.03670, 2021. [Online]. Available: https://arxiv.org/abs/2108.03670 22

[65] G. Li, M. Dong, L. Ming, C. Luo, H. Yu, X. Hu, and B. Zheng, "Deep reinforcement learning based ensemble model for rumor tracking," *Information Systems*, vol. 103, p. 101772, 2022. 22

[66] A. Al-Laith and M. Shahbaz, "Tracking sentiment towards news entities from arabic news on social media," *Future Generation Computer Systems*, vol. 118, pp. 467–484, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X2100025X 22

[67] Walker, Christopher, et al. (2006) Ace 2005 multilingual training corpus ldc2006t06. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2006T06 23, 31

[68] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel, "The automatic content extraction (ACE) program – tasks, data, and evaluation," in *Proceedings of the Fourth International Conference on*

Language Resources and Evaluation (LREC'04). Lisbon, Portugal: European Language Resources Association (ELRA), May 2004. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf 23

[69] X. Wang, Z. Wang, X. Han, W. Jiang, R. Han, Z. Liu, J.-Z. Li, P. Li, Y. Lin, and J. Zhou, "Maven: A massive general domain event detection dataset," in *EMNLP*, 2020. 23, 31, 34, 45

[70] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The Berkeley FrameNet project," in *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*, 1998. [Online]. Available: https://aclanthology.org/C98-1013 23

[71] T. Ge, L. Cui, B. Chang, Z. Sui, F. Wei, and M. Zhou, "Eventwiki: A knowledge base of major events," in *LREC 2018*. LREC 2018, May 2018. [Online]. Available: https://www.microsoft.com/en-us/research/publication/eventwiki-knowledge-base-major-events/ 23

[72] E. Kuzey, J. Vreeken, and G. Weikum, "A fresh look on knowledge bases: Distilling named events from news," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, ser. CIKM '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 1689–1698. [Online]. Available: https://doi.org/10.1145/2661829.2661984 23

[73] R. Samuels, J. Taylor, and N. Mohammadi, "Silence of the tweets: incorporating social media activity drop-offs into crisis detection," *Natural Hazards*, vol. 103, 08 2020. 24

[74] L. C. Wold, Henning Moberg; Vikre, "Online news detection on twitter," in *Master thesis*. NTNU, 2015. [Online]. Available: http://hdl.handle.net/11250/2353650" 25

[75] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Science in China E: Technological Sciences*, vol. 63, no. 10, pp. 1872–1897, Oct. 2020. 25

[76] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *CoRR*, vol. abs/2106.04554, 2021. [Online]. Available: https://arxiv.org/abs/2106.04554 25

[77] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805 26

[78] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," *CoRR*, vol. abs/1609.08144, 2016. [Online]. Available: http://arxiv.org/abs/1609.08144 26

[79] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, "Dive into deep learning," *arXiv preprint arXiv:2106.11342*, 2021. 27, 32

[80] A. Yamaguchi, G. Chrysostomou, K. Margatina, and N. Aletras, "Frustratingly simple pretraining alternatives to masked language modeling," *CoRR*, vol. abs/2109.01819, 2021. [Online]. Available: https://arxiv.org/abs/2109.01819 27

[81] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692 27

[82] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *CoRR*, vol. abs/1901.02860, 2019. [Online]. Available: http://arxiv.org/abs/1901.02860 28

[83] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019. 28