UNIVERSITÉ D'ORLÉANS

ÉCOLE DOCTORALE Mathématiques, Informatique, Physique Théorique et Ingénierie des Systèmes

UR 4022 - LIFO

THÈSE présentée par :

Badreddine FARAH

soutenue le : **8 Janvier 2025**

pour obtenir le grade de : **Docteur de l'Université d'Orléans**

Discipline/ Spécialité : **Informatique**

Information extraction from social networks through text/image representation learning - Application to natural disaster management

Extraction d'Informations dans les réseaux sociaux par apprentissage de représentations textes/images

application à la gestion des catastrophes naturelles

JURY :

Mme TOUGNE RODET Laure (Présidente du jury)	Professeure des Universités, Université LYON 2
Mme BENAMARA ZITOUNE Farah (Rapporteure)	Professeure des Universités, Université de Toulouse
M. PIWOWARSKI Benjamin (Rapporteur)	Directeur de recherche, CNRS, ISIR, Sorbonne Université
Mme GRACIANNE Cécile (Examinatrice)	Ingénieure, BRGM
M. CLEUZIOU Guillaume (Directeur de thèse)	Professeur des Universités, Université d'Orléans
M. HAFIANE Adel (Co-directeur de thèse)	Maître de conférences - HDR, INSA-CVL
Mme HALFTERMEYER Anaïs (Membre invitée)	Maîtresse de conférences, Université d'Orléans
M. CANALS Raphaël (Membre invité)	Maître de conférences, Université d'Orléans

UNIVERSITY OF ORLÉANS

DOCTORAL SCHOOL Mathematics, Computer Science, Theoretical Physics and Systems Engineering

UR 4022 - LIFO

THESIS presented by:

Badreddine FARAH

defended on: **January 8, 2025**

to obtain the degree of:

Doctor of Philosophy at the University of Orleans in **Computer Science**

Information extraction from social networks through text/image representation learning

Application to natural disaster management

JURY:

Mrs. TOUGNE RODET Laure (Jury President)	University Professor, University of LYON 2
Mrs. BENAMARA ZITOUNE Farah (Reviewer)	University Professor, University of Toulouse
Mr. PIWOWARSKI Benjamin (Reviewer)	Research Director, CNRS, ISIR, Sorbonne University
Mrs. GRACIANNE Cécile (Examiner)	Engineer, BRGM
Mr. CLEUZIOU Guillaume (Thesis Supervisor)	University Professor, University of Orléans
Mr. HAFIANE Adel (Co-supervisor)	Associate Professor - HDR, INSA-CVL
Mrs. HALFTERMEYER Anaïs (Invited Member)	Associate Professor, University of Orléans
Mr. CANALS Raphaël (Invited Member)	Associate Professor, University of Orléans

Badreddine FARAH

Extraction d'Informations dans les réseaux sociaux par apprentissage de représentations textes/images : application à la gestion des catastrophes naturelles

Résumé : L'essor des réseaux sociaux, comme Twitter, a rendu ces plateformes essentielles pour diffuser des informations en temps réel lors de crises. Des méthodes automatisées de filtrage et de catégorisation de ces publications, basées sur le texte ou les images, ont été développées pour exploiter ces données. Ce travail propose d'intégrer simultanément texte et images des tweets, en fusionnant ces deux modalités. Dans un premier temps, cette thèse explore l'utilisation de plusieurs encodeurs unimodaux et techniques de fusion, tout en procédant à une analyse des tweets afin de mieux comprendre les relations entre les modalités visuelle et textuelle. Cette thèse introduit ensuite une méthode qui convertit les images dans un espace de représentation compatible avec le texte, rendant ainsi la fusion des deux modalités plus efficace et améliorant la robustesse et la performance du modèle. Cette approche est également hybride, permettant au modèle de traiter aussi bien des tweets unimodaux que multimodaux. Cette étude met en évidence un problème courant en apprentissage multimodal : le déséquilibre des modalités, où l'une d'elles domine le processus d'apprentissage. Une méthode d'adaptation a été proposée pour réguler cette dynamique et permettre une progression d'apprentissage plus équilibrée entre les modalités. En plus des avancées méthodologiques, cette thèse présente M-CATNAT, un jeu de données multimodales français qui comble un manque linguistique et multimodal dans les ressources liées aux crises. M-CATNAT offre des annotations détaillées pour le texte, les images et le contenu combiné, fournissant une base pour le traitement des tweets de crise multimodaux en français.

Mots clés : Apprentissage multimodal, Apprentissage de representations, Extraction d'Informations, Gestion des catastrophes naturelles

Information extraction from social networks through text/image representation learning: Application to natural disaster management

Abstract : With the increasing use of social media, platforms like Twitter have become pivotal for real-time crisis information. Posts on these platforms serve as critical data sources for situational awareness and emergency response. To efficiently exploit this information, previous works have focused on developing automated methods to filter and categorize social media posts, primarily processing either text or images (unimodal data). The present work aims to integrate both images and text from tweets by fusing information from both modalities. Firstly, this thesis explores multiple unimodal encoders with various fusion techniques, along with a deep analysis of multimodal crisis tweets, highlighting the uniqueness of multimodal tweets in terms of modality relations. Secondly, it introduces a caption-based method that translates images into a text-compatible representation space through captioning, allowing for more efficient fusion and thus achieving better and more robust performance. This method also supports hybrid usage, enabling the model to function in both unimodal and multimodal settings without compromising performance. Our experiments revealed the presence of a well-known problem in multimodal learning, namely modality imbalance, where one modality dominates the learning process. We tackle this issue with a data-driven method that adapts to the learning dynamics and allows the dominated modality to catch up. In addition to methodological advancements, this thesis introduces M-CATNAT, a French multimodal dataset that fills a linguistic and multimodal gap in crisis-related resources. M-CATNAT provides detailed annotations for text, images, and combined content, offering a nuanced foundation for French multimodal crisis-related tweet processing.

Keywords : Multimodal learning, Representations learning, Information Extraction, Natural disaster management

LABORATOIRE D'INFORMATIQUE FONDAMENTALE D'ORLEANS

Laboratoire d'Informatique Fondamentale d'Orléans (LIFO)

6 Rue Léonard de Vinci, 45067 Orléans, France

Géosciences pour une Terre durable brgm

Declaration

« With the completion of my doctorate in Computer Science, in my quest for knowledge, I have carried out demanding research, demonstrated intellectual rigour, ethical reflection, and respect for the principles of research integrity. As I pursue my professional career, whatever my chosen field, I pledge, to the greatest of my ability, to continue to maintain integrity in my relationship to knowledge, in my methods, and in my results. »

En français :

« Parvenu à l'issue de mon doctorat en Informatique, et ayant ainsi pratiqué, dans ma quête du savoir, l'exercice d'une recherche scientifique exigeante, en cultivant la rigueur intellectuelle, la réflexivité éthique et dans le respect des principes de l'intégrité scientifique, je m'engage, pour ce qui dépendra de moi, dans la suite de ma carrière professionnelle quel qu'en soit le secteur ou le domaine d'activité, à maintenir une conduite intègre dans mon rapport au savoir, mes méthodes et mes résultats. »

Acknowledgments

This thesis marks the culmination of a long and transformative journey, and I owe deep gratitude to many who walked this path with me.

First and foremost, I want to thank my family, who did everything in their power to help me reach this point. Your unwavering support, encouragement, and belief in me have been my foundation throughout this endeavor.

To my fiancée — thank you for your endless patience, love, and support through all these years. You stood by me through the stressful deadlines, and the quiet victories. I couldn't have done this without you.

A heartfelt thanks to my friends, who cheered me on, lifted my spirits, and reminded me to laugh even when things got tough.

I am sincerely grateful to all my advisors for their guidance, both academically and personally. Your insights, encouragement, and mentorship have played a crucial role in shaping this work — and me.

I also wish to thank the members of the jury for generously taking the time to review and evaluate my thesis. Your feedback and perspectives are deeply appreciated.

I am especially grateful for the time spent at LIFO, where I had the opportunity to work alongside fellow doctoral students and reaserchers. The discussions and shared moments contributed meaningfully to both my research and overall experience.

Special thanks to the French National Research Agency (ANR), the University of Orléans and the BRGM for providing the financial support that made this work possible. I'm also grateful to the CaScIModOT federation for granting access to the regional computing center, which powered much of the research behind these pages.

To everyone who played a part in this journey — thank you.

Contents

List of Figures

List of Tables

Chapter 1

General Introduction

In recent years, social media usage has surged, becoming an integral part of daily life for hundreds of millions of people worldwide. People engage on these platforms for various purposes: staying connected with friends and family, sharing current news, expressing views, collaborating on projects, showcasing creativity, and seeking support. Collectively, creating value across sectors including public health, economic forecasting, and political analysis. During times of crisis, such as natural disasters, data generated from social media can become a source of real-time situational awareness. Surveys of emergency response staff from Europe and the United States [8, 9] have shown that practitioners consider social media, including Twitter (X today), as a valuable source of information. For humanitarian organizations, this information can enhance preparedness, improve mitigation strategies, enable efficient responses, and support recovery efforts [10]. In France, two-thirds of municipalities are now considered vulnerable to natural disasters like floods and earthquakes, a situation expected to worsen over the coming decades. Consequently, analyzing social media posts shared rapidly and spontaneously — particularly on platforms like Twitter — has become a promising method for efficiently assessing the scale of disasters, providing valuable insights to guide crisis management strategies. However, the vast volume of user-generated data on social media requires automated processing, as there is considerable noise within this data. Early efforts to use tweets for gathering information during crisis events mainly focused on text, applying machine learning to filter noise and classify relevant tweets for effective crisis response. Recent research, on the other hand, has highlighted that images shared on social media during disasters also play a critical role in supporting humanitarian efforts. For example, studies of flood-related content on platforms like Twitter, Flickr, and Instagram show that posts containing images are often more relevant to the disaster, helping responders identify and prioritize urgent needs[11]. Similarly, research by Jing et al. [12] found that both images and text shared during disaster events provide complementary, valu-

able information. In addition, as part of our collaboration with the BRGM, we are contributing to the RéSoCIO [1] project—a research initiative focused on crisis management during natural disasters. RéSoCIO aims to explore the potential of using real-time Twitter data analysis during fast-developing crises, such as flash floods and earthquakes, to enhance situational awareness and improve response strategies. Our discussions with French crisis management professionals underscore the need for on-the-ground images, as they deliver actionable insights and context that support rapid, informed decision-making. While previous research has typically focused on either text classification or image classification independently, many disaster-related tweets contain both text and images. Information in one modality can enhance or complement the information in the other, and together, these two modalities can create more effective models for filtering valuable disaster-related information [13]. This has led to the development of multimodal disaster tweet datasets that combine tweet text and images[14, 15] and an increase in studies focusing on multimodal models within disaster response context. Most of the research in this area focused on fusion methods applied to the largely used CrisisMMD [14] dataset. While this dataset is a valuable resource it comes with limitations and challenges as we will show through this thesis.

1.1 Challenges and limitations

Recent research has shown the efficiency of multimodal approaches over unimodal ones for social media crisis classification, yet several limitations and challenges remain. First, social media data presents a unique challenge due to the varying relationships between text and image modalities. These relationships can range from conveying similar information, where the text describes the image, to being complementary, where each modality enhances the other. In some cases, tweets may contain text and images that are entirely unrelated [16]. This variety, combined with the limited size of annotated datasets, makes it challenging to learn complex interactions between modalities, as the standard methods use two distinct pre-trained encoders to learn each modality and then fuse the representations in a late fusion setting. Secondly, our experiments reveal an imbalance in the multimodal learning process, where the model demonstrates a tendency to rely more heavily on textual information, overshadowing the image modality. This imbalance suggests that the model may prioritize textual cues over visual ones, which can limit its ability to fully leverage the information provided by images. Lastly, as most research relies on the CrisisMMD dataset, this dataset presents an inherent limitation: the two modalities (text and images) are labeled independently, resulting in instances that lack an overall multimodal label. Consequently, the common practice is to filter out examples

[1] **https://resocio.brgm.fr/fr**

with differing modality labels, which excludes more than half of the dataset's instances overlooking the variety of content we can have in real-time situations. Additionally, for studies in the French context, this dataset poses another limitation, as it is in English and covers disasters that differ in nature and impact on the population from the events that occur in France.

1.2 Research questions and contributions

Within the context of multimodal crisis tweet classification, this research is guided by several key questions that address the limitations and challenges inherent to multimodal datasets and model performance. These questions are listed below:

- **RQ 1**: ***How do distinct models, and fusion techniques, impact the performances of crisis tweet classification ?***

  To gain a comprehensive understanding of how various pre-trained models and fusion techniques affect performance in crisis tweet classification using the CrisisMMD dataset, Chapter 3 begins with an in-depth analysis of the dataset's class distribution and multimodal structure, focusing on how text and image labels vary and contribute individually to the different classes. Then we establish baseline performance through unimodal classification, processing text and image data separately. Building on this, we explore multimodal classification by applying various fusion techniques to combine both text and image inputs for improved performance. Advanced attention mechanisms and pre-trained vision-language models are explored to enhance performance. Through extensive fine-tuning on the CrisisMMD dataset, we evaluate and compare each model and fusion method, analyzing their distinct impacts on classification accuracy. Finally, by comparing image-text similarities on CrisisMMD with traditional multimodal datasets, we identify unique challenges in fusing Twitter-based text and images, providing insights into optimal modality integration for crisis tweet classification.

- **RQ 2**: ***Can a captions-based approach simplify and improve crisis tweet classification by translating image information into a text-based representation space, enabling better fusion of multimodal data?***

  While transfer learning is widely used in crisis tweet classification, its application in multimodal settings faces challenges. Typically, separate pre-trained encoders are used for each modality (text and image), making it difficult for fusion methods to capture complex inter-modal relationships, especially given the limited dataset sizes. In Chapter 4, we introduce a caption-based method, Caption-based Multi-

modal BERT (CMB), which translates image data into text captions to create a shared semantic space with tweet text. By experimenting with multiple captioning models, this approach aims to evaluate whether using captions to convert images into text representations can simplify multimodal fusion and enhance classification performance. This approach has the added flexibility to handle both unimodal and multimodal instances, thanks to its architectural design. To leverage this flexibility, we explore mixed training strategies where the model is trained with different proportions of unimodal and multimodal data. The goal is to achieve competitive performance across both settings without compromising the other setting. Additionally, we assess the fusion effectiveness of the caption-based method in a controlled setting to ensure that the results are independent of the specific unimodal encoders used, enabling a clearer comparison of fusion capabilities.

- **RQ 3**: ***How can modality imbalance, wherein one modality dominates during model training, be effectively mitigated to support more balanced learning in multimodal crisis tweet classification?***

  Multimodal classification suffers from a well-known *imbalance* problem, while it is observed mainly in audio-video classification, this challenge is also reflected in our experiments where the models rely more on text to perform classification. To address this limitation, Chapter 5 introduces Dynamic Stochastic Shared Embeddings (D-SSE), a method for addressing modality imbalance in multimodal learning systems. By dynamically regulating dominant modalities, D-SSE provides an approach to maintain balanced representation and enhance overall model performance.

  Through extensive empirical analysis, we assess D-SSE's effectiveness in reducing modality imbalance and enhancing classification accuracy across four different datasets. Additionally, we test the method in controlled imbalance settings to evaluate its robustness in different imbalance degrees. Beyond classification performance, we further analyze the representation spaces of each modality individually compared with the multimodal space, providing a deeper analysis of how D-SSE balances and integrates diverse modalities.

- **RQ 4**: ***How can we address the current limitations of the CrisisMMD dataset to effectively apply it to crisis situations in the French context?***

  Chapter 6 presents M-CATNAT, a French multimodal crisis tweet dataset, which aims to overcome limitations in the CrisisMMD dataset by providing gold-standard manual annotations for each modality and the multimodal instance as a whole. By including labels across text, image, and multimodal data, M-CATNAT offers a robust foundation for model training and evaluation that better captures the complexities and relationships between modalities (e.g., redundancy, complementarity

or contradiction). Aligned with CrisisMMD, this dataset supports the development of nuanced, context-sensitive models for French crisis tweet classification, filling a gap in multilingual crisis response resources.

1.3 Thesis structure

Chapter 2 – This chapter provides a comprehensive overview of social media's role in crisis management within the field of crisis informatics. It begins by discussing various unimodal and multimodal datasets, highlighting how different types, and annotations of data align with the diverse needs of crisis management practitioners. The second part of the chapter dives into the methods, models, and techniques used in crisis informatics, tracing the evolution of natural language processing and machine learning within this domain. It covers advancements from early approaches relying on handcrafted features to the current application of pre-trained models and transfer learning.

Chapter 3 – We begin here by examining the CrisisMMD dataset's class distribution and multimodal structure, focusing on how text and image labels differ and contribute to each class. First, baseline performance is established by performing unimodal classification, where text and image data are analyzed separately. The chapter then moves on to multimodal classification, using various fusion techniques to combine text and image inputs for better results. Advanced attention mechanisms and pre-trained vision-language models are introduced to further improve performance. Through fine-tuning on the CrisisMMD dataset, each model and fusion method is evaluated, revealing their specific effects on classification performances. Finally, by comparing image-text similarities on CrisisMMD with other multimodal datasets, the chapter identifies unique challenges of combining Twitter text and images, offering insights into effective methods for classifying crisis tweets.

Chapter 4 – The aim of this chapter is to address the challenges of multimodal crisis tweet classification, where traditionally text and image data require separate pre-trained encoders. This separation makes it difficult for fusion methods to capture the complex interactions between modalities, especially with limited data. To tackle this, the chapter begins by introducing Caption-based Multimodal BERT (CMB), a novel approach that converts image data into text captions, creating a unified semantic space with tweet text. Various captioning models are then tested to determine if this method improves fusion. The chapter further exploits CMB's design, which allows it to process both unimodal and multimodal inputs. Leveraging this flexibility, mixed training strategies are examined, using different ratios of unimodal and multimodal data to achieve strong performance

across both types. Finally, the chapter evaluates the fusion effectiveness of CMB in a controlled setting to ensure encoder-independent results, providing a clear comparison of the fusion capabilities without the influence of specific unimodal encoders.

Chapter 5 – This chapter focuses on addressing learning imbalance in multimodal classification. It begins by analyzing this phenomenon using the well-known CREMA-D dataset. Following this, the chapter introduces Dynamic Stochastic Shared Embeddings (D-SSE), a novel approach to manage modality imbalance in multimodal learning. By adaptively regulating dominant modalities, D-SSE promotes a more balanced training process, enhancing overall model performance. Extensive empirical analysis then evaluates D-SSE's impact on reducing modality imbalance and improving classification performance across four datasets. Additional tests under controlled imbalance conditions further assess the method's robustness. Beyond classification performances, the chapter examines the similarity between each modality's representation space and the shared multimodal space, showing how D-SSE effectively balances and integrates diverse modalities.

Chapter 6 – The M-CATNAT dataset, a French multimodal crisis tweet resource created to address limitations in current datasets for crisis situations, is presented here. The chapter unfolds by first explaining the data's origins and collection methods, detailing the sources and processes used to gather French crisis tweets. It then describes the annotation methodology, clarifying how each modality—text, image, and multimodal instances—was labeled. Following this, an analysis of class distributions is provided, highlighting how these distributions vary across two types of disasters. The chapter then examines annotation quality, presenting inter-annotator scores and alignment metrics with the CrisisMMD dataset to assess consistency and compatibility. Finally, it offers an analysis of unimodal and multimodal labels, exploring their relationships and interactions to illustrate the dataset's multimodal complexity.

Conclusion – The conclusion brings together the key findings from each chapter, emphasizing advancements in multimodal classification for crisis management through combining text and image data from social media. It highlights contributions that support more effective use of multimodal tweets in crisis response and discusses the challenges faced in working with multimodal datasets and fusion techniques. The conclusion also points to potential directions for future research to further strengthen multimodal learning in crisis informatics.

Chapter 2

Overview of Social Media Use in Crisis Management

2.1 Introduction

Social media is defined by [17] as a collection of internet-based applications that are built upon the technological and ideological foundations of Web 2.0, facilitating the creation and exchange of User Generated Content. The primary function of social media lies in its capacity to empower users to generate and disseminate various types of content, including text, images, videos, and audio. Rooted in the internet's original purpose as a platform for information exchange, the evolution of technology has enhanced this capability, enabling a dynamic and interactive medium for content sharing that was previously not possible.

Over the last decade, the usage of social media has seen a large increase, with user numbers reaching beyond four billion as of 2021.[1] Those platforms serve various purposes ranging from maintaining personal relationships, sharing news, expressing opinions, collaborating on knowledge, displaying creative works, to seeking guidance and advice. Such activities are essential for the development and interaction within online communities [18]. The proliferation of social media has also made accessible an immense pool of data generated by users, creating new possibilities in various sectors. In marketing, this data is employed to understanding consumer behaviors, perceptions of brands, and market dynamics. Techniques such as sentiment analysis are applied to gauge customer sentiments, forecast brand loyalty, and measure marketing campaign impacts. For instance, [19] investigates how well a brand on Twitter aligns with its followers. In the domain of hospitality and tourism, insights from social media analytics help in assess-

[1] https://backlinko.com/social-media-users

ing customer experiences, pinpointing popular destinations, and elaborating marketing tactics [20]. In business intelligence, social media data supports predictive models to anticipate market trends and consumer behaviors [21]. In political sciences, it aids in monitoring public sentiment, analyzing communication effects, and evaluating policy stakeholders' opinions, offering critical insights to policymakers [22]. Further, the utility of social media spans several fields, including agriculture [23], banking[24], communication [25], healthcare and public health[26], journalism [27], and even counter-terrorism [28]. A detailed review of these applications is provided in [29]. In crisis management, the role of internet technologies has evolved significantly since the late 1990s as noted in [30]. Early 2000s marked the rise of dedicated crisis websites as primary information hubs. The 2004 Indian Ocean Tsunami and later events such as Hurricane Katrina and the 2007 California wildfires highlighted the critical role of user-generated content in crisis response, with platforms like MySpace and Twitter becoming crucial for real-time information sharing and coordination [31–33]. The adoption of social media in disaster scenarios has notably increased, providing innovative ways for the dissemination of information that surpass traditional methods used by emergency response bodies. These platforms allow affected individuals and observers to share updates in real-time, making social media a fundamental component for situational awareness and community support during emergencies.

2.2 Social media and crisis management

As discussed earlier, the acknowledgment of social media as a vital information source during disasters has grown since the early 2000s, with a clear consensus among various stakeholders involved in disaster management [30, 34]. Key players such as the public, response agencies, and both local and international aid organizations recognize the importance of social media for prompt information gathering and dissemination throughout the disaster lifecycle, including preparation, impact, response, and recovery phases [30].

Social media platforms, especially Twitter, have become indispensable tools in crisis management due to their ability to rapidly disseminate information and facilitate communication among affected populations and emergency responders. During real disasters, the volume of social media activity can be overwhelming. For instance, during Hurricane Sandy, millions of tweets were generated, with peaks of thousands of tweets per minute during critical moments of the disaster [30]. Such high volumes of data provide both opportunities and challenges for crisis management. First, the information needs of stakeholders vary depending on their roles and the nature of the disaster. For example, local emergency services require real-time updates on the fire's progression and evacuation routes during wildfires, while humanitarian organizations dealing with large-scale

disasters like earthquakes focus on gathering data on infrastructure damage and affected communities. The public, on the other hand, might seek updates about the safety of loved ones or information on essential resources like shelters. Specialized organizations may focus on specific aspects such as water quality or healthcare services. To meet these needs, the ability to access and filter pertinent information becomes crucial due to the vast amount of data shared on these platforms. One of the primary challenges in utilizing social media during crises is managing the sheer volume and diversity of data. Social media platforms are flooded with messages, not all of which are useful. A significant proportion of posts may be redundant, irrelevant, or even incorrect. As noted by Castillo [35], "most social media posts do not include new and useful information," but among this noise, there are often crucial updates that can significantly enhance situational awareness[30]. The challenge lies in filtering out the noise to identify actionable insights.

In response to these challenges, researchers have focused on creating specialized datasets and classification tasks that enable the effective use of social media data across different phases of disaster management. Since automated methods, particularly those based on machine learning, require labeled data to train on, the development of these datasets has been crucial. These processes include filtering algorithms to separate relevant from irrelevant data, sentiment analysis to gauge public emotions, and geolocation techniques to map out affected areas. By transforming the large flow of social media posts into structured, actionable insights, these automated methods empower stakeholders to make informed decisions during crises.

In the upcoming sections, we will discuss the specific tasks, datasets, and methods developed by researchers to manage the extensive data on social media efficiently through automated processes in both unimodal and multimodal settings.

2.3 Datasets and tasks

As the needs differ for each stakeholder in crisis management, researchers have developed a variety of tasks and datasets to address these specific requirements [36, 37]. These needs demand the use of different modalities, reflecting the diverse sources of information available during crises, including text, images, and multimodal data. The effective use of these datasets is critical for understanding and responding to the complex dynamics of disaster situations. In the research community for instance, this has resulted in a particular focus on the development of tools adapted for processing tweets, mirroring the emphasis on Twitter's role in crisis management. A 2018 review indicated that over 60% of studies on social media's use in crises specifically examined Twitter [38]. Furthermore,

research has shown Twitter's unique capacity to reveal audience needs and concerns during risk events[39]. This emphasis on Twitter underscores its significance as a data source for understanding and responding to crises, and will be mirrored in the focus of datasets used for training and evaluating crisis management models.

2.3.1 Unimodal datasets

Text

In the domain of crisis informatics, most research primarily focuses on Twitter as a data source [38]. However, even with the use of keyword-based APIs, irrelevant tweets often infiltrate the collected data, necessitating a filtering step [40]. While the concept of filtering may seem straightforward, the definition of a "relevant" or "informative" tweet can vary depending on stakeholder needs. One commonly used dataset, CrisisLexT6 [40], comprises around 60,000 tweets related to six crisis events. These were selected from a larger pool and labeled as "related" or "not related," offering a more objective categorization compared to the subjective taxonomies proposed in other works [41, 42]. These studies aimed to identify tweets enhancing situational awareness, categorizing them as Personal Only, Informative, or Other [43]. Focusing on five categories (Caution and Advice, Casualties and Damage, Donations, Missing/Found People, and Information Source), these approaches provide a systematic way to gather relevant tweets, either manually or automatically. Subsequent research has explored other classification tasks. For instance, some have focused on identifying eyewitness tweets [44], while others have addressed the need for datasets in languages other than English [34, 45–47]. A notable example is the work focusing on French stakeholders, introducing a three-level classification task (relatedness, urgency, and intent to act)[34].

Table 2.1 summarizes some of the efforts made in this area,[2] showcasing the diversity of tasks, and languages involved in crisis-related tweet classification. Despite these advancements, the challenge of filtering irrelevant tweets remains a crucial step in harnessing the power of social media for crisis informatics. Some works extends the use of tweets beyond classification, [50] uses the tweet content to geolocate the user. While the supervised methods need labeled datsets, other works applied unsupervised learning for topic modeling [51]. Despite the rich informations contained in the text, some applications and needs, require to take in account other data modalities from social media. In the next section we will dive into the use of the image in the context of crisis managment.

[2]For a substantial survey of existing datasets for English, see [36].

Table 2.1: Crisis related published text datasets.

	Modality	Tasks	Platform	Size	Language
[48]	Text	informativeness, humanitarian, source	Twitter	28,000	English
[49]	Text	informativeness, humanitarian	Twitter	166,098 (inf.), 141,533 (hum.)	English (94%)
[34]	Text	relatedness, urgency, intent to act	Twitter	12,826	French
[45]	Text	relevancy	Twitter	2,187	Spanish
[46]	Text	relatedness, damage	Twitter	5,642	Italian
[47]	Text	informativeness, humanitarian	Twitter	4,037	Arabic
[44]	Text	eyewitness	Twitter	14,000	English

Image

In the domain of crisis management, the integration machine learning technologies is increasingly reliant on diverse data forms, particularly images. This is due to the fact that images provide a rich, contextual medium for understanding and responding to crises. However, one significant challenge in this domain is the scarcity of publicly available image datasets. For instance, most research on Twitter-related crisis data tends to rely on self-collected datasets [52]. These datasets are often limited by the number of images they contain and the variety of natural disaster events they encompass. Despite these challenges, several efforts have been made to compile more extensive and varied datasets. A notable example is the dataset presented by Nguyen et al. [53], which includes a collection of images sourced from Twitter during four major natural disasters between 2014 and 2016. This dataset is enriched by an additional 20,000 images obtained from Google search. Each image in this dataset is annotated with labels such as 'severe damage,' 'mild damage,' and 'little-to-no damage,' resulting in a total of 25,750 annotated images. This dataset serves as a valuable resource for developing models capable of automatically assessing disaster damage. Another significant dataset is the HumAID dataset [54], which focuses on the humanitarian aspects of crisis response. It comprises approximately 24 million tweets with embedded images, of which about 77,000 are labeled across 11 predefined humanitarian categories ranging from 'infrastructure damage' to 'rescue volunteering.' This dataset plays a crucial role in the development of systems

that can categorize and prioritize tweets based on their urgency and relevance to disaster response. Further enriching the field, the MEDIC dataset [55] offers a multi-task image classification dataset aimed at humanitarian responses. It includes 71,198 images and supports multi-task learning across four different tasks: disaster type, informativeness, humanitarian task, and damage severity. This dataset is particularly valuable for research into multi-task learning and its applications in disaster management. Lastly, the Incidents1M dataset [56], introduced by Weber et al., contains 977,088 images categorized into 43 incident types and 49 place categories. Sourced from platforms like Flickr and Twitter, this dataset is instrumental in training models to perform incident detection and image filtering, thereby advancing the capabilities of computer vision applications in humanitarian aid.

Collectively, these datasets enable a more nuanced and rapid machine learning-driven information gathering for disaster management, significantly enhancing the capabilities of machine learning technologies in crisis management.

2.3.2 Multimodal datasets

As this thesis focuses on multimodal learning and its application in the crisis management field, and *CrisisMMD*[14] is the largest and the most used dataset in this field, more emphasis will be given to describe this dataset. *CrisisMMD* comprises multimodal data collected from Twitter during seven major natural disasters that occurred in 2017. These disasters include Hurricane Irma, Hurricane Harvey, Hurricane Maria, the California wildfires, the Mexico earthquake, the Iraq-Iran earthquake, and the Sri Lanka floods. The data collection involved filtering tweets that contained relevant keywords, ensuring a comprehensive capture of tweets and images shared during these crises. Data collection for each event was conducted over a specific period, using event-specific keywords and hashtags. The *CrisisMMD* dataset contains 16,097 English tweets related to these natural disasters. Each tweet in the dataset includes annotations for both the text and image modalities, making it suitable for training models on tasks such as informativeness assessment, humanitarian aid classification, and damage severity estimation. The dataset's focus on multimodal data allows for the development of models that leverage both textual and visual information to gain a deeper understanding of crisis situations [14]. The initial dataset consists of approximately 14.2 million tweets and 576,294 images. However, to prepare the data for meaningful analysis, several filtering steps were applied: (1) Only tweets containing images were retained to support multimodal analysis. (2) Non-English tweets were excluded. (3) Tweets with fewer than two words or hashtags were discarded. (4) Duplicate tweets were removed based on textual content similarity. After filtering, a random sample of tweets was selected for manual annotation. The *CrisisMMD* dataset

was annotated for three primary humanitarian tasks: (1) **Informative vs. Not Informative:** Classifying tweets and images based on their relevance to humanitarian aid. (2) **Humanitarian Categories:** Categorizing informative content into specific types such as infrastructure damage, rescue efforts, and reports of injured or dead individuals. (3) **Damage Severity Assessment:** Classifying images (only) depicting infrastructure damage by the severity of the damage. Annotations were performed using the Figure Eight crowdsourcing platform, [3] with three different annotators reviewing each item to ensure reliability. This rich, annotated corpus is suitable for training and evaluating models in crisis informatics and multimodal analysis. However, *CrisisMMD* has certain limitations. It annotates images and text independently, without providing a unified label for the entire multimodal tweet (text + image). This necessitates filtering out instances with mismatched labels during model training, potentially discarding valuable data and ignoring the complex interplay between modalities. Additionally, *CrisisMMD* focuses solely on English tweets, limiting its applicability for multilingual scenarios and hindering the development of models capable of processing crisis information in diverse languages. Despite these limitations, the *CrisisMMD* dataset offers valuable applications in situational awareness, disaster response planning, and multimodal information retrieval. It supports research in both natural language processing and computer vision by providing aligned textual and visual data with human-verified annotations.

While *CrisisMMD* is the largest and most used multimodal Crisis tweets datatset, another multimodal dataset [15] is introduced, this resource focuses on damage identification in social media posts using multimodal deep learning. This dataset comprises 5,879 posts collected from Instagram, Twitter, and Google during various crisis events, including posts related to fires, floods, infrastructure damage, natural landscape damage, and human casualties. The dataset is annotated across these damage categories to aid in resource allocation during crisis situations. It features both text and images, making it suitable for training multimodal models. To create this dataset, the authors employed a filtering pipeline to exclude irrelevant or non-informative content, ensuring that only actionable posts were retained. The dataset includes 35,785 total samples, with 5,879 multimodal instances (captioned images).

2.4 Social media content classification methods in crisis contexts

In the domain of crisis management, processing social media data primarily focuses on filtring irrelevant information. This involves classifying the data into categories such as

[3] www.figure-eight.com

informative versus non-informative content, or more detailed classifications such as the intention to act [34]. This section explores various methodologies applied to text, image, and multimodal data derived from social media.[4] We will discuss the main techniques and approaches applied to unimodal and multimodal data.

2.4.1 Unimodal classification methods

Text modality

Early approaches to processing social media content, dating back to the early 2000s, utilized pattern-matching techniques. For example, Fohringer et al. [57] employed predefined keywords to extract social media posts pertinent to flood events. Similarly, Mandel and Culotta [58] analyzed public sentiment during Hurricane Irene using keyword-based filters. In forest fire scenarios, the same keyword-based methodology was used to filter out irrelevant tweets [59]. The method's effectiveness in binary classification scenarios was enhanced by To et al. [60], who developed a refined framework to better utilize keywords. However, this approach has significant limitations, such as missing contextually relevant tweets that do not contain the predefined keywords, and the dynamic nature of language in crisis situations which can affect the retrieval of relevant information.

As Natural Language Processing (NLP) technology advanced, it provided more robust solutions for the extraction and classification of disaster-related information from social media. Initial methods involved feature engineering, using attributes such as word n-grams, text length, hashtag count, user mentions, URLs, tweet metadata [61], and part-of-speech tags [62]. The term frequency-inverse document frequency (tf-idf) technique [63], a specialized form of word n-grams, has been also utilized for text classification [34, 64]. The introduction of word embeddings, such as Word2vec [65] and GloVe [66], marked a significant evolution in NLP. These embeddings have proven effective in enhancing social media analytics, outperforming traditional statistical features in various classification tasks [67]. Training word embeddings on crisis-specific corpora yielded superior results compared to pre-trained models as shown in [68]. These features facilitate the training of classifiers using supervised machine learning techniques, including Support Vector Machines (SVM) [69], random forests [70], and naive Bayes algorithms [71]. Recently, deep learning models such as convolutional neural networks (CNNs) and long short-term memory networks (LSTM) [72] have been employed to further refine classification processes [73, 74].

The recent trend in NLP involves utilizing transfer learning, leveraging pre-trained models to enhance performance on new tasks. The transformer architecture [75] has

[4] More extensive review can be found in [36]

enabled the training of extensive networks on vast text corpora, which are subsequently adapted for specific tasks. A notable instance is BERT [1], which serves as a foundational model for further adaptations. Specialized models like CrisisBERT [76] have been developed to target crisis-related classification tasks, demonstrating enhanced performance and robustness across various metrics. Further Specialised Models are pretrained on specific data to make the transfer learning to the target task more efficient. TweetBert [77] used tweets as training data making it more suitable forTwitter-related tasks and showed better performance compared to the traditional BERT models in Twitter-related NLP tasks. [76] introduced CrisisBERT, an end-to-end transformer-based model for two crisis classification tasks, namely crisis detection and crisis recognition, which showed promising results and demonstrated superior robustness over various benchmarks.

Image modality

Early research in disaster event analysis primarily depended on human interpretation to extract observations from images. With the advent of Deep Learning in computer vision, similar success has been observed in crisis management. Initially, techniques like Scale-invariant Feature Transform (SIFT) [78] and Speeded Up Robust Features (SURF) [79] were utilized to detect floods [12] and fires [80] in social media images. However, the focus shifted towards employing Deep Convolutional Neural Networks (DCNNs) for disaster-related image classification tasks. Most research now leverages transfer learning from models trained on large datasets such as ImageNet [81] or Places 365 [82]. This typically involves modifying the final softmax layer of a pre-trained model (e.g., ResNet [83], VGG [84], InceptionV3 [85], DenseNet [86]) to fit the desired categories.

Specific applications include [53], which utilized a pre-trained VGG16 model [84] on ImageNet [81] for detecting levels of damage. Similarly, Li et al. [87] employed Domain Adversarial Neural Networks (DANN) as a domain adaptation technique for damage image identification. Alam et al. [88] used several pretrained models including ResNet101, VGG16, DenseNet, and EfficientNet, which were fine-tuned for disaster type detection, informativeness, humanitarian categories classification, and damage severity assessment. Recent advancements have seen the introduction of the Incident1M dataset [56], enabling the training of larger networks. For instance, the development of CrisisViT, a new transformer-based image classification model, has significantly improved performance in emergency type, image relevance, humanitarian category, and damage severity classifications [89].

2.4.2 Multimodal classification methods

One of the early efforts to address multimodal classification of social media posts for natural disasters was by Yang et al. (2011) [90], who utilized feature fusion to combine engineered features from both text (word frequency) and image (color and location features) using Simultaneous Partition and Class Parameter Estimation (SPCPE) algorithm [91]. Building on this concept, Chen et al. (2013) [92] examined the relationship between tweets and associated images, designing classifiers that integrate textual features with image characteristics and contextually relevant social features (e.g., posting time, and follower ratio). Their findings demonstrated that multimodal approaches outperform those based solely on text. Expanding on these multimodal methodologies, Mouzannar et al. [15] compared different fusion techniques, namely feature fusion and decision fusion. Their study utilized visual features extracted from pre-trained Deep Convolutional Neural Networks (DCNNs) and textual features from word2vec models. They explored various multimodal setups in a controlled study using a dataset of multimodal social media posts, that were categorized into six distinct disaster-related themes: infrastructural damage, natural landscape damage, fires, floods, human injuries, and scenarios with no apparent damage. This work highlighted the complexities and challenges of accurately categorizing multimodal data in real-world scenarios. In 2019, Gautam et al. [93] presented a detailed comparison of unimodal and multimodal methods for crisis-related multimodal classification. They adopted decision fusion techniques to classify combinations of tweet text and corresponding images into categories of 'informative' and 'non-informative'. This study further validated the superiority of multimodal methods over unimodal approaches, confirming the trend observed in earlier research. Further advancing multimodal analysis Ofli et al. [13], used both textual and visual modalities to learn a joint representation using advanced deep learning techniques. Specifically, they employed convolutional neural networks to develop a multimodal deep learning architecture with a modality-agnostic shared representation. Their comprehensive experiments on real-world disaster datasets showcased that the proposed multimodal architecture outperformed models trained on a single modality (text or image alone). The field has seen an increasing focus on advanced fusion strategies in recent years. Abavisani et al. [94] introduced cross-attention fusion techniques that process inputs separately through two distinct encoders—Bert [1] for text and DenseNet [86] for images. The resulting features are then fused using a cross-attention mechanism, which helps the model filter out irrelevant information from each modality. This method has shown to be superior to basic feature concatenation techniques used in earlier studies. Moreover, early fusion techniques have also evolved. Liang et al. [95] introduced the Multimodal Information Injection Plug-in units, integrating modalities right from the initial layers of the model. This technique maintains the robust intra-modal processing capabilities of large pre-trained unimodal models. Their research

compared these early fusion methods against large pre-trained image-text models such as Vilt [96] and PixelBert [97] and demonstrated the superiority of their approach. In an other work [16] used a new dataset Disrel where a set of tweets were labeled to capture the relation between the image and text. This dataset is then used alongside CrisisMMD [14] in a Multitask-training fashion. The goal is to help the model capture the relation between image and text. A recurrent challenge in the field has been the scarcity of annotated multimodal datasets specifically tailored for crisis scenarios. This was evident in the frequent utilization of the CrisisMMD [14] dataset by most cited studies in this section. Addressing this challenge, Sirbu et al. [98] utilized semi-supervised techniques to incorporate unlabeled data, thereby enhancing the classification performance of models on multimodal crisis datasets.

2.5 Conclusion

This chapter explored the increasing importance of social media in crisis management, emphasizing its role as a key tool for real-time communication and information sharing during emergencies. It highlighted how platforms like Twitter have become valuable sources of data for emergency responders, affected communities, and humanitarian organizations. However, the vast amount of content generated on social media during crises presents challenges in identifying relevant and accurate information. To address this, researchers have developed various datasets and classification tasks aimed at filtering and analyzing social media data effectively. The chapter also discussed the use of machine learning techniques, both for text and images, as well as multimodal approaches, to enhance the ability to extract useful insights from social media content.

Chapter 3

Multimodal Classification for Natural Disaster Management : First Experiments

3.1 Introduction

In recent years, the role of social media platforms in crisis management has grown exponentially, providing real-time information that can be important during emergencies. Platforms like Twitter, where users share both textual content and images related to ongoing disasters, offer a rich source of data for crisis response teams. This data, when efficiently processed, can help authorities make informed decisions about resource allocation, rescue operations, and humanitarian aid. However, due to the large volume of the data, filtering and classification techniques need to be developed. In this chapter, we will dive into Twitter post classification, precisely unimodal and multimodal classification methods using various models and fusion techniques. The CrisisMMD dataset [14], a large collection of tweets consisting of both text and images, has become a cornerstone in unimodal and multimodal classification research for crisis management. It supports tasks like determining the informativeness of a tweet and classifying it into different humanitarian categories, such as identifying damage severity or rescue efforts. This dataset poses several challenges, particularly in how to fuse the text and image modalities effectively for better classification accuracy. This chapter aims to explore a variety of state-of-the-art models and fusion techniques to address these challenges.

In this chapter, we unfold our research methodology in several stages, starting with an in-depth exploration of the CrisisMMD dataset and its structure. We begin by analyzing the class distribution and the dataset's unique multimodal characteristics, focusing on

how the text and image labels might differ in certain instances. Following the dataset analysis, the chapter moves into a detailed exploration of classification techniques. We first examine unimodal methods, where text and image data are processed separately to establish baseline performance for each modality. Next, we shift focus to multimodal classification, where both text and image inputs are fused to enhance the overall performance. The chapter investigates different fusion techniques, which are key to integrating the two modalities. To enhance the fusion process further, the chapter explores advanced attention mechanisms, which allow the model to prioritize the most relevant aspects of the text and images. We also explore pretrained vision-language models, which have demonstrated success in similar tasks, and adapt them to the crisis tweet classification domain. To evaluate the proposed models and fusion methods, we fine-tune them on multimodal tasks using the CrisisMMD dataset and report the results. The goal of our experiments is to compare and assess the performance of each model. Lastly, we compare the relation between image and text between CrisisMMD and classical multimodal datasets to show the challenges of image-text fusion in this specific context of Twitter data.

3.2 CrisisMMD

The CrisisMMD [14] dataset is the largest and the most used dataset in multimodal social media post classification in the context of crisis management. It is a multimodal collection of tweets, where each entry comprises two distinct modalities: text and image. Each modality is independently annotated with a corresponding label. This dataset can be formally represented as $\mathcal{D} = (X_i, Y_i)_{i=1}^{N}$, where $X_i = (X_i^{(\text{text})}, X_i^{(\text{image})})$ denotes the inputs from both modalities, and $Y_i = (y_i^{(\text{text})}, y_i^{(\text{image})})$ represents the respective labels. In this context, $X_i^{(\text{text})}$ refers to the textual content of the tweet, while $X_i^{(\text{image})}$ corresponds to the image associated with the tweet. The labels assigned to each modality are represented as $y_i^{(\text{text})}$ and $y_i^{(\text{image})}$, where $y_i^{(\text{text})} \in \mathcal{C}_{\text{task}}$ is the label for the text modality, and $y_i^{(\text{image})} \in \mathcal{C}_{\text{task}}$ is the label for the image modality. It is crucial to note that these labels are independently assigned for each modality, resulting in some instances where $y_i^{(\text{text})} \neq y_i^{(\text{image})}$, which we have named *discordant instances*.

As outlined in Section 2.3.2, the CrisisMMD dataset is designed to support three distinct tasks: *Informativeness*, *Humanitarian*, and *Damage Severity*. The *Damage Severity* task being a unimodal task focusing only on the image modality ($y_i^{(\text{image})}$), thus our work concentrates only on the two multimodal tasks, *Informativeness* and *Humanitarian*.

- **Informativeness**: a binary classification task identifying whether or not a given tweet (text or image) is informative for humanitarian aid purposes, i.e., useful for

providing assistance to people in need.

- **Humanitarian**: Identifying whether a given tweet (text or image) belongs to one of the following eight categories : *infrastructure and utility damage*; *vehicle damage*; *rescue, volunteering, or donation efforts*;*affected individuals*; *injured or dead people*; *missing or found people*; *other relevant information*; *not humanitarian.*

It is worth noting that there is an association between the two tasks, *Informativeness* and *Humanitarian*. Specifically, as showed in Figure 3.4, the "*Informative*" class in the *Informativeness* task encompasses all the categories in the *Humanitarian* task except for the "not humanitarian" category. This latter category is equivalent to the "not informative" class in the *Informativeness* task. This relationship highlights the alignment between the two tasks, where the *Humanitarian* classification provides a more granular classification.

3.2.1 Data collection

The data was collected using the Twitter API, focusing on various disasters that occurred in 2017. The events covered include Hurricane Irma, Hurricane Harvey, Hurricane Maria, California wildfires, the Mexico earthquake, the Iraq-Iran earthquake, and the Sri Lanka floods. For each disaster, specific time frames were selected for data collection, during which relevant keywords were used to extract tweets. For example, data for Hurricane Irma was collected from September 6th to 21st, 2017, using keywords such as "Hurricane Irma," "Irma storm," and "Irma Hurricane." Similarly, data for other disasters were collected over specific periods, using keywords tailored to each event.

Table 3.1: Crisis Data: Tweets, Images, and Samples

Crisis name	**# tweets**	**# images**	**# filtered tweets**	**# sampled tweets**	**#sampled images**
Hurricane Irma	3,517,280	176,972	5,739	4,041	4,525
Hurricane Harvey	6,664,349	321,435	19,967	4,000	4,443
Hurricane Maria	2,953,322	52,231	6,597	4,000	4,562
California wildfires	455,311	10,130	1,488	1,486	1,589
Mexico earthquake	383,341	7,111	1,241	1,239	1,382
Iraq-Iran earthquake	207,729	6,307	501	499	600
Sri Lanka floods	41,809	2,108	870	832	1,025
Total	**14,223,141**	**576,294**	**36,403**	**16,097**	**18,126**

Table 3.1 provides an overview of the number of texts, images and multimodal tweets annotated for each disaster. It is important to highlight that a single tweet may contain multiple images. In such cases, each image is treated as a separate instance while maintaining the same textual content, effectively creating multiple instances from a single

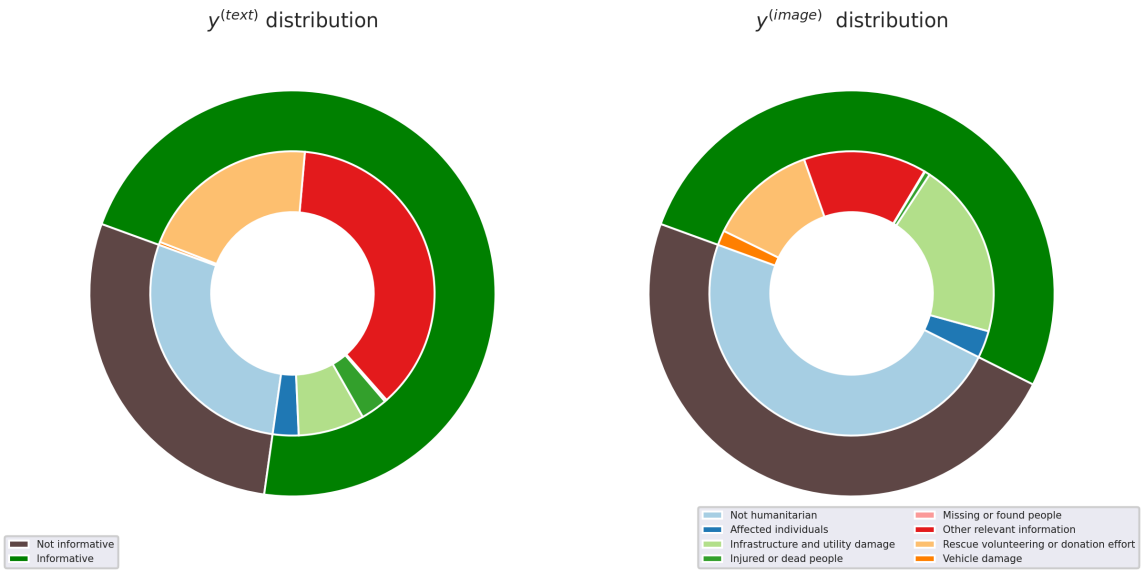


Figure 3.1: Data distribution for Texts ($y^{(\text{text})}$) and Images ($y^{(\text{image})}$) for both $Humanitarian$ (Inner circles) and $Informativeness$ (outer circles)

tweet. For instance, 14,913 tweets have one image, 570 tweets have two images, 271 tweets have three images, and 304 tweets have four images.

In the next section, we will dive into a more detailed analysis of the data including the distributions and examples of the relation between image and text.

3.2.2 Data distribution

In this section, we will first examine the class distribution within each modality—text and images—before exploring the combined multimodal class distribution. We will provide examples of tweets to illustrate the unique contributions and roles that each modality plays in conveying information.

Unimodal distributions. Figure 3.1 shows the text and image data distributions across informative and humanitarian categories. The outer circles in each diagram represent the informativeness class distribution, while the inner circles detail the humanitarian distribution. It is evident from the figure that text data is more informative than image data, with 71.66% of the text categorized as informative compared to 51.84% for images. This suggests that texts bring a higher proportion of useful information. Additionally, damages, particularly infrastructure and utility damage, are more prominently represented in images (20.05%) than in texts (7.53%), indicating that images are more effective in visualizing humanitarian damage. Other humanitarian aspects, such as rescue volunteering or donation efforts, are also better represented in the text modality (20.50%) compared

Figure 3.2: Example tweet text and image (*concordant*) pairs with different annotation labels from different disaster events

to image modality (12.34%), further underscoring the differences in content between the two modalities. For instance, as illustrated in Figure 3.2, tweets that describe detailed rescue operations are more effectively communicated through text, while those showing visual damage are better captured in the image. These examples demonstrate the varied roles that text and image play in presenting information. The diagrams highlight that while each modality provides valuable insights, they differ in what and how they represent information. This underlines the importance of multimodal processing, as combining text and images can leverage the strengths of both to enhance information extraction and analysis.

Multimodal distribution As specified above, the two modalities in CrisisMMD are annotated separately, resulting in *discordant instances* where $y_i^{(\text{text})} \neq y_i^{(\text{image})}$. Table 3.2 shows the proportion of *discordant* and *concordant* instances for each task. In *Humanitarian* task more than half of the instances are *discordant*, this is primarily due to the larger

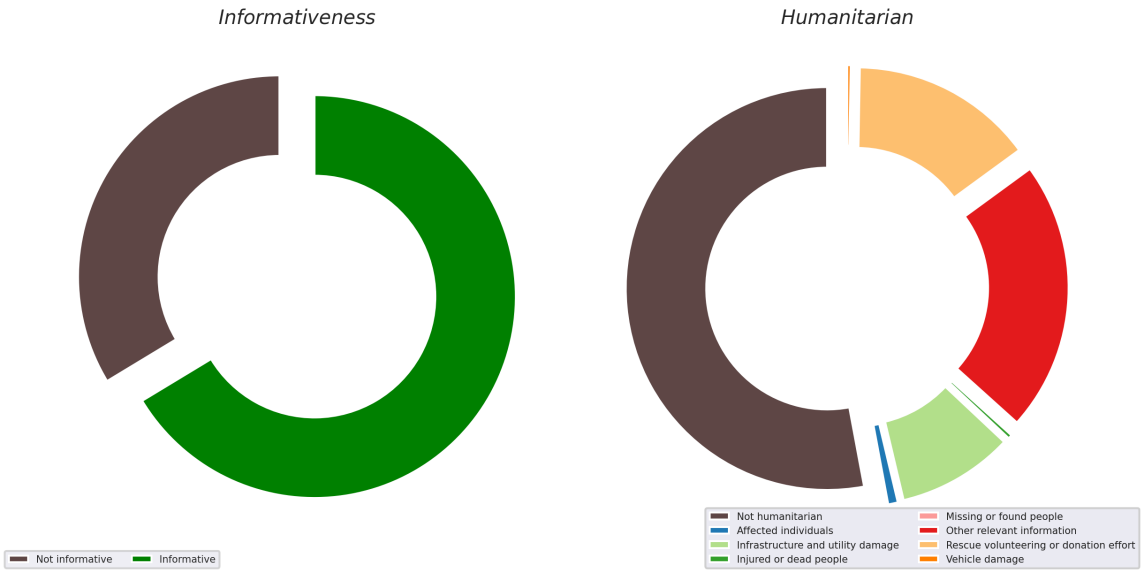


Figure 3.3: Multimodal label distribution for *concordant* instances on *informativenss* and *Humanitarian* tasks

Table 3.2: Concordant and Discordant instances.

	#discordant instances (%)	*#concordant instances* (%)	Total
Informativeness	5374(29.72%)	12708 (70.28%)	18082 (100%)
Humanitarian	10003 (55.32%)	8079 (44.68%)	18082 (100%)

number of classes. Figure 3.3 illustrates the multimodal distribution for *concordant* instances across both tasks. The multimodal distribution reveals that 66.34% of concordant instances are annotated as *informative*, while 33.66% are not. In the *Humanitarian* task, 52.94% of the data is categorized as *not humanitarian*, with significant representation in categories like infrastructure and utility damage (9.30%), *rescue volunteering or donation efforts* (14.69%), and *other relevant information* (21.70%).

Smaller categories, such as *affected individuals* (0.77%), *injured or dead people* (0.31%), and *vehicle damage* (0.27%), show lower representation. This might be due to the reliance of these categories on a single modality, such as vehicle damage primarily being captured in images. These underrepresented categories are **merged** into larger categories, for instance, the damage-related tweets (*infrastructure and utility damage* and *vehicle damage*) are merged into the *infrastructure and utility damage* class, while *affected individuals*, *injured or dead people* and *missing or found people* are merged into *affected individual* class resulting in a **five classes** *Humanitarian* task instead of the original eight classes. Figure 3.4 shows and overview on the CrisisMMD tasks and how the humanitarian tasks are merged.

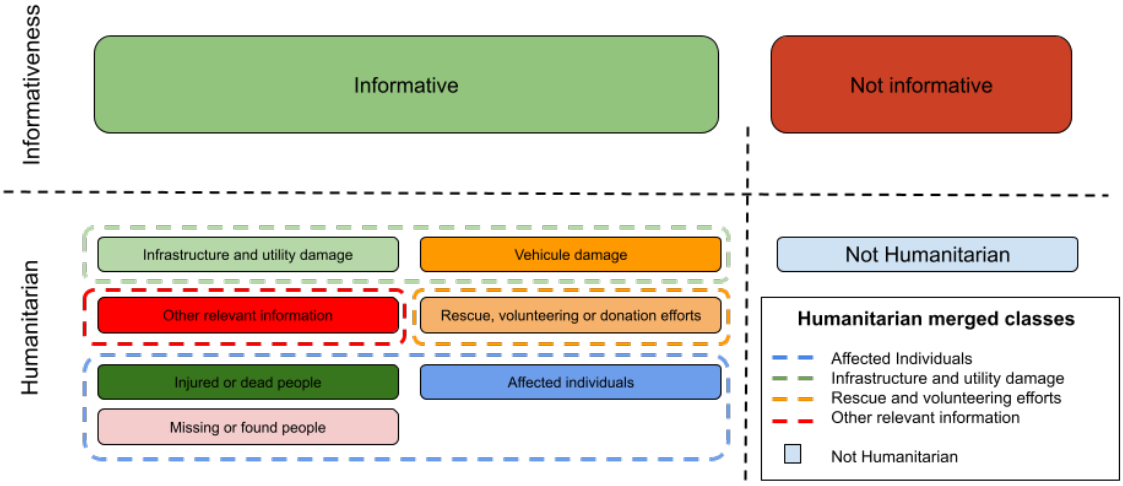


Figure 3.4: Overview of CrisisMMD multimodal tasks.

Figure 3.3 and Table 3.2 highlights a key limitation of the CrisisMMD dataset: a label for the multimodal instance $X_i = (X_i^{(\text{text})}, X_i^{(\text{image})})$ can only be determined when the labels are concordant ($y_i^{(\text{text})} = y_i^{(\text{image})}$). There is no established method to assign a multimodal label when the labels are discordant. While the standard practice on CrisisMMD is to filter out the *discordant* instances [13, 94],other studies have explored the complex relationships between text and images in Twitter data. For instance, [99] categorized these relationships into text-centric (where the text is represented or not in the image) and image-centric (where the image adds or does not add information to the text) categories. Similarly, [16] created a crisis-related dataset that classifies image-text relationships into three categories: unrelated, similar, and complementary, highlighting the diversity of these relationships. Unfortunately, the only image-text relationship in concordant instances is based on similar (task oriented) information, as both modalities contains the same information needed to discriminate the classes.

A question arises: can we derive a unified label from two discordant labels? One might hypothesize that a heuristic could be developed to select a label from the unimodal labels. However, the examples in Figure 3.5 demonstrate the difficulty of finding such a method. For instance, in Figure 3.5(a), the text label is *Affected individuals*, but when the image is considered, the label changes to *Not humanitarian* because the image depicts an animal. Conversely, in Figure 3.5(b), the correct label should be *Rescue volunteering or donation effort*, as the text provides more information about the image. Similarly, in Figures 3.5(c) and 3.5(d), the challenges of label assignment become evident. In 3.5(c), the text reveals that the image is not related to a disaster, changing the context, while in 3.5(d), the text lacks sufficient context and is labeled as *Not humanitarian*, but the image adds

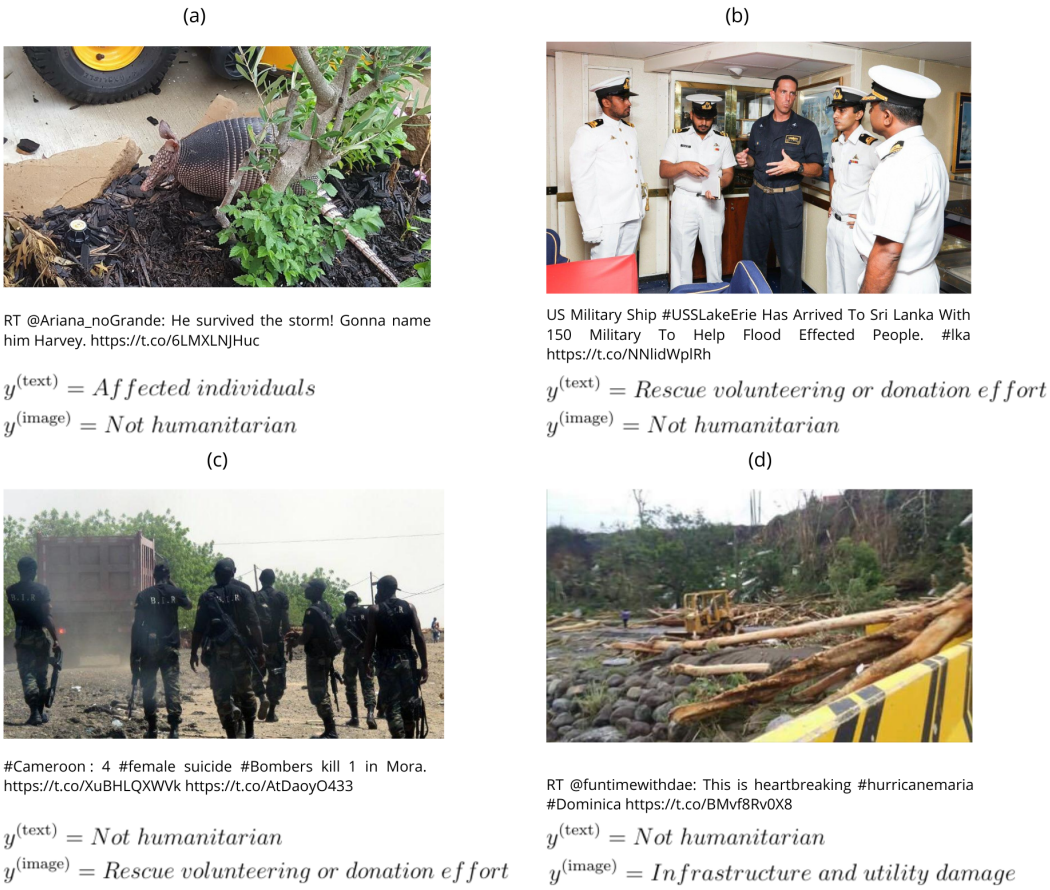


Figure 3.5: Examples of (*discordant*) tweets

context, leading to the final label of *Infrastructure damage*. These examples underscore the inherent difficulty in assigning a single label to a multimodal instance when the text and image labels are discordant. In Chapter 6, we will further analyse the *concordant* and *discordant* image-text pairs on our manually label dataset of tweets.

3.3 Tweets classification

One of the primary tasks in crisis-related social media data analysis is classification, which mainly involves filtering information into relevant and irrelevant categories and further categorizing tweets into more specific classes. In the *CrisisMMD* dataset, these tasks correspond to *Informativeness* and *Humanitarian* classification, respectively. Since this dataset is multimodal, each instance includes two input modalities, such that $X_i = (X_i^{(\text{text})}, X_i^{(\text{image})})$, associated with two labels $Y_i = (y_i^{(\text{text})}, y_i^{(\text{image})})$. Classification can be performed using only the image, only the text, or a combination of both. In the uni-

modal case, where the goal is to train a model that associates either a text or an image with its respective label, all data—both *concordant* and *discordant*—can be used. However, in the multimodal setting, where the objective is to associate a label Y_i with the image-text pair X_i, a challenge arises in choosing between the labels from $y_i^{(\text{text})}, y_i^{(\text{image})}$. To address this, [13] proposed filtering out *discordant* instances and retaining only those where $y_i^{(\text{text})} = y_i^{(\text{image})}$. In the following sections, we will explore this questions in mor details. Most recent works on tweet classification utilize deep learning methods, particularly *transfer learning*. *Transfer learning* has been instrumental in achieving significant advances in model performance across various tasks. Mathematically, *transfer learning* is defined as follows: Let $\mathcal{D}_S = (X_S, Y_S)$ denote the source domain, where X_S and Y_S represent the feature space and label space, respectively. Similarly, let $\mathcal{D}_T = (X_T, Y_T)$ denote the target domain, with feature space X_T and label space Y_T. The objective of *transfer learning* is to utilize knowledge from the source domain $\mathcal{D}_S$ to improve performance in the target domain $\mathcal{D}_T$, even when $\mathcal{D}_S \neq \mathcal{D}_T$ or $Y_S \neq Y_T$. This is typically achieved by pre-training a model on a large, general dataset (source data) and then **fine-tuning** the model on a smaller, task-specific dataset (target data). In our case, we use the CrisisMMD dataset as the target data for fine-tuning.

Previous research has highlighted the effectiveness of this approach. For instance, Ofli et al. [13] employed Word2Vec embeddings combined with convolutional neural networks (CNNs) for text classification, while using a pre-trained CNN, VGG16 [84], for image classification. Similarly, Abavisani et al. [94] utilized more advanced models such as BERT [1] for text and DenseNet [86] for images. In the following sections, we will present various pretrained unimodal models and explore different fusion techniques. These models will be fine-tuned, and we will report the results, including a comparison between each method.

3.4 Unimodal classification

Our experiments aim to achieve two key objectives. First, we seek to reproduce the results from previous studies by following similar methodologies and techniques. Second, we aim to introduce and explore the potential of more recent architectures, such as the Vision Transformer (ViT) [100] for visual data and RoBERTa [101] for textual data. These state-of-the-art models have already shown strong performance in their respective fields, making them promising candidates for enhancing multimodal tasks.

In the subsequent sections, we will first detail the specific architectures and configurations we employ in our experiments. This will be followed by an analysis of the training results and a comparison with existing approaches.

3.4.1 Text classification

In recent years, the Transformer architecture has become the dominant model in natural language processing (NLP). Introduced by [75], the Transformer architecture consists of two primary modules: an encoder and a decoder, both relying on the self-attention mechanism. Architectures like BERT and RoBERTa utilize only the encoder component of the Transformer, focusing on generating rich, context-aware representations of text. These models have achieved state-of-the-art performance across various NLP tasks, including text classification.

The Transformer Encoder

The Transformer encoder is designed to process sequences of tokens (e.g., words in a sentence) and generate contextualized embeddings for each token by attending to every other token in the sequence. Each token from the input is processed (through an embedding layer or a convolution layer for images) resulting in a sequence of embeddings $E = e_1, e_2, \ldots, e_n$, where each embedding e_i is a fixed-dimensional vector. The encoder consists of two main components: multi-head self-attention and a feed-forward network. The self-attention mechanism calculates attention scores for each token in the sequence by projecting the input tokens into three vectors: query (Q), key (K), and value (V). These vectors are used to compute the attention weights, which determine how much focus each token should give to the other tokens in the sequence:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{3.4.1}$$

where d is the dimensionality of the embedding vectors, and $softmax$ ensures that the attention scores sum to 1. The multi-head attention mechanism applies this process multiple times in parallel (using different learned projections) and concatenates the results, allowing the model to capture various aspects of the relationships between tokens. After the self-attention mechanism, a feed-forward neural network is applied to each token's representation, followed by layer normalization and residual connections. The output of the Transformer encoder is a sequence of hidden states $H = (h_1, h_2, \ldots, h_n)$, where each h_i is a contextualized embedding that incorporates information from the entire input sequence.

BERT

As shown in Figure 3.6, BERT (Bidirectional Encoder Representations from Transformers) [1] is a model that uses the Transformer encoder to learn bidirectional representations

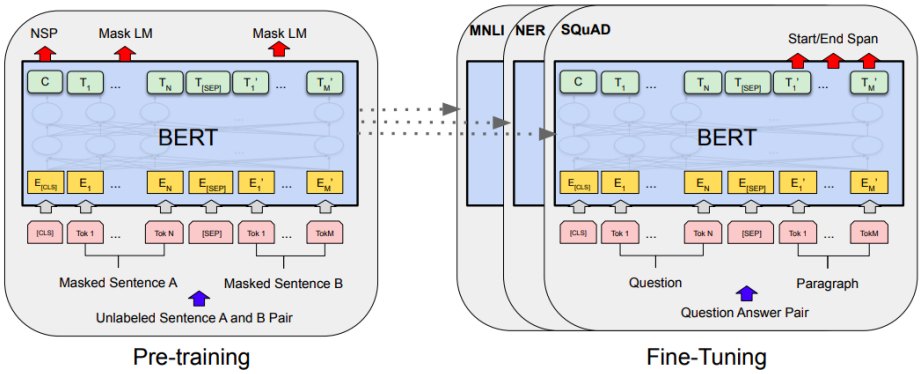


Figure 3.6: BERT pretraining and fine-tuning [1]

of text. Unlike unidirectional models like GPT, BERT captures context from both the left and right sides of a word in a sentence by pre-training on Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) tasks. The model is pretrained on the *BooksCorpus* (800M words) [102] and *English Wikipedia* (2,500M words). For text classification, BERT uses a special token [CLS] that is prepended to the beginning of every input sequence. The input sequence for text classification is represented as:

$$X = [CLS], t_1, t_2, \ldots, t_n \tag{3.4.2}$$

BERT first projects the input sequence into an embedding space with the embedding layer and then processes this input through its stacked transformer encoder layers, resulting in a sequence of hidden states $H = h_{[CLS]}, h_1, h_2, \ldots, h_n$. The hidden state corresponding to the [CLS] token, $h_{[CLS]}$, serves as a representation of the entire sequence and is used for classification. The final classification layer projects $h_{[CLS]}$ into the output space using a linear transformation followed by a $softmax$ function:

$$\hat{y} = \text{softmax}(W h_{[CLS]} + b) \tag{3.4.3}$$

where W and b are learned parameters, and $\hat{y}$ is the predicted probability distribution over the classes. BERT is fine-tuned for text classification by minimizing the cross-entropy loss:

$$\mathcal{L} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \tag{3.4.4}$$

where C is the number of classes, and y_c is the true label for class c.

RoBERTa

RoBERTa (Robustly Optimized BERT Pretraining Approach) [101] is a variant of BERT that improves the pre-training procedure by training on more data and using dynamic masking during the Masked Language Modeling task. RoBERTa also removes the Next Sentence Prediction task, focusing solely on token-level predictions. In addition to the BERT training dataset, the authors used *CommonCrawl News* [103], *OpenWebText* [104] and *Stories* dataset [105] . Like BERT, RoBERTa uses a softmax on top of a Transformer encoder and processes the input in the same format.

3.4.2 Image classification

Image classification is a task in computer vision that involves assigning a label to an image from a predefined set of categories. Given an input image, the goal is to learn a model that can accurately predict the class of the image based on its visual features. Deep learning models, particularly Convolutional Neural Networks (CNNs), have historically been the standard approach for image classification. However, recent advancements like Vision Transformers (ViT) can further improve the accuracy and efficiency of image classification tasks.

Vision Transformer (ViT)

The Vision Transformer (ViT) [100] adapts the Transformer architecture for image classification by dividing an image into a sequence of fixed-size patches, flattening them, and projecting them into embeddings. The input sequence is:

$$X = [CLS], p_1, p_2, \ldots, p_n \tag{3.4.5}$$

where p_i represents the embedding of the i-th image patch, and [CLS] is a special classification token prepended to the sequence. This sequence is processed through multiple layers of the Transformer encoder, consisting of multi-head self-attention and feed-forward networks. ViT computes a set of hidden states $H = h_{[CLS]}, h_1, h_2, \ldots, h_n$, where each h_i corresponds to the representation of a specific image patch, and $h_{[CLS]}$ represents the classification token. The final hidden state associated with the [CLS] token, $h_{[CLS]}$, is used as a summary representation of the entire image. This hidden state $h_{[CLS]}$ is then passed through a linear projection and softmax function to predict class probabilities:

$$\hat{y} = \text{softmax}(W h_{[CLS]} + b) \tag{3.4.6}$$

The use of the final hidden state of the [CLS] token allows ViT to capture global information about the image, enabling it to perform image classification tasks, note that the training is done with cross-entropy in the same way as Bert (see Equation 3.4.4)

DenseNet

DenseNet (Densely Connected Convolutional Networks) [86] is a type of Convolutional Neural Network (CNN) designed to improve the efficiency of information and gradient flow through the network. Unlike traditional CNNs, where each layer receives input only from the previous layer, DenseNet introduces dense connections between layers. In a DenseNet, each layer is connected to every other layer in a feed-forward fashion. This design ensures that feature maps learned by earlier layers are directly reused by subsequent layers, promoting feature reuse and reducing the vanishing gradient problem.

Given an input image I, DenseNet computes feature maps at each layer l as follows:

$$h_l = F_l([h_0, h_1, \ldots, h_{l-1}])$$

where h_l represents the output of the l-th layer, F_l is a composite function consisting of batch normalization, ReLU activation, and convolution, and $[h_0, h_1, \ldots, h_{l-1}]$ represents the concatenation of all feature maps from previous layers. This dense connectivity pattern allows DenseNet to be more parameter-efficient and helps the network learn richer representations. After several dense blocks, a global average pooling layer is applied, followed by a fully connected layer and softmax function to predict the class probabilities:

$$\hat{y} = \text{softmax}(W h_{\text{final}} + b) \tag{3.4.7}$$

where h_{final} is the output of the global average pooling layer, and W and b are the learned parameters for the final linear layer. DenseNet's structure improves gradient flow, reduces the number of parameters, and enhances performance on image classification tasks, making it effective on both small and large datasets.

3.5 Multimodal classification

In multimodal classification, the challenge lies in effectively integrating information from different modalities such as text and images to make accurate predictions. One of the key aspects of multimodal learning is the process of *fusion*, where information from multiple modalities is combined. This can be done at various stages of the model pipeline, leading to different fusion strategies, including *decision fusion*, *late fusion*, and *early fusion*.

Decision fusion, involves performing independent predictions for each modality and then combining the final decisions. This strategy aggregates the outputs from unimodal classifiers through techniques such as voting, averaging, or picking the maximum prediction to form the final decision. While this approach can be computationally efficient and leverages the strengths of unimodal models, it might overlook interdependencies between modalities during feature extraction. *Late fusion*, aims to combine the learned representations from each modality after their respective feature extraction processes but before the final classification step. In this technique, separate models extract features from each modality, which are then fused to train a joint classifier. Late fusion allows each modality to contribute its learned features while preserving some level of cross-modal interaction. On the other hand, *early fusion* integrates raw data from different modalities at the input stage, where features from all modalities are combined and fed into a single model. Pretrained vision-language models commonly use this fusion technique, as they can handle both text and image inputs simultaneously. Early fusion captures correlations and dependencies across modalities low level features.

In this chapter, we will explore these three fusion techniques for multimodal classification. First, we will investigate *decision fusion*, where predictions from unimodal models are aggregated. Next, we will explore *late fusion*, which leverages feature representations from each modality. Lastly, we will employ pretrained vision-language models to fine-tune them in an *early fusion* manner, combining raw data inputs directly for joint learning.

3.5.1 Decision fusion

The goal in this section is to explore the use of already trained unimodal models to perform multimodal classification. Since we train the models independently on text and images using y^{text} and y^{image}, we can use *discordant* instances to train our unimodal models. At test time, multimodal labels are needed to evaluate our models, consequently we use only *concordant* instances to test the models. At inference, time we receive the output distributions from both models, denoted as $\hat{y}^{\text{text}}$ and $\hat{y}^{\text{image}}$, which are the predicted probability distributions over all classes for the text and image models, respectively. Let $\hat{y}^{\text{text}} = [p_1^{\text{text}}, \ldots, p_n^{\text{text}}]$, where n is the number of classes and p_i is the probability associated with class i. The final prediction $\hat{y}$ for each instance is determined by applying different fusion methods to these distributions. Notably, the *Or* and *And* methods are applied only to the binary $informativeness$ task, while the Mean and Max methods can be applied to multi-class tasks as well.

"Or" fusion method

In the *Or* decision fusion method, we first compute the predicted class for each modality using arg max. If either the text or image model predicts the class as informative (class 1), the fused decision will also classify it as informative. Formally, let:

$$\hat{c}^{\text{text}} = \arg\max(\hat{y}^{\text{text}}), \quad \hat{c}^{\text{image}} = \arg\max(\hat{y}^{\text{image}}) \tag{3.5.1}$$

Then, the *Or* fusion decision is applied to these predictions:

$$\hat{y}_{\text{or}} = \max\left(\hat{c}^{\text{text}}, \hat{c}^{\text{image}}\right) \tag{3.5.2}$$

Here, the class is predicted as informative if either the text *or* image model predicts it as such, otherwise the model predicts *not informative (0)*.

"And" fusion method

In the *And* fusion method, both the text and image models must predict the class as informative for the fused decision to also predict informative. Again, we apply arg max to both distributions to obtain the predicted classes:

$$\hat{c}^{\text{text}} = \arg\max(\hat{y}^{\text{text}}), \quad \hat{c}^{\text{image}} = \arg\max(\hat{y}^{\text{image}}) \tag{3.5.3}$$

The fused decision $\hat{y}_{\text{and}}$ is then given by:

$$\hat{y}_{\text{and}} = \min\left(\hat{c}^{\text{text}}, \hat{c}^{\text{image}}\right) \tag{3.5.4}$$

In this case, the class is predicted as informative only if both models predict it as informative.

Mean fusion method

In the Mean fusion method, we do not apply arg max immediately. Instead, we directly combine the probability distributions from both the text and image models by averaging them. The fused decision $\hat{p}_{\text{mean}}$ for class i is given by:

$$\hat{p}_{\text{mean}}(i) = \frac{1}{2}\left(\hat{p}_i^{\text{text}} + \hat{p}_i^{\text{image}}\right) \tag{3.5.5}$$

After averaging the distributions, the final class prediction is obtained by taking the arg max of the fused distribution:

$$\hat{y}_{\text{mean}} = \arg\max(\hat{p}_{\text{mean}}) \tag{3.5.6}$$

Max fusion method

Similar to the Mean fusion method, the Max fusion method operates directly on the probability distributions without first applying arg max. The fused decision $\hat{p}_{\max}$ for class i is computed as:

$$\hat{p}_{\max}(i) = \max\left(\hat{p}_i^{\text{text}}, \hat{p}_i^{\text{image}}\right) \tag{3.5.7}$$

After combining the distributions by taking the element-wise maximum, the final class prediction is given by:

$$\hat{y}_{\max} = \arg\max(\hat{p}_{\max}) \tag{3.5.8}$$

The idea is to select the prediction with the highest confidence (probability).

3.5.2 Feature fusion

Let $\pi(h^{\text{text}}, h^{\text{image}})$ be a fusion module that takes the image representation h^{text} and the text representation h^{image} and produces a distribution of probabilities $\hat{y}$ on the classes. The goal of the fusion module is to combine the information from both modalities to generate a more robust and accurate prediction. For each fusion method, the function π is defined differently, depending on how the image and text representations are combined. Below, we describe how π is computed for each fusion method. Cross-Attention, Co-Attention, and Feature Concatenation methods are applied to two image models alongside BERT as the text encoder. Specifically, $VL_{DenseNet}$ uses $DenseNet$ as the image encoder, and VL_{ViT} uses ViT, while Vanilla Attention Fusion and Bottleneck Attention methods use ViT exclusively as the image encoder.

In contrast with the decision fusion methods where unimodal classifiers are trained independently, the fusion methods aim to train the hole architecture the of two encoders and the fusion module π using a joint cross-entropy loss :

$$\mathcal{L} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \tag{3.5.9}$$

where C is the number of classes, y_c is the true label for class c. and $\hat{y}_c$ is the probability assigned to class c from the overall distribution $\hat{y} = \pi(h^{\text{text}}, h^{\text{image}})$

Feature concatenation fusion

In the Feature Concatenation method, the fusion function π simply concatenates the image and text representations directly without applying any attention masks. The fused representation is given by:

$$h_{fused} = [h^{\text{text}}|h^{\text{image}}] \tag{3.5.10}$$

The fusion function π for Feature Concatenation is defined as:

$$\pi(h^{\text{text}}, h^{\text{image}}) = \text{softmax}(W_{\text{fc}}^{\top} h_{fused} + b_{\text{fc}}) \tag{3.5.11}$$

This method does not involve any attention mechanism. It directly concatenates the image and text representations to form the fused representation, which is then passed through a fully connected layer and softmax to produce the final probability distribution $\hat{y}$.

Cross-Attention fusion

Intoduced by [94], Cross-Attention fusion mechanism selectively combines h^{text} and h^{image} by applying attention masks based on the features of the opposite modality. The attention masks α^{text} and α^{image} are computed as follows:

$$\alpha^{\text{text}} = \sigma(W_{\text{a}}^{\top} h^{\text{image}} + b_{\text{a}}) \tag{3.5.12}$$

$$\alpha^{\text{image}} = \sigma(W_{\text{b}}^{\top} h^{\text{text}} + b_{\text{b}}) \tag{3.5.13}$$

The fused representations are then obtained by applying the attention masks:

$$h_{fused}^{\text{text}} = \alpha^{\text{text}} \cdot h^{\text{text}} \tag{3.5.14}$$

$$h_{fused}^{\text{image}} = \alpha^{\text{image}} \cdot h^{\text{image}} \tag{3.5.15}$$

The final fused representation is:

$$h_{fused} = [h_{fused}^{\text{text}} | h_{fused}^{\text{image}}] \tag{3.5.16}$$

The fusion function π in the Cross-Attention method is then defined as:

$$\pi(h^{\text{text}}, h^{\text{image}}) = \text{softmax}(W_{\text{fc}}^{\top} h_{fused} + b_{\text{fc}}) \tag{3.5.17}$$

In this method, each modality's attention mask is influenced by the opposite modality, allowing selective focus on the most relevant features from both image and text and filter out the irrelevant information from the opposite modality. The fused representation is the concatenation of the masked representations, and the final probability distribution $\hat{y}$ is obtained through a fully connected layer and softmax.

Co-Attention fusion

In the Co-Attention method, the fusion function π computes attention masks for both the image and text representations based on their joint features. The attention masks α^{image} and α^{text} are computed using the concatenated representation of both modalities:

$$\alpha^{\text{text}} = \sigma(W_{\text{a}}^{\top} [h^{\text{text}} | h^{\text{image}}] + b_{\text{a}}) \tag{3.5.18}$$

$$\alpha^{\text{image}} = \sigma(W_{\text{b}}^{\top} [h^{\text{text}} | h^{\text{image}}] + b_{\text{b}}) \tag{3.5.19}$$

The fused representations are obtained by applying these attention masks:

$$h_{fused}^{\text{text}} = \alpha^{\text{text}} \cdot h^{\text{text}} \tag{3.5.20}$$

$$h_{fused}^{\text{image}} = \alpha^{\text{image}} \cdot h^{\text{image}} \tag{3.5.21}$$

The final fused representation is:

$$h_{fused} = [h_{fused}^{\text{text}} | h_{fused}^{\text{image}}] \tag{3.5.22}$$

The fusion function π for Co-Attention is then:

$$\pi(h^{\text{text}}, h^{\text{image}}) = \text{softmax}(W_{\text{fc}}^{\top} h_{fused} + b_{\text{fc}}) \tag{3.5.23}$$

In this method, attention masks are computed jointly from both modalities by concatenating their representations. This ensures that both modalities influence the attention scores simultaneously, leading to the fused representation. The final probability distribution $\hat{y}$ is produced using a fully connected layer followed by softmax.

Transformer layer-based fusion

Vanilla attention For this method we use two transformer-based encoders, BERT and ViT, the output is a sequence of contextualized embeddings for text tokens and image patches. The goal of this method is to use all the embeddings, differentiating itself from the previous methods where only $h_{[\text{CLS}]}$ is used. This method enables token level intercation as each text token interacts with each patch of the image and vice versa.

Given a text input $X^{(\text{text})} = ([CLS], t_1, t_2, \ldots, t_n)$, BERT processes the sequence and produces contextualized embeddings for each token, represented as:

$$H^{(\text{text})} = (h_{[CLS]}^{(\text{text})}, h_1^{(\text{text})}, \ldots, h_n^{(\text{text})})$$

Simultaneously, an image input $X^{(\text{image})}$ is divided into patches and passed through a Vision Transformer (ViT), generating a sequence of N patch embeddings:

$$H^{(\text{image})} = (h_{[CLS]}^{(\text{image})}, h_1^{(\text{image})}, \ldots, h_N^{(\text{image})})$$

To perform cross-modal fusion, we concatenate the outputs from BERT and ViT, forming a combined sequence:

$$H = (h_{[CLS]}^{(\text{text})}, h_1^{(\text{text})}, \ldots, h_n^{(\text{text})}, h_{[CLS]}^{(\text{image})}, h_1^{(\text{image})}, \ldots, h_N^{(\text{image})})$$

This sequence is then passed through a Transformer encoder with L layers (see Section 3.4.1, which performs multimodal attention. In this layer, the self-attention mechanism computes interactions between each text token and each image patch. The overall architecture is depicted in Figure 3.7 (a).

The output of the Transformer encoder layer is a new sequence of embeddings which can be feed to another layer:

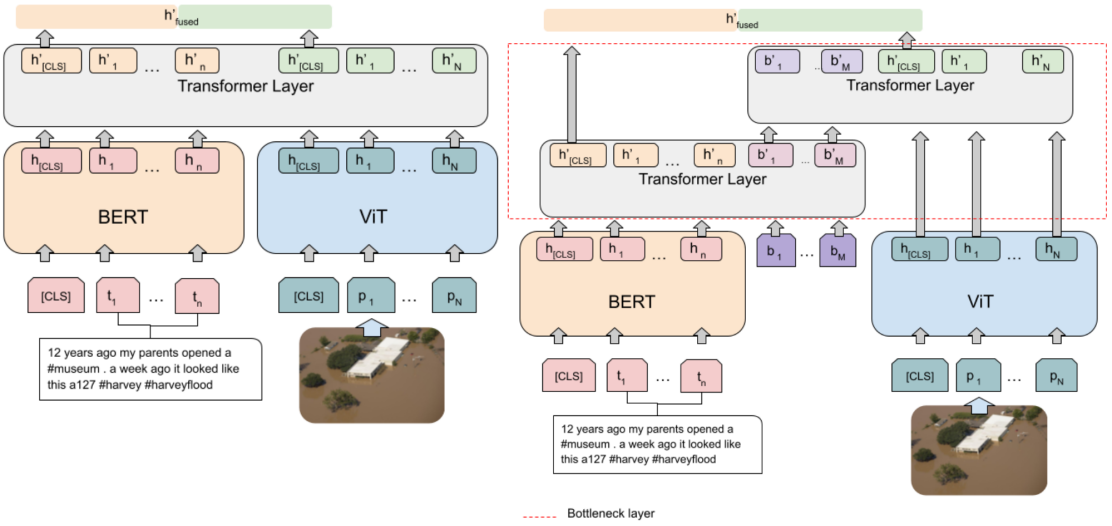
$$H' = (h_{[CLS]}^{\prime(\text{text})}, h_1^{\prime(\text{text})}, \ldots, h_n^{\prime(\text{text})}, h_{[CLS]}^{\prime(\text{image})}, h_1^{\prime(\text{image})}, \ldots, h_N^{\prime(\text{image})})$$

These embeddings contain multimodal information, where each text token and image patch is enriched by its interaction with the other modality. The final hidden states $h_{[CLS]}^{\prime(\text{text})}$ and $h_{[CLS]}^{\prime(\text{image})}$ are concatenated to serves as a global representation of the fused text-image input :

$$h'_{fused} = [h_{[CLS]}^{\prime(\text{text})} | h_{[CLS]}^{\prime(\text{image})}]$$

For multimodal classification tasks, this representation is passed through a linear classification layer:

$$\pi(H^{(\text{image})}, H^{(\text{text})}) = \text{softmax}(W h'_{fused} + b)$$



(a) Vanilla Attention (b) Bottleneck Attention

Figure 3.7: Transformer-based fusion methods

Bottleneck Attention Unlike the vanilla attention method introduced in the previous section, which allows free information exchange between different data types (modalities), this method uses bottleneck tokens to control the flow of information. These tokens force the model to gather, condense, and share only the most important information from each modality. The Bottleneck Attention Fusion method offers an alternative to traditional multimodal approaches like full pairwise attention. Instead of allowing unrestricted attention between all text tokens and image patches, this method limits cross-modal interactions to a small set of bottleneck tokens. These tokens act as intermediaries, condensing and transmitting only the most relevant information between text and image modalities. By doing so, the method enhances computational efficiency while maintaining performance, reducing the complexity typically associated with full pairwise interactions. Formally, after generating the sequence of contextual embeddings for text, $H^{(\text{text})}$, and image patches, $H^{(\text{image})}$, bottleneck tokens, denoted as $B = (b_1, \ldots, b_M)$, are introduced into the combined multimodal sequence. As depicted in Figure 3.7 (b) the bottleneck tokens are concatenated with the text (or the image) into a transformer layer encoder input, the resulting $B' = (b'_1, \ldots, b'_M)$ is then combined with the other modality into another layer input, resulting in:

$$H' = ({h'_{[CLS]}}^{(\text{text})}, {h'_1}^{(\text{text})}, \ldots, {h'_n}^{(\text{text})}, b'_1, \ldots, b'_M, {h'_{[CLS]}}^{(\text{image})}, {h'_1}^{(\text{image})}, \ldots, {h'_N}^{(\text{image})})$$

The layer output H' can be fed to another bottleneck attention layer, this structured bottleneck acts as a filter, concentrating the multimodal interaction into fewer latent

representations, thus lowering the computational cost.

The final hidden states h'_{fused} are obtained by concatenating the output from the text and image bottleneck-attended sequences.

$$h'_{fused} = [h'^{(\text{text})}_{[CLS]} | h'^{(\text{image})}_{[CLS]}]$$

These states are then passed through a classification layer for multimodal tasks, as follows:

$$\pi(H^{(\text{text})}, H^{(\text{image})}) = \text{softmax}(W h'_{\text{fused}} + b)$$

3.5.3 Pretrained Vision-Language models

The success of unimodal transfer learning, along with the impressive performance of pretrained models on various vision-language tasks such as Visual Question Answering (VQA), Natural Language for Visual Reasoning (NLVR2), and Grounded Question Answering (GQA), motivated us to fine-tune these models on the CrisisMMD dataset. In this section, we introduce two pretrained models: LXMERT and ViLT, which we fine-tuned for our specific use case.

LXMERT

LXMERT (Learning Cross-Modality Encoder Representations from Transformers) [106] is a vision-and-language model designed to capture cross-modality interactions between visual and textual inputs. It employs a large-scale Transformer-based architecture with three main encoders: (1) a language encoder, (2) an object relationship encoder for vision, and (3) a cross-modality encoder that bridges vision and language features. LXMERT is pretrained on five diverse tasks, including masked language modeling, object prediction, cross-modality matching, and visual question answering (VQA). This pretraining allows the model to learn both intra-modality (within vision or language) and cross-modality (between vision and language) relationships. LXMERT has demonstrated state-of-the-art performance on several vision-language tasks, particularly on VQA and NLVR2.

ViLT

ViLT (Vision-and-Language Transformer Without Convolution or Region Supervision) [96] is a more recent model that simplifies the vision-language interaction process by entirely removing the convolutional backbone commonly used in other vision-language

models like LXMERT. Instead of using region-based features extracted from object detectors, ViLT processes images as patch embeddings, similar to the approach used in Vision Transformers (ViTs). This results in a much lighter and faster model, while maintaining competitive performance across a variety of vision-language tasks. ViLT relies on the transformer architecture to handle both visual and textual information, making it significantly more efficient compared to models that involve heavy visual feature extraction. The key difference between ViLT and LXMERT lies in ViLT's minimalistic approach to image processing, which enables faster computation without compromising downstream performance.

3.6 Experimental settings

In this section, we describe the experimental setup, including the hyperparameters for each model and fusion method, as well as the dataset splits used in the following experiments. We follow a standard train-test split approach, where the CrisisMMD dataset is divided into training, validation, and test sets. In all experiments, we ensure no data leakage between the splits.

3.6.1 Hyperparameters

BERT and RoBERTa. We use the Huggingface Transformers library [107] to fine-tune both BERT and RoBERTa models. For training, we use a batch size of 32 and a learning rate of 2×10^{-5}, optimized with the AdamW optimizer [108] and a weight decay of 0.01. The models are trained for up to 20 epochs, with the best-performing model on the validation set selected (mainly after 3 to 5 epochs of training).

ViT and DenseNet. For image classification, we employ Vision Transformer (ViT) and DenseNet models. The ViT model is fine-tuned with a learning rate of 2×10^{-5} using the AdamW optimizer. The input image size is set to 224×224, with training conducted using a batch size of 32 for 10 epochs. In contrast, the DenseNet model is trained with the stochastic gradient descent (SGD) optimizer, using a learning rate of 1×10^{-4} and the same input size and batch size as ViT.

Decision Fusion. In decision fusion methods (Mean, Max, Or, And), each modality (text and image) is trained independently using the parameters described above. For testing, only concordant instances are used for evaluation. The final predictions from the unimodal models are combined using the specified fusion technique. No additional

training is required for these methods, as they directly leverage the fine-tuned unimodal models.

Feature fusion. As detailed in Section 3.5.2, the concatenation, co-attention, and cross-attention methods use ViT and DenseNet as visual encoders. For VL_{ViT}, the models are trained with a learning rate of 2×10^{-5}, the AdamW optimizer, and a batch size of 32, for 20 epochs, selecting the best model on the validation set. For $\text{VL}_{\text{DenseNet}}$, we use the SGD optimizer with a learning rate of 2×10^{-3}, while other parameters remain the same. For the Transformer-based fusion models (*Vanilla* and *Bottleneck*), we use a learning rate of 5×10^{-5}, the AdamW optimizer, and 4 Transformer layers for the fusion mechanism. The batch size is set to 32, and models are trained for 20 epochs.

Pretrained Vision-Language Models. The LXMERT and ViLT models are fine-tuned with a batch size of 32. The ViLT model uses a learning rate of 1×10^{-5}, while the LXMERT model uses a learning rate of 2×10^{-5}. Both models are optimized using the AdamW optimizer with a weight decay of 0.01, and trained for 10 epochs.

3.6.2 Dataset splits

We divide the CrisisMMD dataset into three splits for all experiments: training, validation, and test sets. We used the splits from [13]. Concordant and discordant instances are carefully handled, especially in multimodal experiments: in unimodal experiments, we used only concordant instances in order to compare our results with previous works. For multimodal fusion experiments, only concordant instances are used during both training and testing, except in the case of decision fusion methods, in which all the data were used (both concordant and discordant) to train the unimodal models. Note that for each experiment we will detail how the data is handled.

3.7 Evaluation metrics

The performance of the models is evaluated using several key metrics, as described below:

Accuracy. Accuracy is the ratio of correctly predicted instances to the total number of instances. It is calculated as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. While accuracy provides a general overview of model performance, it can be misleading in imbalanced datasets where the majority class dominates.

Precision. Precision measures the accuracy of the positive predictions, defined as the ratio of true positives to all predicted positives:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

High precision indicates that the model has a low false positive rate and makes correct positive predictions.

Recall. Recall, also known as sensitivity or true positive rate, is the ratio of true positives to the actual number of positives:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

A high recall indicates that the model correctly identifies most of the positive instances, minimizing false negatives. This metric is particularly important for the *Informativeness* task, where the goal is to ensure that informative tweets are not missed.

F1-Score. The F1-score is the harmonic mean of precision and recall, and it balances the trade-off between the two. It is particularly useful in situations where there is an uneven class distribution, as it provides a single metric that accounts for both false positives and false negatives:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1-score is preferred when the cost of false positives and false negatives is similar a balance between precision and recall is needed.

F1-Weighted. In the case of imbalanced datasets like CrisisMMD, the F1-weighted score is used to account for class imbalance by weighting the F1-score of each class according to its support (the number of true instances for each class). It is defined as:

$$\text{F1-weighted} = \frac{\sum_{c=1}^{C} \text{F1}(c) \times \text{support}(c)}{\sum_{c=1}^{C} \text{support}(c)}$$

where C is the number of classes, and $\text{F1}(c)$ represents the F1-score for class c. The weighted F1 score is used to ensure a fair comparison with other works, as this metric is

largely used on CrisisMMD, mainly to ensure that the performance on smaller classes is properly accounted.

Due to the imbalanced nature of the CrisisMMD dataset, particularly for the Humanitarian task, it is essential to use metrics that account for class imbalance. Accuracy alone may not provide a meaningful measure of performance in such scenarios, as the model could achieve high accuracy by predicting only the majority class. Therefore, the F1-weighted score is used to ensure that performance is balanced across all classes. Additionally, precision and recall help further analyze the trade-off between false positives and false negatives, providing a more comprehensive evaluation of the model's effectiveness.

3.8 Results and discussion

3.8.1 Unimodal classification

Table 3.3 presents the results on the Informativeness and Humanitarian tasks and compared with two notable and recent[1] works. The first, by [13], examines the performance of CNN-based models, such as VGG16 [84], on image classification, as well as word embedding models like Word2Vec paired with CNNs. The second, by [109], conducts extensive experiments on a variety of text and image classification models, including Word2Vec [65], BERT [1], ResNet152 [83], and VGG16 [84]. This comparison serves two purposes: first, to confirm the results consistency with prior research, and second, to explore state-of-the-art models such as Vision transformer [100] (ViT) for image classification and RoBERTa [101] for text classification.

In this experiment, we compare the performance of text-based and image-based models, with all models trained and tested on concordant instances. Several key trends are observed from these results, which align with findings from previous works, including those by [13] and [109].

For the text-based models, RoBERTa slightly outperforms BERT across both tasks. In the Informativeness task, RoBERTa achieves an F1 score of 86.91%, compared to BERT's 86.78%. Similarly, in the Humanitarian task, RoBERTa's F1 score is 83.03%, which is again higher than BERT's 80.92%. Word2Vec, as expected, performs significantly lower than both RoBERTa and BERT, with F1 scores of 80.90% and 67.70% on the Informativeness and Humanitarian tasks, respectively. This is consistent with the findings of [109], who also reported lower performance for Word2Vec compared to more modern transformer-based models. Furthermore, our results on BERT align with the findings of [109].

[1] At the time of these experiments

Table 3.3: Performance on Informativeness and Humanitarian Tasks, models trained and tested on *concordant instances*

Task	Modality	Model	Accuracy (%)	F1-weighted (%)	Recall (%)	Precision (%)
Informativeness	Text	BERT	86.70	86.78	86.70	86.92
		RoBERTa	86.83	86.91	86.78	87.14
		Word2Vec [13]	80.80	80.90	81.00	81.00
	Image	ViT	87.09	86.96	87.09	86.94
		DenseNet	85.14	85.05	85.14	85.00
		ResNet152 [109]	85.73	84.40	84.37	84.43
		VGG16 [13]	83.30	83.20	83.30	83.10
Humanitarian	Text	BERT	80.84	80.92	80.84	81.10
		RoBERTA	82.94	83.03	82.23	83.25
		Word2Vec [13]	70.40	67.70	70.00	70.00
	Image	ViT	83.98	83.75	83.98	83.92
		DenseNet	76.63	74.80	76.63	73.47
		ResNet152 [109]	78.30	76.37	79.60	73.40
		VGG16 [13]	76.80	76.30	76.80	76.40

For the image-based models, ViT consistently outperforms traditional CNN architectures. In the Informativeness task, ViT achieves the highest F1 score of 86.96%, surpassing DenseNet (85.05%), ResNet152 (84.40%), and VGG16 (83.20%). The same pattern is observed in the Humanitarian task, where ViT leads with an F1 score of 83.75%, while DenseNet, ResNet152, and VGG16 follow with slightly lower performances.

In summary, our results are consistent with previous works. We can note that RoBERTa shows a slight advantage over BERT in both tasks, and while CNN models perform well, ViT consistently delivers better results. These findings are consistent with the results from [13, 89, 109], underlying the strength of transformer-based models over traditional CNN and word embedding models in both classification tasks. In order to keep consistency over all experiments and be able to compare our results with previous works, we will use BERT as text encoder in all the following multimodal experiments.

3.8.2 Decision fusion

Decision fusion techniques enable us to exploit the full extent of training data by training models independently (see Section 3.5.1). In these results, models are trained on both *concordant* and *discordant* data, but are tested solely on the concordant test set. This approach is used because, although the models are trained independently, we require the multimodal label during testing to evaluate our predictions effectively. Results are shown in Table 3.4.

In terms of unimodal results, BERT and ViT models show different behaviors when trained on all data (see Tables 3.3 and 3.4). BERT maintains a stable performance in

Table 3.4: Performance of Unimodal and decision fusion method, models trained on all training set and tested on *concordant* subset from the test set.

Trained on all data (concordant + discordant)							
				Tested on concordant data			
Task	**Modality**	**Type**	**Model**	**Accuracy (%)**	**F1-weighted (%)**	**Recall (%)**	**Precision (%)**
Informativeness	Unimodal	Text	BERT	86.77	86.47	86.77	86.65
		Image	ViT	85.85	86.09	85.85	86.79
	Multimodal	Decision Fusion	Mean	92.44	92.41	92.44	92.40
			Max	92.44	92.41	92.44	92.40
			Or	87.29	86.53	87.29	88.38
			And	85.33	85.73	85.33	88.15
Humanitarian	Unimodal	Text	BERT	72.88	73.52	72.88	80.47
		Image	ViT	83.87	83.55	83.87	83.91
	Multimodal	Decision Fusion	Mean	90.99	90.99	90.99	91.13
			Max	88.80	88.83	88.80	89.01

the informativeness task with a F1 score of approximately 86.47% but shows a decline in performance in the humanitarian task, dropping to 73.52%. This decrease highlights BERT's sensitivity to shifts in class distribution seen between the training and testing datasets. In contrast, ViT's performance remains more consistent in the humanitarian task but decreases in the informativeness task, suggesting sensibility to the class distribution changes. Those results are coherent with the change in data distributions from Figures 3.1 and 3.3.

The multimodal decision fusion methods like Max and $Mean$ generally improve performance across tasks, benefiting from integrating inputs from multiple modalities. Notably, the $Mean$ model excels in the humanitarian task with an accuracy of 90.99%, outperforming the Max fusion method. The Or method's relatively superior performance compared to And in the informativeness task can be attributed to the prevalence of the informative class in the test set, making it advantageous for at least one modality to predict this dominant class accurately. Conversely, the And model shows lower performance due to its restrictive requirement for agreement among all modalities, which is less adaptive in cases of divergent or conflicting modal information. Overall, the use of both modalities helps to improve the classification performances underlying the usefulness of multimodal methods. These results should be interpreted with caution. The nature of the test set, consisting only of *concordant* data, inherently favors predictions based on modality agreement such as Max, $Mean$, and Or. This configuration might not fully represent real-world complexities where modalities may diverge as seen in Figure 3.5 underscoring a limitation in of those methods. In the following section, we will explore the results of the feature fusion techniques where the fusion is made on the representation level in contrast with the decision level.

Table 3.5: Performance of The different features fusion Methods for Informativeness and Humanitarian Tasks

Task	Model	Fusion Method	Accuracy	F1-weighted	Recall	Precision
Informativeness	$VL_{DenseNet}$	concat	90.61	90.60	90.61	90.59
		co-attention	90.48	90.50	90.48	90.53
		cross-attention	90.35	90.28	90.35	90.27
	VL_{ViT}	concat	91.46	91.41	91.46	91.40
		co-attention	91.40	91.28	91.40	91.37
		cross-attention	91.07	90.95	91.07	91.03
	Transformer attention	vanilla	91.66	91.70	91.66	91.77
		Bottleneck	**92.05**	**92.03**	**92.05**	**92.02**
Humanitarian	$VL_{DenseNet}$	concat	86.81	86.64	86.81	86.79
		co-attention	86.39	86.04	86.39	85.83
		cross-attention	86.39	86.03	86.39	85.90
	VL_{ViT}	cross-attention	87.64	87.69	87.64	87.79
		concat	87.23	87.24	87.23	87.26
		co-attention	86.81	86.83	86.81	87.37
	Transformer attention	vanilla	88.38	88.32	88.38	88.30
		Bottleneck	**88.59**	**88.59**	**88.59**	**88.68**

3.8.3 Feature fusion

The performance comparison of various models using different fusion methods, as presented in Table 3.5, reveals several key insights.

Firstly, models based on ViT, including the vanilla Transformer, Bottleneck, and VL_{ViT} architectures, consistently outperform the $\text{VL}_{\text{DenseNet}}$ models. For instance, in the Informativeness task, the VL_{ViT} model with concatenation achieves an F1 score of 91.41%, compared to the highest score of 90.60% for $\text{VL}_{\text{DenseNet}}$ with the same fusion method. A similar trend is seen in the Humanitarian task, where VL_{ViT} with cross-attention reaches an F1 score of 87.69%, while $\text{VL}_{\text{DenseNet}}$ achieves 86.64% using concatenation. These results suggest that ViT-based models perform better, mainly due to the superior performance of ViT on image-only classification tasks. Secondly, the table shows that there is no significant difference between fusion methods—concatenation, co-attention, and cross-attention—for both the VL_{ViT} and $\text{VL}_{\text{DenseNet}}$ models. For example, in the Informativeness task, VL_{ViT} achieves F1 scores of 91.41%, 91.28%, and 90.95% for concatenation, co-attention, and cross-attention, respectively—a difference of less than 0.5%. Similarly, for $\text{VL}_{\text{DenseNet}}$, concatenation (90.60%), co-attention (90.50%), and cross-attention (90.28%) perform very similarly. This suggests that the choice of fusion method has a minimal effect on the overall performance of these models. Thirdly, Transformer attention-based fusion methods (Vanilla and Bottleneck attention) outperform the other fusion methods. In the Informativeness task, the vanilla Transformer achieves an F1 score of 91.70%, while Bottleneck fusion further improves this to 92.03%,

Table 3.6: Performance of Models Trained on Non-Balanced and Balanced Data for Non-Balanced and Balanced Test Sets (Informativeness Task)

Train Data	**Model**	**Test: Non-Balanced**		**Test: Balanced**	
		Accuracy	**F1-weighted**	**Accuracy**	**F1-weighted**
Non-Balanced	Mean	**92.44**	**92.41**	**91.77**	**91.75**
	Bottleneck	92.05	92.03	91.37	91.35
	Vanilla	91.66	91.70	91.67	91.66
Balanced	Mean	87.74	87.90	88.10	88.09
	Bottleneck	90.68	90.76	90.77	90.77
	Vanilla	**91.13**	**91.23**	**91.57**	**91.57**

outperforming concatenation (91.41%) and cross-attention (90.95%). Similarly, in the Humanitarian task, Bottleneck fusion (88.59%) surpasses concatenation (87.24%) and cross-attention (87.69%). This highlights the effectiveness of Transformer attention layers in capturing more detailed interactions between modalities compared to late fusion techniques, this results are consistent with state of the art models. Importantly, all the fusion-based methods outperform the unimodal models in both the Informativeness and Humanitarian tasks, highlighting the benefit of combining text and image data. This confirms the utility of multimodal approaches in crisis tweet classification, as they leverage complementary information from both modalities, providing more complete information and improving decision-making in crisis management. Lastly, while these fusion methods perform well, it is important to note that the *Mean* fusion method from Table 3.4 outperforms all other methods. We hypothesize that the train and test distribution favored the *Mean* method, leading to superior performance. In the next section, we will compare the decision fusion methods with feature fusion methods in different training and test distributions.

3.8.4 Decision and feature fusion

As underlined in the previous section, although the good performance of feature fusion methods, it still outperformed by the *Mean* fusion methods, firstly as underlined in Section 3.5.1, the unimodal training is done on all the data (*concordant* and *discordant*) while the feature fusion models are trained on *concordant* data only, resulting in much larger training data set for the first models. In Table 3.6, we compared the best-performing models from the decision and fusion methods in different training and testing settings for the *Informativeness* task.

A key observation is that when trained on non-balanced data, the *Mean* model achieves the highest performance across both test settings, with 92.44% accuracy and

92.41% F1-weighted score on the non-balanced test set, and 91.77% accuracy and 91.75% F1-weighted on the balanced test set. This superior performance is partly due to the fact that the *Mean* model was trained on a larger dataset. However, all models experience a slight performance drop when tested on the balanced set, indicating a sensitivity to distribution shift. In the balanced training setting, where all models are trained on equal-sized datasets, feature fusion models (Bottleneck and Vanilla) outperform the *Mean* model. For example, the Vanilla model achieves 91.57% accuracy and F1-weighted score on the balanced test set, compared to 88.10% and 88.09% for the *Mean* model. The impact of data distribution shift is minor for these feature fusion models, suggesting that feature-level interactions contribute to more robustness on distribution shifts. Conversely, the *Mean* model, which relies on decision-level fusion, is heavily impacted by the training data size and distribution, especially when evaluated on a different test distribution (e.g., a drop from 92.44% to 87.74% accuracy). This further supports the hypothesis that feature-level interactions enable models to achieve more consistent performance across various data distributions, mitigating the effect of distribution shifts on model generalization.

3.8.5 Pretrained Vision-language models

Table 3.7 compares the performance of the LXMERT and ViLT models on the Informativeness and Humanitarian tasks. Both models exhibit relatively lower performance compared to previously reported models, as demonstrated in the previous Table 3.6. In the Informativeness task, LXMERT achieves an accuracy of 89.50% and an F1 score of 89.42%, while ViLT performs slightly better with 89.96% accuracy and 89.92% F1 score. Despite these decent results, both models fall short of the *Mean* (92.44%) and Bottleneck Fusion (92.05%) methods. Even the Vanilla Transformer surpasses LXMERT and ViLT, with an accuracy of 91.66% and an F1 score of 91.70%, highlighting the competitive edge of late fusion methods over these pretrained vision-language models. Similarly, for the Humanitarian task, LXMERT reaches an accuracy of 85.65%, while ViLT scores 87.23%. These results, although comparable, are still lower than the majority of the techniques introduced earlier, further emphasizing the superior performance of late fusion methods.

In conclusion, while LXMERT and ViLT have shown to be effective for general multimodal tasks, they underperform in comparison to late fusion models, such as Bottleneck Fusion and Vanilla Transformer Fusion, which demonstrate better adaptation to the nature of Twitter data. We hypothesize that this performance gap stems from the difference and more challenging nature of the image-text relationships in the CrisisMMD dataset compared to the datasets used for pretraining LXMERT and ViLT, which are primarily

Table 3.7: Best Performance of Pretrained Vision Language Models on Informativeness and Humanitarian Tasks

Task	Model	Accuracy	F1-weighted	Recall	Precision
Informativeness	LXMERT	89.50	89.42	89.50	89.41
	ViLT	89.96	89.92	89.96	89.90
Humanitarian	LXMERT	85.65	85.70	85.65	86.08
	ViLT	87.23	87.26	87.23	87.33

image-caption datasets like MSCOCO [110] and Conceptual Captions [111]. The complex relationship between images and text in Twitter data remains a significant challenge for current multimodal models. In the next section, we will analyze this relationship further using CLIP [2] similarity scores to compare CrisisMMD with other datasets.

3.9 Image-Text relationship in CrisisMMD

To better understand the failure of pretrained image text models presented in the previous section and for better understanding the Twitter crisis related tweets, we analyse in this section the CLIP similarity score differences between a variety of datasets. We use this score to analyze the diversity of relationships between image and text pairs in *CrisisMMD* and compare it to other datasets like *MSCOCO* [110] and baselines like *DisRel* [16].

3.9.1 CLIP similarity score

The CLIP similarity score is based on cosine similarity between the image and text embeddings generated by the CLIP [2] model. CLIP (Contrastive Language-Image Pre-training) is designed to map both text and image data into a shared embedding space resulting in a representation space where an image and its description have a high cosine similarity. Given an image representation $v_i \in \mathbb{R}^d$ and a text representation $t_i \in \mathbb{R}^d$, both projected into the same dimensional space $\mathbb{R}^d$, the cosine similarity between them is calculated as:

$$\text{Similarity}(v_i, t_i) = \frac{v_i \cdot t_i}{\|v_i\|\|t_i\|} \tag{3.9.1}$$

This similarity score lies between -1 and 1, where a score of 1 indicates perfect alignment between the image and text (i.e., they share a high level of semantic similarity), while a score closer to -1 implies a complete mismatch.

3.9.2 Analysis of image-text relationships

To better understand the variation in similarity scores, we introduce *Random CMMD* as a baseline. In this dataset, images are randomly assigned to texts, breaking any meaningful relationship between the two modalities. The similarity score distributions for *CrisisMMD*, *MSCOCO* [110] and *Random CMMD* are shown in Figure 3.8 (1). We also report results for *concordant* and *discordant* instances in (2), and results for different classes in the *DisRel* dataset in (3).

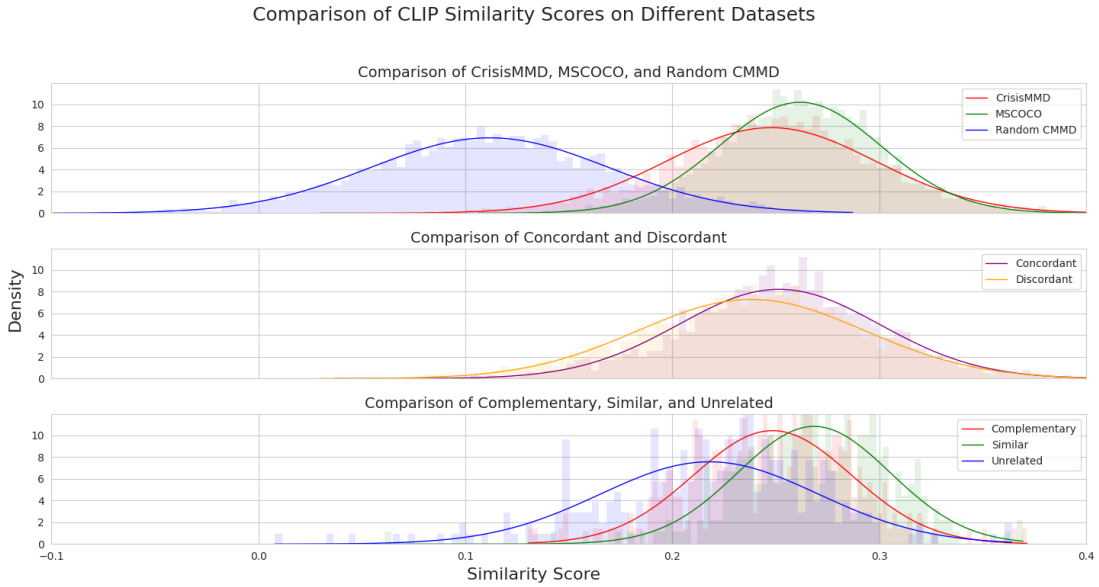


Figure 3.8: Comparison of CLIP Similarity Scores across different datasets and pair types.

MSCOCO, which pairs images with constructed captions, shows a distribution with relatively high similarity scores and a narrow variance suggesting a strong alignment between images and their corresponding texts, as expected from a dataset designed to describe visual content precisely. The resulting distribution is centered around high similarity values. In contrast, *CrisisMMD* displays slightly lower scores with broader distribution with higher variance, highlighting the varied relationships between text and images in crisis-related tweets. In this dataset, image and text are not always directly correlated; they may provide complementary information, describe different aspects of a situation, or sometimes be entirely unrelated. This variability is reflected in the lower and more dispersed similarity scores. While some pairs show high alignment, similar to *MSCOCO*, many others have lower scores, indicating weaker semantic alignment. This variability helps explain the lower performance of pretrained vision-language models on this dataset, as discussed in the previous section. The *Random CMMD* dataset, where

images are randomly paired with unrelated text, has the lowest similarity scores, confirming that without a meaningful connection between image and text, the alignment is minimal. This dataset serves as a lower bound for comparison. We also analyze *concordant* and *discordant* instances. *Concordant* instances are instances where both image and text share the same label, while discordant instances have differing labels between the two modalities. The comparison reveals that concordant instances have slightly higher similarity scores. Interestingly, the difference between concordant and discordant distributions in *CrisisMMD* mirrors the difference between the overall distributions of *CrisisMMD* and *MSCOCO*. Underlying the limitation of CrisisMMD where *discordant* instances are disregarded, ignoring a large range of cases where image and text are further deccorelated. Finally, the *DisRel* dataset, where 4,600 multimodal tweets collected during various natural disasters in the USA in 2017 and annotated into three categories of image-text relations: *similar*, *complementary*, and *unrelated*, this dataset further illustrates the nuances of multimodal relationships present in *CrisisMMD*. Similar pairs exhibit high similarity scores, comparable to the high-alignment pairs in *MSCOCO*. Complementary pairs, where image and text provide distinct yet related information, display more moderate similarity scores. Unrelated pairs, as expected, show very low similarity scores. This indicates that the CLIP similarity scores are consistent and closely reflect the relationships between image and text.

Overall, the results show that Twitter's crisis-related multimodal data encompasses a wide range of relationships between image and text, from closely aligned (similar) to complementary to completely unrelated. This diversity is similar to what is seen in the *DisRel* dataset, which stands in contrast to the relatively more uniform relationships in popular image-text datasets like *MSCOCO* that are used to train large multimodal models. The variety of the relationships in Twitter data and the limited size of datasets present a challenge for multimodal methods to effectively capture the diverse relations between modalities present in this data type. In the next chapter, we will approach this challenge by using captions to project both image and text into a shared embedding space. This approach should reduce the complexity in capturing the relationships between the two modalities and improve the model's ability to handle the diverse relationships in *CrisisMMD*, ultimately leading to better performance.

3.10 Conclusion

In this chapter, we conducted an in-depth exploration of multimodal tweet classification within the context of crisis management, utilizing the CrisisMMD dataset. We began by analyzing the dataset's structure and distribution, highlighting the challenges posed by discordant instances where text and image labels differ. Our analysis revealed that while

text and images individually provide valuable information, their combination can lead to a richer representation of the content, which is crucial for effective crisis response.

We implemented and evaluated a range of state-of-the-art unimodal models, including BERT and RoBERTa for text classification, and ViT and DenseNet for image classification. The results confirmed that transformer-based models outperform traditional methods, with ViT and RoBERTa achieving the highest accuracies in their respective modalities. These findings are consistent with prior research and underscore the effectiveness of transformer architectures in handling crisis-related data. Building on the unimodal results, we investigated various multimodal fusion techniques to integrate textual and visual information. We explored decision fusion methods, feature fusion strategies, and advanced attention mechanisms, including transformer-based fusion and bottleneck attention. Our findings indicated that feature-level fusion methods, particularly those utilizing transformer attention mechanisms, outperformed decision-level fusion techniques. The bottleneck attention-based fusion method demonstrated superior performance. We also evaluated pretrained vision-language models such as LXMERT and ViLT. Although these models have shown remarkable performance on general multimodal tasks, they underperformed in the context of crisis-related tweet classification. This underperformance can be attributed to the unique and complex relationships between images and text in the CrisisMMD dataset, which differ significantly from the datasets used to pretrain these models. Our analysis of the image-text relationships using CLIP similarity scores revealed that CrisisMMD encompasses a wide range of relationships—from closely aligned to completely unrelated pairs as reflected from the different similarity distributions. This diversity presents a challenge for models pretrained on datasets with more straightforward image-text alignments, such as MSCOCO. The varied relationships highlight the need for models that can adapt to the unique characteristics of social media data during crises.

In conclusion, our exploration of multimodal tweet classification demonstrated the effectiveness of combining text and image data, with transformer-based fusion methods showing the best performance. Pretrained vision-language models struggled with the variety of modality relationships in CrisisMMD. The diverse relations in Twitter data and limited dataset sizes pose challenges for multimodal models. In the next chapter, we will address this by using captions to project both modalities into a shared embedding space, aiming to simplify the inter-modal interactions and improve performance.

Chapter 4

Caption based tweets classification

4.1 Introduction

Traditional approaches to multimodal fusion often use separate unimodal encoders for each modality—typically one for text and another for images—followed by a fusion of their respective embeddings. Techniques such as simple concatenation [13] and more advanced methods like cross-attention [94] have been employed to enhance the interaction between modalities. However, in Chapter 3, we showed that while these multimodal methods generally outperform unimodal models, the differences in performance between various techniques, including concatenation, cross-attention, and co-attention (see 3.5), are minimal and do not result in substantial improvements. Additionally, we showed that general-purpose self-supervised vision-language models [96, 106] are sub-optimal on the CrisisMMD dataset as they are primarily trained on homogeneous data, such as images and their captions.They, therefore, do not take into account the heterogeneous informational content that may exist between an image and its associated text in social media data as showed in Section 3.9.2. Furthermore, previous studies [112, 113] have shown that different modalities tend to overfit and converge at different rates when using two encoders. This indicates that optimizing the same objective for various modalities results in sub-optimal and imbalanced learning.

To address those challenges posed by the variety of relationships and the scarcity of annotated data, which make learning more difficult, this chapter draws inspiration from recent advancements in captioning methods, and explores a captions-based approach for vision-language classification tasks related to crisis tweets. We hypothesize that integrating text and image embeddings into a shared semantic space, using a single trainable encoder, can help simplify and improve the learning process. For this, we propose to use a caption-based method, which consists of translating a modality (e.g.

image) towards another modality (e.g. text). We, therefore, present the CMB method ***C**aption-based **M**ultimodal **B**ERT*, that leverages captioning models to translate image into text and concatenate resulting captions with the tweet's text before inputting the combined data into a text encoder.

To thoroughly evaluate this approach, we conduct experiments with a variety of captioning models that have been pre-trained on different datasets. We also propose a training strategy that allows to use our method in unimodal settings without significant performance loss thus allowing the method to deal with both unimodal and multimodal tweets. Our experiments focus on crisis context, in particular on the well-studied multimodal CrisisMMD dataset [14]. The examples in Figures 4.5 and 4.6 illustrate the challenges of this dataset in terms of the relationship between modalities (complementarity, heterogeneity or partial redundancy [99]). In this chapter, we focus on several key developments. First, we introduce CMB, a method aimed at homogenizing the input space of image and text data, which enhances information processing in BERT-like architectures. Following this, we conduct a comparison of three captioning models and demonstrate that this method produces highly competitive results when applied to crisis-related tweets. Additionally, we outline a training strategy that enables the use of our models in unimodal settings with minimal loss of performance. A qualitative analysis is then provided to explore the success and failure of our approach. Lastly, we conduct an error analysis that contrasts the performance of multimodal models, highlighting their strengths and weaknesses.

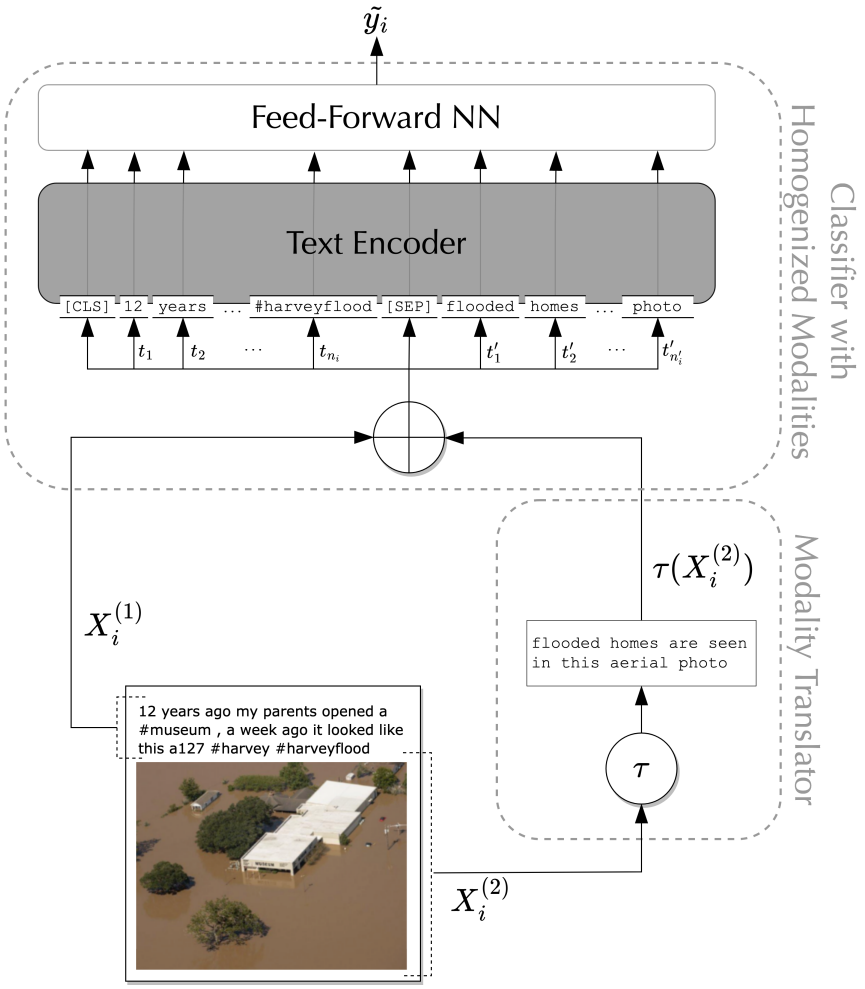


Figure 4.1: Overview of the CMB method structured in two main parts: a non-learnable process represented by the modality translator (image captioning model) and the homogenized modalities classifier as trainable module.

4.2 Image captioning for multimodal fusion

In previous works, captions have been used to represent images for multimodal interaction, such as for Multimodal Named Entity Recognition (MNER) [114]. It has been argued that models trained for semantic understanding, like image captioning models, may provide better image representations. The use of captions as image representations allows for the incorporation of textual descriptions of the image content, providing additional context and useful information to understand and interpret the image. Moreover, [6] uses captions as a translated version of image in addition to the tweet's image and text to improve performance in classifying multimodal tweets. Contrary to [6], the method CMB proposed in this chapter uses - through captions - modality translation as a substitute and not as a complement to the image to improve inter-modality interactions while keeping the modal simple to train and use in real-world scenarios (with limited resources). Moreover, we argue that the use of one encoder helps to overcome the problem of different convergence rates of each modality encoder and lead to more efficient learning.

4.3 Methodology

4.3.1 Problem statement

As mentioned in Section 3.2, $\mathcal{D}$ is a multimodal dataset with N instances, denoted by $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^{N}$. Each $X_i \in \mathcal{X}$ represents a multimodal data point with M modalities, represented as $X_i = \{X_i^{(p)}\}_{p=1}^{M}$. Associated with each instance is a label $y_i \in \mathcal{C}$, where $\mathcal{C}$ is the set of possible labels. The multimodal classification task aims at learning a classifier $h : \mathcal{X} \rightarrow \mathcal{C}$ that combines information from different modalities to make accurate predictions. In the following, as described in Section 3.2 we consider only bimodal instances ($M = 2$) where the two modalities are different in nature, typically textual and visual modalities. Thus, $X_i = (X_i^{(\text{text})}, X_i^{(\text{image})})$, $X_i^{(text)} = (t_1, \ldots, t_{n_i})$ is a sequence of tokens while $X_i^{(image)} \in \mathbb{R}^{C \times H \times W}$ is an image (where C, H, and W are the number of channels, the height, and the width of the image respectively). We specifically consider only the *concordant* instances.

4.3.2 Model overview

Since modalities of different natures (typically text and image) are hard to combine in multimodal fusion strategies, we propose a two-step categorization approach that leverages modality translation to homogenize the inputs (Figure 4.1). The first step,

Modality Translator consists in translating one modality into a space compatible with the space of the other modality using a translator function $\tau(.)$. In our particular context of text-image tweets processing, the image modality $X_i^{(image)}$ is translated into a text (sequence of tokens) by means of an image captioning process $\tau(X_i^{(image)}) = (t'_1, \ldots, t'_{n'_i})$. The second step *Classifier with Homogenized Modalities* takes as input the concatenation of the homogenized modalities $X_i^{(text)} \oplus \tau(X_i^{(image)})$ to train both a text encoder and the final classification layer. Let us notice that the approach allows to process both multimodal ($X_i^{(text)} \oplus \tau(X_i^{(image)})$) or unimodal data ($X_i^{(text)}$ only or $\tau(X_i^{(image)})$ only).

4.3.3 Modality translators

In our multimodal approach, we treat the captioning model as a translator between different modalities. For this investigation, we utilize three publicly available captioning models, each differing in their choice of encoders and the training data used. We start by discussing CLIP, as it is the most commonly used image encoder, and then we describe the details of each captioning model .

CLIP [2] (Contrastive Language-Image Pretraining) as shown in Figure 4.2 is a model that learns joint representations of images and text by leveraging a large dataset of 400 million image-text pairs. It consists of two encoders: an image encoder $f_\theta(I)$ and a text encoder $g_\phi(T)$, which map images and text into a shared embedding space. The resulting feature vectors $\mathbf{v}_I = f_\theta(I)$ and $\mathbf{v}_T = g_\phi(T)$ are normalized, and their similarity is measured using cosine similarity. CLIP is trained using a contrastive loss function that maximizes the similarity between correct image-text pairs while minimizing it for incorrect pairs. This design allows the model to create highly correlated visual and textual representations, significantly reducing both training time and data requirements as showed by [115]. Consequently, the resulting model is capable of effectively transferring its learned representations across various tasks, including zero-shot classification and effectively encoding images for image captioning. Recent advancements in image captioning models have taken advantage of CLIP's robust multimodal embeddings.

Several approaches have shown that CLIP is highly effective as a core component for generating high-quality image captions. Additionally, we use a model based on RCNN [116] as the image encoder. The details of these models are provided below:

- **CLIP Prefix for Image Captioning** [115]. A captioning model using the CLIP [2] (Contrastive Language-Image Pre-training) multimodal encoder, which is trained on a large dataset sourced from the Web. It extracts image features using the CLIP encoder and employs them as a prefix for text generation by a lan-

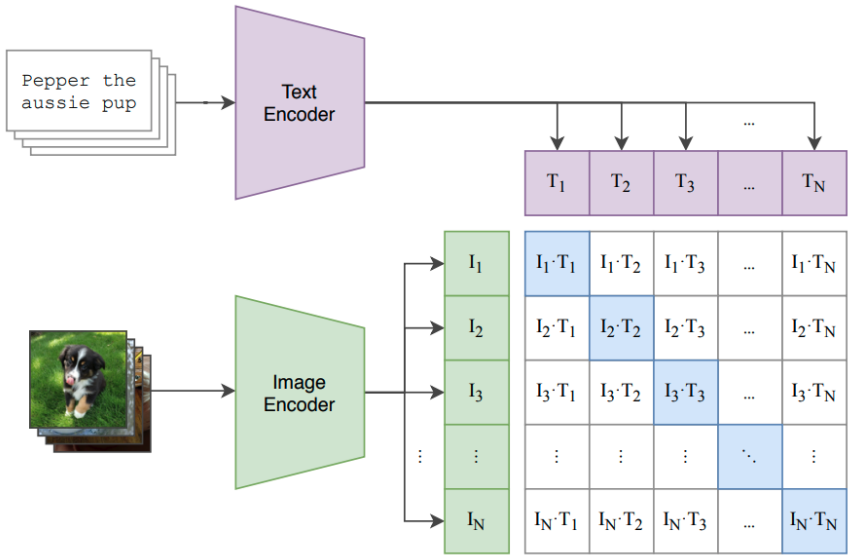


Figure 4.2: Overview of CLIP architecture [2].

guage model. Trained with either the MSCOCO dataset [110] or the Conceptual Caption dataset [111], the authors argue that the rich semantic features of the CLIP encoder, coupled with a pre-trained language model (GPT2 [117]), provide a comprehensive understanding of both visual and textual data, resulting in high quality captions. In the following, these captioning models are referred to as *CLIP_cc* (trained with Conceptual Captions) and *CLIP_coco* (trained with MSCOCO).

- **Fine-grained Image Captioning with CLIP Reward**. In [118] the authors use CLIP [2] in two ways, first as an image encoder, and then CLIP text and CLIP image are used to compute a multimodal similarity that then serves as a reward function for each generated caption to obtain descriptive and distinctive captions. This captioning model is referred to as *CLIP_reward* in the following.

- **Transformer based Captioning**.[75] uses the Transformer architecture and first process the image using Mask-RCNN [119], the features are then fed to the Transformer in order to generate the caption. This captioning model is referred to as *Trans_cap* in the rest of the article.

The CMB approach we propose in this work allows to independently integrate any of the four captioning models presented above. In the following experiments, we will focus on measuring the contribution of the captions generated by each model in our proposed multimodal framework.

4.3.4 Multimodal classifier with homogenized modalities

Transformer based models [1, 75, 117, 120] has achieved state of the art result in large set of natural language processing tasks, by leveraging various self-supervised pre-training techniques. Therefore, without loss of generality, we use BERT [1] as text encoder to extract embeddings from the input $(X_i^{(text)} \oplus \tau(X_i^{(image)}))$, which are then used to perform the classification task using a feed forward neural network layer.

4.4 Experimental Setup

4.4.1 Dataset and tasks

To test our method we use **CrisisMMD** benchmark, as discussed in Section 3.2, the dataset focuses on identifying crisis related tweets, specifically tweets labeled for Informativeness and Humanitarian tasks. Since the dataset assigns separate labels for the text and image components of each tweet, it is not directly applicable to multimodal classification tasks, leading to the adoption of a method that uses only *concordant* tweets with matching labels across both modalities, as described in [13].

4.4.2 Baselines

We compare our approach to both uni- and multimodal baselines in the field of image-text classification. Recently, the trend is to use multiple architectures in this research area. To provide a comprehensive evaluation of our approach, we implemented two popular multimodal architectures: the feature concatenation [109] and the cross-attention [94] methods. Additionally, we included the unimodal methods BERT [1] (text only) and DenseNet [86] (image only) for comparison. Furthermore, we conduct a comparison between our approach and two other methods: CLIP [2] and CLIP-Concat. In the latter, we concatenate features from CLIP image encoder and BERT to accomplish the multimodal classification task. We aim to evaluate the effectiveness of CMB against these well-established methods to illustrate its usefulness in the crisis domain.

4.4.3 Implementation details

In this work we consider the four modality translators described above. Specifically, for each captioning model, we produced five captions for every image using the authors' provided demo for CLIP-based captioning models (CLIP_cc, CLIP_coco and

CLIP_reward) and the demo from a GitHub repository[1] for the Trans_cap model. We keep the most similar caption - from the five generated by each model - to the image in terms of clip similarity score [118].

We used the BERT base model from [107] as a text encoder with token type embeddings different for captions ($\tau(X_i^{(image)})$) and texts ($X_i^{(text)}$). To classify, we computed the mean pooling on embeddings provided by BERT and used a classification layer consisting of a ReLU activation function and a dropout layer between two feed-forward layers. The models were trained on a Nvidia Tesla V100 GPU for 3 to 5 epochs with $batch\ size = 128$ and $learning\ rate = 5e^{-5}$.

4.4.4 Unimodal and multimodal settings

To evaluate the robustness of our methodology across diverse scenarios, we conduct an array of experiments involving the training and evaluation of various iterations of our model, each employing distinct modalities such as Text, Image, or Caption. Table 4.1 uses specific notations to delineate these different experimental setups. The first three rows of the table focus on unimodal approaches (T_T, I_I, and C_C settings), wherein a single modality is exclusively used for both training and evaluation. For text categorization (T_T, C_C), we employ BERT [1], while for image categorization (I_I), DenseNet [86] is utilized. The fourth row shows multimodal approaches (TI_TI setting), involving the simultaneous use of Text and Image during both training and evaluation. This incorporates a basic feature concatenation strategy, along with the previously mentioned cross-attention architecture [94] which uses attention mechanisms to merge representations from the two unimodal architectures (BERT and DenseNet). Finally, the last line corresponds to our multimodal method based on modality translation, whose architecture is generic enough to adapt to multiple settings. We investigate its ability to learn efficient hybrid models for categorizing both multimodal and unimodal data by training it on both text and caption and then evaluate it on different data configurations: Text only (TC_T), Caption only (TC_C) and Text $\oplus$ Caption (TC_TC).

We utilize the same dataset across various configurations. In the TC configurations, both text and caption are used as input during training or testing phases. In the T configuration, only text is used as input to the model, with captions being ignored. Conversely, in the C configuration, only captions are used as input, while text is disregarded. This data split remains consistent across all configurations, encompassing training, validation, and testing phases.

To expand our experiments, we propose a mixed training strategy in which the model

[1] https://github.com/ruotianluo/self-critical.pytorch.

Table 4.1: Training and evaluation settings.

TrainEval	*Text*	*Image*	*Caption*	*Text + Image*	*Text ⊕ Caption*
Text	T_T	-	-	-	-
Image	-	I_I	-	-	-
Caption	-	-	C_C	-	-
Text + Image	-	-	-	TI_TI	-
Text ⊕ Caption	TC_T	-	TC_C	-	TC_TC

is trained on a mixture of data that contains text-only instances ($X_i^{(text)}$) and on instances with Text ⊕ Caption ($X_i^{(text)} \oplus \tau(X_i^{(image)})$). The details of our experiments are presented in the following sections.

4.5 Experimental results

This section presents the results of our experiments on the proposed method using different captioning models compared with unimodal and multimodal baselines. Additionally, we evaluate the robustness of CMB on unimodal data settings.

4.5.1 Captioning models comparison

Table 4.2 displays the results of our experiments on the two multimodal tasks of the CrisisMMD dataset (Informativeness and Humanitarian). The first two rows report unimodal classifiers results, while the next two rows present the performances of multimodal baselines. The last four rows show the performances of our approach using each of the four captioning models described before. Firstly, the results obtained confirm that multimodal methods outperform unimodal ones. Additionally, our method is very competitive with other multimodal baselines, regardless of the modality translator used.

Specially, the performance gap between the different captioning models can be significant, as demonstrated by the 3% performance gap on the informativeness task between the method based on Trans_cap and the best-performing method. Importantly, the CLIP_cc captioning model allows the CMB method to outperform the proposed strong baseline of 1-3% in terms of weighted F1-score on both CrisisMMD tasks. Moreover, employing the identical data splits, our experiments yield competitive results with state-of-the-art [95] findings. In the tasks of informativeness and humanitarian, we scored 91.96 and 94.84 W-F1, respectively, while their scores were 91.3 and 93.6. Furthermore, we can observe the gap between the CLIP_cc and CLIP_coco models despite using the

Table 4.2: Comparisons on CrisisMMD in terms of classification accuracy, Macro F1-score and weighted F1-score.

Models		Informativeness			Humanitarian		
		Acc	F1-m	F1-w	Acc	F1-m	F1-w
Unimodal	BERT	0.8691±0.0031	0.8465±0.0041	0.8667±0.0033	0.8183±0.0068	0.6964±0.0271	0.8179±0.0069
	DenseNet	0.8393±0.0028	0.8186±0.0022	0.8396±0.0024	0.7608±0.0178	0.5366±0.0495	0.7416±0.0272
	CLIP	0.8594±0.0281	0.8407±0.0283	0.8591±0.0265	0.8136±0.0224	0.6105±0.0404	0.8005±0.0325
Multimodal	Features-concat	0.9014±0.0020	0.8870±0.0031	0.9008±0.0023	0.8561±0.0102	0.6772±0.0070	0.8533±0.0098
	Cross-Attention	0.9022±0.0037	0.8885±0.0043	0.9019±0.0037	0.8549±0.0060	0.6718±0.0059	0.8520±0.0058
	CLIP-Concat	0.9020±0.0045	0.8871±0.0054	0.9012±0.0046	0.8608±0.0080	0.7778±0.0146	0.8589±0.0077
CMB	CMB-Trans_cap	0.8920±0.0039	0.8749±0.0045	0.8907±0.0039	0.8527±0.0065	0.7536±0.0274	0.8520±0.0070
	CMB-CLIP-reward	0.9125±0.0033	0.8987±0.0040	0.9115±0.0034	0.8765±0.0080	0.7631±0.0266	0.8755±0.0079
	CMB-CLIP-coco	0.9098±0.0024	0.8954±0.0029	0.9087±0.0025	0.8681±0.0064	0.7156±0.0307	0.8660±0.0066
	CMB-CLIP-cc	**0.9203±0.0036**	**0.9081±0.0043**	**0.9196±0.0037**	**0.8827±0.0033**	**0.7810±0.0176**	**0.8826±0.0032**

same architecture: we hypothesize that the Conceptual Caption dataset [111] has more coverage and similarity with Twitter images, containing figurative representations with different levels of iconicity such as photographs, drawings, charts, or even non-figurative pieces of digital art for instance. Finally, the clear gap between the CLIP-based models and the simple Transformer model confirms the findings of [121], indicating that CLIP generalizes better on human-centric tasks (defined by the authors as involving knowledge of cultural, social, aesthetic and/or affective components of the world). In our case, this fits particularly our need on social media, where humans post for other humans. This result is supported by a recent study [122], where the authors used LLaVA [123], a Multimodal Large Language Model (M-LLM), to generate captions and explanations, which were then fed into a RoBERTa model to classify multimodal tweets. Their results on *Informativeness* showed that using only the caption outperformed other methods. It's worth noting that our model achieved better performance, with a 92% accuracy compared to their 91%.

In order to expand our investigations, we compared our approach with MARMOT [6] : a model incorporating images, captions, and textual elements for tweets classification. Employing two distinct captioning models, namely Trans_cap and CLIP_cc, the experiments outcomes are shown in Table 4.3. Despite the diminished parameter count (approximately one-third) in our method relative to MARMOT, it exhibited competitive performance in both scenarios. Noteworthy disparities were observed, particularly in the Trans_cap scenario, whereas the performance differentials were closely aligned when utilizing CLIP_cc captions. This underlines the influence of employing high-quality captions in this configuration. In the continuation of this paper, we use the best captioning model, namely CLIP_cc.

Table 4.3: Comparison with MARMOT [6], a multimodal model integrating caption, text, and image modalities. Two captioning models, namely *Trans_cap* and *CLIP_cc*, are employed. The parameter count (in millions) is reported, excluding parameters associated with captioning models.

	Informativeness			
	Acc	F1-m	F1-w	Parameter count
MARMOT Trans_cap	0.9133±0.0018	0.9012±0.0021	0.9131±0.0018	309
CMB-Trans_cap	0.8920±0.0039	0.8749±0.0045	0.8907±0.0039	112
MARMOT CLIP_cc	0.9220±0.0022	0.9105±0.0024	0.9215±0.0021	309
CMB-CLIP_cc	0.9203±0.0036	0.9081±0.0043	0.9196±0.0037	112

4.5.2 Unimodal and multimodal results

As discussed previously, the CMB method contains two parts: a modality translator and a trainable module (text encoder + classifier). The input of the trainable part is exclusively text-based, enabling the use of both unimodal ($X_i^{(text)}$ or $\tau(X_i^{(image)})$) and multimodal ($X_i^{(text)} \oplus \tau(X_i^{(image)})$) data as input. This section investigates the unimodal and multimodal capabilities of the model by conducting an experiment with three main configurations, TC_T, TC_C and TC_TC. Results of the experiments are reported in Figure 4.3.

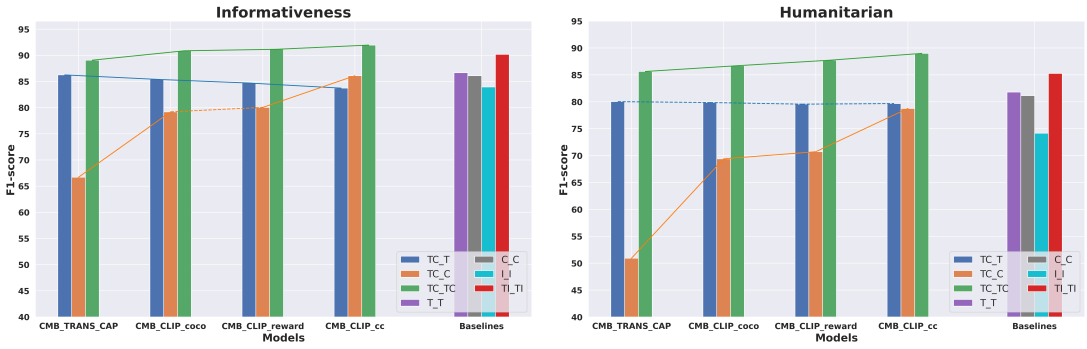


Figure 4.3: Comparison of modality translators on Unimodal and Multimodal settings (see Table 4.1). Y-axis shows mean Weighted F1-score results over 10 runs. The X-axis represents the different modality translators used in this experiment. Dotted lines are used for non-significant differences.[2]

Performances of CMB (for the different captioning models) is compared across the three main configurations as well as to BERT [1] (as T_T configuration), CMB_CLIP_cc (C_C), DenseNet [86] (I_I) and the features concatenation model (TI_TI). As men-

[2]We used significance level = 0.05

tioned in the previous section, the results clearly illustrate the improved performance when using the two modalities sources for classification. Additionally, these experiments reveal a relationship between the captions-only (TC_C) and multimodal (TC_TC) performances, indicating that the model performance depends on the caption quality. The figures also show that captions-only models can outperform the (I_I) DenseNet model. However, the model in the TC_T setting exhibits poor performance compared to BERT (in T_T configuration) with a 2-3% loss. These results suggest a form of *caption dependency* of the models. To address this issue, we discuss a mix training method in the next section.

4.5.3 Mix training analysis

Twitter posts can be either unimodal or multimodal, but in practice, keywords are used as filters to retrieve tweets with the Twitter API, making some tweets containing only text (without image). Rather than using two different models depending on the nature of each tweet (unimodal or multimodal), we suggest that the CMB model can be used to process both unimodal and multimodal tweets. This hybridization requires an appropriate training strategy, called the *mix training* approach. In this study, we propose an experiment to test our method in TC_T and T_T settings. The goal of this mix training approach is to train a hybrid model that can handle both multimodal (*text+caption*) and unimodal (*text-only*) tweets. To this end, we consider training samples with various

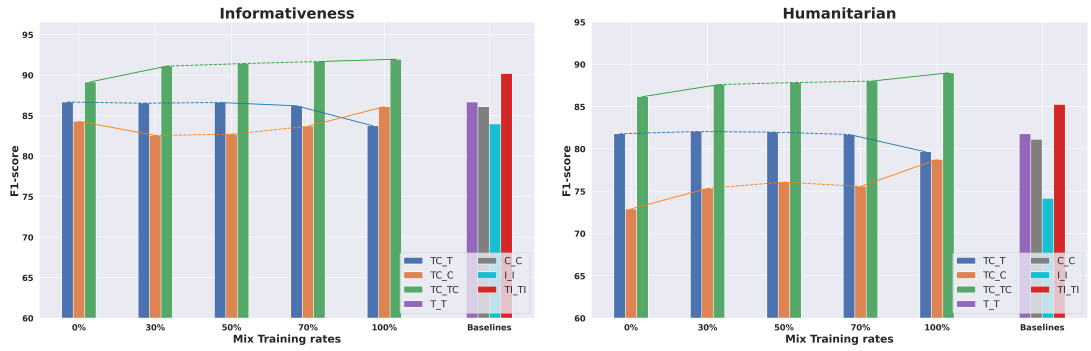


Figure 4.4: Mix training results. In the X-axis, $X\%$ represents the proportion of multimodal instances ($X_i^{(text)} \oplus \tau(X_i^{(image)})$) in training data, the remaining $100 - X\%$ data are text only ($X_i^{(text)}$). Dotted lines are used for non-significant differences.

text-only/*text+caption* balances and observe the ability of the derived models to classify *text-only*, or *text+caption* tweets. In this experiment, the captioning model used is CLIP_cc. The proportions of *text+caption* tweets in the training sample is increased from 0% to 100%, with the remaining examples being *text-only* tweets. The results of

the mix-training strategy, as presented in Figure 4.4, provide interesting insights into the performance of the hybrid model in different scenarios. When evaluated on text-only tweets (TC_T), the model's performance (F1-score) experiences a slight decrease of approximately 2-3% as the proportion of text+caption tweets in the training sample increases from 50% to 100%. This indicates that the introduction of captions has affected the performance of the model on text-only settings, specially when having more than 70% *text+caption* examples. Additionally, it's evident that the performance in caption-only evaluation (indicated by the orange bar in TC_C) improves as the proportion of *text+caption* examples increases. However, there is one unexplained anomaly observed during the informativeness experiment when transitioning from 0% to 30% of *text+caption* examples.

On the other hand, in the multimodal setting (TC_TC), the same change in the training sample, with an increase in the proportion of text+caption tweets from 50% to 100%, results in an improvement of about 0.5-1% in the model's performance. This suggests that the hybrid model benefits from exposure to a higher proportion of multimodal data during training, leading to better performance in the multimodal tweet classification task. The proposed hybrid model demonstrates its competitiveness with BERT (T_T) when trained on a sample with 50% text+caption tweets, achieving an impressive F1-score of 81.98%. Despite a small loss of performance in the TC_TC setting compared to the text-only scenario, the hybrid model remains clearly superior to the TI_TI baseline, showcasing its capability to effectively handle both unimodal and multimodal tweets. Furthermore, the analysis reveals that the best compromise on caption rates, is achieved at 50%. This suggests that a balanced mixture of text-only and text+caption tweets in the training sample is optimal for achieving peak performance in the hybrid model. Finally, this experiment indicates that the mix training strategy makes the learned models more robust and usable for classifying both unimodal and multimodal tweets in real time.

4.5.4 Qualitative analysis

Finally, we provide several examples to illustrate the success and failure cases of our proposed method. Figure 4.5 shows three examples in which the cross-attention baseline and BERT failed, but our method made the correct prediction on the Humanitarian task. In particular, Figure 4.5(a) is an example of the success of using a caption, as the synthetic image is difficult to be processed by a conventional image model pre-trained on ImageNet [124]. In Figure 4.5(b) the successful generation of the words "blocks the road" may have contributed to the prediction of the correct label, highlighting the model's ability to capture discriminative information. Finally, Figure 4.5 (c) is a multimodal example

Figure 4.5: Examples of caption using success cases when baseline BERT and Cross-attention failed

in which the text alone fails to predict the label; in this case, the caption captures information that is merged with the text to achieve the right prediction. In Figure 4.6, we present three examples where the caption model produced inaccurate results. The image in Figure 4.6(a) illustrates the potential of captions to be misleading and result in incorrect predictions; the model captured some information from the image in a poorly structured sentence, resulting in an incorrect prediction. The image in Figure 4.6(b) demonstrates a failure due to an inaccurate caption generated by the captioning model, highlighting the sensitivity of our approach to the modality translator. Finally, the image in Figure 4.6(c) shows a situation in which both the label and the prediction fall into a gray area, where either could be considered a valid label. These examples provide insight into the successes and failures of our proposed method, specifying potential areas for improvement.

Figure 4.6: Examples of caption leading to failure cases when baseline BERT and Cross-attention succeed.

4.6 Error analysis

To provide a more rigorous comparison between the two multimodal fusion methods—namely, caption-based fusion and cross-attention—we evaluate the performance of the models on specific cases where one modality is defective. Specifically, we compare the predicted labels from each fusion method with those from their respective unimodal models used in the fusion process. For the text modality, we use BERT predictions in both cases. For the image modality, we use predictions from Clip_cc (in the C_C setting with BERT as the encoder) for the caption-based fusion method, and DenseNet for the cross-attention model. Our goal is to investigate how the models utilize information from each modality. We suppose that when a unimodal classifier correctly classifies an example, the necessary information is available to the multimodal encoder. Therefore, an optimal fusion method should correctly classify every example where at least one of the unimodal models predicts the correct label. Formally, consider an example X_i with true label y_i. Let: $y(X_i^{(\text{text})})$ be the prediction from the text model (BERT), $y(X_i^{(\text{image})})$ be the prediction from the image model (DenseNet), $y(\tau(X_i^{(\text{image})}))$ be the prediction from the caption-based model.

We aim for the fusion model to satisfy:

$$\text{If } y(X_i^{(\text{text})}) = y_i \text{ or } y(X_i^{(\text{image})}) = y_i, \text{ then } y_{\text{fusion}}(X_i) = y_i.$$

For the caption-based fusion method, the condition becomes:

$$\text{If } y(X_i^{(\text{text})}) = y_i \text{ or } y(\tau(X_i^{(\text{image})})) = y_i, \text{ then } y_{\text{fusion}}(X_i) = y_i.$$

Table 4.4: Performance of Fusion Models when only one modality predicts correctly. 'Text' refers to cases where $y(X_i^{(\text{text})}) \neq y_i$ and $y(X_i^{(\text{image})}) = y_i$; 'Caption' or 'Image' refers to cases where $y(X_i^{(\text{text})}) = y_i$ and $y(X_i^{(\text{image})}) \neq y_i$. Proportions are shown with counts in parentheses.

Method	**Modality**	$y_{\text{fusion}} = y_i$	$y_{\text{fusion}} \neq y_i$	**Mean accuracy**
CMB-CLIP_cc	Text	68.8% (77)	31.2% (35)	**78.4%**
	Caption	88.1% (104)	11.9% (14)	
Cross-Attention	Text	70.9% (78)	29.1% (32)	**73.9%**
	Image	76.9% (130)	23.1% (39)	

In the context of humanitarian tasks, we analyzed the performance of two multimodal fusion methods: Cross-Attention and CMB-CLIP_cc. Our focus is on examples where only one modality predicts the correct label y_i, specifically cases where either the text modality $y(X_i^{(\text{text})})$ or the image modality $y(X_i^{(\text{image})})$ is incorrect.

As shown in Figure 4.3, DenseNet (image modality) exhibits more errors than the caption-based model. This indicates that the image modality processed by DenseNet is less reliable compared to the caption-based approach, this explains the difference results between Cross-attention and CMB-CLIP_cc on the image and caption lines. From Table 4.4, we observe that the CMB-CLIP_cc model classifies correctly more often when the image model predicted a wrong label, achieving an accuracy of 88.1% in this scenario, compared to 76.9% for the Cross-Attention model. This suggests that CMB-CLIP_cc is more effective at compensating for errors in the image modality. Additionally, CMB-CLIP_cc experiences only a negligible drop in accuracy when the text modality is wrong (68.8%) compared to Cross-Attention (70.9%), indicating that it maintains robust performance even when the text modality is unreliable. Overall, the CMB-CLIP_cc model is more efficient when one modality provides incorrect information, achieving 78.4% accuracy (compared to 73.9% of Cross-attention model), in these cases by leveraging a more efficient fusion process that effectively combines the correct predictions from the modality to improve overall accuracy. However, both models show a bias toward the text modality, meaning they are more likely to make incorrect predictions when the text modality fails to predict the correct label. This bias is stronger in the CMB-CLIP_cc model but is also noticeable in the Cross-Attention model. In the next chapter, we will analyse more the problem of imbalanced modalities. This issue is well-known in audio-video multimodal learning and can significantly affect the performance of fusion models. Addressing this imbalance can help improve the effectiveness of multimodal fusion in humanitarian tasks.

4.7 Conclusion

In this chapter, we introduced CMB (Caption-based Multimodal BERT) a multimodal classification approach, designed to address the challenges inherent in visual and textual fusion, particularly in the context of social media posts during crises. Our approach leverages a modality translation mechanism, where an image is transformed into its textual representation using a captioning model, which is then fused with the accompanying tweet text. This method simplifies the fusion process by working within a shared textual space.

Our experiments on the CrisisMMD dataset demonstrated that CMB achieves competitive results, often outperforming traditional multimodal baselines. Specifically, CMB showed a significant improvement over models that rely on basic image-text fusion methods, such as feature concatenation and cross-attention. The results indicate that translating images into text using high-quality captioning models can enhance the interaction between modalities, making it easier to process and classify the data effectively. One of the key findings of our work is the impact of caption quality on model performance. Through comparisons of different captioning models, we observed that the CLIP-based captioning models, particularly CLIP_cc, outperformed others such as Trans_cap. This reinforces the idea that better semantic understanding of images through captioning leads to more effective multimodal fusion. However, while the quality of captions positively influenced results, the performance was still robust across different captioning models, demonstrating the flexibility of CMB in adapting to various pre-trained caption generators. Another significant contribution of our study was the development of a mixed training strategy. In real-world scenarios, social media data is often incomplete, with some posts lacking images. To address this, we introduced a hybrid model that can handle both unimodal and multimodal inputs. Our experiments showed that this hybrid model performs competitively, maintaining high accuracy even when trained with varying proportions of text-only and text-plus-caption data. This adaptability is crucial for deploying models in real-world settings where image availability is inconsistent.

Despite the strong performance of CMB, there are still areas for improvement. For instance, our qualitative analysis revealed that the model is sometimes sensitive to the quality of generated captions. In cases where captions were poorly generated or provided incorrect information, the model struggled to make accurate predictions. This highlights the need for further improvements in captioning models. Additionally, while our method focuses on crisis-related tasks, it would be valuable to investigate its applicability to other domains. Another limitation lies in the tendency of multimodal fusion models to favor one modality over another, particularly the text modality. Our error analysis suggests that both CMB and other fusion models exhibit a bias toward the text modality,

potentially overlooking useful information from images. Addressing this issue, perhaps by improving modality balancing techniques, could further enhance model performance. In the next chapter we will dive deep in this problem of imbalance multimodal learning, we will propose a method to get more balance training dynamics and better overall performances.

Chapter 5

Dynamic Regularization Strategy for Mitigating Modality Imbalance in Multimodal Learning

5.1 Introduction

In recent years, the field of Multimodal Learning (MML) has garnered significant attention due to its potential to emulate human-like perception and interpretation of the world [125, 126]. By integrating information from multiple sensory channels such as vision, audio, and text, MML aims to provide a more comprehensive understanding of the external environment. This approach has led to substantial progress in various domains, including video classification [127, 128], event localization [129], action recognition [130], and audiovisual speech recognition [131]. The real world is inherently multimodal, with objects and events characterized by multiple modalities. For instance, action recognition can integrate data from video, audio, and motion sensors [132–134]. Multimodal data is generally more informative and diverse than single-modal data, offering a richer representation of the world. Consequently, multimodal learning has the potential to surpass single-modal approaches, attracting widespread attention across various domains. Despite the promise of multimodal approaches to outperform their unimodal counterparts, effectively leveraging diverse modalities remains a significant challenge. A key issue in this context is "modality imbalance", where learning efficiencies and convergence rates differ substantially between the modalities involved in training [113, 135, 136]. This imbalance can hinder the optimal exploitation of multimodal capabilities, potentially leading to suboptimal performance or even inferior scores compared to single-modal models in certain situations [113, 135–137]. The modality imbalance problem arises from the

presence of a dominant modality and subordinate modality during training. Due to the inherent greediness of the learning process [135], model updates tend to lean excessively towards the dominant modalities, neglecting the learning of non-dominant ones. This phenomenon has been observed across various multimodal tasks [113, 135], highlighting the inefficiency in leveraging and fusing information from diverse modalities as a significant challenge in the field of multimodal learning. In the context of multimodal crisis-tweet classification, the previous chapters demonstrated the advantages of using multimodal data over unimodal data, as it leads to better performance. However, as discussed in Section 4.6, we also found that these models still suffer from the same modality imbalance problem, with a tendency to rely more on text.

To address this issue, researchers have proposed various methodologies. Some approaches involve using additional classifiers for each modality and employing gradient mixing [113], while others aim to improve unimodal performance through knowledge distillation from well-established models [138]. Alternative techniques include the DropPathway method [139], which randomly omits one modality during training, and on-the-fly gradient modulation to regulate the learning rate of the dominant modality [136]. Another approach involves using class prototypes to stimulate the learning of slower modalities [137], further addressing the imbalance between modalities. Most existing studies [113, 136, 138] have focused on adjusting each modal gradient during backpropagation, either by assigning different learning rates to different modal branches or by introducing additional losses for each modality. These strategies aim to maximize the contribution of each modality, typically by controlling the learning process at the modal level by updating the complete parameters of each modality. In this work, we explore the application of Stochastic Shared Embeddings (SSE) [140] to address the challenge of modality imbalance in multimodal learning. SSE is a data-driven method for regularizing embedding layers by stochastically transitioning between embeddings during stochastic gradient descent. Building upon the success of SSE in multimodal contexts [94], we introduce a novel extension: Dynamic Stochastic Shared Embeddings (D-SSE). Our approach dynamically regularizes the dominant modality, enabling a controlled balance that allows the slower-learning modality to catch up without compromising overall system performance. Through comprehensive empirical evaluations, we show the efficiency of D-SSE in not only mitigating modality imbalance but also enhancing the performance of multimodal learning systems. Figure 5.1(a) for instance reveals a clear imbalance in multimodal training on the CREMA-D dataset [3]. It demonstrates the domination of the *Audio* modality in the training process, with the $Audio - Video$ model closely following the audio model's performance. In contrast, Figure 5.1(b) illustrates a more balanced training scenario, accompanied by an overall rise in $Audio - Video$ performance. This marked difference in model behavior, particularly the improved balance and increased

accuracy observed with the D-SSE approach, motivates further exploration of D-SSE's potential on other datasets to validate its efficacy in enhancing multimodal learning.

This chapter presents several key contributions to addressing the modality imbalance problem in multimodal learning. First, we propose a novel, model-agnostic balancing method called D-SSE, designed to regulate the dominant modality and promote more balanced learning. Our method is compatible with a range of fusion techniques and can be applied in various imbalance scenarios. Furthermore, we conduct a detailed analysis of the imbalance during training by examining the utilization rate of each modality, along with an evaluation of the similarities in the embedding spaces of different modalities. By addressing the modality imbalance problem, we aim to improve the efficiency of multimodal learning and unlock the full potential of integrating information from diverse modalities.

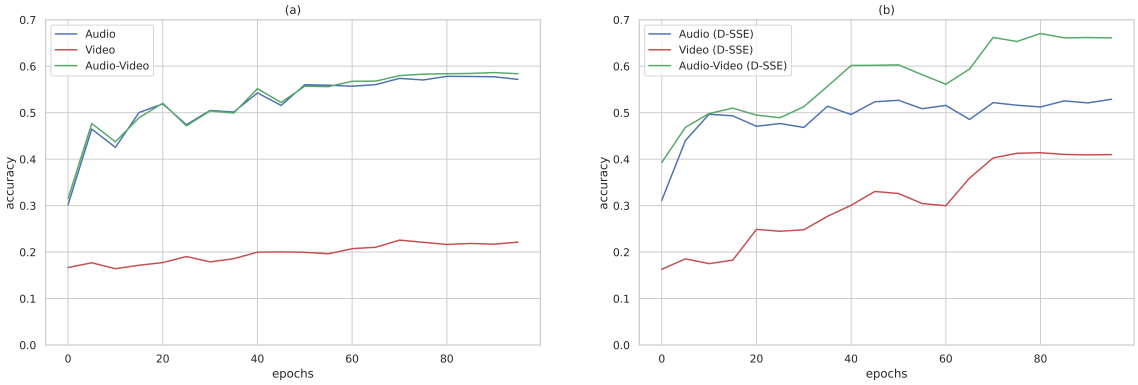


Figure 5.1: Unimodal and multimodal performance, Models trained on CREMA-D [3], we report the results of the Audio, Video and multimodal (Audio-Video) using Film [4] fusion method. fig (a) in the left show the validation accuracies when training on baseline method, in the right fig (b) shows the validation accuracies when trained with the D-SSE method.

5.2 Multimodal learning imbalance

Multimodal deep neural networks have gained considerable attention in recent years for their ability to leverage information from diverse modalities. However, the effectiveness of these models is hindered by modality imbalance, where certain modalities exhibit discrepancies in convergence rates and generalization performance.

Early study by [113] elucidated the varying convergence rates and generalization ca-

pabilities across modalities, it showed that certain modalities perform more effectively in a unimodal context than when integrated with another modality in a multimodal training setup. On the other hand [136] identified the dominance of superior-performing modalities in gradient updates, potentially overshadowing others. Moreover, [135] introduced a conditional rate utilisation of a modality in a multimodal training, this metric allowed to quantify the imbalance and confirmed the problem of modality imbalance. Furthermore it has been noticed that multimodal neural networks can exploit modality bias present in the data, this is particularly evident in applications like Visual Question Answering (VQA), where the model uses text bias to answer the questions avoiding cross-modal interactions [141, 142].

To mitigate this limitation, several strategies have been proposed. For instance, [113] introduced gradient blending techniques to mix gradients based on generalization and overfitting of each modality. [136] proposed on-the-fly gradient modulation, adjusting learning rates dynamically during training based on approximate uni-modal performance. Additionally, [137] suggested using class prototypes in order to accelerate the slow-learning modality by enhancing its clustering toward prototypes, [138] focused on knowledge distillation from well-trained models to enhance uni-modal branches performances. These approaches offer insights and directions for addressing modality imbalance and advancing the efficacy of multimodal learning frameworks.

Building upon the effectiveness of Stochastic Shared Embeddings (SSE) in multimodal frameworks, as highlighted in previous research [94]. Originally developed for unimodal contexts, SSE is used within the CrisisMMD [14] scenario to enhance the dataset with new multimodal examples. it employs a data-driven methodology that regularizes embedding layers by stochastically transitioning between embeddings during the stochastic gradient descent process. Specifically, in the CrisisMMD setting, these transitions are applied to create new multimodal instances by stochastically pairing text and images from different multimodal instances. However, this approach is not designed to balance the learning process, nor does it consider the changing learning dynamics within each modality during training. Recent studies, such as [135], have showed that incorporating adaptive balancing method can lead to more balanced models with improved performances. In this context, we propose a novel method called *Dynamic Stochastic Shared Embeddings (D-SSE)*. Unlike traditional SSE, D-SSE dynamically adapts the regularization rate for each modality based on its performance throughout the training process. This dynamic adaptation allows us to modulate the learning influence of the dominant modality, thereby providing the less dominant modality with an enhanced opportunity to contribute to the model's learning process.

5.3 Problem statement

Let $\mathcal{D}$ be a multimodal dataset with N instances, denoted by $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^{N}$. Each $X_i \in \mathcal{X}$ represents a multimodal example with M modalities, represented as $X_i = \{X_i^{(p)}\}_{p=1}^{M}$. Associated with each instance is a label $y_i \in \mathcal{C}$, in a set of K possible labels $\mathcal{C} = \{1, \ldots, K\}$.

The multimodal classification task aims at learning a classifier $h : \mathcal{X} \rightarrow \mathcal{C}$ that combines information from different modalities to make accurate predictions. In the following, we consider only bimodal instances ($M = 2$) where the two modalities are different in nature. Generally in the case of two modalities, we use a multimodal Deep Neural Network (DNN) with two unimodal branches, denoted by ϕ_0 and ϕ_1, taking $X_i^{(0)}$ and $X_i^{(1)}$ as input. A fusion module f is added on top of the encoders, such that

$$\hat{Y}_i = f(Z_i^{(0)}, Z_i^{(1)}) \tag{5.3.1}$$

where $Z_i^{(0)} = \phi_0(X_i^{(0)})$ and $Z_i^{(1)} = \phi_1(X_i^{(1)})$.

Here, $\hat{Y}_i \in \mathbb{R}^K$ where K is the number of classes. The result is then fed to a softmax function (σ) to predict the label of the example :

$$\hat{y}_i =_{k=1\ldots K} \sigma_k(\hat{Y}_i) \tag{5.3.2}$$

In order to track the performance of each unimodal branch independently of the fusion mechanism, we use two linear probes Pr_0 and Pr_1, one for each unimodal branch. The two probes are applied to unimodal branch outputs, such that

$$\hat{Y}_i^{(0)} = Pr_0(Z_i^{(0)}) \tag{5.3.3}$$

$$\hat{Y}_i^{(1)} = Pr_1(Z_i^{(0)}) \tag{5.3.4}$$

The probes are then independently trained from the multimodal DNN, such that the gradient from Pr_0 and Pr_1 is not taken into account in ϕ_0 and ϕ_1 weights update (see Figure 5.2).

To quantify modality imbalance we adopt the methodology outlined in [136]. We compute the scores as follows

$$s_i^{(0)} = \sigma_{y_i}(\hat{Y}_i^{(0)}) \tag{5.3.5}$$

$$s_i^{(1)} = \sigma_{y_i}(\hat{Y}_i^{(1)}) \tag{5.3.6}$$

This scores computes the probabilities assigned by each probe to the right label, this gives us insight of discriminative quality of each modality representations.

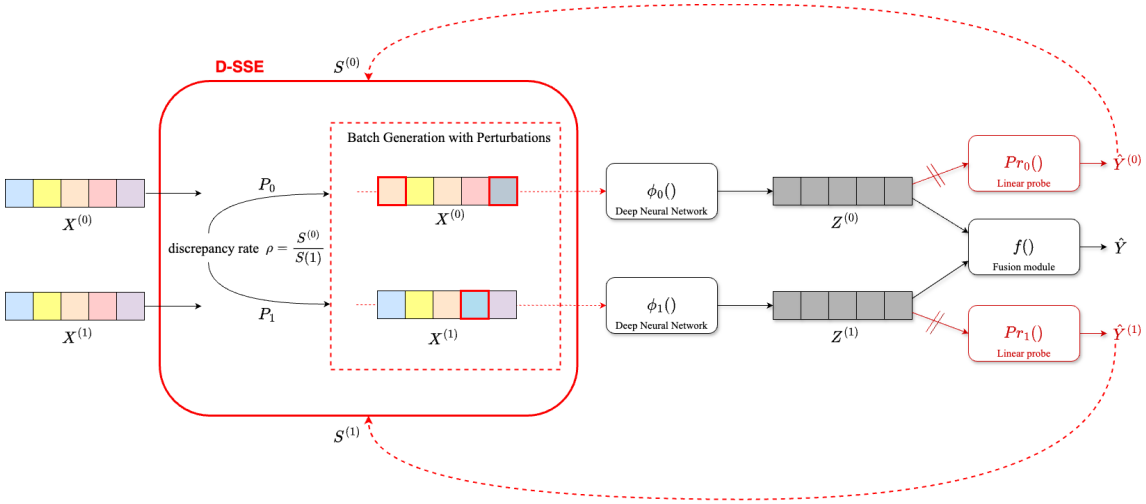


Figure 5.2: Overall architecture of D-SSE.

5.4 Dynamic SSE Regularization (D-SSE)

In this study, we introduce a novel approach to address the challenge of imbalance in multimodal learning, building upon the foundational concept of Stochastic Shared Embeddings (SSE) [140]. Our method aims to enhance performance and mitigate the inherent imbalance present in multimodal datasets through a data-driven strategy. While SSE [140] originally proposed utilizing a knowledge graph to regulate embedding layers, and its application to multimodal learning was explored in [140], our work takes this approach further. We present a unique adaptation that specifically targets the balancing of multimodal training. Our contribution lies in the processing of feature maps from various modalities as embeddings, employing class labels to construct the set of potential transitions. Crucially, we introduce a more rigorous regularization scheme by enabling transitions exclusively between embeddings of instances belonging to different classes. This means that during training, based on the labels assigned to each sample, there is a probability of embeddings being exchanged between different samples, but only if they belong to different classes, this choice is backed by our preliminary experiments (see Section 5.5.3).

We define a transition as an artificial perturbation, where we stochastically change one modality from a multimodal instance. This change is governed by different parameters:

- P_c: The probability of applying a transition to a given example.
- $\{P_m\}_{m=1}^{M}$: The probabilities of changing modality (m).

The P_m are dynamically computed after each training epoch, while P_c is a predetermined hyper-parameter controlling the regularization degree during training, it allows us to fix the proportion of examples where transitions between embeddings can occur.

Let $(X_i, y_i) \in \mathcal{D}$, where $X_i = (X_i^{(0)}, X_i^{(1)})$ represents a multimodal instance with two modalities. We define the transition probabilities as follows:

$$P_c(X_i) = \text{Probability of a transition occurring on } X_i$$

This transition involves switching one of its modalities with a modality from another data point X_j belonging to a different class ($y_i \neq y_j$). The two authorized transitions are:

$$(X_i^{(0)}, X_i^{(1)}) \to (X_j^{(0)}, X_i^{(1)}), (X_i^{(0)}, X_i^{(1)}) \to (X_i^{(0)}, X_j^{(1)})$$

Given that a transition occurs, we introduce two additional probabilities:

$$P_0(X_i) = P((X_i^{(0)}, X_i^{(1)}) \to (X_j^{(0)}, X_i^{(1)}))$$
$$P_1(X_i) = P((X_i^{(0)}, X_i^{(1)}) \to (X_i^{(0)}, X_j^{(1)}))$$

Where:

- P_0 represents the probability of transitioning modality m_0 (the first modality (0))
- P_1 represents the probability of transitioning modality m_1 (the second modality (1))
- $P_1 + P_0 = 1$ (complementary probabilities)

To dynamically re-balance the training, P_0 (and P_1 as a complimentary probability) is adjusted for each epoch of training based on a modality discrepancy rate, computed as follows:

$$\rho = \frac{S^{(0)}}{S^{(1)}} \tag{5.4.1}$$

and

$$S^{(0)}(\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N} s_i^{(0)} \quad , \quad S^{(1)}(\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N} s_i^{(1)} \tag{5.4.2}$$

$$P_0 = \begin{cases} \frac{1}{2} + \frac{1}{2}\tanh(a \cdot (\rho - 1)) & \text{if } \rho \geq 1 \\ 1 - \left(\frac{1}{2} + \frac{1}{2}\tanh(a \cdot (\rho^{-1} - 1))\right) & \text{if } \rho < 1 \end{cases} \tag{5.4.3}$$

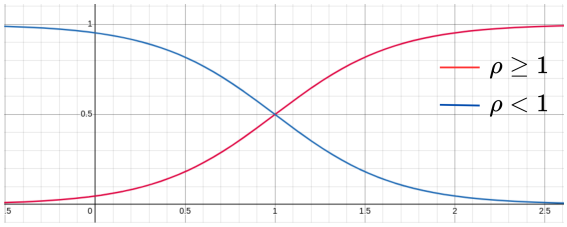


Figure 5.3: Graphic of P_0 function from Equation 5.4.3 in both cases, $\rho \geq 1$ and $\rho < 1$

where a is a hyper parameter.

From this definition and Figure 5.3 we can see that P_0 is computed from $S^{(0)}$ and $S^{(1)}$ scores such that if $\rho > 1$ the first modality (0) is dominating thus P_0 will be > 0.5 making modality (0) more regularized forcing the model to use the second modality (1) to make predictions.

We summarize the overall approach in Algorithm 1 and Algorithm 2. Algorithm 1 outlines the Mini-batch Generation Algorithm, which details the process of generating mini-batches with controlled stochastic transitions. To create a perturbated set of batches $\mathcal{B}$ in respect to the given set of parameters P_c and $\{P_m\}$. This algorithm iterates through a dataset $\mathcal{D}$, where each data sample consists of two modalities and a label, applying stochastic perturbations based on P_c to determine whether to swap modalities between samples with different labels. Within the to be swapped samples, the choice of modality to swap is based on $\{P_m\}$, this algorithm results in a set of batches containing perturbed samples. Algorithm 2 presents the Training Algorithm with Dynamic Probability Update, illustrating how dynamic probability updates are incorporated during training. This algorithm iterates through a dataset $\mathcal{D}$ for a specified number of epochs E, dynamically updating the probabilities $\{P_m\}$. The perturbations are only applied from epoch 2 since the computation of $\{P_m\}$ are derived from the model performances.

Algorithm 1 Batch Generation with Perturbations

Require: Dataset $\mathcal{D} = \{(X_i^{(0)}, X_i^{(1)}, y_i)\}_{i=1}^{N}$, P_c, P_0, B (batch size)

$\mathcal{B} \leftarrow \{\}$
$B_t \leftarrow \{\}$
for each $(X_i^{(0)}, X_i^{(1)}, y_i)$ in $\mathcal{D}$ **do**
    Sample $n \sim U(0, 1)$
    if $n < P_c$ **then**
        Sample $(X_j^{(0)}, X_j^{(1)}, y_j) \in \mathcal{D}$ with $y_j \neq y_i$
        Sample $n' \sim U(0, 1)$
        if $n' < P_0$ **then**
            Add $(X_j^{(0)}, X_i^{(1)}, y_i)$ to B_t
        else
            Add $(X_i^{(0)}, X_j^{(1)}, y_i)$ to B_t
        end if
    else
        Add $(X_i^{(0)}, X_i^{(1)}, y_i)$ to B_t
    end if
    if $\text{len}(B_t) == B$ **then**
        Add B_t to $\mathcal{B}$
        $B_t \leftarrow \{\}$
    end if
end for
return $\mathcal{B}$

Algorithm 2 Training Algorithm with Dynamic Probability Update

Require: Dataset $\mathcal{D} = \{(X_i^{(0)}, X_i^{(1)}, y_i)\}_{i=1}^{N}$, P_c, a, E: number of epochs

for e in $\{1 \dots E\}$ **do**
  $S^{(0)}, S^{(1)}, \{P_m\} \leftarrow 0$
  if $e > 1$ **then**
    Construct the set of perturbated batches $\mathcal{B}$ with Algorithm 1
  else
    $\mathcal{B} \leftarrow \mathcal{D}$
  end if
  for B_t in $\mathcal{B}$ **do**
    Feed-forward B_t to the model
    Calculate $S^{(0)}(B_t)$, $S^{(1)}(B_t)$ with Eq 5.4.2,
    $S^{(0)}$ += $S^{(0)}(B_t)$
    $S^{(1)}$ += $S^{(1)}(B_t)$
    Backward pass
  end for
  Update $\{P_m\}$ from $S^{(0)}$, $S^{(1)}$ using Eqs 5.4.1, 5.4.3 and a
end for

5.5 Experiments

In this section, we detail experiments conducted to assess the D-SSE method. We begin by describing the datasets employed, which include multimodal datasets containing audio-video and text-image pairs. We explain the experimental setup, detailing the algorithms and hyper-parameters used. We then show the performance of our method on multimodal datasets using various metrics. Subsequently, we examine its performance under different imbalance scenarios. Lastly, we investigate how D-SSE mitigates multimodal imbalance compared to baseline methods.

5.5.1 Datasets

To evaluate our approach we used three widely used datasets to analyse its effect on modality learning imbalance.

CREMA-D [3] is a dataset created specifically for recognizing emotions in speech. It contains 7,442 short video clips, each lasting 2-3 seconds, featuring 91 actors uttering brief words. The dataset covers six primary emotions: anger, happiness, sadness, neutrality, disgust, and fear. Emotion labels were assigned through crowd-sourcing involving 2,443

raters, as detailed in [137].

The dataset is divided into a training set and a validation set, comprising a total of 6,698 samples with a 9:1 ratio, and a separate testing set containing 744 samples. We utilize this dataset for emotion classification by utilizing both the audio (*modality* 0) and video (*modality* 1). Various studies have highlighted the modality imbalance in this dataset [136, 137]. Despite videos outperforming audio in a unimodal setting, it has been observed that the performance of the video modality decreases drastically when trained in a multimodal setting underlying the modal imbalance towards audio modality.

ModelNet40 dataset is part of the Princeton ModelNet collection [7], comprising 3D objects across 40 categories, with 9,483 training samples and 2,468 test samples. The classification task involves identifying a 3D object based on its front (*modality* 0) and back (*modality* 1) 2D views [7]. Each sample consists of a set of 2D images (224×224 pixels) representing a 3D object. For the unimodal encoders, we employ ResNet18 [83] architecture.

The Colored-and-Gray MNIST [5], referred to as CG-MNIST, is a synthetic dataset derived from the original MNIST dataset [143]. In the training set, comprising 60,000 instances, each instance contains two images: a gray-scale image and a monochromatic image. The color of the monochromatic images in the training set is strongly correlated with their corresponding digit labels. In the test set, consisting of 10,000 instances, each example also contains gray-scale and monochromatic images, although with a much lower correlation between color and label compared to the training set. We designates the monochromatic image as the first modality (*modality* 0) and the gray-scale image as the second modality (*modality* 1). The CG-MNIST dataset is primarily used to demonstrate the effectiveness of multimodal balancing training methods beyond traditional audio-visual datasets [135, 137]. To generate the monochromatic images, we adhere to a method outlined in [5]. Initially, we designate ten distinct colors, each assigned to a digit category as its mean color. We then sample a color from a normal distribution centered around the corresponding mean color $\boldsymbol{\mu}$, [1] with a provided variance $\boldsymbol{\sigma}_{\text{train}}$. The variance parameter of this distribution offers control over the degree of color bias present in the data. For the training set, color sampling occurs according to the equation:

$$\text{Color} \sim \mathcal{N}(\boldsymbol{\mu}_{\text{color}}, \boldsymbol{\sigma}_{\text{train}})$$

where $\boldsymbol{\mu}_{\text{color}} \in \boldsymbol{\mu}$ and $\boldsymbol{\mu} = (\mu_r, \mu_g, \mu_b)$. Here, $\boldsymbol{\mu}_{\text{color}}$ represents the mean color, and $\boldsymbol{\sigma}_{\text{train}}$ represents the variance.

This dataset serves as a framework for examining model performance in different im-

[1] $\boldsymbol{\mu}$ and σ are used after normalization.

balanced setups, where the biased modality may be favored during training due to the color-label correlation. Our primary objective in using this dataset is to analyze the effect of varying $\boldsymbol{\sigma}_{\text{train}}$ on the performance of our method. For the test set, we fix $\boldsymbol{\sigma}_{\text{test}} = 1$, which ensures that the color will be sampled from a high-variance distribution, thereby eliminating any potential bias.

CrisisMMD [14] As presented in Section 3.2 this benchmark aims to identify crisis events that require emergency response using social media posts as a basis. The dataset for this benchmark consists of tweets (image-text pairs) obtained through searching for specific hashtags on Twitter and labeled for three tasks : Informativeness, Humanitarian, and Damage severity assessment. In this work, we only consider the first two multimodal tasks: *Informativeness* and *Humanitarian*.

5.5.2 Experimental settings

To train our models, we adopted the methodologies described by [137] and [135] for our study on the CREMA-D and ModelNet datasets, utilizing ResNet18 [83] as the encoder backbone. The input data was mapped into 512-dimensional vectors and processed with several fusion methods detailed in Table 5.2. In the audio modality of the CREMA-D dataset, we converted data into spectrograms sized 257×1004. The visual modality training dataset was constructed by randomly sampling three frames from each video. Training on the ModelNet40 and CREMA-D datasets involved mini-batches of 64, using an SGD optimizer with 0.9 momentum and a weight decay of 1×10^{-4}. We started with a learning rate of 1×10^{-3}, reducing it gradually to 1×10^{-4}. The Colored-and-gray MNIST dataset featured a neural network with two convolution layers and one max-pooling layer. For the CrisisMMD dataset, Bert [1] and DenseNet [86] served as the encoders for text and images, respectively. A two-layer MLP was used to fuse outputs from each encoder. Both datasets operated under a learning rate of 2×10^{-3}, with parameters $P_c = 0.1$ and $a = 1.5$ set across all experiments. Training was conducted on an Nvidia Tesla V100 GPU, lasting 100 epochs for CREMA-D and ModelNet40, and 20 and 50 epochs for CrisisMMD and Colored-and-gray MNIST, respectively.

5.5.3 Preliminary experiments

To investigate the impact of transition probabilities on our method's performance, we conducted experiments on the CREMA-D dataset [3] using various values for P_c and P_{class}, as shown in Figure 5.4. P_c represents the probability of applying D-SSE to a training example, with $P_c = 0$ indicating no D-SSE application (baseline) and $P_c = 0.5$

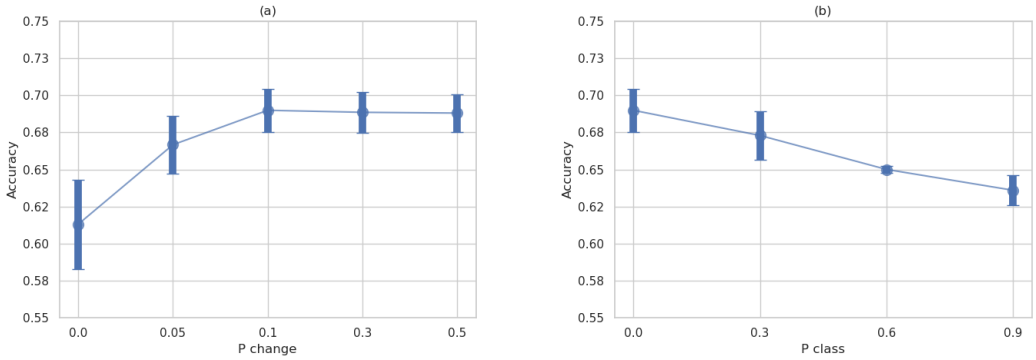


Figure 5.4: Effect of $P_{change}(P_c)$ and P_{class} on performance, accuracies are reported on 3 runs on CREMA-D dataset.

meaning D-SSE is applied to half of the training examples. Results in Figure 5.4(a) demonstrate that D-SSE ($P_c > 0$) consistently outperforms the baseline, with optimal performance achieved at $P_c = 0.1$ (mean accuracy 69%), followed by a slight decrease for higher values. In comparison with [94], our method restrains transitions on instances belonging to different classes as underlined in Section 5.4, with a transition defined as :

$$(X_i^{(0)}, X_i^{(1)}) \rightarrow (X_j^{(0)}, X_i^{(1)}) \; or \; (X_i^{(0)}, X_i^{(1)}) \rightarrow (X_i^{(0)}, X_j^{(1)})$$

and

$$y_i \neq y_j$$

To validate this choice experimentally we introduce P_{class} which is the probability of $(y_i = y_j)$ meaning a transition occurs in the same class, while $1 - P_{class}$ is the probability of the transition occurring in different classes $(y_i \neq y_j)$. We fixed $P_c = 0.1$ based on previous results.

Results reported in Figure 5.4(b) show that lower values of P_{class} (more cross-class transitions) yield better performance, with peak accuracy at $P_{class} = 0.0$ (mean accuracy 69%). Performance decreases as P_{class} increases, suggesting that cross-class transitions are more beneficial. These findings demonstrate that our method's performance is sensitive to both P_c and P_{class}, with optimal values occurring at moderate levels of P_c probability and a preference for cross-class transitions with $P_{class} = 0$. In the following experiments we fix $P_c = 0.1$ and $P_{class} = 0$ originally defined for D-SSE.

Table 5.1: Accuracy score comparison on five multimodal tasks using concatenation fusion. The results are the mean and the standard deviation over 3 runs runs with different random seeds.

	CREMA-D	ModelNet40	CG-MNIST	Crisis-MMD	
				Informativeness	Humanitarian
Modality 0	59.50±1.2	73.44±0.5	97.80±0.05	86.91±0.03	81.83±0.07
Modality 1	62.84±2.6	69.66±0.46	29.54±0.64	83.93±0.03	76.08±0.01
Baseline	61.95±0.32	75.3±0.32	30.04±2.35	90.14±0.20	85.61±0.10
R-SSE	66.37±1.23	75.69±0.61	88.17±4.90	90.17±0.30	87.87±0.51
D-SSE	**69.34±0.84**	**77.0±0.78**	**91.88±1.14**	**90.42±0.19**	**88.24±0.05**

5.5.4 Effectiveness of D-SSE on multimodal datasets

To evaluate the effectiveness of our proposed D-SSE method, we performed a comprehensive comparative analysis against baseline methods on five multimodal tasks derived from four datasets. Additionally, we implemented R-SSE (Random-SSE), which maintains a constant P_c while P_m is randomly sampled from a uniform distribution $U(0, 1)$, to assess the impact of our dynamic P_m modulation strategy.

The experimental results, summarized in Table 5.1, using concatenation as a fusion technique, indicate the superior performance of the D-SSE method across five diverse multimodal tasks: CREMA-D, ModelNet40, CG-MNIST and Crisis-MMD (Informativeness and Humanitarian). On the CREMA-D dataset, D-SSE achieved the highest accuracy of 69.34%, significantly surpassing both individual modalities and the baseline. For ModelNet40, D-SSE showed a modest improvement with an accuracy of 77.0%. In the CG-MNIST dataset, D-SSE demonstrated substantial improvement, achieving 91.88% accuracy compared to the baseline's 30.04%, underscoring its effectiveness in scenarios with highly imbalanced modalities. On the Crisis-MMD dataset, D-SSE consistently outperformed other methods on both tasks (Informativeness and Humanitarian) with relatively smaller margins. Notably, R-SSE also showed improvements over the baseline across all datasets; however, D-SSE consistently achieved the best results, demonstrating the efficacy of dynamic modulation of P_m. These findings underscore the robustness and adaptability of the D-SSE method across varied multimodal learning tasks, suggesting its potential as a powerful tool for enhancing multimodal fusion and improving overall model performance.

To further investigate the effectiveness of D-SSE, we implemented it across four fundamental fusion methodologies: Concatenation, Summation, Gated [144], and Film [4]. The first two fusion strategies are simpler, utilizing only one layer on top of unimodal

		Baselines			R-SSE			D-SSE		
Dataset	Fusion Method	Multimod.	Modal.0	Modal.1	Multimod.	Modal.0	Modal.1	Multimod.	Modal.0	Modal.1
CREMA-D	Unimodal	-	59.5±1.2	62.8±2.6	-	-	-	-	-	-
	Concatenation	62.0±0.3	53.5±2.9	23.8±2.4	66.4±1.2	52.3±3.3	30.6±4.2	**69.3±0.8**	49.5±1.6	39.7±3.6
	Sum	61.4±0.8	54.5±1.1	22.1±3.3	66.8±1.7	53.8±1.6	28.8±4.1	**70.5±1.0**	51.2±1.3	41.1±3.5
	Gated	61.5±0.4	56.3±1.3	27.8±3.8	67.0±2.4	57.5±2.1	32.6±1.6	**70.1±0.6**	55.1±1.2	41.6±1.1
	Film	58.7±0.5	57.7±0.9	17.3±1.0	65.8±1.9	57.3±1.0	29.0±1.8	**68.8±0.9**	55.1±1.2	40.3±2.3
ModelNet40	Unimodal	-	73.4±0.5	69.7±0.5	-	-	-	-	-	-
	Concatenation	75.3±0.3	45.4±2.1	39.3±3.1	75.7±0.6	57.3±4.0	49.0±6.8	**77.0±0.8**	61.5±1.9	56.1±3.4
	Sum	75.2±0.4	44.7±4.9	46.0±1.8	74.3±0.3	59.7±3.7	53.1±6.6	**76.0±0.9**	57.7±7.8	60.2±4.6
	Gated	73.1±0.6	49.9±1.1	68.0±0.8	**74.4±0.5**	57.5±0.7	67.5±1.0	74.3±1.0	51.8±2.3	69.2±0.4
	Film	73.6±0.1	61.6±2.0	51.8±3.8	73.3±0.6	65.5±1.4	58.3±1.2	**74.8±0.2**	66.1±0.8	62.8±0.2

Table 5.2: Performance Comparison on CREMA-D [3] and ModelNet40 [7] datasets. The results show the mean accuracy and the standard deviation over 3 runs with different random seeds.

encoders, with modality 0 and modality 1 results computed by splitting the weight of the last linear layer. In contrast, the Gated and Film configurations employ more complex computations to fuse representations from the two encoders, as detailed in Section 3. Results are reported in Table 5.2. On the CREMA-D dataset, our findings corroborate previous studies [113, 137], indicating that modality 1 (video) outperforms the multimodal approach, highlighting the limitations of simple multimodal training. Additionally, our results align with the trends depicted in Figure 5.1, where audio performance (modality 0) tends to dominate multimodal learning. Moreover, we observed that different fusion strategies have limited impact on the final results on both datasets; however, incorporating R-SSE enhances modal performance, confirming the effectiveness of the SSE method in multimodal contexts. Furthermore, our approach generally outperforms both baseline methods and the R-SSE approach across all configurations, demonstrating its efficiency in addressing the imbalance issue observed in these datasets. Our proposed method (D-SSE) also mitigates the performance gap between unimodal approaches, as illustrated in Figure 5.1(b) on the CREMA-D dataset, while it helps the lesser imbalance dataset ModelNet40 (See Section 5.5.7 for more details) to achieve better unimodal performance. Overall, these results validate the effectiveness of D-SSE in multimodal learning, showcasing its potential to enhance performance across various multimodal datasets and fusion strategies.

5.5.5 D-SSE in different imbalance setups

To assess the efficacy of our technique under various imbalanced conditions, we utilize a synthetic dataset known as CG-MNIST. This dataset includes two modalities : the first is the standard MNIST dataset presented in gray-scale, the second is a colored version of the MNIST dataset, where the color is directly correlated with the label. This correlation

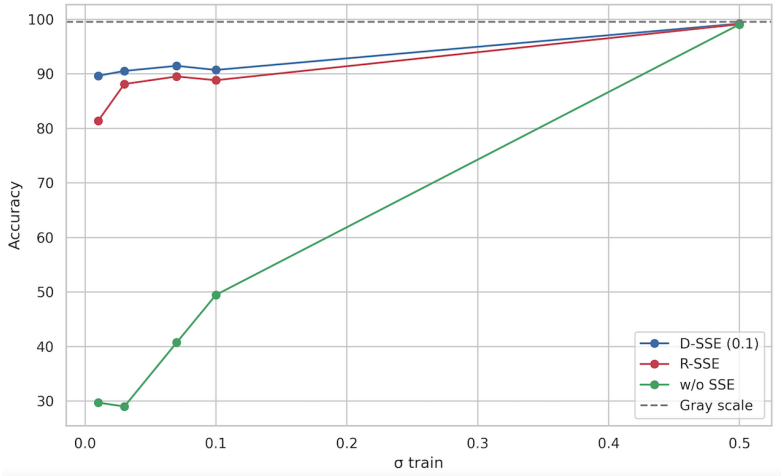


Figure 5.5: Performance comparison on CG-MNIST [5] for different σ_{train} values

introduces an imbalance during the training process, causing the model to favor the biased modality (colored image). This setup provides us with a robust framework to examine our method's performance in different imbalanced setups.

To show the effect of this bias we test on unbiased dataset ($\sigma_{test} = 1$). Figure 5.5 shows the results of our method over 5 values of σ_{train} (from 0.01 where the correlation is high to 0.5 where almost no correlation. The results shows consistent performance of D-SSE method ($> 88\%$) outperforming both the base method and the R-SSE method and approaching the gray-scale results shown in gray. Additionally the performance are mainly the same for the 4 first values (0.01-0.1) while the naive method (without regularization) is impacted by this changes going from under 30% to 50% in accuracy. Moreover, this experience shows that our method still yields good performance in balanced setups, it yields comparable performance with the base method when $\sigma_{train} = 0.5$. Finally the same trends are observed with Table 5.1 where D-SSE outperforms R-SSE in spite of its good performance.

5.5.6 Comparison with balancing methods

Evaluating our method against existing approaches presents several challenges due to the variations in parameters, preprocessing steps, and the availability of code across different works. These factors make it difficult to perform a direct and fair comparison. In this work, we focused our evaluation on the CREMA-D dataset and selected two popular balancing methods: OGM-GE [136] and PMR [137]. These methods were chosen because the code and preprocessing parameters were accessible, ensuring a consistent and fair comparison framework. Unlike our previous experiment with the CREMA-D dataset,

	audio-video	*audio*	*video*
Unimodal	-	59.53	46.59
Imbalanced	58.87	56.39	18.60
Modality Drop	59.37	**58.80**	17.89
OGM-GE[136]	61.41	42.87	26.47
PMR[137]	59.20	56.34	19.46
R-SSE	61.93	56.53	26.99
D-SSE (Ours)	**62.78**	57.38	**28.12**

Table 5.3: Comparison with other balancing methods, the accuracy scores are reported on CREMA-D dataset using 1 sampled frame from the video modality

where we utilized 3 frames per video as shown in Table 5.2, we followed the processing proposed by [136] and used only one frame per video in this experiment.

Results from Table 5.3 shows a comparison with popular balancing methods using concatenation as a fusion method. OGM-GE uses on the fly gradient modulation to slow the dominant modality while PMR uses prototypes of classes to help the dominate modality to yields more discriminatory representations. Furthermore, inspired by [139] we implemented Modality Drop, where the modality chosen for the transition is dropped out by changing its embedding with zeros. First we clearly see that the balancing methods outperform the baseline in *audio-video* setting. OGM-GE improves both the overall performance and the *video* performance, while affecting negatively the dominating modality *audio*. PMR on the other hand improved slightly over multimodal performance and *video* while keeping roughly the same performance on the *audio* modality. Finally our method yields better performances in *audio*, *video* and multimodal *audio-video* settings compared to all methods, confirming its effectiveness.

5.5.7 Imbalance mitigation

Imbalance with d_{util} metric

Given a bimodal model f, following [135], we define the degree of imbalance (or discrepancy) in the utilization of the two modalities by f as:

$$d_{\text{util}}(f) = u(m_1 \mid m_0) - u(m_0 \mid m_1).$$

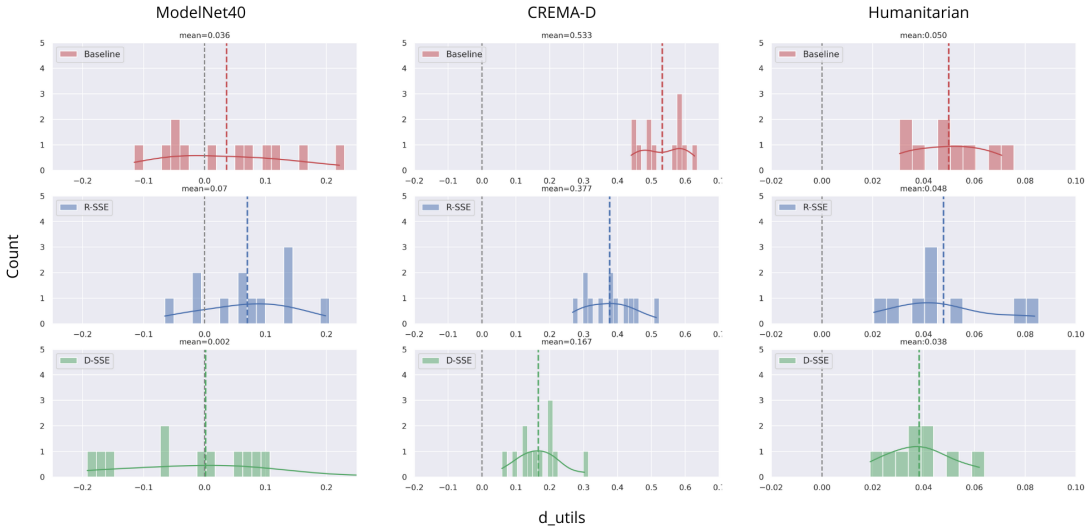


Figure 5.6: Histograms of d_{util} values across different datasets The results are on ModelNet40, CREMA-D, and CrisisMMD Humanitarian with respectively 12, 12, and 9 runs for each dataset. Baseline (no balancing), R-SSE, and D-SSE represent the balancing method. The dashed lines represent the mean d_{util} for each configuration. Higher absolute values of d_{util} indicate greater imbalance in modality utilization.

where $u(.,.)$ is the conditional utilisation rate that quantifies the relative change in accuracy when dropping out a modality in the test phase:

$$u(m_1 \mid m_0) = \frac{Acc(m_0, m_1) - Acc(m_0^{\text{zeros}}, m_1)}{Acc(m_0, m_1)}$$
$$u(m_0 \mid m_1) = \frac{Acc(m_0, m_1) - Acc(m_0, m_1^{\text{zeros}})}{Acc(m_0, m_1)}$$

Here, $Acc(m_0, m_1)$ denotes the accuracy on the test set using both modalities, while $Acc(m_0^{\text{zeros}}, m_1)$ and $Acc(m_0, m_1^{\text{zeros}})$ represent the accuracies when the respective modality is replaced by a zero vector.

The value of $d_{\text{util}}(f)$ lies within the interval $[-1, 1]$.[2] When $d_{\text{util}}(f)$ is close to -1 or 1, it indicates that the model predominantly benefits from one modality given the other, but not vice versa. Thus, a large $|d_{\text{util}}(f)|$ suggests a significant imbalance in the utilization of the two modalities by the model f. We conducted 12 experiments on the CREMA-D and ModelNet40 datasets, and 9 experiments on the Crisis-MMD dataset (Humanitarian task), using primarily the concatenation, sum, and Film fusion methods. Note that gated fusion was omitted due to the impact of zero vector on the fusion module, which zeros out the gate. From Figure 5.6, we observe that the most imbalanced dataset is CREMA-D, followed by Humanitarian task. The mean d_{util} for the CREMA-D

[2]With the assumption: $Acc(m_0, m_1) > \max(Acc(m_0^{\text{zeros}}, m_1), Acc(m_0, m_1^{\text{zeros}}))$

dataset is highest for the Baseline (mean = 0.533), indicating significant reliance on a single modality. In contrast, ModelNet40 exhibits less imbalance when trained without balancing strategy (mean = 0.036). The R-SSE method appears to mitigate modality imbalance in the CREMA-D and Humanitarian datasets, but not in ModelNet40. D-SSE consistently helps balance the model's reliance on both modalities across all datasets. These findings suggest that D-SSE methods can effectively address modality imbalance in certain datasets, enhancing the model's performance and utilization of both modalities. Finally, our observation on CrisisMMD confirms our early findings in Section 4.6, on *Humanitarian* we observe that the models trained generally leans to the *text* modality (modality 0) with mean $d_{\text{util}} = 0.05$.

Imbalance with Nearest Neighbor Similarity Scores and confidence scores

To further assess the impact of our method on addressing the imbalance, we introduce the Sim^k score to quantify the similarity between representation spaces. We adopt the metric proposed in [145], where for each batch B_t, we compute the similarity between the representation spaces of the multimodal (after fusion) and a specific modality m (before fusion) as follows:

$$Sim^k_{mm,m}(X_i) = \frac{|nn^k_{mm}(X_i) \cap nn^k_m(X_i)|}{k} \tag{5.5.1}$$

Here, $X_i = (X_i^{(0)}, X_i^{(1)})$, and nn^k_{mm} represents the k nearest neighbors to the instance $X_i \in \mathcal{D}$ after fusion, while $nn^k_m(X_i)$ denotes the k nearest neighbors to the feature map $Z_i^{(m)} = \phi_m(X_i^{(m)})$. We compute the overall mean over the dataset for each epoch and present the results in Figure 5.7.

In Figure 5.7 (a), we observe the evolution of representation similarities during training. For the CREMA-D dataset, the *audio* modality shows greater similarity with the multimodal representations, as indicated by the red curve, starting at 0.37 and stabilizing at 0.2. The visual-multimodal similarity, outlined in blue, stabilizes at 0.1. This confirms early observations about the dominance of the audio modality in training. However, when we apply our method, the training similarities stabilize at roughly the same value of 0.15 (green and yellow curves) for both modalities, demonstrating the effect of D-SSE in rebalancing the training process and achieving a more balanced multimodal representation space. Similarly, for the ModelNet40 dataset, the nearest neighbor similarity shows that the modality m_0 (the front view, depicted in red) has a higher initial similarity with the multimodal representations compared to m_1 (depicted in blue). Over the course of training, m_1 stabilizes around 0.35, while m_0 stabilizes around 0.2. When applying D-SSE, the gap between the two similarities narrows to a smaller range, indicating a balanced

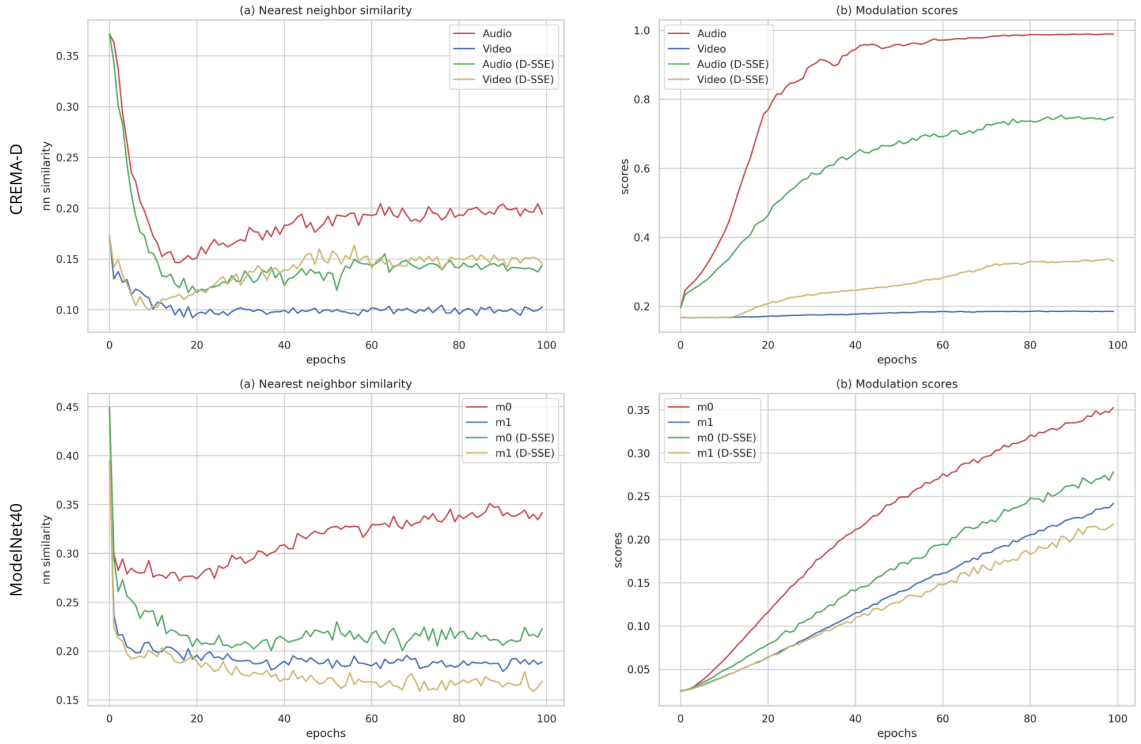


Figure 5.7: (a) Nearest neighbor similarity scores for CREMA-D and ModelNet40 datasets, illustrating the similarity between unimodal embeddings and the multimodal embeddings. (b) Modulation scores, $S^{(0)}$ and $S^{(1)}$, showing the confidence of predictions for each modality over the training process.

representation space. Furthermore, Figure 5.7 (b) illustrates the modulation scores $S^{(0)}$ and $S^{(1)}$ as described in Section 5.3. For CREMA-D, the *audio* modality yields more confident predictions compared to *video*, but with D-SSE, the gap between the modalities narrows. In the ModelNet40 dataset, m_1 consistently shows higher confidence than m_0. However, D-SSE reduces the disparity, leading to more balanced confidence scores between the modalities. These results support our findings in Figure 5.1, where the potential of our method to close the accuracy gap between modalities while improving overall multimodal performance is clearly showed. By applying D-SSE, we achieve a more balanced training process, resulting in improved and more balanced utilization of both modalities.

5.6 Conclusion

In this chapter, we demonstrated that multimodal training can often converge to suboptimal performance due to modality imbalance, underscoring the limitations of basic joint multimodal training approaches. Our study introduced the Dynamic Stochastic Shared Embedding (D-SSE) technique, which leverages stochastic transitions between embeddings guided by dynamically computed probabilities P_m. We conducted comprehensive evaluations using three different datasets: CREMA-D (audio and video), ModelNet40 (two views of 3D objects), and CrisisMMD (image and text pairs), our method showed significant effectiveness in enhancing multimodal fusion and improving overall model performance. Specifically, D-SSE not only boosted the overall accuracy of multimodal models but also mitigated the imbalance between modalities, leading to more balanced multimodal learning process. This was revealed by consistent improvements in both unimodal and multimodal accuracy, as well as the imbalance mitigating effect showed by our method on nearest neighbor similarities and confidence scores. Additionally, our experiments on the CG-MNIST dataset highlighted D-SSE's effectiveness in different scenarios from highly to moderate imbalanced modalities, further validating its robustness.

However, the current implementation of D-SSE is restricted to two modalities scenarios. Future work should aim to extend the applicability of D-SSE to configurations involving three or more modalities. This expansion requires redefining specific elements of the method, particularly the modulation components such as P_m by using a softmax to get probabilities P_m for each modality, and support complex multimodal environments. Another critical point for improvement is hyperparameter tuning. The dynamic modulation of P_m has shown potential, suggesting that further exploration into dynamic updates for P_c and a could lead to more stable training processes and enhanced performance. By addressing these limitations and extending the framework to accommodate more com-

plex multimodal interactions, D-SSE can be adapted for a broader range of applications, thereby enhancing the robustness and efficacy of multimodal models in diverse real-world scenarios.

Chapter 6

Towards Multimodal French Tweets Classification

6.1 Introduction

The widespread use of social media, particularly on platforms like X (formerly Twitter), has provided massive amounts of real-time data that can be applied in fields such as public health, economics, and politics. During natural disasters, tweets can offer critical information to humanitarian organizations, helping them assess the disaster's scope and prioritize relief efforts. However, processing the vast amount of data from social media requires automated systems, which rely on annotated datasets for training. The automation of social media information processing has recently become a hot research topic, with a majority of existing approaches relying heavily on supervised learning methods. Their primary objective is to categorize each post into a specific class based on the targeted task, such as informativeness [14] or relatedness [34]. Annotated datasets being the core element in supervised learning, considerable efforts have been made to manually construct such datasets. While most resources are text-based and predominantly in English, it's worth noting that resources in other languages also exist [34, 45–47, 146] (see the following section for more details).

In the previous chapters, we explored various fusion methods and introduced techniques to enhance modality fusion, resulting in improved classification performance. Our findings corroborate earlier studies that demonstrate the superior performance of multimodal approaches over unimodal ones. Despite these promising results, the use of social media in crisis scenarios still faces significant limitations. A key issue is the reliance on the widely-used CrisisMMD dataset, which, like in many previous studies, forms the basis of our experiments. This scarcity of multimodal annotated data poses a major

challenge for applying these methods to the French crisis context, where social media posts are primarily in French, and the nature and impact of disasters differ from those represented in the existing dataset. In collaboration with the BRGM, this thesis aims to close this gap and to lay the foundation for multimodal French tweet classification in the context of crisis situations. To support this effort, we present in this chapter the M-CATNAT dataset with a perspective to use it for training models capable of classifying French tweets. M-CATNAT is a multimodal dataset of 1,356 French tweets about various natural disasters with annotations aligned with CrisisMMD tasks. However, the CrisisMMD dataset has the following limitation : it contains annotations for each of the two modalities of a tweet (text and image), but does not propose gold (i.e. manually annotated) annotations for the multimodal tweet as a whole (text+image). Since learning a multimodal model requires such a global label, tweets whose annotations for the two modalities do not match are generally filtered out [13]. Unlike CrisisMMD, the M-CATNAT dataset includes not only gold annotations for the image modality and the text modality but also the gold label of the tweet as a whole, i.e the multimodal instance (text+image). Interestingly, the full annotation model proposed makes it possible to offer a greater representativeness of tweets in terms of the diversity of relationships between modalities (e.g. redundancy, complementary or contradiction in the information). We present in this chapter the detailed framework defined for the multimodal annotation process of the tweets contained in M-CATNAT.[1]

Seven annotators from the research team have fully annotated 1,356 tweets, out of the 1,430 planned.[2] This process has resulted in the creation of 4,691 manual annotations tasks, as each instance entails three labels (one for the image, one for the text, and one for the multimodal instance with some tweets having multiple images). The CrisisMMD humanitarian classes definition guided our annotation process, ensuring consistency and reliability in the labeled data. In addition, the annotation process was regularly evaluated to check both its consistency (inter-annotator agreement) and its alignment with the CrisisMMD dataset.

6.2 Unimodal and multimodal crisis-related datatsets

The surge in the use of social media has led to an increased interest in processing their content with numerous studies focusing, in particular, on crisis-related tweets. As reported in Table 6.1, a lot of emphasis has been placed on text-based analysis, given that the majority of annotated datasets primarily feature labels on text alone. These datasets have introduced various tasks, such as assessing the *relatedness* of tweets to a

[1]The dataset is available here : https://github.com/badreddineFarah/M-CATNAT

[2]Some of the selected tweets were deleted.

specific disaster [34], determining the *informativeness* of tweets [48, 49], classifying *crisis types*, and detecting *eyewitnesses* [44]. In the domain of image processing for tweets, tasks include estimating the *damage severity* [14, 53] and *damage type* [15], evaluating *informativeness* [14]. Most resources are in English, although resources also exist in other languages like Spanish [45], Italian [46], German [146] and Arabic [47]. For the French language, [34] proposed a text only disaster tweet dataset with three classification levels including *relatedness*, *urgency* and *intent to act*. One should note that the various tasks addressed lead the community to produce datasets more or less compatible with the central resource that is CrisisMMD. Each dataset whose task mentions *humanitarian* is either an implementation of the CriseMMD guidelines or a mapping to a simplified set of classes, keeping it compatible with CrisisMMD.[3]

Despite recent studies highlighting the effectiveness of multimodal datasets in improving the performance of machine learning models through multimodal training [13, 94, 95, 109], limited attention has been given to the creation of new multimodal crisis-related datasets and CrisisMMD remains the most widely used dataset despite its inherent limitations. A first notable limitation is that CrisisMMD independently annotates images and text, necessitating practitioners to filter instances with *discordant* labels (different labels for image and text) and retain instances with consistent labels for training. This approach does not take into account the various relationships between modalities in social media data [99**?**]. In response to this gap, [16] introduced Disrel, a dataset for classifying the relationship between image and text. Their work demonstrated that incorporating such tasks can enhance the performance of models trained on CrisisMMD.

Our work aims to enhance CrisisMMD in two dimensions. Firstly, we augment the corpus by incorporating French tweets, in order to provide data for multilingual and multimodal training. Secondly, we provide three annotations for each instance, hypothesizing that this enriched dataset should enable models to better capture the relationships between images and texts, thereby improving overall system performance. Unfortunately, due to time constraints, we were unable to test this hypothesis or conduct extensive experimentation within the frame of this thesis. Importantly, we adhered to the CrisisMMD guidelines to ensure alignment between our annotations and the original annotation guide.

[3]For a substantial survey of existing datasets for English, see [36].

Table 6.1: Crisis related published resources.

	Modal. labels	Tasks	Platform	Size	Language
[48]	Text	informativeness, humanitarian, source	Twitter	28,000	English
[49]	Text	informativeness, humanitarian	Twitter	166,098 (inf.), 141,533 (hum.)	English (94%)
[34]	Text	relatedness, urgency, intent to act	Twitter	12,826	French
[45]	Text	relevancy	Twitter	2,187	Spanish
[46]	Text	relatedness, damage	Twitter	5,642	Italian
[47]	Text	informativeness, humanitarian	Twitter	4,037	Arabic
[44]	Text	eyewitness	Twitter	14,000	English
[147]	Image	damage severity, humanitarian, disaster type	Twitter, Google, ...	71,198	-
[53]	Image	damage severity	Twitter, Google	25,758	-
[14]	Image, text	informativeness, humanitarian, damage severity	Twitter	16,097	English
[15]	Multimodal	damage type	Instagram, Twitter, Google	5,879	English
M-CATNAT (Ours)	**Image, text, multimodal**	**Informativeness, humanitarian**	**Twitter**	**1,430**	**French**

6.3 Data collection

To initiate the annotation process, it is essential to gather data. For this purpose, we used datasets collected as part of the RéSoCIO project by the French Geological Survey (BRGM) via its SURICATE-Nat platform [148]. The data collection process relies on the interrogation of Twitter's "academic" API on the basis of a carefully chosen keyword search, exploring French-speaking lexical fields related to floods on the one hand, and earthquakes on the other. Below are the specific queries used for each type of disaster, with the aim of retrieving a maximum number of posts describing the effects of the phenomena, as well as a minimum number of off-topic posts.

Flood key-words The keywords used are the following (taking spelling errors into account) : "*inondation*" (flood), "*inondé*" (flooded / inundated), "*sous l'eau*" (under water), "*rivière en crue*" (swollen river), "*crue*" (flood upward trend / freshet), "*décrue*" (flood downward trend), "*onde de crue*" (flood wave), "*sort de son lit*" (rise above the banks), "*torrentiel*" (torrential), "*emporté par les eaux*" (washed away). The query excludes retweets and tweets with flood lexical fields related to politics, sex or migration, while focusing only on French language tweets by specifying "lang:fr".

Earthquake key-words "*Séisme*" or "*tremblement de terre*" (earthquake), "*magnitude*" (magnitude), and "*terre tremble*" (ground shaking) are the keywords used to retrieve the tweets (taking spelling errors into account as well). The query excludes retweets and tweets with earthquake lexical fields related to sex, migration or politics, while focusing only on French language tweets by specifying "lang:fr".

Table 6.2: Disaster events considered.

Crisis type	Affected areas	Search start date	Search end date	# collected tweets	# sampled tweets (images)
Floods	Alpes-Maritimes et Var	10/2/2015 12:00:00 AM	10/4/2015 11:59:00 PM	14846	108 (124)
	Carcassonnais	10/14/2018 12:00:00 AM	10/15/2018 11:59:00 PM	14341	171 (205)
	Secteurs de Béziers / Narbonnais	10/22/2019 12:00:00 AM	10/24/2019 11:59:00 PM	7550	104 (126)
	Pays Basque / Béarn/ Pyrénées	12/12/2019 12:00:00 AM	12/15/2019 11:59:00 PM	6838	132 (158)
	Cote d'Azur 83/06	11/22/2019 12:00:00 AM	11/24/2019 11:59:00 PM	9983	164 (206)
	Cote d'Azur 83/06	12/1/2019 12:00:00 AM	12/1/2019 11:59:00 PM	6353	89 (103)
	Valleraugue, Saumane	9/19/2020 12:00:00 AM	9/20/2020 11:59:00 PM	5721	72 (83)
	Alex storm (south of France)	10/1/2020 12:00:00 AM	10/3/2020 11:59:00 PM	7804	110 (129)
		Date	Time		
Earthquakes	Barcelonnette	07/04/2014	07:26:59 PM	11085	66 (72)
	La Rochelle	28/04/2016	06:46:53 AM	3802	87 (91)
	Thouars	21/06/2019	06:50:57 AM	2668	56 (58)
	Le Teil	11/11/2019	10:52:45 AM	6448	192 (224)
	Strasbourg	12/11/2019	01:38:00 PM	3123	79 (86)
Total				100562	**1430 (1665)**

6.3.1 Data sampling

As illustrated in Table 6.2, we sampled tweets from various events of natural disaster, taking one-third from earthquake events and two-thirds from flood events, in order to respect the representativeness of each disaster while keeping the number of tweets sufficient for each disaster. The earthquakes and flash floods selected correspond to significant events for mainland France, from a phenomenological and/or a crisis management point of view. To maintain data quality and relevance, we applied additional criteria, exclusively selecting tweets with text lengths between 5 to 40 words to ensure sufficient textual content for comprehensive analysis. Additionally, we selectively included tweets with at least one associated image, as the presence of visual content is pivotal for preserving the multimodal nature of the dataset. Each tweet was divided into three instances: text alone, image alone, and the combination of text and image, resulting in a total of 4,961 annotated instances (since some tweets contained multiple images). For tweets with multiple images, we annotated the text, each individual image, and each multimodal combination of the text with each image.

Since starting the annotation process, approximately 10% of the data has been deleted from X (formerly Twitter), reducing the dataset size from the foreseen 1,500 tweets to 1,430, and later down to 1,356 as more tweets were removed. Additionally, there is still a possibility that more tweets may be deleted, further reducing the total number of tweets available for exploitation.

6.4 Tasks descriptions

As said before, we use the *humanitarian* and *informative* tasks (from CrisisMMD) to annotate the dataset.[4] However, we also adapt the task to the practises in the community, as its was already done within several works [13, 94, 95, 109] which have merged classes relative to damages (“infrastructure_and_utility_damage” and “vehicle_damage") into “infrastructure_and_utility_damage” and classes relative to human casualties (“affected_individuals",”injured_or_dead_people",”missing_or_found_people”) into “affected_individuals", resulting in a five classes task described bellow:

Affected individuals If the tweet/image reports or shows individuals affected by the disaster event, such as people sitting outdoors, individuals standing in lines for assistance, people in need of shelter facilities, missing or found individuals, or deceased individuals.

[4]The annotation guide is delivered with the resource (https://github.com/badreddineFarah/M-CATNAT).

Infrastructure and utility damage If the tweet/image reports or shows a damaged structure or one whose use is affected by an earthquake, fire, heavy rains, floods, strong winds, gusts, etc., such as damaged houses, roads, buildings; flooded houses, streets, highways; blocked roads, bridges, paths; collapsed bridges, power lines, cars, boats, communication poles, etc.

Rescue volunteering or donation effort If the tweet/image reports or shows any type of rescue, volunteering, or donation effort, such as transporting people to safe locations, evacuating people from the hazardous area, individuals receiving medical or food aid, people in shelter facilities, monetary donations, blood donations, or services, etc.

Other relevant information If the tweet/image does not fit into any of the three categories above but still contains important information useful for humanitarian aid.

Not related or not relevant information If the tweet/image is not useful for humanitarian aid.

As described in the CrisisMMD paper, only informative tweets have been annotated into one of the *humanitarian* classes. Thus, considering other tweets as not informative (not related or not relevant) the annotated dataset can also be used to train models on the *informativeness* task.

6.5 Annotation methodology

In the annotation methodology, we implemented a four-phase process to ensure high-quality annotation of the dataset. At the end of each phase, an assessment of the annotations was made according to different indicators, followed by a consultation meeting focusing on examples with substantial disagreement occurred. A clarification, accompanied by illustrative examples was then incorporated into the annotation guide. Also, to mitigate the bias in labelling the modalities from each (multimodal) tweet, the order of presentation of the instances to be annotated needs to be carefully considered. An annotator who is first presented with a complete multimodal tweet (text + image) will undoubtedly be influenced in his annotation (a posteriori) of each of the modalities. We have therefore ensured that the text and image modalities are presented first and in a separate slot (groups of annotators) or phases from the multimodal tweet (text+image).

This sequential presentation aimed to ensure independent labeling decisions for both unimodal and multimodal instances, minimizing potential bias in annotations.

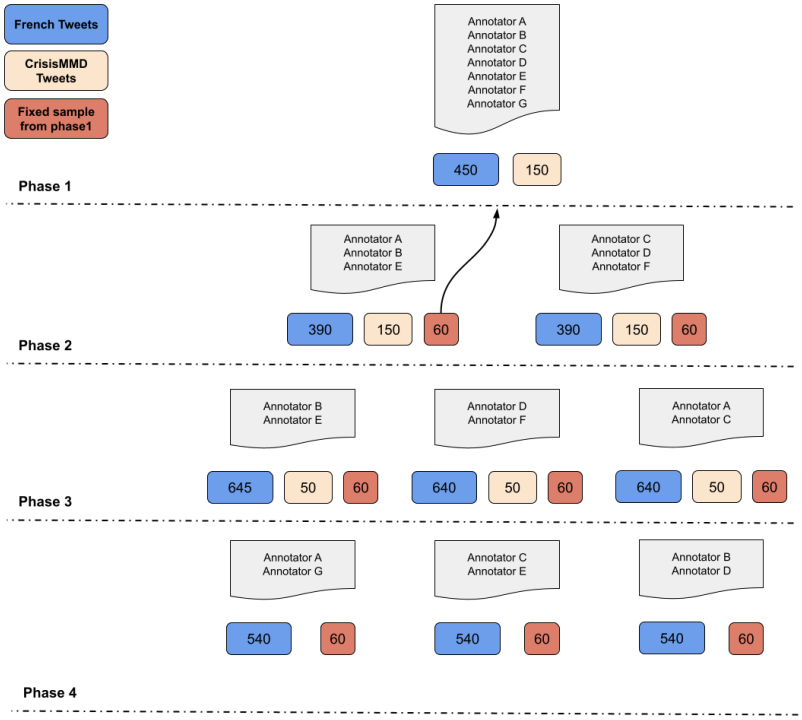


Figure 6.1: Overview of the annotation protocol. This diagram illustrates our multi-phase annotation methodology. In each phase, we annotate a specific number of new instances (image, text and multimodal instances), highlighted in blue. To maintain alignment with the CrisisMMD, a small portion of this dataset is also annotated. Lastly, to ensure the self-consistency of each annotator, some instances are selected for re-annotation.

As presented in Figure 6.1, Phase 1 involved seven annotators collectively annotating one shared dataset of 600 instances constituted of 150 instances from CrisisMMD and 450 instances from new French tweets ($\frac{1}{3}$ images, $\frac{1}{3}$ texts and $\frac{1}{3}$ multimodal instances). In this phase, when an instance is annotated identically by at least 5 annotators (out of 7), this annotation is definitively used as a label. The other instances (for which the annotation is not consensual) are discussed and collectively labelled at the end of the phase during the consultation meeting. This process allowed us to clarify and adjust the annotation guide. This first phase is an agreement (or training) phase for the annotators. Subsequent phases gradually "industrialise" the annotation process, with annotators organised into parallel pools.

Phase 2 comprised two pools, each pool was made up of 3 annotators responsible for annotating 600 instances. Among the 600 instances, we maintained a set of 150 CrisisMMD instances and introduced 60 instances from phase 1 in order to compute the consistency of annotations through the phases for each annotator involved in the process. At the end of this phase, the instances with two identical annotations are definitively labelled, while we proceed to a correction phase during which each disagreement is resolved by an annotator from the other pool.

In phase 3, we decomposed our team into three annotation pairs, for each annotation pool we reduced the number of CrisisMMD instances to 50 while keeping the 60 instances from phase 1. The disagreement is resolved by a third annotator from another pool.

Lastly, Phase 4 involves newly arranged pairs from the annotators team, each annotating 540 instances, along with the 60 instances from Phase 1. This process divided into phases resulted in 1,358 tweets annotated with 4,691 annotation task completed.

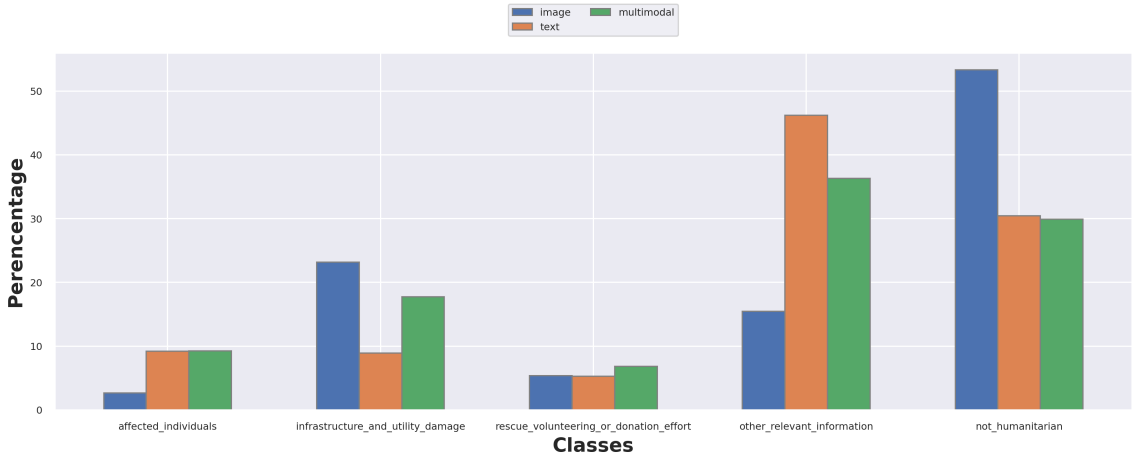


Figure 6.2: Comparison of label distribution for each modality, the distributions are calculated on a 1,558 fully annotated instances.

6.6 Resource analysis

In this section, we present a detailed analysis of the M-CATNAT resource. Our initial goal was to annotate 1,430 tweets, but we were able to annotate only 1,356 tweets because some tweets were deleted during the annotation process. This resulted in a total of 1,558 multimodal instances, as some tweets contained multiple images. Firstly, we outline the distribution of labels for each modality and for each type of disaster (earthquakes and floods). Secondly, we present the results of annotation scores, including the Fleiss Kappa score [149] and the CrisisMMD alignment score. Finally, we show the distribution of class concordance/discordance according to modalities.

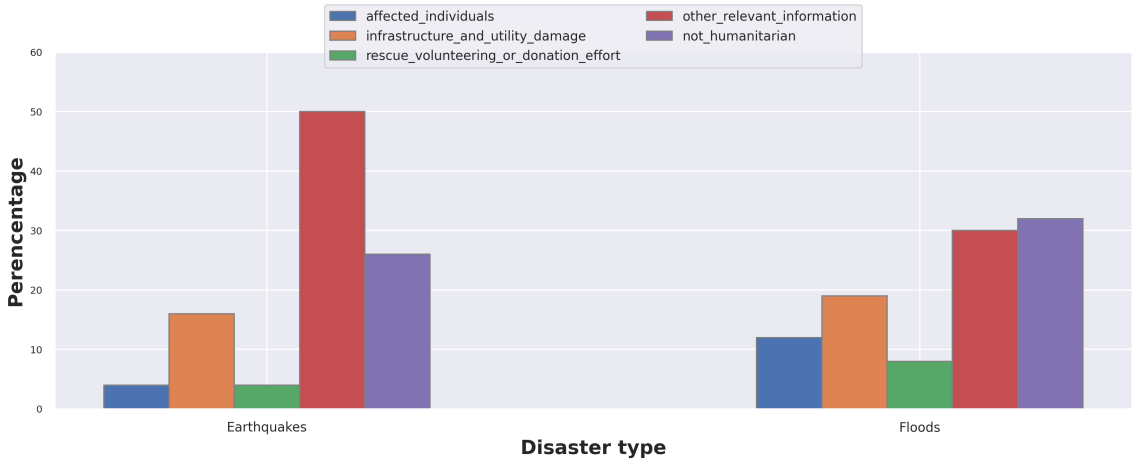


Figure 6.3: Comparaison of label distribution for earthquakes and floods, the result are on 496 earthquake related tweets and 1063 floods related tweets.

6.6.1 Label distribution

Figure 6.2 illustrates the label distribution across different classes. Notably, the class "Affected Individuals" exhibits the lowest representation. The "Not Humanitarian" class, equivalent to "Not informative" in the *informativeness* task of CrisisMMD, constitutes 30% in text and multimodal instances but rises to approximately 50% in image labels. Conversely, "Other Relevant Information" represents 46% in texts and 36% in multimodal instances, contrasting with its lower representation at 16% in images. This difference is attributed to the enriching role of textual information in multimodal instances, aiding in disaster characterization regarding intensity and location. The distribution of multimodal labels closely aligns with the text distribution, highlighting the importance of textual information in crisis datasets in general and in the M-CATNAT dataset in particular. The emphasis on textual information stems from its reflective of tweet intentions and

Table 6.3: Fleiss kappa scores on each phase.

Phase	1	2	3	4
Mean (over pools)	0.64	0.70	0.61	0.61
Common examples	0.64	0.68	0.66	0.63

Table 6.4: CrisisMMD alignment scores on each phase.

Phase	1	2	3
CrisisMMD alignment	79%	74%	77%

provision of more precise details. Classes such as "Rescue Volunteering and Donation Efforts" are more prominent in CrisisMMD (14.5%), indicating that the disasters in our data are comparatively less severe.

Figure 6.3 reports the distribution of multimodal annotations according to the type of disaster in our dataset. "Other Relevant Information" exhibits higher representation in earthquake-related tweets, due to the lower magnitude of these earthquakes where tweets often provide basic information about the incident, such as location and magnitude. Conversely, "Affected Individuals" and “Rescue volunteering or donation effort” are more represented in flood-related tweets, reflecting the extensive damage caused by such events.

6.6.2 Annotator scores

In accordance with the annotation methodology outlined in the "Annotation Methodology" section, our process involves four phases, each subdivided into annotator pools. In consequence we propose to measure inter-annotator agreement on each phase using the Fleiss Kappa [149]. Table 6.3 reports the results obtained: first row shows the mean Fleiss Kappa score overs pools for each phase, while in the second row Fleiss Kappa scores were calculated specifically for the 60 common examples (highlighted as red samples in Figure 6.1). The variations observed are mainly due to the changes made to the annotation guide over the phases; it is also significant to note that the inter-annotator agreements remained above 0.60 in each of the three phases. To be complete, we noticed the agreement between annotators varied across modalities. A higher agreement is observed for images (up to 0.85) compared to texts which exhibited a poorer annotator agreement, particularly in Phase 1 (0.56). In addition, as underlined in Section 6.5 at every phase we made annotations for a sample of 60 fixed instances (in red in Figure 6.1) in order to monitor the intra-annotator agreement along the whole process. The percentage scores obtained vary between 67% and 78% for the transition from phase 1 to phase 2 and between 82% and 89% for the transition from phase 2 to phase 3 while

the scores reached 87% to 91% in the last phase suggesting a relatively high annotators consistency.

6.6.3 CrisisMMD alignment

Moreover, we assessed alignment with the CrisisMMD dataset (see Table 6.4). In Phase 1, alignment was computed on 150 examples after disagreement resolution, yielding a score of 79% (with a notable increase to 90% for multimodal instances). In Phase 2, the alignment score was computed on 150 examples for each pool and the mean score obtained is 74%, while it reached 77% in phase 3 on 50 examples for each pool (mean score of the three pools is reported). These results are encouraging given our will to use the proposed resource in addition to CrisisMMD in a multilingual configuration.

6.6.4 Modality label analysis

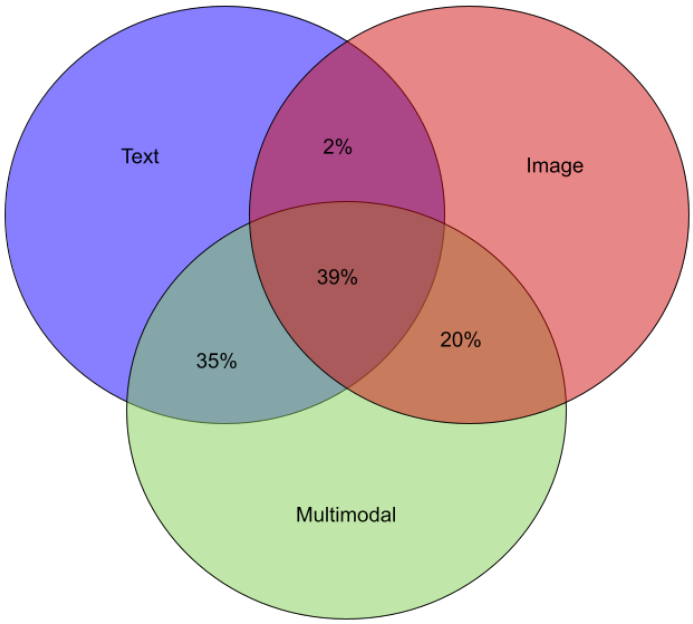


Figure 6.4: Label Combinations for Image, Text, and Multimodal on 1,558 fully annotated instances.

Figure 6.4 shows an analysis of the labelling of the 1,558 [5] examples in the current M-CATNAT dataset. It reveals the relation in the distribution of labels across the three

[5]We can get 1,558 image-text instances from the 1356 annotated tweets as a tweet can have multiple images.

labels assigned to one Tweet : the *image* label, the *text* label and the *multimodal* label. A significant portion, accounting for 39% of the examples, demonstrates consistency across all three labels ; this suggests a strong correlation among the different types of data and their respective labels. In 35% of cases, although the text label differs from the image label, it aligns with the multimodal label ; this implies that the textual content plays a crucial role in shaping the combined interpretation. Conversely, 20% of the examples have an image label distinct from the text label but aligning with the multimodal one, highlighting the influence of visual elements in these instances. Following these results, one may consider the need to train models to spot instances where the information looked for may be found in a specific modality. Lastly in a small proportion, just 2%, the image and text labels match but differ from the multimodal label, and lastly, 4% of cases shows complete disparity across all three labels. These findings confirm our findings in Section 3.9.2 on the variety of relations between different modalities and emphasize the importance of considering multimodal annotation rather than an independent annotation of the text and the image.

It's important to note that most works based on CrisisMMD exploitation typically use *concordant* instances where labels match across modalities. However, as shown in Figure 6.4, this approach overlooks more than 50% (57% in our case) of the data, disregarding the diverse image-text relationships present in social media data.

6.7 Conclusion

As part of this thesis, our results showed the interest of a refined methodology responding to the challenges of multimodality using deep learning techniques. In the context of natural disaster and crisis management, the use of multimodal data holds significant promise for improving the efficiency of social media usage, but the lack of fully multimodal annotated instances is a pitfall that this contribution attempts to overcome. In this chapter we introduced M-CATNAT, a French multimodal dataset aligned with CrisisMMD, featuring three labels for each instance. This alignment facilitates multilingual training, thereby broadening the applicability of deep learning techniques. We provide annotator scores and other analyses to support deep learning practitioners in utilizing the dataset effectively.

In future work, we intend to train models using this dataset in both monolingual and multilingual settings. Even with the rise of zero-shot and few shot classification through in context learning using multimodal large language models, M-CATNAT will continue to be valuable for evaluating these models and their prompts. Additionally, the dataset offers a more detailed framework for assessing model performance across different

image-text relationships, providing deeper insights into their capabilities.

Chapter 7

General Conclusion

Throughout this thesis, we investigated methods to advance multimodal classification for crisis management, focusing on the integration of text and image data from social media posts during emergency events. By addressing the challenges of combining multimodal data, each chapter contributed to the understanding and enhancement of deep learning models that leverage both textual and visual information for more effective crisis response.

In Chapter 3, we initiated our study with a detailed analysis of the CrisisMMD dataset, a foundational dataset for multimodal crisis research. This analysis revealed limitations and challenges on this largely used dataset. We applied state-of-the-art transformer-based models including BERT and RoBERTa for text, and ViT for image classification, demonstrating that these architectures are well-suited to handle crisis-related data, significantly outperforming traditional approaches. Building on these unimodal findings, we experimented with multimodal fusion techniques, including feature fusion and decision fusion methods, as well as more advanced transformer-based fusion mechanisms like bottleneck attention. Our findings indicated that feature-level fusion using attention-based mechanisms achieved the highest accuracy. We also examined pretrained vision-language models like LXMERT and ViLT, which performed below expectation on CrisisMMD, likely due to the dataset's complex image-text relationships. A CLIP similarity score analysis confirmed that CrisisMMD includes a wider range of relationships than datasets like MSCOCO. This diversity highlights the need for fusion models specifically adapted to this type of data.

In Chapter 4, we addressed the complexity of image-text fusion by introducing Caption-based Multimodal BERT (CMB). This model translates images into captions, creating a shared textual representation to facilitate fusion with the accompanying text in a simplified manner. Through this modality translation mechanism, CMB demonstrated

superior performance, often outperforming traditional multimodal baselines, such as feature concatenation and cross-attention, which handle raw image data directly. One key finding was the impact of caption quality on model performance, as captions with higher semantic accuracy led to more effective fusion. By comparing various captioning models, we found that CLIP-based models, suggesting that an accurate semantic translation of images into text is essential for effective multimodal learning. To account for real-world social media scenarios, where some posts lack images, we developed a mixtraining strategy that makes CMB model capable of processing both unimodal (text-only) and multimodal (text and caption) inputs. This hybrid approach proved effective, maintaining high accuracy on both text only and multimodal senarios. However, CMB showed sensitivity to poor quality captions, which sometimes led to misclassifications, highlighting an area for improvement in future work. Moreover by analysis the errors of the models we showed an imbalance towards text, specifically the models relied more on the text ignoring in some cases the image modality making the multimodal learning suboptimal.

Chapter 5 addressed the challenge of modality imbalance in multimodal classification, a phenomenon where one modality tends to dominate, limiting the effectiveness of the overall model. Although commonly seen in audio-video classification, this issue also emerged in our experiments with text-based multimodal models, where reliance on text often overshadowed other modalities. To counteract this, Chapter 5 introduced Dynamic Stochastic Shared Embeddings (D-SSE), a novel approach designed to dynamically regulate dominant modalities, ensuring a more balanced representation and enhancing model accuracy. Through extensive empirical analysis, D-SSE was shown to effectively reduce modality imbalance and improve classification accuracy across four diverse datasets. Further tests in controlled imbalance scenarios confirmed D-SSE's robustness in handling different degrees of imbalance. Beyond boosting classification performance, Chapter 5 provided insights into how each modality's representation space aligned with the shared multimodal space. This deeper analysis demonstrated D-SSE's ability to balance and integrate diverse modalities within the model, enhancing its overall efficacy.

In Chapter 6, we addressed a crucial limitation found in crisis-related use of social media, particularly within the context of humanitarian aid. While previous chapters primarily focused on model development, this chapter introduced M-CATNAT, a French multimodal dataset created to overcome gaps in existing resources like CrisisMMD. M-CATNAT not only expands CrisisMMD into the French language but also enriches it with comprehensive multimodal labeling. Each tweet is annotated with separate labels for text, image, and the combined multimodal content, offering a nuanced view of content relationships. To support effective use of M-CATNAT, we provided thorough analyses, including class distribution, disaster type tweets analysis, and annotator scores to ensure dataset reliability and alignement with CrisisMMD. Additionally, we examined the in-

teractions between unimodal and multimodal labels, which underscored the complexity of image-text relationships in crisis-related content, confirming initial insights into the intricate ways these modalities interact. This resource is a step forward in broadening the scope for multilingual and multimodal crisis response models.

The findings of this thesis have laid the groundwork for several promising research directions that can extend our contributions to multimodal crisis response. First, in the area of captioning, future research could focus on developing task-oriented captioning specifically tailored to crisis scenarios. Such captions would enhance both model accuracy and interpretability, providing outputs that highlight essential contextual details needed in crisis situations, ultimately making model predictions more informative and actionable. As video continues to grow as a communication medium on social media, integrating video data into multimodal crisis response models presents a valuable research avenue. Future studies could explore how video can deepen situational awareness by capturing the real-time, dynamic aspects of events. This addition could provide crisis management personnel with a more comprehensive view of unfolding situations, enabling faster and more contextually rich decision-making. Given the reliance on a single dataset in this thesis, future research should also prioritize the creation and annotation of additional datasets to fully harness the potential of social media for crisis management. A larger pool of annotated data, especially in multilingual contexts, would allow for a more robust model performance across diverse crisis scenarios. With large language models (LLMs) showing promising potential in other domains, exploring their application in crisis management is a compelling direction. However, specific benchmarks are needed to assess the reliability and adaptability of LLMs in this domain. Expanding datasets, such as M-CATNAT, could facilitate more thorough evaluations and strengthen the effectiveness of LLMs in real-world crisis applications. By pursuing these directions, future research can build on the advancements made in this thesis, driving meaningful progress in multimodal learning and enhancing the capabilities of crisis response models.

Chapter 8

Résumé substantiel en français

8.1 Introduction

Ces dernières années, l'utilisation des réseaux sociaux a connu une augmentation exponentielle, devenant un élément central de la vie quotidienne pour des millions de personnes à travers le monde. Ces plateformes permettent de rester connecté, de partager des informations, d'exprimer des opinions, et de collaborer sur divers projets. Leur impact s'étend à de nombreux secteurs, notamment la santé publique, la prévision économique et l'analyse politique. Dans les situations de crise, comme les catastrophes naturelles, les réseaux sociaux peuvent devenir une source précieuse d'informations en temps réel. Plusieurs études ont confirmé que les acteurs de la gestion de crise considèrent ces plateformes, notamment Twitter (X aujourd'hui), comme des outils d'aide à la prise de décision. En France, où les catastrophes naturelles comme les inondations sont de plus en plus fréquentes, l'analyse des posts partagés sur ces plateformes permet d'évaluer l'ampleur des désastres et d'orienter les stratégies de gestion de crise. Cependant, le volume massif de données générées par les utilisateurs nécessite des méthodes de traitement automatisées, car une grande partie de ces informations est considérée comme du bruit. Les premières approches pour exploiter Twitter dans la gestion de crise se sont principalement concentrées sur le traitement du texte, utilisant des techniques d'apprentissage automatique pour filtrer et classifier les tweets pertinents. Cependant, des recherches récentes ont démontré que les images partagées pendant ces crises jouent un rôle crucial en complément des textes, permettant une meilleure compréhension de la situation sur le terrain. Notre travail s'inscrit dans le cadre du projet RéSoCIO, en collaboration avec le BRGM, qui vise à exploiter l'analyse des données issues de Twitter en temps réel pour améliorer la gestion des crises. Nos discussions avec des professionnels français de la gestion de crise soulignent le besoin d'images de terrain pour compléter les informations textuelles et faciliter la prise de décisions rapides et informées. Si les travaux précé-

dents se sont principalement concentrés sur la classification du texte ou des images de manière indépendante, il est essentiel d'intégrer ces deux modalités pour une analyse plus complète et efficace des tweets de crise. L'utilisation de modèles multimodaux présente cependant plusieurs défis auxquels cette thèse tente de répondre.

8.2 Réseaux sociaux et gestion de crise

L'essor des médias sociaux a transformé la gestion de crise, offrant un flux d'informations en temps réel essentiel lors des catastrophes. De ce fait, les organisations doivent désormais intégrer ces plateformes dans leurs stratégies d'intervention. Pour ce faire, il est crucial de développer des méthodes efficaces de filtrage et d'analyse du contenu généré par les utilisateurs. Dans cette optique, les chercheurs ont mis au point des ensembles de données spécifiques, conçus pour entraîner des modèles d'apprentissage automatique à distinguer les informations pertinentes du bruit ambiant. Cette démarche implique souvent de classer les données en fonction de leur utilité et pertinence. En pratique, les méthodes de classification varient en fonction du type de données à traiter. Pour le texte, on observe une évolution des simples correspondances de mots-clés vers des modèles de traitement du langage naturel plus sophistiqués, tels que BERT. Parallèlement, pour les images, réseaux de neurones convolutifs sont utilisés pour analyser et catégoriser le contenu visuel. Par ailleurs, la combinaison de texte et d'images, ou données multimodales, offre une compréhension plus riche des situations de crise comme divers travaux l'ont montré. Ainsi, des techniques de fusion de données comme la concaténation et la sommation de représentations sont employées pour exploiter pleinement cette richesse d'informations. Pour entraîner ces modèles de fusion des ensembles de données comme CrisisMMD ont été introduits. L'usage des deux modalités a montré des résultats prometteurs en terme de précision de classification, une précision qui permettra au gestionnaire de crise d'avoir des informatisons plus fiables permettant une gestion de catastrophe naturelle plus efficace.

8.3 Classification multimodale de tweets: premières expériences

Dans le but de mieux explorer la thématique de classification multimodale de tweets de catastrophes naturelles, nous avons mené dans un premier temps une exploration approfondie de la classification multimodale des tweets dans le contexte de la gestion de crise, en utilisant le dataset CrisisMMD. Nous avons commencé par analyser la structure et la distribution du dataset, mettant en évidence les défis posés par les instances

discordantes où les labels du texte et de l'image diffèrent. Notre analyse a révélé que, bien que le texte et les images fournissent individuellement des informations précieuses, leur combinaison permet d'obtenir une représentation plus riche du contenu, ce qui est essentiel pour une réponse efficace aux crises. Nous avons implémenté et évalué plusieurs modèles unimodaux, notamment BERT et RoBERTa pour la classification du texte, ainsi que ViT et DenseNet pour la classification des images. Les résultats ont confirmé que les modèles basés sur les Transformers surpassent les méthodes traditionnelles, avec ViT et RoBERTa atteignant les meilleures précisions dans leurs modalités respectives. Ces résultats sont cohérents avec les recherches antérieures et soulignent l'efficacité des architectures de Transformers pour traiter les données liées aux crises. En nous appuyant sur les résultats unimodaux, nous avons exploré diverses techniques de fusion multimodale afin d'intégrer les informations textuelles et visuelles. Nous avons examiné des méthodes de fusion décisionnelle et des stratégies de fusion de caractéristiques ("features") ,qui sont catégorisées en deux principales classes, la fusion tardive (co-attention, concaténation et cross-attention) ainsi que des mécanismes d'attention avancés, notamment la fusion basée sur les couches de Transformers et l'attention par goulot d'étranglement. Nos résultats ont indiqué que les méthodes de fusion au niveau des caractéristiques, en particulier celles utilisant des mécanismes d'attention basés sur les Transformers, surpassent les techniques de fusion au niveau décisionnel. La méthode de fusion basée sur l'attention par goulot d'étranglement a démontré des performances supérieures. Nous avons également évalué des modèles préentraînés de vision et langage tels que LXMERT et ViLT. Bien que ces modèles aient montré des performances remarquables sur des tâches multimodales générales, ils se sont révélés moins performants dans le contexte de la classification de tweets liés aux crises. Cette contre-performance peut être attribuée aux relations uniques et complexes entre les images et le texte dans le dataset CrisisMMD, qui diffèrent significativement des datasets utilisés pour préentraîner ces modèles. Notre analyse des relations image-texte à l'aide des scores de similarité CLIP a révélé que CrisisMMD englobe une large gamme de relations, allant d'une correspondance étroite à des paires complètement indépendantes, comme en témoignent les différentes distributions de similarité. Cette diversité constitue un défi pour les modèles préentraînés sur des datasets avec des alignements image-texte plus directs, tels que MSCOCO. Les relations variées soulignent la nécessité de modèles capables de s'adapter aux caractéristiques uniques des données issues des réseaux sociaux en temps de crise. Ces premières expérimentations sur la classification multimodale des tweets ont démontré l'efficacité de la combinaison des données textuelles et visuelles, les méthodes de fusion basées sur les Transformers offrant les meilleures performances. Les modèles préentraînés de vision et langage ont rencontré des difficultés face à la diversité des relations entre modalités dans CrisisMMD.

8.4 Classification multimodale basée sur les légendes d'images

Comme exposé dans le chapitre précédent, bien que les approches multimodales surpassent généralement les modèles unimodaux, les différences de performance entre les différentes techniques de fusion (telles que la concaténation, la cross-attention et la co-attention) demeurent relativement faibles et n'apportent pas de gains substantiels. Par ailleurs, nous avons observé la grande diversité des relations pouvant exister entre une image et son texte associé, notamment sur les réseaux sociaux. Cette diversité, combinée à la limitation des données, complexifie la tâche des modèles pour capturer les relations entre l'image et le texte, rendant ainsi la fusion multimodale plus difficile. En partant de l'hypothèse selon laquelle l'intégration des deux modalités (image et texte) dans un même espace de représentation pourrait faciliter l'interaction inter-modale et ainsi améliorer les performances de classification, nous proposons dans ce chapitre la méthode CMBx (Caption-based Multimodal BERT). Contrairement aux approches classiques qui utilisent des encodeurs distincts pour le texte et l'image, notre méthode repose sur un mécanisme de traduction de modalité : une image est d'abord convertie en texte à l'aide d'un modèle de génération de légendes (captioning model), puis cette légende est fusionnée avec le texte du tweet. Cette approche permet une fusion plus simple et plus efficace en opérant dans un espace textuel commun. Nos expériences réalisées sur le jeu de données CrisisMMD montrent que CMB offre des performances compétitives, surpassant fréquemment les approches traditionnelles basées sur la concaténation ou la cross-attention. L'un des résultats clés de notre étude concerne l'impact de la qualité des légendes générées sur les performances du modèle. En comparant différents modèles de génération de légendes, nous avons constaté que ceux basés sur CLIP offrent de bonnes performances. De plus, CMB reste robuste, même avec diverses variantes de modèles de légendes, ce qui démontre sa flexibilité face à différentes sources de données. Un autre apport significatif de notre travail réside dans le développement d'une stratégie d'entraînement mixte. En effet, de nombreux tweets sur les réseaux sociaux ne comportent pas d'images, ce qui représente un défi pour les modèles multimodaux. Pour y remédier, nous avons proposé une stratégie d'entraînement permettant à CMB de traiter aussi bien des entrées unimodales (texte seul) et multimodales (texte et légende d'image). Nos résultats expérimentaux montrent que ce modèle hybride parvient à maintenir une précision élevée dans les deux cas.

Toutefois, CMB présente certaines limites. Nous avons observé que le modèle reste sensible à la qualité des légendes générées : lorsque celles-ci sont mal formulées ou contiennent des informations incorrectes, les prédictions deviennent moins précises. Cela souligne la nécessité d'améliorer les modèles de génération de légendes pour mieux exploiter les données visuelles. Enfin, comme c'est souvent le cas pour de nombreux mod-

èles de fusion multimodale, nous avons constaté que CMB tend à privilégier le texte au détriment des informations visuelles, ce qui peut limiter l'impact des modalités visuelles.

8.5 Mitigation du déséquilibre multimodal

Dans ce chapitre, nous avons mis en évidence que l'entraînement multimodal peut souvent aboutir à une performance sous-optimale en raison d'un déséquilibre entre les modalités, révélant ainsi les limites des approches classiques d'entraînement conjoint. Pour remédier à ce problème, nous avons introduit la technique Dynamic Stochastic Shared Embedding (D-SSE), qui repose sur des transitions stochastiques entre les embeddings, guidées par des probabilités dynamiques P_m, un paramètre qui se calcule en utilisant les dynamiques dominant/dominées des modalités durant l'entraînement. Nos expériences, menées sur trois ensembles de données distincts — CREMA-D (audio et vidéo), ModelNet40 (deux vues d'objets 3D) et CrisisMMD (paires image-texte) — ont démontré l'efficacité de notre approche. D-SSE améliore significativement la fusion multimodale et optimise la performance globale des modèles. Plus précisément, cette méthode permet non seulement d'augmenter la précision des modèles multimodaux, mais aussi de réduire les déséquilibres entre les modalités, favorisant ainsi un apprentissage plus équilibré. Cette amélioration se traduit par une progression constante des performances en mode unimodal et multimodal, ainsi qu'une meilleure homogénéité des similarités entre les représentations des deux modalités comparé à la représentation multimodale. Par ailleurs, nos tests sur le jeu de données CG-MNIST ont confirmé la robustesse de D-SSE dans divers scénarios, allant de déséquilibres modérés à sévères entre les modalités. Toutefois, notre implémention actuelle de D-SSE se limite aux cas impliquant uniquement deux modalités. Un travail futur devra se concentrer sur son extension à des configurations intégrant trois modalités ou plus. Un autre axe d'amélioration concerne le réglage des hyper-paramètres : la modulation dynamique de P_m a montré un fort potentiel, suggérant que l'exploration de mises à jour dynamiques pour P_c (le paramètre qui contrôle le nombre de transitions entre embeddings durant l'entraînement) pourrait stabiliser davantage l'entraînement et améliorer encore les performances. En surmontant ces limitations et en élargissant le cadre d'application de D-SSE, cette approche pourrait être adaptée à une large gamme de scénarios multimodaux, renforçant ainsi sa robustesse et son efficacité dans des contextes réels variés.

8.6 M-CATNAT un dataset de tweets en français

Dans le domaine de la gestion des catastrophes naturelles et des crises, l'exploitation des données multimodales présente un potentiel considérable pour améliorer la réactivité et la précision de l'analyse des informations issues des réseaux sociaux. Dans le cadre de cette thèse, nos résultats ont mis en évidence l'intérêt d'une méthodologie affinée répondant aux défis de la multimodalité grâce à des techniques d'apprentissage profond ciblé pour cette tache de classification de tweets multimodaux. Cependant, la rareté de bases de données annotées de manière exhaustive et alignées sur des scénarios réels constitue un obstacle majeur. Cette contribution vise à pallier cette limitation en proposant un jeu de données spécifiquement conçu pour la recherche dans ce domaine. Dans ce chapitre, nous introduisons M-CATNAT, un jeu de données multimodal en français. Il comprend des annotations détaillées avec trois étiquettes par instance, une étiquette étant dédiée à la portion du texte de l'instance, une seconde à sa partie image et la dernière étant une annotation globale de l'instance, rendant le corpus réellement multimodal et propice à l'étude de l'influence de chaque modalité. De plus M-CATNAT est conçu pour être aligné avec CrisisMMD, ce qui favorise l'entraînement de modèles en contextes multilingues, élargissant ainsi la portée et l'applicabilité des approches d'apprentissage profond. De plus, nous avons inclus des scores d'annotateurs pour assurer une évaluation rigoureuse de la qualité des annotations et fournir aux chercheurs des indications précises sur la fiabilité des données. Nous avons également réalisé une série d'analyses exploratoires afin de mieux comprendre les corrélations entre les modalités. Ces analyses ont révélé des tendances significatives dans l'association des éléments textuels et visuels, mettant en lumière les défis inhérents à la classification automatique dans un cadre multimodal. Dans les perspectives futures, nous prévoyons d'explorer l'entraînement de modèles de classification en exploitant pleinement M-CATNAT, aussi bien dans des configurations monolingues que multilingues.

8.7 Conclusion

Cette thèse se concentre sur l'amélioration de la classification multimodale pour la gestion de crise, en intégrant les données textuelles et visuelles provenant des réseaux sociaux. Chaque chapitre aborde des défis spécifiques, avec l'objectif commun de développer des modèles plus fiables et permettre aux gestionnaires de crises un accès à une information plus riche et fiable. Dans un premier temps, le Chapitre 3 analyse le jeu de données CrisisMMD, un élément central pour la recherche sur les crises multimodales. Nous y appliquons des modèles basés sur des Transformers (BERT, RoBERTa, ViT) pour le traitement du texte et des images. Les résultats obtenus montrent que la fusion des carac-

téristiques ("features") à l'aide de mécanismes d'attention améliore la précision globale. Cependant, les modèles vison-langage pré-entraînés comme LXMERT ne parviennent pas à capturer correctement les relations complexes entre l'image et le texte, mettant en évidence les limites de l'approche traditionnelle. Pour surmonter ces limitations, le Chapitre 4 introduit un nouveau modèle : Caption-based Multimodal BERT (CMB), qui traduit les images en légendes textuelles. Ce mécanisme simplifie la fusion avec le texte et permet d'obtenir des performances supérieures aux méthodes classiques. Toutefois, cette approche reste sensible à la qualité des légendes générées, ce qui peut nuire à la précision du modèle. Une stratégie hybride, capable de traiter aussi bien des entrées textuelles seules que multimodales, est mise en place pour améliorer la robustesse du modèle dans des scénarios variés. Le Chapitre 5 aborde un autre défi majeur de la classification multimodale : le déséquilibre des modalités. En effet, dans nos expériences, le texte a souvent dominé les autres modalités, ce qui a limité l'efficacité du modèle. Pour pallier ce problème, nous proposons Dynamic Stochastic Shared Embeddings (D-SSE), une méthode innovante qui régule dynamiquement les modalités dominantes. Ce mécanisme permet d'équilibrer l'importance des modalités, améliorant ainsi la précision du modèle et sa capacité à intégrer harmonieusementles données textuelles et visuelles. Enfin, le Chapitre 6 présente M-CATNAT, un jeu de données multimodal français, développé pour combler les lacunes de CrisisMMD, notamment en termes d'annotations où on a annoté les images, les textes et les contenus multimodaux. Ce jeu de données enrichit les ressources existantes et permet de mieux modéliser les interactions complexes entre les modalités, tout en soutenant l'amélioration des modèles multilingues.

Les résultats de cette thèse ouvrent plusieurs perspectives pour les recherches futures. Parmi celles-ci, le développement de légendes orientées "tâche" spécifiquement adaptées aux crises, l'intégration des données vidéo pour une meilleure prise de décision en temps réel, ainsi que l'extension des jeux de données multilingues pour renforcer la généralisation des modèles. Enfin, l'application des grands modèles de langage (LLM) dans la gestion de crise représente une direction prometteuse, nécessitant toutefois des évaluations spécifiques pour mesurer leur fiabilité et leur efficacité dans ce domaine. Ces pistes de recherche devraient permettre d'approfondir les progrès réalisés et d'élargir les capacités de ces systèmes aidant ainsi les gestionnaires de catastrophes naturelles à mieux récupérer l'information issue des réseaux sociaux.

Bibliography

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

[3] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

[4] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[5] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9012–9020, 2019.

[6] Patrick Y Wu and Walter R Mebane Jr. Marmot: A deep learning framework for constructing multimodal representations for vision-and-language tasks. *Computational Communication Research*, 4(1), 2022.

[7] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, 2015.

[8] Christian Reuter, Thomas Ludwig, Therese Friberg, Sylvia Pratzler-Wanczura, and Alexis Gizikis. Social media and emergency services?: Interview study on current and potential use in 7 european countries. *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)*, 7(2):36–58, 2015.

[9] Linda Plotnick, Starr Roxanne Hiltz, Jane A Kushma, and Andrea H Tapia. Red tape: Attitudes and issues related to use of social media by us county-level emergency managers. In *ISCRAM*, 2015.

[10] C Castillo, M Imran, P Meier, JK Lucas, J Srivastava, H Leson, F Ofli, and P Mitra. Together we stand—supporting decision in crisis response: Artificial intelligence for digital response and micromappers. *by OCHA and partners. Istanbul: Tudor Rose, World Humanitarian Summit*, pages 93–95, 2016.

[11] Robin Peters and Joao De Albuquerque. Investigating images as indicators for relevant social media messages in disaster management. 01 2015.

[12] Min Jing, Bryan W Scotney, Sonya A Coleman, Martin T McGinnity, Xiubo Zhang, Stephen Kelly, Khurshid Ahmad, Antje Schlaf, Sabine Gründer-Fahrer, and Gerhard Heyer. Integration of text and image analysis for flood event image recognition. In *2016 27th Irish signals and systems conference (ISSC)*, pages 1–6. IEEE, 2016.

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

[14] Firoj Alam, Ferda Ofli, and Muhammad Imran. Crisismmd: Multimodal twitter datasets from natural disasters. In *Twelfth international AAAI conference on web and social media*, 2018.

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

[16] Tiberiu Sosea, Iustin Sirbu, Cornelia Caragea, Doina Caragea, and Traian Rebedea. Using the image-text relationship to improve multimodal disaster tweet classification. In *The 18th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2021)*, 2021.

[17] Andreas M Kaplan and Michael Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68, 2010.

[18] Pritam Gundecha and Huan Liu. Mining social media: a brief introduction. *New directions in informatics, optimization, logistics, and production*, pages 1–17, 2012.

[19] Joseph T. Yun, Utku Pamuksuz, and Brittany R. L. Duff. Are we who we follow? computationally analyzing human personality and brand following on twitter. *International Journal of Advertising*, 38(5):776–795, July 2019.

[20] Roberto Micera and Raffaele Crispino. Destination web reputation as "smart tool" for image building: the case analysis of naples city-destination. *International Journal of Tourism Cities*, 3(4):406–423, December 2017.

[21] Jaewoong Choi, Janghyeok Yoon, Jaemin Chung, Byoung-Youl Coh, and Jae-Min Lee. Social media analytics and business intelligence research: A systematic review. *Information Processing Management*, 57(6):102279, November 2020.

[22] Prabhsimran Singh, Yogesh K. Dwivedi, Karanjeet Singh Kahlon, Ravinder Singh Sawhney, Ali Abdallah Alalwan, and Nripendra P. Rana. Smart monitoring and controlling of government policies using social media and cloud computing. *Information Systems Frontiers*, April 2019.

[23] Nathaniel Whittingham, Andreas Boecker, and Alexandra Grygorczyk. Personality traits, basic individual values and gmo risk perception of twitter users. *Journal of Risk Research*, 23(4):522–540, April 2020.

[24] Francis M Manzira and Felix Bankole. Application of social media analytics in the banking sector to drive growth and sustainability: A proposed integrated framework. In *2018 Open Innovations Conference (OI)*, pages 223–233. IEEE, 2018.

[25] Aaron W. Baur. Harnessing the social web to enhance insights into people's opinions in business, government and public administration. *Information Systems Frontiers*, 19(2):231–251, April 2017.

[26] Soulakshmee D Nagowah and Ravesh Joaheer. A model for classifying people at risk of diabetes mellitus using social media analytics. In *Smart and Sustainable Engineering for Next Generation Applications: Proceeding of the Second International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering (ELECOM 2018), November 28–30, 2018, Mauritius 2*, pages 195–204. Springer, 2019.

[27] Wasim Ahmed and Sergej Lugovic. Social media analytics: analysis and visualisation of news diffusion using nodexl. *Online Information Review*, 43(1):149–160, February 2019.

[28] Annette Ranko, Justyna Nedza, and Nikolai Röhl. A common transnational agenda? communication network and discourse of political -salafists on twitter. *Mediterranean Politics*, 23(2):286–308, April 2018.

[29] Cécile Zachlod, Olga Samuel, Andrea Ochsner, and Sarah Werthmüller. Analytics of social media data – state of characteristics and application. *Journal of Business Research*, 144:1064–1076, May 2022.

[30] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys*, 47(4):1–38, July 2015.

[31] Irina Shklovski, Moira Burke, Sara Kiesler, and Robert Kraut. Technology adoption and use in the aftermath of hurricane katrina in new orleans. *American Behavioral Scientist*, 53(8):1228–1246, 2010.

[32] Kate Starbird, Leysia Palen, Amanda L. Hughes, and Sarah Vieweg. Chatter on the red: what hazards threat reveals about the social life of microblogged information. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, page 241–250, Savannah Georgia USA, February 2010. ACM.

[33] Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 1079–1088, Atlanta Georgia USA, April 2010. ACM.

[34] Diego Kozlowski, Elisa Lannelongue, Frédéric Saudemont, Farah Benamara, Alda Mari, Véronique Moriceau, and Abdelmoumene Boumadane. A three-level classification of french tweets in ecological crises. *Information Processing & Management*, 57(5):102284, 2020.

[35] Carlos Castillo. *Big crisis data: social media in disasters and time-critical situations*. Cambridge University Press, 2016.

[36] Yu Feng, Xiao Huang, and Monika Sester. Extraction and analysis of natural disaster-related vgi from social media: review, opportunities and challenges. *International Journal of Geographical Information Science*, 36(7):1275–1316, 2022.

[37] Anna Kruspe, Jens Kersten, and Friederike Klan. Detection of actionable tweets in crisis events. *Natural Hazards and Earth System Sciences*, 21(6):1825–1845, 2021.

[38] Mats Eriksson. Lessons for crisis communication on social media: A systematic review of what research tells the practice. *International Journal of Strategic Communication*, 12(5):526–551, 2018.

[39] Kenneth A Lachlan, Patric R Spence, and Xialing Lin. Expressions of risk awareness and concern through twitter: On the utility of using the medium as an indication of audience needs. *Computers in Human Behavior*, 35:554–559, 2014.

[40] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *Pro-*

ceedings of the international AAAI conference on web and social media, volume 8, pages 376–385, 2014.

[41] Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. Practical extraction of disaster-relevant information from social media. In *Proceedings of the 22nd International Conference on World Wide Web*, page 1021–1024, Rio de Janeiro Brazil, May 2013. ACM.

[42] Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. Extracting information nuggets from disaster-related messages in social media. *Iscram*, 201(3):791–801, 2013.

[43] Sarah Elizabeth Vieweg. *Situational awareness in mass emergency: A behavioral and linguistic analysis of microblogged communications*. PhD thesis, University of Colorado at Boulder, 2012.

[44] Kiran Zahra, Muhammad Imran, and Frank O Ostermann. Automatic identification of eyewitness messages on twitter during disasters. *Information processing & management*, 57(1):102107, 2020.

[45] Alfredo Cobo, Denis Parra, and Jaime Navón. Identifying relevant messages in a twitter-based citizen channel for natural disaster situations. In *Proceedings of the 24th international conference on world wide web*, pages 1189–1194, 2015.

[46] Stefano Cresci, Maurizio Tesconi, Andrea Cimino, and Felice Dell'Orletta. A linguistically-driven approach to cross-event damage assessment of natural disasters from social media messages. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1195–1200, 2015.

[47] Alaa Alharbi and Mark Lee. Crisis detection from arabic tweets. In *Proceedings of the 3rd workshop on Arabic corpus linguistics*, pages 72–79, 2019.

[48] Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. What to expect when the unexpected happens: Social media communications across crises. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 994–1009, 2015.

[49] Firoj Alam, Hassan Sajjad, Muhammad Imran, and Ferda Ofli. Crisisbench: Benchmarking crisis-related social media datasets for humanitarian information processing.

[50] Gaëtan Caillaut, Cécile Gracianne, Nathalie Abadie, Guillaume Touya, and Samuel Auclair. Automated construction of a french entity linking dataset to geolocate

social network posts in the context of natural disasters. In *19th International Conference on Information Systems for Crisis Response and Management*, 2022.

[51] Volodymyr V Mihunov, Navid H Jafari, Kejin Wang, Nina SN Lam, and Dylan Govender. Disaster impacts surveillance from social media with topic modeling and feature extraction: case of hurricane harvey. *International Journal of Disaster Risk Science*, 13(5):729–742, 2022.

[52] Naina Said, Kashif Ahmad, Michael Riegler, Konstantin Pogorelov, Laiq Hassan, Nasir Ahmad, and Nicola Conci. Natural disasters detection in social media and satellite imagery: a survey. *Multimedia Tools and Applications*, 78:31267–31302, 2019.

[53] Dat T Nguyen, Ferda Ofli, Muhammad Imran, and Prasenjit Mitra. Damage assessment from social media imagery data during disasters. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, pages 569–576, 2017.

[54] Firoj Alam, Umair Qazi, Muhammad Imran, and Ferda Ofli. Humaid: Human-annotated disaster incidents data from twitter with deep learning benchmarks. In *Proceedings of the International AAAI Conference on Web and social media*, volume 15, pages 933–942, 2021.

[55] Firoj Alam, Tanvirul Alam, Md. Arid Hasan, Abul Hasnat, Muhammad Imran, and Ferda Ofli. Medic: A multi-task learning dataset for disaster image classification. *Neural Computing and Applications*, 35:2609–2632, 2023.

[56] Ethan Weber, Dim P Papadopoulos, Agata Lapedriza, Ferda Ofli, Muhammad Imran, and Antonio Torralba. Incidents1m: a large-scale dataset of images with natural disasters, damage, and incidents. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4768–4781, 2022.

[57] Joachim Fohringer, Doris Dransch, Heidi Kreibich, and Kai Schröter. Social media as an information source for rapid flood inundation mapping. *Natural Hazards and Earth System Sciences*, 15(12):2725–2738, 2015.

[58] Benjamin Mandel, Aron Culotta, John Boulahanis, Danielle Stark, Bonnie Lewis, and Jeremy Rodrigue. A demographic analysis of online sentiment during hurricane irene. In *Proceedings of the second workshop on language in social media*, pages 27–36, 2012.

[59] Bertrand De Longueville, Robin S Smith, and Gianluca Luraschi. " omg, from here, i can see the flames!" a use case of mining location based social networks to

acquire spatio-temporal data on forest fires. In *Proceedings of the 2009 international workshop on location based social networks*, pages 73–80, 2009.

[60] Hien To, Sumeet Agrawal, Seon Ho Kim, and Cyrus Shahabi. On identifying disaster-related tweets: Matching-based or learning-based? In *2017 IEEE third international conference on multimedia big data (BigMM)*, pages 330–337. IEEE, 2017.

[61] Venkata Kishore Neppalli, Cornelia Caragea, and Doina Caragea. Deep neural networks versus naive bayes classifiers for identifying informative tweets during disasters. In *Proceedings of the 15th Annual Conference for Information Systems for Crisis Response and Management (ISCRAM)*, 2018.

[62] Xiaodong Ning, Lina Yao, Xianzhi Wang, and Boualem Benatallah. Calling for response: automatically distinguishing situation-aware tweets during crises. In *Advanced Data Mining and Applications: 13th International Conference, ADMA 2017, Singapore, November 5–6, 2017, Proceedings 13*, pages 195–208. Springer, 2017.

[63] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[64] Yang Xiao, Beiqun Li, and Zaiwu Gong. Real-time identification of urban rainstorm waterlogging disasters based on weibo big data. *Natural Hazards*, 94(2):833–842, 2018.

[65] T Mikolov. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[66] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[67] Kevin Stowe, Michael J. Paul, Martha Palmer, Leysia Palen, and Kenneth Anderson. Identifying and categorizing disaster-related tweets. In Lun-Wei Ku, Jane Yung-jen Hsu, and Cheng-Te Li, editors, *Proceedings of the Fourth International Workshop on Natural Language Processing for Social Media*, pages 1–6, Austin, TX, USA, November 2016. Association for Computational Linguistics.

[68] Hongmin Li, Xukun Li, Doina Caragea, and Cornelia Caragea. Comparison of word embeddings and sentence encodings as generalized representations for crisis tweet classification tasks. *Proceedings of ISCRAM Asia Pacific*, 2018.

[69] C Cortes. Support-vector networks. *Machine Learning*, 1995.

[70] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[71] Karl-Michael Schneider. A comparison of event models for naive bayes anti-spam e-mail filtering. In *10th Conference of the European Chapter of the Association for Computational Linguistics*, 2003.

[72] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[73] Yu Feng, Sergiy Shebotnov, Claus Brenner, and Monika Sester. Ensembled convolutional neural network models for retrieving flood relevant tweets. *Image*, 10(1), 2018.

[74] Dat Nguyen, Kamela Ali Al Mannai, Shafiq Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. Robust classification of crisis-related data on social networks using convolutional neural networks. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 632–635, 2017.

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

[76] Junhua Liu, Trisha Singhal, Lucienne TM Blessing, Kristin L Wood, and Kwan Hui Lim. Crisisbert: a robust transformer for crisis classification and contextual crisis embedding. In *Proceedings of the 32nd ACM conference on hypertext and social media*, pages 133–141, 2021.

[77] Mohiuddin Md Abdul Qudar and Vijay Mago. Tweetbert: a pretrained language representation model for twitter text analysis. *arXiv preprint arXiv:2010.11091*, 2020.

[78] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

[79] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, pages 404–417. Springer, 2006.

[80] Shannon Daly and James A Thom. Mining and classifying image posts on social media to analyse fires. In *ISCRAM*, pages 1–14, 2016.

[81] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[82] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

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

[84] Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[85] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[86] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[87] Xukun Li, Doina Caragea, Cornelia Caragea, Muhammad Imran, and Ferda Ofli. Identifying disaster damage images using a domain adaptation approach. In *Proceedings of the 16th International conference on information systems for crisis response and management*, 2019.

[88] Firoj Alam, Ferda Ofli, Muhammad Imran, Tanvirul Alam, and Umair Qazi. Deep learning benchmarks and datasets for social media image classification for disaster response. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 151–158. IEEE, 2020.

[89] Zijun Long, Richard Mccreadie, and Muhammad Imran. Crisisvit: A robust vision transformer for crisis image classification. *arXiv preprint arXiv:2401.02838*, 2024.

[90] Yimin Yang, Hsin-Yu Ha, Fausto Fleites, Shu-Ching Chen, and Steven Luis. Hierarchical disaster image classification for situation report enhancement. In *2011 IEEE international conference on information reuse & integration*, pages 181–186. IEEE, 2011.

[91] ShuChing Chen, Srinivas Sista, Mei-Ling Shyu, and Rangasami L Kashyap. Indexing and searching structure for multimedia database systems. In *Storage and Retrieval for Media Databases 2000*, volume 3972, pages 262–270. SPIE, 1999.

[92] Tao Chen, Dongyuan Lu, Min-Yen Kan, and Peng Cui. Understanding and classifying image tweets. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 781–784, 2013.

[93] Akash Kumar Gautam, Luv Misra, Ajit Kumar, Kush Misra, Shashwat Aggarwal, and Rajiv Ratn Shah. Multimodal analysis of disaster tweets. In *2019 IEEE Fifth international conference on multimedia big data (BigMM)*, pages 94–103. IEEE, 2019.

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

[95] Tao Liang, Guosheng Lin, Mingyang Wan, Tianrui Li, Guojun Ma, and Fengmao Lv. Expanding large pre-trained unimodal models with multimodal information injection for image-text multimodal classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15492–15501, 2022.

[96] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

[97] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.

[98] Iustin Sirbu, Tiberiu Sosea, Cornelia Caragea, Doina Caragea, and Traian Rebedea. Multimodal semi-supervised learning for disaster tweet classification. In *Proceedings of the 29th international conference on computational linguistics*, pages 2711–2723, 2022.

[99] Alakananda Vempala and Daniel Preoţiuc-Pietro. Categorizing and inferring the relationship between the text and image of twitter posts. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, pages 2830–2840, 2019.

[100] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[101] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[102] Yukun Zhu. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *arXiv preprint arXiv:1506.06724*, 2015.

[103] Joel Mackenzie, Rodger Benham, Matthias Petri, Johanne R Trippas, J Shane Culpepper, and Alistair Moffat. Cc-news-en: A large english news corpus. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3077–3084, 2020.

[104] Aaron Gokaslan and Vanya Cohen. Openwebtext corpus. *Skylion007.github.io*, 2019. http://web.archive.org/save/http://Skylion007.github.io/OpenWebTextCorpus.

[105] Trieu H Trinh and Quoc V Le. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*, 2018.

[106] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[107] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. ACL.

[108] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[109] Zijun Long and Richard McCreadie. Is multi-modal data key for crisis content categorization on social media? In *ISCRAM 2022 Conference Proceedings 226 19th International Conference on Information Systems for Crisis Response and Management*, pages 1068–1080, 2022.

[110] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conf. on computer vision*, pages 740–755. Springer, 2014.

[111] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image cap-

tioning. In *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2556–2565, 2018.

[112] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20029–20038, 2023.

[113] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12695–12705, 2020.

[114] Shuguang Chen, Gustavo Aguilar, Leonardo Neves, and Thamar Solorio. Can images help recognize entities? a study of the role of images for multimodal ner. *arXiv preprint arXiv:2010.12712*, 2020.

[115] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

[116] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[117] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[118] Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. Fine-grained image captioning with clip reward. *arXiv preprint arXiv:2205.13115*, 2022.

[119] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proc. of the IEEE int. conf. on computer vision*, pages 2961–2969, 2017.

[120] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[121] Romain Bielawski, Benjamin Devillers, Tim Van De Cruys, and Rufin Vanrullen. When does CLIP generalize better than unimodal models? when judging human-centric concepts. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 29–38, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[122] Pablo Giaccaglia, Carlo A Bono, Barbara Pernici, et al. Enhancing emergency post classification through image information amplification via large language models. In *Proc. Conf. on Information Systems for Crisis Response and Management (ISCRAM 2024)*, pages 1–14, 2024.

[123] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[124] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[125] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[126] Nitish Srivastava and Russ R Salakhutdinov. Multimodal learning with deep boltzmann machines. *Advances in neural information processing systems*, 25, 2012.

[127] Yu-Gang Jiang, Zuxuan Wu, Jinhui Tang, Zechao Li, Xiangyang Xue, and Shih-Fu Chang. Modeling multimodal clues in a hybrid deep learning framework for video classification. *IEEE Transactions on Multimedia*, 20(11):3137–3147, 2018.

[128] Haiman Tian, Yudong Tao, Samira Pouyanfar, Shu-Ching Chen, and Mei-Ling Shyu. Multimodal deep representation learning for video classification. *World Wide Web*, 22:1325–1341, 2019.

[129] Yan Xia and Zhou Zhao. Cross-modal background suppression for audio-visual event localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19989–19998, 2022.

[130] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2015.

[131] Gerasimos Potamianos, Chalapathy Neti, Juergen Luettin, and Iain Matthews. Audio-visual automatic speech recognition: An overview. *Issues in visual and audio-visual speech processing*, 22:23, 2004.

[132] Dong Wang, Di Hu, Xingjian Li, and Dejing Dou. Temporal relational modeling with self-supervision for action segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 2729–2737, 2021.

[133] Javed Imran and Balasubramanian Raman. Evaluating fusion of rgb-d and inertial sensors for multimodal human action recognition. *Journal of Ambient Intelligence and Humanized Computing*, 11(1):189–208, 2020.

[134] Amir Shahroudy, Tian-Tsong Ng, Yihong Gong, and Gang Wang. Deep multimodal feature analysis for action recognition in rgb+ d videos. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1045–1058, 2017.

[135] Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, pages 24043–24055. PMLR, 2022.

[136] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8228–8237, 2022.

[137] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20029–20038, June 2023.

[138] Chenzhuang Du, Tingle Li, Yichen Liu, Zixin Wen, Tianyu Hua, Yue Wang, and Hang Zhao. Improving multi-modal learning with uni-modal teachers. *arXiv preprint arXiv:2106.11059*, 2021.

[139] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[140] Liwei Wu, Shuqing Li, Cho-Jui Hsieh, and James L Sharpnack. Stochastic shared embeddings: Data-driven regularization of embedding layers. *Advances in Neural Information Processing Systems*, 32, 2019.

[141] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems*, 32, 2019.

[142] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[143] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[144] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. Efficient large-scale multi-modal classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[145] Bénédicte Pierrejean and Ludovic Tanguy. Predicting word embeddings variability. In *Proceedings of the seventh joint conference on lexical and computational semantics*, pages 154–159, 2018.

[146] Sabine Gründer-Fahrer, Antje Schlaf, Gregor Wiedemann, and Gerhard Heyer. Topics and topical phases in german social media communication during a disaster. *Natural language engineering*, 24(2):221–264, 2018.

[147] Firoj Alam, Tanvirul Alam, Md Arid Hasan, Abul Hasnat, Muhammad Imran, and Ferda Ofli. Medic: a multi-task learning dataset for disaster image classification. *Neural Computing and Applications*, 35(3):2609–2632, 2023.

[148] Samuel Auclair, Faïza Boulahya, Babiga Birregah, Robin Quique, Rachid Ouaret, and Eddie Soulier. Suricate-nat: Innovative citizen centered platform for twitter based natural disaster monitoring. In *2019 International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, pages 1–8. IEEE, 2019.

[149] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical methods for rates and proportions*. john wiley & sons, 2013.